UNIVERSIDAD COMPLUTENSE DE MADRID
FACULTAD DE EDUCACIÓN
Departamento de Métodos de Investigación y Diagnóstico en Educación

EL VALOR AÑADIDO EN EDUCACIÓN : CUESTIONES TEÓRICAS Y METODOLÓGICAS

MEMORIA PARA OPTAR AL GRADO DE DOCTOR
PRESENTADA POR

Enrique Navarro Asencio

Bajo la dirección de la doctora

María Castro Morera

Madrid, 2014

UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE EDUCACIÓN

DEPARTAMENTO DE MÉTODOS DE INVESTIGACIÓN Y

DIAGNÓSTICO EN EDUCACIÓN

EL VALOR AÑADIDO EN EDUCACIÓN: CUESTIONES TEÓRICAS Y METODOLÓGICAS

Tesis doctoral realizada por:

Enrique Navarro Asencio

Bajo la dirección de:

María Castro Morera

Madrid, 2013

A Sara y mis padres

“todo necio confunde valor y precio”

Proverbios y Cantares

Antonio Machado

Agradecimientos

Durante el tiempo que he estado preparando esta tesis doctoral he escuchado en repetidas ocasiones la comparación de este trabajo con un embarazo, incluso yo mismo he hecho este paralelismo, pese a haberme preguntado en qué sentido exactamente se pueden comparar los dos procesos. Pienso que la similitud puede ser estar en la alegría con que recibes la concepción, aunque se te quede cara de susto; en las primeras semanas de gestación, cuando se producen esos vómitos tan frecuentes; en cuando, pasados unos meses, no consigues verte los pies, incluso en el mismo día del parto. Sin embargo creo que ya sé sobre todo dónde está el parecido: unos padres primerizos y un doctorando siempre hablan de lo mismo: de su futuro bebé, que para el que escribe es su tesis doctoral. Y también el entorno del doctorando, como el de los padres primerizos, está pendiente de su evolución. Preguntan millones de veces (la cantidad es directamente proporcional al tiempo que tardes en defenderla) ¿cómo va tu tesis?, que equivale al ¿para cuándo? a una embarazada. Durante la elaboración de mi tesis no solo he contado con el interés de muchas personas, también con su ayuda y apoyo, por ello quiero mencionarlas aquí.

Por supuesto a mi familia, a mis padres (Enriqueta y Armando), sin ellos yo no estaría aquí y lo que soy se lo debo a ellos. Y a mi compañera de viaje Sara, que ha vivido todo el proceso conmigo y ha tenido la capacidad para aguantarme.

A mi directora María Castro, sin su ayuda esto hubiera sido imposible. Ella ha sido mi guía espiritual y terrenal.

A toda la gente del departamento MIDE de la Universidad Complutense de Madrid. De todos he podido aprender mucho y no solo cuestiones de metodología. Empezando por José Luís Gaviria, que me dio la oportunidad de empezar en la investigación educativa y al que considero un modelo de cómo debe ser un profesor-investigador universitario, todos los profesores y profesoras con los que he podido trabajar (Arturo de la Orden, Ángeles Blanco, Chantal Biencinto, Coral González, Covadonga Ruíz, Eduardo López, Elvira Carpintero, Joseph Mafokozi, José Manuel García, Inmaculada Asensio, Mª José García de la Barrera, Mª José Fernández, Mercedes García, Narciso García, Rafael Carballo y Xavier Ordoñez) y terminando por Miguel A. Serra.

También he contado con el apoyo de profesores del departamento MIDE de otras universidades, como Jesús Jornet (Universidad de Valencia), Luis Lizasoain y Luis Joaristi (Universidad del País Vasco) y Javier Tourón (Universidad de Navarra).

A todo el grupo de investigación MESE (Medida y Evaluación de Sistemas Educativos), la mayoría de ellos también forman parte del departamento MIDE de la Universidad Complutense. La evaluación educativa de la que se nutre esta tesis fue realizada por este equipo.

Al equipo E (Eva Expósito, Eva Jiménez y Esther López), y también Bianca Thoilliez que aunque, en teoría, pertenece al dpto. de teoría e historia de la educación, siempre le ha atraído el lado oscuro. Ellas han sido compañeras de camino en este doctorado con quienes he compartido alegrías y marrones. Y aunque ahora sigamos cauces distintos, estoy seguro de que volveremos a encontrarnos.

A la gente de la UNIR (Universidad Internacional de la Rioja) que me ha acogido en mi presente etapa de desarrollo profesional, que valoró estricta y objetivamente mi currículum y me ha dado la oportunidad de aplicar todo lo que he aprendido estos años.

Tengo que acordarme aquí también de la UNED (Universidad Nacional de Educación a Distancia). En parte porque he recibido el apoyo de profesores que trabajan allí, como algunos miembros del equipo E, Ángel de Juanas, Arturo Galán o el grupo de investigación ESPYD. Y en parte por otros motivos que no vienen el caso pero que conservo en mi memoria y que también me han servido de aprendizaje.

Fuera del ámbito universitario también hay personas que merecen mi agradecimiento. A toda la gente del antiguo CIDE (Centro de Investigación y documentación Educativa), sitio donde disfruté de mi primera beca de postgrado.

Tampoco quiero olvidarme de los dos tutores que he tenido en mis estancias predoctorales en EE.UU, la primera en Portland (Oregón) y la segunda en Boston (Massachusetts). Son el Dr. Yeow Meng Thum y el Dr. Henry Braun respectivamente y ha sido un privilegio poder trabajar y aprender de ellos.

Y a mis amigos de Alicante que no me perdonarán que Camps haya defendido la tesis antes que yo.

Espero no olvidarme de nadie pero si es así, aunque no os haya puesto aquí, tenéis mi agradecimiento. De verdad, a todos, gracias por haber estado y seguir estando ahí.

Abstract

Large scale student assessments from different courses of compulsory education, which are carried out externally and have the purpose of compiling information about the results of the educational system, have acquired a growing importance in Spain. These evaluations are planned and implemented by external agents to the educational centre and they assess great student samples, or the entire population, at the same time.

The Organic Law of Education (LOE, 2006) binds the evaluation to the accountability, that is to say, information must be useful to value the running of the educational system or schools which are supported with public investment. If a great investment of public money is being carried out, it is necessary to know how this amount is being invested and which the outcome is. This kind of assessment systems considers the evaluated agents as main heads of the results obtained by the students. They are assumed as the heads because the academic results of their students are used to carry out the assessment and take decisions supported by this information.

In Spain, with the first diagnosis assessment in 1997, (INECSE, 1998), general assessments were carried out in order to obtain a global diagnosis of the educational system. In the first one, data from 14 and 16 years old students were gathered, but there have been some more, for example, one assessment of the last course of Primary Education in the years 1999, 2003 and 2007 (Instituto de Evaluación, 2007). In the year 1999 it was also carried out an assessment of the last course of Compulsory Secondary Education (INECSE, 2003), which continued with the first 16 years old student diagnosis evaluation. Currently, Diagnosis general Assessments stipulated in the LOE are being carried out (Instituto de Evaluación, 2010; 2011).

Spain also takes part in external international assessments organised by different organisations. For example, the OCDE is in charge of the PISA assessment (*Program for International Student Assessment*) which is carried out every three years since 2000 (INECSE, 2002; 2004; Instituto de Evaluación, 2007b; 2010b); or the IEA (*International Association for the Evaluation of Educational Achievement*)

which is in charge of assessments such as TIMSS (*Trends in International Mathematicas and Science Study*) (INCE, 2002) and PIRLS (*Progress in International Reading Literacy Study*) (INECSE, 2006).

Assessments in Spain collect information about the same academic courses in different years, that is, they analyze transversely different student cohorts. The results of each assessment are studied separately to describe the status of a particular school in a particular moment, as in the case of national assessments, or from a country, in international assessments.

Another possibility is the study of the change produced in schoolchildren results, carrying out different measurements of the academic achievement throughout time on the same cohort of students. This kind of study is not very common in Spain. The first contribution of this kind is the one made by Marchesi, Martínez y Martín (2004), who carried out a longitudinal study of all the information collected, by means of measurement tools developed ad hoc, on a sample of 31 Secondary schools in Madrid during the course 1996-1997.

There is another contribution, which also uses a school sample from Madrid, carried out during the courses 2005-2006 and 2006-2007, whose features and results are summarized in the monograph by Martínez, Gaviria y Castro (2009). The main feature of this assessment is that collects information about the same students at the beginning and at the end of two academic years, and it has as an objective the study of the Value Added (VA hereon) of schools. And with these data several studies of different aspects have been carried out, mainly related to the measurement of academic performance and the study of growth in learning. For instance, it was carried out the study of the dimensionality of scoring used as result measurements and estimated with the Item Answer Theory (TRI) (Lizasoain & Joaristi, 2009), the relationship patterns among those marks and the differences which are produced among cohorts (Gaviria, Biencinto & Navarro, 2009) and the study of the growth shape using longitudinal multi-level patterns in order to analyse the results (Castro, Ruíz & López, 2009).

In addition to these studies, this kind of data demands the need of carrying out empirical verifications of other methodological aspects which are indispensable if we try to obtain estimations of schools AV with the results of the

assessment. Trying different methodologies to obtain the comparability of all different result measurements carried out on the same student or a comparison of the different approaches to analyze the VA are indispensable. This thesis brings about a new study of the data of the aforementioned assessment.

In the international outlook, studying change in learning has a long tradition which starts with the earlier works done, mainly, by Willet y Rogosa (Rogosa & Willett, 1983; Willett, 1989a; 1989b). Analyzing the change involves an advance regarding transversal studies which observe school results in a certain moment, but it is necessary to consider the following question: Is change measure important for educational research? The answer is obvious, just by measuring individual change it is possible to report about the progress of each individual and, consequently, assess the effectiveness of educational systems (Willett, 1994). As people acquire new skills, learn something new, grow physically and intellectually and attitudes and interests are developed, a change is being produced, therefore, we need to know and measure that learning change in order to know about the growth process of individuals. This analysis on learning change, along with the movement of school effectiveness and the study of school effects, and the approximation, from an economic point of view, of the educational production function, are joined to the current need of great scale assessments with accountability purposes in order to obtain AV analytical models in education.

In a general sense, it is possible to distinguish between, on the one hand, educational assessments based on VA which use two measurements of school results to create models. When just two marks are used, it is analyzed the learning profit between two different courses (Demie, 2003; Ray, 2006; Jakubowski, 2008). And, on the other hand, those assessments which use more than two measurements and analyze the growth in learning (Sanders & Horn, 1994; McCaffrey, Lockwood, Doretz & Hamilton, 2003; Zvoch & Stevens, 2003; Singer & Willett, 2003; Ponisciak & Bryk, 2005; Zvoch & Stevens, 2006; Stevens & Zvoch, 2006; Castro, Ruíz & López, 2009).

Many of the states in the U.S. use the VA test of the teacher or schools as a methodology to analyze the data supplied by the assessments. The different states vary as much in the incentives provided as in the analysis models of AV used. For

example, the VA analysis in Tennessee (Sanders & Horn, 1994; Sanders, Saxton & Horn, 1997), the ones developed in Dallas (Webster & Mendro, 1997; Webster, 2005), Memphis (Potamites, Chaplin & Isenberg, 2009) or California (Doran & Izumi, 2004) are some of the most outstanding approaches.

Other countries also carry out VA studies with marks from educational assessments, although with an informative purpose or aimed at the research rather than the estimation of specific results of schools or teachers. For example, in England (Demie, 2003; Ray, 2006), Malta (Hutchison & Misfud, 2005), Australia (Younk, 1999), Poland (Jakubowski, 2008) or the previously mentioned Spanish VA analyses (Martinez, Gaviria & Castro, 2009).

This thesis has the final purpose of carrying out AV estimations of the first cycle of compulsory secondary education of schools in Madrid community. These estimations are the final result of a process marked by the decisions which are taken in different methodological aspects in the elaboration of a model for analyzing the AV. Therefore, the general objective is:

Creating an Value Added Model methodologically appropriate for data from the longitudinal assessment carried out in 2006 and 2007 on a sample of first cycle of compulsory secondary education students in Madrid Community

From the proposed general objective, the following specific objectives are derived:

a. Studying the origin of Value Added analysis in Education.
b. Defining Value Added in Education
c. Analyzing methodological features characteristic of Value Added analysis.
d. Creating a performance scale with assessment data which makes possible to identify the student evolution in mathematical learning.
e. Selecting the appropriate way of measuring change in learning from assessment data.
f. Comparing different Value Added estimation models of schools.

The first three objectives are dealt with in the theoretical part of the work, and the next three in the empirical section. The achievement of our proposed objectives carries out, above all, an empirical comparison work of the processes done to get AV final estimations. The specific characteristics of employed data make possible to carry out the process from different methodological approaches. Therefore, this process of study must be useful for taking a decision regarding which is the most appropriate VA analysis methodology for assessment data.

This thesis, through its different chapters, tries to answer the different methodological aspects which concern educational research in the responsibility of making and studying general assessments and its results. In particular, the thesis is structured in nine chapters and three annexes. The first five chapters create the theoretical body of the work and describe aspects related to the educational and social concern produced by assessments, appearance and development of VA analysis, methodological questions related to its measurement and different approaches used to get that purpose; the next three chapters deal with the empirical aspects of the thesis, such as the description of the sample and the design of measure tools, the development of a vertical scale or comparing the results produced by different Value Added Models (VAM hereon); the last chapter, the ninth, deals with conclusions, limitations and prospects of our work. The three annexes are related to the empirical analyses of the thesis. The first one adds adjustment results from the theory of answer to item and the classic theory of tests, of all items which took part of all measure tools employed. Along with a study of the characteristics of the vertical scale developed. The second one shows an alternative to percentiles to carry out the calculation of horizontal distances between accumulative distribution curves. And the third one deals with the syntax employed to estimate the vertical scale with BILOG and Added Value models with SPSS.

The empirical studies done, described in detail, are the following:

- The characteristics of performance data from Madrid Community assessment, with a longitudinal character, require carrying out a comparison process which establishes four achievement scores in a common scale. Assessments are done about the same construct or

scope, but are increased progressively in complexity and difficulty, to adjust to vertical scale characteristics. The specificity of the design, with two measurements in each academic course, allows carrying out the process in different ways. By changing certain parameters during the comparison and estimation of achievement scores it is possible to get different results. A comparison of these aspects it is carried out in this first empirical study.

- The second empirical work is directly connected with making an added value model. Using a multilevel growth model for the analysis and counting on four performance measurements, are factors which permit carrying out the process from different perspectives. For example, it is possible to use only two measurements to carry out an analysis using the first one as main covariable and the last one as dependent variable. It is also possible to create a growth model with the last three measurements as guideline and the first data gathering as covariable. The growth trajectory can be modified by using the real distance in months among implementations or by making adjustments based on the observed empirical reality, such as the passing of summer among implementations or the third implementation of data gathering when two months have passed from the beginning of the course. Another aspect is the choice of the first status, its modification has effects on the relationship between the initial point and growth.

The main results obtained begin with VA conceptual delimitation, which is understood as the estimation of the school or teacher contribution to the student learning, trying to isolate it from other possible factors beyond its control. The term VA is used to make reference to that estimation, the specific datum, but also to the methodology of information analysis and, in a global view, to the assessment design.

Using VA results generates a debate about if they must be an informative tool used for identifying schools condition, finding the best practices or diagnosing problematic situations. Or, however, they must have a goal closer to public services accountability, with the purpose of proposing guidance about budgetary allocation,

providing tools for parents to choose a school or rewarding or penalizing schools according to their results.

AV improves other ways of school or teacher assessment such as transversal studies or simple comparison of averages at different moments but, as Doran points out (2003), they are not a panacea. In consequence, it should not be the only source of information which accountability systems rely on, moreover those which try to judge educational institutions with the intention of penalizing or reinforcing them.

Regarding the first empirical study, in the light of the results observed in horizontal calibration, Separate Calibration without transformation creates feature estimations with the highest similarity among shapes; therefore the design of equivalent groups fulfills its objective. Nevertheless, employing Whole Calibration and Fixed Calibration tends to create feature estimations similar among shapes too, although with slight differences among applications.

The vertical linking that employ Conjoint Calibration methodology produces more stable results than Fixed Calibration or Separate Calibration. Fixed Calibration procedure seems to have problems at the same time that improves the applications, without producing a growth between the third and fourth application. The four types of calibration separately show a tendency to decrease the growth as the feature increases, this tendency softens in Conjoint Calibration.

Regarding the second empirical study, employing a time measure in months is the most adequate, by reducing the distance between the second and third application it is also softened the growth among applications. The results of the study point out that if it is divided by the correction factor when the application number is employed as time measure, school estimations are almost the same to the ones produced by the model which uses the number of months.

The model which changes the starting point from the first to the second application is which produces a better adjustment. This change in the starting point eliminates the possible Regression To the Mean effect and confirms the arguments of Rogosa (1995). Therefore, the biggest measure error of the first application may be the source of that negative covariance in the model which uses the first application as initial status. An adjustment made a posteriori of status and

growth remainders through a new simple regression analysis, provides very similar results to the model which changes the starting point. And, therefore, it can be also a good option to eliminate the effect which initial status has over growth. However, models which include an adjustment predictor in multilevel model can change drastically school estimations.

Finally, model comparison reveals that estimations produced by gain and growth models show a certain degree of similarity, with correlation coefficients, around 0.5. Estimated profit and gross profit models produce similar remainders, reaching a correlation coefficient over 0.9.

Mixed Linear Model is one of the most interesting for analyzing assessment data. Its versatility allows analyzing changes in each of the last two assessed courses or estimating all the profit between the first and last application. This indicator of global profit has a great similarity with the growth remainder obtained through longitudinal multilevel model, with Pearson values above 0.95.

In Mixed Linear Model there is no significant correlation between status and profit scoring of each course, therefore the possible regression effect is avoided. Besides, summer period can be omitted in the analyses, since it has turned out to be problematic. Data characteristics of this assessment and different results obtained reveal that mixed linear model can be the most appropriate for this kind of situations. This methodology, in addition to allowing the estimation of the profit for each course separately, is adapted to the characteristics of vertical scale which reduces scoring variance as it advances in the scale. .

This work proves the importance of carrying out educational assessments with a high methodological accuracy. Simply status assessment is not enough to create a basis for taking decisions about teachers, schools or educational policies. It is necessary to analyze the change produced in learning and Value Added Models are one of the most powerful tools to do it. Nevertheless, many of its aspects (vertical scales, models, etc.) are in constant revision.

Key words: Educational Assessment, Value Added, Accountability, Vertical Scale, Equating, Regression To the Mean, Gain, Growth, Mixed Linear Models.

Índice

Índice de Tablas

Índice de Gráficos

Índice de Figuras

Introducción

Las evaluaciones a gran escala sobre estudiantes de diferentes cursos de la enseñanza obligatoria, que se realizan de forma externa y tienen el propósito de recopilar información de los resultados del sistema educativo, han adquirido una importancia creciente en España. Estas evaluaciones son planificadas y ejecutadas por agentes externos al centro educativo y evalúan a grandes muestras, o todo el censo de alumnos al mismo tiempo.

La Ley Orgánica de Educación (LOE, 2006) vincula la evaluación con la rendición de cuentas (*accountability*), es decir, la información debe servir para valorar el funcionamiento del sistema educativo o de las escuelas que se mantienen con inversión pública. Si se lleva a cabo una gran inversión de dinero público en el sistema educativo, es necesario conocer cómo se está invirtiendo ese dinero y qué resultados produce.

Este tipo de sistemas de evaluación considera a los agentes evaluados como principales responsables de los resultados que obtienen los estudiantes. Se asume que son los responsables porque se utilizan los resultados académicos de sus estudiantes para llevar a cabo la evaluación y tomar decisiones basadas en esa información.

Existen dos versiones del sistema de rendición de cuentas denominadas "*high-stakes*" (alto impacto) y "*low-stakes*" (bajo impacto), puede decirse que son dos puntos opuestos en el mismo sistema. En el primero, en función de los resultados obtenidos en las evaluaciones, las autoridades competentes otorgan

diferentes tipos de incentivos (premios o castigos) a las escuelas. En el segundo, los resultados de las evaluaciones tienen una finalidad más informativa, es decir, conocer cómo se encuentra el sistema educativo pero sin un carácter sancionador.

La rendición de cuentas en España tiene un carácter informativo (*low-stakes*) y de diagnóstico de la situación, con propósitos de mejora de la escuela pero, aun así, se necesita una información fiable y ajustada a la realidad educativa. Esta rendición de cuentas se distancia de, por ejemplo, el sistema de alto impacto estadounidense que, desde la promulgación en 2001 de la ley *No Child Left Behind* (NCLB) (ningún niño dejado atrás), permite que se utilice la información proporcionada por las evaluaciones con fines sancionadores, de modo que las escuelas que no consigan que sus estudiantes alcancen unos niveles de logro anuales determinados serán penalizadas. También es posible premiar a las que lo consigan.

En España, con la primera evaluación de diagnóstico en 1997[1] (INECSE, 1998), se comenzaron a realizar evaluaciones generales con el fin de llevar a cabo un diagnóstico global del sistema educativo. En la primera se recogieron datos de estudiantes de 14 y 16 años pero ha habido algunas más, por ejemplo, una evaluación en el último curso de Educación Primaria los años 1999, 2003 y 2007 (Instituto de Evaluación, 2007). En el año 1999 también se evaluó el último curso de Educación Secundaria Obligatoria (INECSE, 2003) que continuó con la evaluación diagnóstica iniciada sobre estudiantes de 16 años. Actualmente se llevan a cabo las Evaluaciones generales de Diagnóstico estipuladas en la LOE (Instituto de Evaluación, 2010; Instituto de Evaluación, 2011).

España también participa en evaluaciones externas internacionales organizadas por distintos organismos. Por ejemplo, la OCDE se encarga de la evaluación PISA (*Program for International Student Assessment*) que se realiza cada

[1]Aunque en esta fecha se iniciaron las evaluaciones periódicas sobre las etapas obligatorias de la educación en España, la primera evaluación global del sistema educativo español tuvo lugar en el año 1976. En esta fecha el gobierno encargó a una comisión de expertos la evaluación de los resultados obtenidos tras la implantación de la Ley General de Educación y Financiamiento de la Reforma Educativa (LGE) de 1970.

En la década de los ochenta, el Centro de Investigación, Documentación y Evaluación (CIDE), actual Centro Nacional de Investigación e Innovación Educativa (CNIIE), llevó a cabo varios estudios de carácter valorativo sobre diversos elementos parciales del sistema educativo en sus niveles no universitarios. Uno de ellos se dedicó al estudio sobre los factores que determinan el rendimiento de los estudiantes a través de un modelo causal de análisis (CIDE, 1990).

tres años desde el 2000 (INECSE, 2002; INECSE, 2004; Instituto de Evaluación, 2007b; Instituto de Evaluación, 2010b); o la IEA (*International Association for the Evaluation of Educational Achievement*) que se encarga de evaluaciones como TIMSS (*Trends in international Mathematicas and Science Study*) (INCE, 2002) y PIRLS (*Progress in International Reading Literacy Study*) (INECSE, 2006).

El objetivo principal de estas evaluaciones es describir los niveles medios de rendimiento de los distintos países participantes. Es posible extraer dos características principales de las evaluaciones mencionadas:

- La incorporación de factores de contexto que permita el ajuste de los resultados empleando técnicas estadísticas
- Y el carácter periódico de las evaluaciones que pretende analizar la evolución de los resultados.

La primera característica se vincula directamente con la realidad educativa que analizan las distintas evaluaciones. Esta realidad está influenciada por factores del contexto que pueden determinar los resultados que se obtienen. Las primeras evaluaciones realizadas en España con propósitos diagnósticos recopilaban información del contexto de los estudiantes y las escuelas, con la finalidad de llevar a cabo comparaciones de resultados escolares en función de determinados factores como, por ejemplo, el sexo o la titularidad de las escuelas (INECSE, 2003), e incluso también actitudes hacia los estudios (Instituto de Evaluación, 2007). Por tanto, las puntuaciones que se utilicen para reflejar el logro de las escuelas deben, de alguna forma, considerar esas variables que son ajenas al control escolar y pueden determinar sus resultados.

La segunda característica es su carácter periódico, tienen un carácter plurianual y su objetivo es analizar la evolución de los agentes evaluados, por ejemplo, cómo están cambiando los resultados de determinados países en el caso de la evaluación PISA o como evolucionan los resultados de un curso académico concreto en el caso de las evaluaciones generales de diagnóstico realizadas en España. Las primeras evaluaciones generales de diagnóstico perseguían ese propósito (INECSE, 2003; Instituto de Evaluación, 2007) y las actuales, iniciadas en 2009, también buscan esa comparabilidad entre los resultados de las evaluaciones (Instituto de Evaluación, 2009). Por tanto, si las evaluaciones de diagnóstico que se

realizan cada año pretenden llevar a cabo una comparación de los resultados anualmente, es necesario que cumplan con ciertos requisitos metodológicos.

Las evaluaciones en España recogen información de los mismos cursos académicos en años distintos, es decir, analizan de forma transversal cohortes distintas de alumnos. Los resultados de cada evaluación se estudian por separado para describir el estatus de una determinada escuela en un momento concreto, como ocurre con las evaluaciones nacionales, o de un país, con las evaluaciones internacionales. No obstante, existe la posibilidad de analizar de forma global la evolución de los resultados de un mismo curso conseguidos mediante evaluaciones transversales. Estos análisis permiten evaluar el cambio que se produce de cohorte a cohorte[2], estudiando los cambios en las proporciones de alumnos que se sitúan en los diferentes niveles de logro académico de cada evaluación.

Otra posibilidad es el estudio del cambio que se produce en los resultados escolares de los estudiantes, llevando a cabo diferentes mediciones del logro académico a lo largo del tiempo sobre una misma cohorte de estudiantes. Este tipo de estudio no es muy común en España. La primera aportación de este tipo es la realizada por Marchesi, Martínez y Martín (2004), que llevan a cabo un estudio longitudinal de la información recogida, mediante instrumentos de medida elaborados ad-hoc, sobre una muestra de 31 escuelas de Madrid que impartían Educación Secundaría Obligatoria (ESO) en el curso 1996-1997. Se evalúan diferentes materias académicas (lengua, matemáticas, ciencias sociales, biología y física y química) al inicio del primer curso de ESO y al final del segundo y cuarto curso. Los autores elaboran un modelo lineal general de medidas repetidas con el objetivo de valorar los cambios que se producen en los resultados a lo largo del tiempo. Cada materia se estudia por separado y se incluye el contexto socioeconómico también como factor en el modelo.

Hay otra aportación, que también utiliza una muestra de escuelas de Madrid, realizada durante los cursos 2005-2006 y 2006-2007 y cuyas características y resultados se resumen en el monográfico de Martínez, Gaviria y Castro (2009). Esta evaluación fue un estudio piloto vinculado a un proyecto de

[2]Ver *Apartado III.3.1.2* para la definición de los distintos tipos de análisis del cambio.

Investigación y Desarrollo[3] que se incluy,o en los planes generales de actuación de la inspección educativa durante los cursos 2005-2006[4], 2006-2007[5] y 2007-2008[6]. La principal característica de esta evaluación es que recoge información de los mismos estudiantes al inicio y final de dos cursos académicos. Se utilizaron instrumentos de medida, también elaborados ad-hoc, de matemáticas y compresión lectora y se estudiaron tres cohortes diferentes de estudiantes: 5º y 6º de Educación Primaria (EP), 1º y 2º de ESO y 3º y 4º de ESO.

Esta evaluación tiene por objetivo el estudio del Valor Añadido[7] (VA en adelante) de las escuelas. Y con los datos recogidos se han llevado a cabo estudios de varios aspectos, principalmente relacionados con la medida del rendimiento académico y el estudio del crecimiento en aprendizaje. Por ejemplo, se ha realizado el estudio la dimensionalidad de las puntuaciones utilizadas como medidas de resultados y estimadas con Teoría Respuesta al Ítem (TRI) (Lizasoain & Joaristi, 2009), los patrones de relación entre esas puntuaciones y las diferencias que se producen entre cohortes (Gaviria, Biencinto & Navarro, 2009) y el estudio de la forma del crecimiento utilizando modelos multinivel longitudinales para el análisis de los resultados (Castro, Ruíz & López, 2009).

Además de los estudios realizados, este tipo de datos demanda la necesidad de realizar comprobaciones empíricas de otros aspectos metodológicos que resultan imprescindibles si se pretende obtener estimaciones del VA de las escuelas con los resultados de la evaluación. Probar diferentes metodologías para lograr la comparabilidad de las diferentes mediciones de resultados realizadas sobre el mismo estudiante o una comparación de las diferentes aproximaciones para el análisis del VA se hacen indispensables. Esta tesis lleva a cabo un nuevo estudio de los datos de la mencionada evaluación.

[3]Son los resultados del proyecto I+D financiado por el Ministerio de Ciencia y Tecnología (Ref. SEC2003-09742) y titulado "El valor añadido en educación y la función de producción educativa: un estudio longitudinal". Su investigador principal fue el profesor José Luís Gaviria Soto y los desarrolló el grupo de investigación de la Universidad Complutense de Madrid MESE (Medida y Evaluación de Sistemas Educativos).

[4]RESOLUCIÓN de 7 de septiembre de 2005, de la Viceconsejera de Educación, por la que se aprueba el Plan General de Actuación de la Inspección Educativa para el curso 2005-2006.

[5]RESOLUCIÓN de 2 de octubre de 2006, de la Viceconsejera de Educación, por la que se aprueba el Plan General de Actuación de la Inspección Educativa para el curso 2006-2007.

[6]RESOLUCIÓN de 21 de septiembre de 2007, del Viceconsejero de Organización Educativa, por la que se aprueba el Plan General de Actuación de la Inspección Educativa para el curso 2007-2008.

[7]El análisis del Valor Añadido es el tema central de esta tesis y se define en profundidad en el *Capítulo III.*

Otro aspecto que pone de manifiesto una revisión es la falta de crecimiento entre los resultados de matemáticas de las dos últimas aplicaciones de la cohorte de 1º y 2º de ESO. Esta características puede comprobarse en el trabajo de Castro, Ruíz y López (2009) y es un fenómeno que resulta extraño en este tipo de datos. Se ha seleccionado esta cohorte de estudiantes para llevar a cabo los diferentes estudios empíricos de este trabajo.

En el panorama internacional, el estudio del cambio tiene una larga tradición que comienza con los trabajos tempranos de, principalmente, Willet y Rogosa (Rogosa & Willett, 1983; Willett, 1989a; 1989b). Analizar el cambio supone un avance respecto a los estudios transversales que observan los resultados escolares en un momento determinado, pero es necesario plantearse la siguiente cuestión: ¿Es la medida del cambio importante en la investigación educativa? La respuesta está clara, cuando las personas adquieren nuevas habilidades, cuando se aprende algo nuevo, se crece intelectual y físicamente y las actitudes y los intereses se desarrollan se está produciendo un cambio, por tanto, debemos conocer y medir este cambio para tener una idea sobre el progreso y crecimiento de los individuos. Solo midiendo el cambio individual es posible informar sobre el progreso de cada persona y, consecuentemente, evaluar la efectividad de los sistemas educativos (Willett, 1994).

Este análisis del cambio en aprendizaje, junto con el movimiento de eficacia escolar y el estudio de los efectos escolares, y la aproximación, desde una perspectiva más economista, de la función de producción educativa, se unen a la necesidad actual de las evaluaciones a gran escala con propósitos de rendición de cuentas para dar lugar a los modelos de análisis del VA en educación.

De forma general, es posible distinguir entre, por un lado, las evaluaciones educativas basadas en el VA que utilizan dos medidas de resultados escolares para elaborar los modelos. Cuando se emplean solo dos puntuaciones, se analiza la ganancia en aprendizaje entre dos cursos distintos (Demie, 2003; Ray, 2006; Jakubowski, 2008). Y, por otro, aquellas evaluaciones que utilizan más de dos medidas y analizan el crecimiento en aprendizaje (Sanders & Horn, 1994; McCaffrey, Lockwood, Doretz & Hamilton, 2003; Zvoch & Stevens, 2003; Singer &

Willett, 2003; Ponisciak & Bryk, 2005; Zvoch & Stevens, 2006; Stevens & Zvoch, 2006; Castro, Ruíz & López, 2009).

La metodología de análisis del VA tiene una larga tradición en Estados Unidos, sobre todo desde la publicación de la mencionada ley NCLB. Esta ley crea un sistema de evaluación basado en la rendición de cuentas y dispone que todos los estudiantes deben ser evaluados anualmente a lo largo del periodo de escolarización obligatoria para estudiar su evolución hacia unos objetivos definidos previamente. Por tanto, se hace responsables a los docentes y escuelas de los resultados de sus estudiantes.

Muchos de los estados de EE.UU utilizan el análisis del VA del docente o las escuelas como metodología para analizar los datos proporcionados por las evaluaciones. Los distintos estados varían tanto en los incentivos que proporcionan como en los modelos de análisis del VA utilizados. Por ejemplo, el análisis del VA en Tennessee (Sanders & Horn, 1994; Sanders, Saxton & Horn, 1997), el desarrollado en Dallas (Webster & Mendro, 1997; Webster, 2005), en Memphis (Potamites, Chaplin & Isenberg, 2009) o el de California (Doran & Izumi, 2004) son algunas de las aproximaciones más relevantes.

Otros países también llevan a cabo estudios del VA con las puntuaciones de evaluaciones educativas, aunque con una finalidad informativa u orientada a la investigación más que a la estimación de resultados concretos de las escuelas o los docentes. Por ejemplo, en Inglaterra (Demie, 2003; Ray, 2006), Malta (Hutchison & Misfud, 2005), Australia (Younk, 1999), Polonia (Jakubowski , 2008) o los ya mencionados análisis del VA españoles (Martinez, Gaviria & Castro, 2009).

El análisis del VA puede ser una opción adecuada a las necesidades de las evaluaciones basadas en la rendición de cuentas y, aunque su desarrollo conlleva cierto grado de dificultad metodológica, su versatilidad y también variedad de aproximaciones permite su adaptación a distintas situaciones de evaluación. Una de las principales características del VA es el análisis del cambio en aprendizaje mediante el estudio de su ganancia o crecimiento a lo largo de un periodo de tiempo. Se requieren, por tanto, dos o más mediciones de los resultados escolares de un mismo estudiante y, dependiendo del tipo de análisis del VA, es necesario que esos resultados puedan compararse. Otra de las características es que una vez

que se cuenta con datos de cambio en aprendizaje, mediante el tratamiento estadístico deben estimarse puntuaciones que reflejen los efectos de docentes o escuelas sobre ese cambio, libres de la influencia de otros factores que no dependan del control de las escuelas como, por ejemplo, lo que los alumnos ya conocen (rendimiento previo) o factores contextuales del estudiante (nivel de estudios de los padres, nivel socioeconómico, condición de inmigrante, etc.) o del centro (nivel socioeconómico del centro, número de alumnos por aula, etc.). La inclusión o no de predictores de contexto en los análisis del VA es un tema objeto de debate (Ballou, Sanders & Wright, 2004; Tekwe et al., 2004; Keeves, Hungi & Afrassa, 2005; Choi, Goldschmidt & Yamashiro, 2006; Ferrão, 2009).

Esta tesis tiene la finalidad última estimar el VA de las escuelas de la Comunidad de Madrid en el primer ciclo de educación secundaria obligatoria. Estas estimaciones son el producto final de un proceso marcado por las decisiones que se toman en diferentes aspectos metodológicos de la elaboración de un modelo para el análisis del VA. Por consiguiente, el objetivo general es:

Elaborar un modelo de Valor Añadido metodológicamente adecuado a los datos de la evaluación longitudinal realizada en 2006 y 2007 sobre una muestra de estudiantes de primer ciclo de educación secundaria obligatoria de la Comunidad de Madrid

La elaboración de este modelo conlleva necesariamente afrontar y resolver distintas cuestiones metodológicas relacionadas, por un lado, con el tratamiento de la información recogida sobre los resultados escolares y, por otro, con los modelos utilizados para el análisis del VA. Cómo tratar las respuestas de los estudiantes a los distintos instrumentos de medida diseñados para medir un mismo constructo pero con dificultad creciente, o cómo estimar las puntuaciones de rendimiento a partir de esas respuestas, son aspectos que se situarían dentro del primer grupo. Y cuestiones vinculadas al estudio del crecimiento y la estimación final del VA asociado a las escuelas se encontrarían dentro del segundo grupo.

Del objetivo general planteado se derivan los siguientes objetivos específicos:

a. Estudiar el origen del análisis del Valor Añadido en Educación.
b. Definir el Valor Añadido en Educación
c. Analizar las rasgos metodológicos característicos del análisis del Valor Añadido.
d. Elaborar una escala de rendimiento con los datos de la evaluación que permita identificar la evolución en aprendizaje en matemáticas de los estudiantes.
e. Seleccionar la forma adecuada de medir el cambio en aprendizaje con los datos de la evaluación.
f. Comparar diferentes modelos de estimación del Valor Añadido de las escuelas.

Los tres primeros objetivos se tratan en la parte teórica del trabajo y los tres siguientes, en la parte empírica. Conseguir los objetivos planteados conlleva, sobre todo, un trabajo de comparación empírica de los procesos realizados para conseguir las estimaciones finales de VA. Las características concretas de los datos empleados permiten abordar el proceso desde diferentes aproximaciones metodológicas. Por lo que este proceso de estudio debe ayudar a tomar una decisión respecto a cuál es la metodología de análisis del VA adecuada a los datos de la evaluación.

El carácter longitudinal de la evaluación realizada bajo el amparo del mencionado proyecto I+D permitió recoger información del rendimiento de los estudiantes al inicio y final de dos cursos académicos. La primera aplicación se realizó al comienzo del primer curso de educación secundaria obligatoria, y la última, al acabar el segundo curso, que también coincide con el final del primer ciclo de esta etapa de secundaria.

La evaluación abarcaba, además de la cohorte de estudiantes que se analiza en esta tesis (1º ESO – 2º ESO), dos cohortes más. Una que incluye a los estudiantes que comenzaron en 5º de Educación Primaria en la primera aplicación y finalizaron en 6º. Este periodo es el último ciclo de la etapa de Educación Primaria. Y una última cohorte que evaluó a estudiantes durante el último ciclo de ESO, es decir, comenzaron la evaluación en 3º ESO y acabaron al término de 4º de ESO.

Se seleccionó la cohorte de primer ciclo de ESO por los problemas encontrados en la evolución de sus resultados de rendimiento en matemáticas. Como demostró el trabajo de Castro, Ruiz y López (2009), que no detectó crecimiento entre las dos últimas aplicaciones.

Existe una doble preocupación respecto a este tipo de evaluaciones. Por un lado, la preocupación de los agentes sociales y educativos que se ve reflejada en esa mayor participación del país en evaluaciones internacionales, además de organizar las suyas propias. También se refleja en el desarrollo de la legislación que regula estas evaluaciones, que ha iniciado un proceso de rendición de cuentas a partir de los resultados de la educación, y el aumento de artículos de prensa que hablan sobre dichos resultados. Por otro lado, también hay una preocupación de la investigación en educación por los aspectos más metodológicos de la evaluación. Los resultados deben permitir hacer estimaciones adecuadas del funcionamiento del sistema educativo o el agente evaluado, de ahí la preocupación por cómo deben ser los instrumentos de medida, cómo garantizar la comparabilidad de los resultados, cómo medir el cambio o crecimiento en las evaluaciones periódicas o cómo obtener puntuaciones de las escuelas, a partir de datos de los estudiantes son algunas de las preocupaciones.

Esta tesis, a través de sus diferentes capítulos, trata de dar respuesta a los distintos aspectos metodológicos que preocupan a la investigación educativa encargada de la realización y estudio de las evaluaciones generales y sus resultados. Concretamente, la tesis se estructura en nueve capítulos y tres anexos[8]. Los cinco primeros capítulos forman el cuerpo teórico del trabajo y describen aspectos relacionados con la preocupación educativa y social producida por las evaluaciones, la aparición y desarrollo de los análisis del Valor Añadido, las cuestiones metodológicas vinculadas a su medida y las diferentes aproximaciones que se utilizan para conseguir ese propósito. Los tres capítulos siguientes se

[8]Los tres anexos están relacionados con los análisis empíricos de la tesis. El primero incorpora los resultados de ajuste desde la teoría respuesta al ítem y la teoría clásica de los test, de todos los ítems que formaron parte de los instrumentos de medida utilizados. Junto con un estudio de las características de la escala vertical elaborada. Además, también incluye información sobre la reducción muestral sufrida en la tercera aplicación. El segundo presenta una alternativa a los percentiles para llevar a cabo el cálculo de distancias horizontales entre curvas de distribución acumuladas. Y el tercero se dedica a la sintaxis utilizada para la estimación de la escala vertical con BILOG y los modelos de Valor Añadido estimados con SPSS.

dedican a los aspectos empíricos de la tesis, como la descripción de la muestra y el diseño de los instrumentos de medida, la elaboración de una escala vertical de rendimiento o la comparación de los resultados producidos por diferentes Modelos de Valor Añadido (MVA en adelante). El capítulo noveno, se dedica a las conclusiones, limitaciones y prospectiva del trabajo.

De forma más concreta, el primer capítulo estudia la preocupación social y educativa sobre la evaluación educativa en España. Se centra en la descripción del panorama de la evaluación educativa en España, destacando la preocupación social y educativa creciente. Se analiza la legislación actual sobre evaluación y cómo los medios de comunicación se hace eco de los resultados.

En el segundo capítulo se analiza el tipo de datos utilizados como medidas de resultados en las evaluaciones. También aborda el estudio de los efectos escolares desde la aparición de los iniciales trabajos de eficacia escolar.

El tercer capítulo se dedica a la descripción de la aparición y desarrollo del VA en educación, así como su definición y características específicas. Cuál es el origen del concepto de VA como se conoce hoy en día, qué persigue o para qué se utilizan los resultados son algunas de las cuestiones que se tratan.

El cuarto capítulo versa sobre los posibles aspectos metodológicos que pueden resultar problemáticos y que, por tanto, deben considerarse cuando se desarrollan los MVA. Por un lado, las cuestiones vinculadas a la utilización de puntuaciones de rendimiento obtenidas mediante test y la elaboración de escalas verticales de logro. Por otro lado, la importancia del análisis del crecimiento y aspectos relacionados con la causalidad de las puntuaciones de VA estimadas, el efecto de los predictores de contexto y otras cuestiones metodológicas que deben considerarse.

El quinto capítulo, el último con carácter teórico, recorre los MVA más reseñados. Se diferencia entre modelos de cambio cohorte a cohorte, modelos univariantes, que incluyen los distintos análisis de la ganancia con dos tomas de datos, y modelos multivariantes con más de dos mediciones del logro académico.

El sexto capítulo es el primero con carácter empírico y analiza los aspectos relacionados con el diseño de los test utilizados en la evaluación, la recogida de información, la muestra utilizada y las características de los datos.

El séptimo capítulo tiene como objetivo la elaboración de una escala vertical del rendimiento académico. Para ello, analiza diferentes procesos de equiparación horizontal y vertical, así como de estimación de las puntuaciones del rasgo a través de modelos de Teoría Respuesta al Ítem. Se comparan los resultados empleando principalmente distancias horizontales y tamaños del efecto.

El octavo capítulo, analiza la trayectoria de crecimiento de los estudiantes con el objetivo de utilizar la más adecuada a los datos empíricos empleados en el trabajo. El diseño específico de los datos necesita que se prueben diferentes opciones de análisis del crecimiento. Se han utilizado medidas de tiempo distintas y se ha modificado el punto inicial de partida para conocer sus posibles efectos. Además, se analiza el efecto de ese estatus inicial en el crecimiento estimado ya que puede afectar a las estimaciones de VA y se comparan modelos. Finalmente, se llevan a cabo las estimaciones finales de VA asociadas a las escuelas desde diferentes aproximaciones.

Y el noveno, describe las conclusiones extraídas a lo largo de todo el trabajo. Además comenta las posibles limitaciones del trabajo y su prospectiva.

Parte Teórica:

Evaluación Educativa y Valor Añadido en Educación

Capítulo I: Evaluación educativa en España

Las evaluaciones educativas generales realizadas de forma periódica y orientadas a comprobar el funcionamiento de diferentes aspectos del sistema educativo e, incluso destinadas a la evaluación de escuelas, han adquirido una importancia creciente en la educación española. Estas evaluaciones recopilan información sobre diversos aspectos de los centros educativos que varía en función de la finalidad de la evaluación. Una de las principales fuentes de información de las que se nutre este tipo de evaluaciones son los estudiantes y utilizan test estandarizados para conseguir medidas de sus resultados académicos (CIDE, 1990; INECSE, 2002; INECSE, 2004; Instituto de Evaluación, 2007; 2007b; 2009).

España no solo lleva a cabo evaluaciones planificadas y desarrolladas por organismos nacionales, también forma parte de estudios de evaluación internacionales organizados por la OCDE como la evaluación PISA (*Program for International Student Assessment*) que se realiza cada tres años desde el 2000 (INECSE, 2002; INECSE, 2004; Instituto de Evaluación, 2007b; Instituto de Evaluación, 2010b); y la IEA (*International Association for the Evaluation of Educational Achievement*) con evaluaciones como TIMSS (*Trends in international Mathematicas and Science Study*) (INCE, 2002) y PIRLS (*Progress in International Reading Literacy Study*) (INECSE, 2006).

La proliferación de estas evaluaciones a gran escala del sistema educativo, cada vez hay un número mayor de ellas y son desarrolladas por organismos distintos, y la utilización de sus resultados en el procesos de toma de decisiones

sobre educación, solo hay que escuchar las referencias de algunos políticos españoles a los resultados del estudio PISA, son factores que demandan el mayor rigor metodológico posible en su planificación y desarrollo.

Existe una doble preocupación por este tipo de evaluaciones. Por un lado, la preocupación de los agentes sociales y educativos que se ve reflejada en esa mayor participación del país en evaluaciones internacionales, además de organizar las suyas propias. También se refleja en el desarrollo de la legislación que regula estas evaluaciones y el aumento del número de publicaciones en la prensa que hablan sobre sus resultados. Por otro lado, también hay una preocupación de la investigación educativa. Los resultados de las evaluaciones deben permitir hacer estimaciones adecuadas del funcionamiento del sistema educativo o el agente evaluado, de ahí la preocupación metodológica por cómo deben ser los instrumentos de medida, cómo garantizar la comparabilidad de una escala medida, cómo medir el cambio o crecimiento en aprendizaje en las evaluaciones periódicas, o cómo obtener puntuaciones de las escuelas a partir de datos de los estudiantes.

Agentes sociales y educativos como los encargados de realizar las propias evaluaciones, los políticos que toman las decisiones en educación y los medios de comunicación, muestran un gran interés y, al mismo tiempo, preocupación por la evaluación que se lleva a cabo sobre el sistema educativo español. El aumento del número de evaluaciones que se desarrollan en el país, junto con una evolución de la legislación sobre evaluación en la enseñanza obligatoria y las noticias publicadas en la prensa nacional, son los aspectos que demuestran esa preocupación creciente.

I.1 Evaluaciones generales en España

En las evaluaciones generales, la información proporcionada por los resultados de los estudiantes no se utiliza para aprobar o suspender una materia escolar determinada. El objetivo final va más allá.

La finalidad común de las evaluaciones generales en España es el diagnóstico global de la situación del sistema educativo o, más bien, de un curso determinado o etapa educativa de la enseñanza obligatoria. Es el caso de las

primeras evaluaciones generales llevadas a cabo por el Instituto Nacional de Evaluación Educativa[9] (INEE) sobre alumnos de 14 y 16 años en 1995 (INECSE, 1998). Esta evaluación, además de recoger los resultados escolares mediante test estandarizados, analiza otros factores educativos como los planes de estudios y métodos de enseñanza, el funcionamiento de los centros, la función docente y la sociedad y el sistema educativo. Con el propósito de llevar a cabo un diagnostico de una etapa concreta, se realizaron las evaluaciones en Educación Primaria (EP) en los años 1999, 2003 y 2007 (Instituto de Evaluación, 2007) con el objetivo inicial de conocer y valorar los resultados educativos de, por un lado, el "…*grado de adquisición, por parte de los alumnos, de los contenidos señalados para el curso al final de la educación primaria…*" y, por otro, de los "…*aspectos metodológicos, la práctica educativa, el clima escolar, las actitudes y expectativas que poseen los distintos agentes de la comunidad educativa (alumnos, padres, profesores y equipos directivos) con respecto a la EP, los métodos y hábitos de trabajo y de estudio de los alumnos y diferentes aspectos del entorno escolar.*" (pág. 21). Las pruebas se diseñaron para que los resultados de estas tres evaluaciones fueran comparables, como señala el citado informe.

En la misma línea se encuentra la evaluación llevada a cabo sobre el último curso de Educación Secundaria Obligatoria (ESO) en el año 1999 (INECSE, 2003) que, además, equipara sus resultados de redimiento con los de la evaluación general de 1995 que también contaba con una muestra de alumnos de 16 años. Esta evaluación al terminar la ESO también tenía el objetivo de hacer un diagnóstico de lo que saben los estudiantes en Ciencias de la Naturaleza, Ciencias Sociales, Geografía e Historia, Lengua Castellana y Literatura y Matemáticas, al finalizar la enseñanza obligatoria. Y un segundo objetivo que busca relacionar, a

[9]Este instituto encargado de llevar a cabo las evaluaciones a nivel nacional, también coordina la participación de España en estudios internacionales de evaluación como PISA (Instituto de Evaluación, 2010b), TIMMS (INCE, 2002) o PIRLS (INECSE, 2006). Este organismo que depende del Ministerio de Educación ha sufrido cambios desde su creación en 1990 por la Ley Orgánica General del Sistema Educativo (LOGSE). En su etapa inicial se denominó Instituto Nacional de Evaluación y Calidad del sistema educativo (INECSE). Más tarde, en el año 2000, la Ley Orgánica de Calidad Educativa (LOCE) cambio la denominación por Instituto Nacional de Calidad Educativa (INCE). Esta ley fue derogada por la actual Ley Orgánica de Educación (LOE), de 2006, que le otorgó el nombre de Instituto de Evaluación (IE). Actualmente, con el nuevo gobierno, el nombre ha sido actualizado a Instituto Nacional de Evaluación Educativa (INEE). No obstante sus funciones apenas han sufrido variaciones.

través de resultados comparados, el rendimiento de los alumnos con los factores contextuales y los procesos educativos.

El principal aspecto de análisis son los resultados académicos de sus estudiantes, evaluando la adquisición de los contenidos curriculares en diversas materias escolares. Es una evaluación del rendimiento académico. No obstante, al mismo tiempo, se recoge información de otros aspectos relacionados con el entorno del estudiante y de la escuela y también del funcionamiento de la misma para tratar de relacionarlos con esos resultados. Por tanto, se parte de una consideración previa de que existen determinados factores que pueden hacer varíar los resultados y buscan indentificarlos.

Otra de las características de estas primeras evaluaciones es su carácter plurianual, es decir, se planifican con la finalidad de comprobar hacia dónde se dirige el sistema educativo mediante la recogida de datos en otros puntos temporales de los mismos grupos de edad. Bajo el amparo de este marco evaluativo se llevan a cabo dos evaluaciones al final de ESO y tres al final de la EP, en ambos casos cada cuatro años.

Estas evaluaciones periódicas siguen una estructura de tiempo similar a la desarrollada en el Proyecto para la Evaluación Internacional de Estudiantes (PISA), que las lleva a cabo cada tres años desde el 2000 y en las que España participa (INECSE, 2002; 2004; Instituto de Evaluación, 2007b; 2010b). La evaluación PISA no evalúa contenidos curriculares, es decir, no valora lo que se ha enseñado a los alumnos en las escuelas, sino que es una evaluación de conocimientos y destrezas (competencias[10]) que se esperan de un estudiante que se encuentra a punto de acabar la escolaridad obligatoria. De esta forma se facilita la comparación entre los resultados de los diferentes países participantes, independientemente de sus formas de organización educativa y del currículo escolar. Sin embargo, se debe tener mucha precaución a la hora de vincular los resultados en estas pruebas con lo que se lleva a cabo en los centros escolares porque, solo desde hace poco, el

[10]El proyecto PISA, en su evaluación de 2003 define competencia de la siguiente forma:*"...al tiempo que evalúa los **conocimientos** de los estudiantes, el proyecto OCDE/PISA también examina su capacidad para reflejar y aplicar su **conocimiento y experiencia a los asuntos del mundo rea**l. Por ejemplo, para entender y evaluar los consejos científicos sobre seguridad de la alimentación, un adulto no sólo necesitará conocer algunos datos básicos sobre la composición de los nutrientes, sino que también deberá ser capaz de aplicar esa información. Se utiliza el término **competencia** para condensar esta concepción amplia de los **conocimientos y destrezas**."* (INECSE, 2004, pág. 14)

sistema educativo educativo español se dirige hacia la enseñanza por compentencias.

La evaluación PISA recoge información sobre las competencias en matemáticas, ciencias o comprensión lectora. Cada evaluación tiene una de las competencias como eje principal sobre la que se lleva cabo una estudio en mayor profundidad, aunque se evalúan todas las competencias en cada estudio. PISA se inició con Comprensión Lectora como materia principal a evaluar en el año 2000. En la siguiente evaluación fue Matemáticas (2003) y en 2006 fue Ciencias. El ciclo volvió a comenzar en 2009. Una particularidad de los datos es que la información puede compararse a partir de que la competencia evaluada haya sido materia principal, ya que ese estudio en profundidad permite incluir más ítems en las pruebas que se utilizarán en las siguientes evaluaciones como anclaje para llevar a cabo el proceso que permita comparar esos resultados. Por tanto, los resultados en Comprensión Lectora pueden ser comparados desde el inicio de la evaluación PISA en 2000 y actualmente se cuenta con cuatro mediciones. Matemáticas cuenta con tres resultados comparables y Ciencias únicamente con dos.

Junto con las pruebas que miden esas destrezas se aplican cuestionarios para recoger información del contexto socioeconómico del estudiante y la escuela, con el objetivo de relacionarlos con los resultados académicos de los estudiantes. Los datos de contexto recogidos son los siguientes:

- Hábitos de estudios y sus actitudes ante el aprendizaje;
- El entorno y las características familiares de los alumnos y su nivel socio-económico y cultural;
- Las características de los centros educativos, su financiación y gestión, sus procesos de toma de decisiones y su política de personal;
- Y las características de los procesos de aprendizaje de los alumnos tales como las estrategias de aprendizaje auto-regulado de los alumnos, sus motivaciones y orientaciones y los estilos de aprendizaje.

Tomando como referente la evaluación PISA y el marco legislativo establecido por la Ley Orgánica de Educación (LOE) de 2006 en su artículo 144, se planifican las nuevas evaluaciones generales de diagnóstico del sistema educativo español, llevadas a cabo por el Instituto de Evaluación y los organismos

correspondientes de las Comunidades autónomas, que continúan la trayectoria iniciada con las llevadas a cabo al final de la EP y Secundaria Obligatoria, y cuyo objetivo inmediato es obtener datos representativos del grado de adquisición de las competencias[11] básicas del currículo en las etapas educativas mencionadas (artículo 144.1 de la LOE). No obstante, el objetivo no solo incluye el diagnóstico, sino también la mejora del sistema y la búsqueda de calidad y equidad en la educación:

> "*Las evaluaciones generales de diagnóstico del sistema educativo deben tener como finalidad contribuir a la mejora de la calidad y la equidad de la educación, orientar las políticas educativas, aumentar la transparencia y eficacia del sistema educativo y ofrecer información sobre el grado de adquisición de las competencias básicas.*" (Instituto de Evaluación, 2009, pág. 11).

Todas las evaluaciones mencionadas asumen que los estudiantes, las escuelas y, de forma más global, el sistema educativo tienen características específicas que determinan los resultados educativos. En la evaluación general de diagnóstico se recomienda que los resultados se comparen utilizando las variables de contexto de los estudiantes y las escuelas:

> *"Por todo ello es necesario considerar los contextos socioculturales de alumnos y centros para poder explicar debidamente los resultados de la evaluación; solo deben realizarse comparaciones en un marco contextual que contribuya a explicar las diferencias"* (Instituto de Evaluación, 2009, pág. 12)

La evaluación general de diagnóstico tiene una finalidad formativa para el conjunto del sistema. Sus resultados deben ser útiles para conseguir la transformación y mejora del sistema, además de construir una base sólida para la toma de decisiones sobre política educativa. Para conseguir información sobre la adquisición de las competencias básicas del currículo de todo el territorio nacional, la evaluación se lleva a cabo sobre una muestra y no está prevista para extraer

[11]En esta evaluación las competencias quedan definidas como "***capacidades*** *de los sujetos para utilizar sus* ***conocimientos, habilidades y actitudes*** *en la comprensión de la realidad y en la resolución de problemas prácticos planteados en* ***situaciones de la vida cotidiana****; en resumen, la aplicación de los conocimientos en un contexto determinado para la resolución de un problema*." (Instituto de Evaluación, 2009, pág. 11)

resultados de estudiantes o centros concretos. Para llenar este hueco la LOE ,en sus artículos 21 y 26, establece que las distintas administraciones autonómicas lleven a cabo evaluaciones de diagnóstico anuales sobre todos los centros educativos, es decir, una evaluación censal con los mismos objetivos que la evaluación de diagnóstico llevada a cabo a nivel nacional.

Las Comunidades Autónomas también pueden elaborar sus propios planes de evaluación si lo consideran necesario. Por ejemplo, la Comunidad de Madrid lleva a cabo anualmente la Evaluación de Conocimientos y Destrezas Indispensables (CDI)[12]. Esta evaluación se aleja de la finalidad únicamente formativa[13], establecida en la LOE y confirmada en las resoluciones que dictan las instrucciones de la CDI, que debe tener el proceso evaluativo. Uno de los resultados producidos por esta evaluación son las clasificaciones de centros educativos publicadas en la prensa, en función de las puntuaciones que obtienen sus estudiantes en las pruebas. Este resultado, aunque tiene un carácter inicial meramente informativo, repercute sobre las escuelas y no precisamente para su mejora. Los padres, sobre todo aquellos que llevan a sus hijos a las escuelas situadas en la parte baja del ranking, muestran preocupación por los resultados y buscan respuestas en los centros e incluso pueden llegar a cambiarlos de escuela.

Si un sistema de evaluación tiene por objetivo construir rankings comparativos de escuelas, como ocurre en la evaluación CDI, o de países como pasa con PISA, no puede simplemente utilizar las medias de los resultados de los estudiantes que forman parte de cada uno de los centros evaluados. Procediendo de esta forma se presenta una información sesgada que no se ajusta a la realidad, por dos motivos:

- Las escuelas atienden a poblaciones distintas de estudiantes con características específicas. Algunas de estos factores vinculados al

[12]La evaluación CDI está coordinada por las direcciones generales de cada una de las etapas educativas y tiene un carácter censal. Las últimas resoluciones legislativas que dictan las instrucciones de esta evaluación son la Resolución 21/2011 de 28 de marzo de 2011 (BOCM nº 70) para 3º de ESO y la 35/2011, en el mismo BOCM, para 6º de EP.

[13]Las finalidades principales de la CDI, según los decretos mencionados, deben dirigirse hacia la obtención de información sobre el grado de adquisición de los conocimientos y destrezas que se consideran indispensables, la orientación de la Consejería de Educación y de los propios centros respecto de la eficacia en sus planes y acciones educativas y la organización en los centros medidas de refuerzo dirigidas a garantizar que todos los alumnos adquieran los conocimientos y las destrezas indispensables.

contexto económico y social del alumno pueden influir en su resultados académico y, por tanto, deberían considerarse.

- Puede que haya centros que seleccionen únicamente alumnos de altas capacidades y, sus resultados no serían comparables, de forma justa claro, con centros que no tienen esa capacidad de selección.

Por tanto, si la rigurosidad metodológica del proceso de evaluación debe ser alta en aquellos sistemas cuyo objetivo es conseguir información para diagnosticar la situación o mejorar el proceso, todavía lo debe ser más cuando el propósito, o uno de ellos, es establecer comparaciones de centros educativos, a través de rankings. Aunque estas clasificaciones no tengan la finalidad directa de sancionar a las escuelas, sí pueden llegar a hacerlo de forma indirecta.

Otra tendencia que rompe la CDI respecto a las evaluaciones generales de diagnóstico, al estudio PISA y las evaluaciones previas mencionadas, es la vinculación de los resultados académicos con los datos de contexto. Este aspecto hace más vulnerables las clasificaciones porque ¿cómo se pueden comparar centros educativos que tienen poblaciones distintas de estudiantes?. Por tanto, los resultados asociados a las escuelas, sobre todo en las comparaciones, deben reflejar la realidad educativa.

Con la evaluación PISA ocurre un fenómeno similar al de la publicación de los rankings de centros de la CDI. PISA recoge una gran cantidad de información sobre factores de contexto relacionados con el alumno y las escuelas, y recomienda que los posibles análisis que se lleven a cabo en los distintos países tengan en cuenta determinadas variables de contexto que se encuentran relacionadas con las puntuaciones de resultados de los estudiantes (OCDE, 2006). Sin embargo, los medios de comunicación utilizan únicamente la información bruta, sin tomar en consideración esos efectos del contexto.

Emplear este ranking para valorar como se encuentran los sistemas educativos en la formación de las competencias analizadas puede conducir a errores. Las puntuaciones utilizadas como resultados de los distintos países en la evaluación se encuentran en una escala común, pero existen factores diferenciales entre esos estados evaluados que pueden alterar los resultados y que no se consideran en estas clasificaciones (distintas poblaciones de estudiantes, niveles

socioeconómicos diferentes, etc.). Por ejemplo, en España la enseñanza de competencias básicas dentro del sistema educativo formal obligatorio es una tendencia comenzada con la LOE (2006) y todavía se encuentra en un proceso inicial, por lo que considerar las puntuaciones PISA como un elemento de análisis de la situación del sistema educativo español no es del todo correcto. Y, sobre todo, comparar directamente esas puntuaciones brutas con los resultados de otros países con condiciones y sistemas educativos distintos al español.

I.2 Legislación sobre evaluación educativa

El gran impulso a la actividad evaluadora del sistema educativo comenzó en 1990 con la Ley Orgánica de Ordenación General del Sistema Educativo (LOGSE), que creó el Instituto Nacional de Calidad y Evaluación (INCE), actualmente denominado Instituto Nacional de Evaluación Educativa (INEE), y cuya tarea principal es la evaluación periódica general de los niveles educativos no universitarios, así como la coordinación de la participación de España en los proyectos internacionales de evaluación.

La evaluación de los centros docentes, impulsada desde las administraciones educativas mediante un plan sistemático, se produce por primera vez durante el curso 1991/1992. El Ministerio de Educación puso en marcha una experiencia piloto de evaluación externa de los centros docentes denominada Plan de Evaluación de Centros (Plan EVA), que llevó a cabo el servicio de inspección educativa. Uno de los principales objetivos de este plan fue difundir la cultura evaluadora en el ámbito educativo e impulsar procesos de evaluación interna en los centros a través de la evaluación formativa externa.

En 1995, la Ley Orgánica de la Participación, la Evaluación y el Gobierno de los Centros Docentes (LOPEG) establecía los distintos contenidos y modalidades de evaluación y las competencias de las diferentes instituciones en esta materia y también regulaba las funciones de la Inspección Educativa. Según esta Ley, la evaluación debe aplicarse tanto a los alumnos como a los procesos educativos, el profesorado, los centros y la propia Administración. Del mismo modo, señala que corresponde a las distintas administraciones educativas la elaboración y puesta en

marcha de planes de evaluación periódica de los centros escolares sostenidos con fondos públicos y que el INEE es el encargado de realizar la evaluación general del sistema educativo y de dar apoyo a las administraciones en sus respectivos planes y programas de evaluación.

Por su parte, la Ley Orgánica de Calidad de la Educación (LOCE), de 2002, dedicaba sus títulos VI y VII a la evaluación y a la inspección del sistema educativo, respectivamente. Esta Ley establecía que la evaluación debía extenderse a todo el ámbito educativo regulado en la misma y aplicarse sobre los procesos de aprendizaje de los alumnos, los procesos educativos, la actividad del profesorado, los centros docentes, la Inspección de Educación y la propia administración educativa. Asimismo, cambia la denominación del hasta entonces Instituto Nacional de Calidad y Evaluación (INCE), por la de Instituto Nacional de Evaluación y Calidad del Sistema Educativo (INECSE). Además, esta Ley señalaba, entre otros aspectos, la necesidad de especialización de la Inspección Educativa, atribuyendo a las Comunidades Autónomas la facultad de regular su organización y funcionamiento.

El 3 de mayo de 2006 se aprueba la Ley Orgánica de Educación (LOE) que, en un esfuerzo por simplificar el complejo panorama normativo existente, deroga las leyes anteriores (LOGSE, LOPEG y LOCE) y se establece como norma básica de ordenación general del sistema educativo español.

La Ley Orgánica de Educación (LOE, 2006) reconoce la importancia de la transparencia en el funcionamiento del sistema educativo y, en este sentido, subraya la necesidad de ofrecer una información pública del uso que se hace de los medios y recursos puestos a disposición del mismo, así como de los resultados que por medio de ellos se alcanzan. Además, establece el marco global tanto de la evaluación general del sistema educativo como de la evaluación de los propios centros educativos. Por su parte, las administraciones educativas de las Comunidades Autónomas, en el marco de sus competencias, pueden elaborar y realizar planes de evaluación de los centros educativos, que deben tener en cuenta las situaciones socioeconómicas y culturales de las familias y alumnos que acogen, el entorno del propio centro y los recursos de que dispone. Las administraciones

educativas también deben apoyar y faciliar la auto-evaluación de los centros educativos.

Como dispone la LOE:

> *"La existencia de un marco legislativo capaz de combinar objetivos y normas comunes con la necesaria autonomía pedagógica y de gestión de los centros docentes obliga, por otra parte, a establecer* ***mecanismos de evaluación y de rendición de cuentas****. La importancia de los desafíos que afronta el sistema educativo demanda como contrapartida una información pública y transparente acerca del uso que se hace de los medios y recursos puestos a su disposición, así como una valoración de los resultados que con ellos se alcanza. La* ***evaluación*** *se ha convertido en un valioso instrumento de seguimiento y de evaluación de los resultados obtenidos y de la mejora de los procesos que permiten obtenerlos. Por ese motivo, resulta imprescindible establecer procedimientos de evaluación de los distintos ámbitos y agentes de la actividad educativa, alumnado, profesorado, centros, currículo, Administraciones, y comprometer a las autoridades correspondientes a* ***rendir cuentas*** *de la situación existente y el desarrollo experimentado en materia de educación"* (LOE, 2006, p. 17161)

La evaluación se vincula a la rendición de cuentas además de al seguimiento y mejora del sistema educativo. La situación legislativa en España refleja la importancia de establecer mecanismos de evaluación y rendición de cuentas que permitan obtener información pública y transparente tanto del uso de los recursos educativos públicos como de los resultados que con ellos se alcanzan. Por tanto, la evaluación es un instrumento valioso para el seguimiento y valoración de los resultados y de mejora de los procesos necesarios para alcanzarlos.

El uso que debe hacerse de los resultados obtenidos mediante la evaluación de los centros y del sistema educativo, queda establecido en el artículo 140 de la LOE, de 2006, y señala que deben orientarse hacia las siguientes finalidades:

a. Contribuir a mejorar la calidad y la equidad de la educación.
b. Orientar las políticas educativas.
c. Aumentar la transparencia y eficacia del sistema educativo.
d. Ofrecer información sobre el grado de cumplimiento de los objetivos de mejora establecidos por las Administraciones educativas.

e. Proporcionar información sobre el grado de consecución de los objetivos educativos españoles y europeos, así como del cumplimiento de los compromisos educativos contraídos en relación con la demanda de la sociedad española y las metas fijadas en el contexto de la Unión Europea.

Estas evaluaciones no han de servir para valoraciones individuales de los alumnos, ni para realizar clasificaciones de los centros. En estos apartados se ve claro el carácter no sancionador de la evaluación en el sistema educativo español. Los principios de la evaluación se centran fundamentalmente en el diagnóstico y la mejora escolar, y también se estudia el gasto público que se dirige a educación. Esta es una de las principales diferencias con otros sistemas de evaluación basados en la rendición de cuentas que otorgan premios o castigos en función de los resultados obtenidos en el proceso evaluativo. No obstante, dentro de este artículo (144.2), se especifica que a pesar de las finalidades de la evaluación concretadas en la LOE, no garantiza que las Comunidades Autónomas puedan utilizar estos datos para llevar evaluaciones concretas de estudiantes o clasificaciones de centros.

De acuerdo con la actual legislación, el Gobierno presentará anualmente al Congreso de los Diputados un informe sobre los principales indicadores del sistema educativo español, los resultados de las evaluaciones de diagnóstico españolas o internacionales y las recomendaciones planteadas a partir de ellas.

La responsabilidad de la evaluación general del sistema educativo recae en el Ministerio de Educación, Cultura y Deporte a través del INEE. Este organismo, que depende de la Secretaría General de Educación, formación profesional y universidades, actúa en colaboración con los organismos de evaluación correspondientes de las Comunidades Autónomas, siendo éstas últimas las encargadas de llevar a cabo la evaluación del sistema educativo en su respectivo territorio en lo que respecta a las etapas no universitarias. Las acciones que lleva a cabo el INEE, en colaboración con las administraciones educativas de las Comunidades Autónomas, son, por un lado, los planes plurianuales de evaluación general del sistema educativo, haciendo públicos previamente a su realización los criterios y procedimientos de evaluación. Por otro lado, el Sistema Estatal de Indicadores de la Educación, que contribuirá al conocimiento del sistema

educativo y a orientar la toma de decisiones de las instituciones educativas y de todos los sectores implicados en la educación.

La LOE determina que el INEE y los organismos correspondientes de las Comunidades Autónomas, deberán colaborar en la realización dos tipos de evaluaciones: Por un lado, las evaluaciones generales de diagnóstico, de carácter muestral, que permitan obtener datos representativos tanto del alumnado y de los centros de las Comunidades Autónomas, como del conjunto del Estado. Estas evaluaciones han de versar sobre las competencias básicas del currículo y realizarse en la enseñanza primaria y secundaria. Por otro lado, la Ley establece que las Comunidades Autónomas, dentro del marco de referencia de las evaluaciones generales de diagnóstico[14], deben realizar en todos los centros una evaluación de diagnóstico, de carácter censal, de las competencias básicas alcanzadas por sus alumnos al finalizar el segundo ciclo de la EP y el segundo curso de la ESO. Dichas evaluaciones tendrán un carácter formativo y orientador para los centros e informativo para las familias y para el conjunto de la comunidad educativa. Estas evaluaciones ya se están llevando a cabo anualmente en algunas Comunidades Autónomas y con carácter trianual a nivel nacional. Las primeras evaluaciones generales de diagnóstico a nivel nacional, desde esta perspectiva, se llevaron a cabo en EP en 2009 (Instituto de Evaluación, 2010) y en ESO en 2010 (Instituto de Evaluación, 2011).

Finalmente, otra tarea del INEE, en colaboración con las instituciones equivalentes en las Comunidades Autónomas, es la coordinación de la participación del Estado español en las evaluaciones internacionales como las mencionadas PISA o TIMMS.

I.3 Resultados de las evaluaciones en la prensa

La importancia creciente de la evaluación en el sistema educativo provoca un interés por parte de la prensa[15], sobre todo desde la participación en estudios

[14]Más información en la publicación del INEE sobre el marco de la evaluación (Instituto de Evaluación, 2009)

[15]En esta tesis se comentan algunas noticias y en el blog titulado Evaluación y Valor añadido en Educación (www.evaluaymide.wordpress.com) pueden consultarse otras más recientes.

internacionales, que se hace eco de los resultados de estas evaluaciones y publica noticias relacionadas con este tipo de temática. Algunos de los titulares, que se pueden encontrar en la prensa, relacionados con la prueba de conocimientos y destrezas indispensables (CDI) que lleva a cabo anualmente la Comunidad de Madrid:

- "Los alumnos de 6º de Primaria se quedan en un Bien. Los alumnos de 6º de Primaria que este año han realizado la prueba de Conocimientos y Destrezas Indispensables (CDI) en la Comunidad de Madrid sacaron un 6,79 de media, más de un punto por encima de lo obtenido el pasado año, aunque los estudiantes de 3º de la ESO, que también realizan la prueba de nivel, rozan el suficiente, con un 5,32" (31/05/2010 en Que)[16]
- "IU acusa a la Comunidad de facilitar un 'ranking' con la prueba de primaria" (29/03/2010 en El Mundo)[17]
- REPORTAJE: Clasificación según la prueba de 6º de Primaria Colegio Estilo (privado) Número 3: 'No hemos cambiado nuestra forma de trabajar' (28/03/2010 en El País)[18]
- REPORTAJE: Clasificación según la prueba de 6º de Primaria Colegio Joaquín Blume Número 6:'Que se vea que la educación pública funciona'. El colegio de Torrejón escala 707 puestos en un año (28/03/2010 en El País)[19]
- "La escuela pública mejora su nota. La calificación de los centros gratuitos crece pero aún sigue por detrás del resto - Los privados sacan un 6,34 de media, los concertados un 5,6 y los públicos un 5,25" (28/03/2010 en el país)[20]

[16] http://www.que.es/madrid/201005311606-alumnos-primaria-quedan-bien.html
[17] http://www.elmundo.es/elmundo/2010/03/29/madrid/1269871486.html
[18] http://www.elpais.com/articulo/madrid/hemos/cambiado/forma/trabajar/elpepiespmad/20100328elpmad_5/Tes
[19] http://www.elpais.com/articulo/madrid/vea/educacion/publica/funciona/elpepiespmad/20100328elpmad_6/Tes
[20] http://www.elpais.com/articulo/madrid/escuela/publica/mejora/nota/elpepiespmad/20100328elpmad_2/Tes

- "Los colegios de Torrejón, entre los primeros de la Comunidad en la Prueba de Conocimientos y Destrezas" (07/04/2010 en Global.com)[21]

Estos titulares reflejan la importancia que pueden adquirir los resultados que se obtienen de estas evaluaciones generales. Cualquier persona puede extraer sus propias conclusiones sin conocer realmente cómo se ha llevado a cabo el proceso en sí. Además, si llevan a cabo clasificaciones de escuelas en función de los resultados de la evaluación puede incluso dejar perplejos a los propios centros, que pueden extraer conclusiones totalmente distintas.

En el caso de la CDI, dos centros que han ascendido vertiginosamente en el ranking desde el año 2009 al 2010 tienen opiniones contradictorias respecto a esta situación. La directora de uno de ellos opina que en su centro no ha cambiado nada la forma de llevar a cabo la enseñanza, continúan con el mismo sistema y profesorado (es un centro privado y podría llevar a cabo remodelaciones en su plantilla, por ejemplo). En cambio, el otro centro, de titularidad pública, asegura que el ascenso puede deberse a la implantación del bilingüismo hace unos 4 años. No obstante, resulta anecdótico que el curso anterior estuviera en la parte final del ranking. En mi opinión, estos cambios tan drásticos pueden ser debidos a un posible error en la medición del curso anterior, suponiendo claro que los datos de un año y otro puedan ser comparados.

Lo curioso es que, más allá de criticar la publicación de rankings de centros educativos o conocer qué aspectos han provocado los ascensos o descensos dentro de ellos, nadie se pregunta si el proceso de evaluación ha sido el adecuado ¿son estas comparaciones de los resultados de las escuelas adecuadas? ¿son los resultados de los distintos años equiparables entre sí?

Todas las cuestiones mencionadas plantean que el proceso de evaluación y sobre todo la elaboración de las puntuaciones que se utilizan como resultados para llevar a cabo comparaciones que pueden afectar a las escuelas, debe realizarse de la forma más rigurosa posible. Es decir, empleando la metodología adecuada desde la elaboración de los instrumentos de medida hasta el proceso de análisis de la información que produzca las puntuaciones finales. Además, si la rendición de

[21] http://www.globalhenares.com/noticia/75633/TORREJÓN---Actualidad/colegios-torrejón-siguen-entre-primeros-comunidad-prueba-conocimientos-destrezas.html

cuentas está presente como objetivo de la evaluación, aunque solo sea con propósitos de bajo impacto para las escuelas, debe tenerse en cuenta que factores del entorno socioeconómico del estudiante y la escuela pueden desvirtuar los resultados, enmascarando los logros reales conseguidos por los centros educativos.

A nivel nacional, en primavera del año 2009 se puso en marcha la ya mencionada evaluación general de diagnóstico, con la finalidad de evaluar las competencias básicas de los alumnos al final del segundo ciclo de EP (4º EP) y al final del primer ciclo de ESO (2º ESO). Estas evaluaciones a cargo del Ministerio de Educación y llevadas a cabo a través del INEE también han generado titulares de prensa:

- "Los alumnos con una sola materia en valenciano salen peor evaluados" (24/06/2010 en El país edición C. Valenciana)[22]
- "La Comunidad Valenciana, a la cola en la evaluación de alumnos de Primaria. La prueba se hizo a niños de nueve y 10 años siguiendo la metodología del informe Pisa" (15/06/2010 en el País)[23]
- "El examen a 180.000 alumnos revela la caída del rendimiento en Secundaria (10/09/2010 en El país edición Andalucía)[24]
- "Cataluña y País Vasco, en los puestos 12 y 14 de la Evaluación de Diagnóstico; Asturias, segunda" (19/05/2010 en Magisnet)[25]
- "Las pruebas de evaluación educativa realizadas a casi 179.000 alumnos andaluces arrojan un nivel medio-alto" (21/09/2010 en Teleprensa.es)[26]

No solo las evaluaciones a nivel nacional han aumentado a lo largo de esta última década, también los estudios de evaluación internacionales han adquirido una mayor relevancia. El estudio PISA (INECSE, 2002; INECSE, 2004; OCDE, 2006)

[22] http://www.elpais.com/articulo/Comunidad/Valenciana/alumnos/sola/materia/valenciano/salen/peor/evaluados/elpepiespval/20100624elpval_10/Tes

[23] http://www.elpais.com/articulo/sociedad/Comunidad/Valenciana/cola/evaluacion/alumnos/Primaria/elpepusoc/20100615elpepusoc_16/Tes

[24] http://www.elpais.com/articulo/andalucia/examen/180000/alumnos/revela/caida/rendimiento/Secundaria/elpepiespand/20100910elpand_2/Tes

[25] http://www.magisnet.com/noticia/6030/INFORMACION/cataluña-país-vasco-puestos-12-14-evaluación-diagnóstico-asturias-segunda.html

[26] http://www.teleprensa.es/andalucia-noticia-242655-Las-pruebas-de-evaluaci26oacute3Bn-educativa-realizadas-a-casi-179000-alumnos-andaluces-arrojan-un-nivel-medio-alto.html

o los estudios TIMMS (INECSE, 2002) y PIRLS (INECSE, 2006) son los principales representantes. Estos estudios también son fuente de noticias en la prensa nacional:

- Mates: necesita mejorar. El último informe PISA sobre educación concluye que el nivel de los alumnos españoles en matemáticas es "ligeramente inferior".- Las otras dos áreas analizadas, lectura y ciencias, también revelan datos negativos (04/12/2007 en el País)[27]
- El Informe PISA revela que el nivel de lectura baja de forma acusada en España (05/12/2007 en el diario Información)[28]
- La educación española está estancada desde 2003, según el nuevo informe PISA. La mitad norte del país tiene un nivel más que aceptable, mientras que el sur está muy por debajo de la media (29/11/2007 en cadenaser.es)[29]

En resumen, en el sistema educativo español, el número de evaluaciones educativas que se llevan a cabo con grandes muestras de estudiantes se ha incrementado en las últimas dos décadas. En la legislación educativa actual, la evaluación se considera un elemento fundamental para la mejora de la calidad de la educación y el aumento de la transparencia del sistema educativo, al tratarse de un valioso instrumento de seguimiento y valoración del funcionamiento y de los resultados del sistema educativo y de mejora de los procesos que permiten alcanzarlos. Demostración de este hecho es que el marco normativo hace hincapié en la necesidad de que se evalúen todos los elementos que conforman el sistema educativo, es decir, la evaluación no debe llevarse a cabo únicamente sobre los resultados, sino que resulta imprescindible establecer procedimientos de evaluación de todos los elementos que forman el sistema educativo: los procesos de aprendizaje de los alumnos, los resultados educativos, el currículo, la actividad del profesorado, los procesos de enseñanza, la función directiva, el funcionamiento de los centros educativos, la inspección y las propias administraciones educativas.

[27] http://www.elpais.com/articulo/sociedad/Mates/necesita/mejorar/elpepusoc/20071204elpepusoc_1/Tes

[28] http://www.diarioinformacion.com/cultura/2249/cultura-informe-pisa-revela-nivel-lectura-baja-forma-acusada/699517.html

[29] http://www.cadenaser.com/sociedad/articulo/educacion-espanola-estancada-2003-nuevo/csrcsrpor/20071129csrcsrsoc_1/Tes

La rendición de cuentas, poco a poco, va teniendo mayor importancia en el sistema educativo español como consecuencia de la creciente autonomía pedagógica y financiera que demanda un mayor control de los usos de estas competencias. Una mayor autonomía debe acompañarse de procesos evaluativos sistemáticos que proporcionen información a las escuelas sobre sus puntos fuertes y sus carencias y, por tanto, qué aspectos deben mejorar (Martínez Arias, Gaviria, & Castro, 2009). Esta finalidad de rendir cuentas, que asume la evaluación en España, no se dirige hacia la sanción o recompensa de las escuelas evaluadas, tiene un carácter informativo.

A pesar de esta tendencia, la evaluación CDI realizada en la Comunidad de Madrid utiliza los resultados para llevar esa rendición de cuentas a unos niveles mayores. La publicación de rankings de escuelas sin ningún tipo de información que pueda ayudar a la interpretación de esos resultados, es una forma de sancionarlas.

Si los resultados de las evaluaciones van a emplearse con finalidades cada vez más cercanas a la rendición de cuentas de alto impacto, es necesario contar con estimaciones de resultados que reflejen lo que realmente pasa en las escuelas o sistemas educativos evaluados. La investigación en educación refleja ese interés social y educativo creciente por le evaluación y sus resultados a través del estudio y desarrollo de distintas metodologías, el VA es una de ellas.

El interés y preocupación social y educativa por la evaluación y sus resultados también despierta un interés de los investigadores especializados en metodología de evaluación y análisis de resultados. Las preocupaciones principales están relacionadas, por un lado, con el proceso de medida, como la obtención de puntuaciones de resultados a partir de test estandarizados y lograr la comparabilidad de los resultados de evaluaciones plurianuales, sean o no sobre los mismos estudiantes. Y, por otro, con el análisis de la información, cómo conseguir estimaciones que permitan diagnosticar las situación del sistema, centro o docente evaluado, lo que implica la utilización de puntuaciones de cambio que reflejen el proceso de aprendizaje que se produce en las aulas y la consideración de factores ajenos a la escuela que pueden influir en los resultados.

Capítulo II: Resultados de las evaluaciones, eficacia escolar y efectos escolares

Las herramientas fundamentales que se utilizan en una evaluación general para obtener sus resultados son, en primer lugar, los instrumentos de evaluación utilizados para recoger la información. Pueden diferenciarse dos tipos de instrumentos. Por un lado, las pruebas, normalmente test estandarizados, encargadas de recoger la información de los resultados académicos de los estudiantes y, por otro, los cuestionarios encargados de recoger información de contexto en caso de ser necesario. Este trabajo se centra en los primeros. Y, en segundo lugar, las puntuaciones de los resultados académicos obtenidas a partir de las respuestas de los estudiantes a los distintos instrumentos.

Las pruebas utilizadas para evaluar los resultados son la herramienta visible de todo el proceso de evaluación. Si se pretende evaluar a grandes poblaciones de estudiantes, incluso censos completos, es necesario elaborar instrumentos de medida que permitan la comparación de la información recogida. Y para que sea comparable, todos los estudiantes evaluados deberían responder a las mismas pruebas (o equivalentes) y en las mismas condiciones.

Las respuestas de los estudiantes a esos instrumentos de evaluación son la materia prima que permite la estimación de puntuaciones que determinan la situación del sujeto en el aspecto evaluado. Este trabajo se centra en el desarrollo de esta fase de la evaluación. Esa información bruta, recogida con las pruebas, debe

tratarse para conseguir unos resultados que reflejen, de la forma más real posible, la situación de los estudiantes.

Ambos aspectos se encuentran relacionados y forman parte del proceso de medida. La obtención de unas puntuaciones de rendimiento válidas (miden lo que pretenden medir y no otra cosa) y fiables (un sujeto debe obtener la misma puntuación o equivalente siempre que responda a la misma prueba de evaluación, en el caso teórico de que no se produzca aprendizaje) depende, en gran medida, de cómo se ha llevado a cabo la elaboración de los instrumentos. La utilización de una medida de resultados u otra puede hacer variar las estimaciones finales de los resultados de las escuelas (McCaffrey, Lockwood, Doretz & Hamilton, 2003; Goldschmidt, Choi & Martinez, 2004; Lockwood, McCaffrey, Hamilton, Stecher Le & Martínez, 2007; Ballou, 2009; Briggs & Betebenner, 2009).

II.1 Resultados de las evaluaciones

Los instrumentos para la evaluación de resultados de los estudiantes en evaluaciones generales se basan en las mismas premisas que los test empleados en psicología para medir la inteligencia, la motivación, el autoconcepto, etc. Estas pruebas tratan de medir variables que no son directamente observables, por ejemplo, cuánto sabe un alumno en matemáticas. Hay una teoría psicométrica que respalda la medición de estas variables no observables (latentes) a partir de las respuestas de los sujetos a los test (Muñiz, 1990; Muñiz, 1994; Martinez & Hernández, 2006). El proceso por el que se consigue se denomina medida.

Cuando se miden aspectos académicos los contenidos que van a formar parte de las pruebas deben ser cuidadosamente seleccionados. Las primeras evaluaciones generales desarrolladas en España partían de los contenidos curriculares para elaborar las preguntas que formaron parte de los instrumentos de evaluación y tenían por objetivo el estudio del rendimiento académico (INECSE, 2003; Instituto de Evaluación, 2007). Los diferentes contenidos curriculares se vinculan con distintos procesos cognitivos para dar lugar a ítems que evalúan lo que el alumno sabe en una materia académica, su rendimiento. Todas las preguntas eran objetivas, es decir, ítems de opción múltiple con una única

respuesta correcta. En definitiva, las preguntas de una prueba de evaluación dependen de dos factores:

- Los contenidos curriculares o conocimientos que el alumnado debe adquirir de la materia evaluada.
- Los procesos cognitivos que se ejecutan en las resolución de las preguntas. No es suficiente evaluar conceptos que únicamente requieran la memorización, por ejemplo, en matemáticas, que un estudiante reconozca el teorema de Pitágoras entre otros teoremas, también debe saber resolver esa fórmula y cómo aplicarla en problemas sobre los lados de un triángulo rectángulo.

La evaluación evoluciona acorde con el tipo de enseñanza. Las primeras evaluaciones generales medían el rendimiento en base a una enseñanza curricular basada en contenidos y el desarrollo de diferentes procesos cognitivos. No existe una definición única y válida del concepto de rendimiento académico.

Bloom (1972) considera que el rendimiento es fruto del trabajo escolar, en la medida que el estudiante sea capaz de aplicar los conocimientos aprendidos a otras situaciones. Para Pérez (1981) el rendimiento es producto únicamente de la voluntad del alumno en el aula, dejando de lado factores individuales del estudiante como su propia capacidad.

El concepto de rendimiento académico está inacabado, se ha ido construyendo a partir de distintas definiciones que van integrando los diferentes elementos que conforman el carácter multidimensional del término. Esta afirmación pone de manifiesto la complejidad del término rendimiento, como un concepto que todavía sigue en construcción.

Un grupo de autores ven el rendimiento como un resultado, es decir, un producto. Para Tourón (1985) el rendimiento es un resultado del aprendizaje producido por el alumno, pero no es el producto de una única capacidad, sino el resultado de una suma de factores que actúan en y desde la persona que aprende. Para González (1975) el rendimiento escolar es también un resultado de diversos factores derivados del sistema educativo, la familia y del propio alumno. El rendimiento académico también es un constructo, resultado de los procesos

cognitivos del propio estudiante y de la influencia de distintas variables sobre el alumno y, por tanto, el rendimiento también es un producto (De la Orden, 1985) que no depende únicamente de lo que ocurre dentro de las escuelas. Estas variables pueden estar relacionadas con la escuela y su entorno, con características del aula, de los docentes y de sus compañeros de clase, con aspectos del contexto socio-cultural y económico del estudiante y, por supuesto, con características del propio alumno.

El concepto de rendimiento académico no está acabado, ha ido evolucionado hasta conseguir ese carácter multidimensional. Un claro ejemplo es la definición actual en términos de competencias educativas. En la primera evaluación PISA llevada a cabo define el rendimiento académico en unos nuevos términos:

> *"La meta del proyecto consiste en la* ***evaluación del rendimiento*** *de los sistemas educativos en relación con sus objetivos subyacentes (tal como los define la sociedad) y no en relación con la enseñanza y aprendizaje de un cuerpo de conocimientos. Esta medición de los resultados auténticos es necesaria si se pretende animar a los centros y a los sistemas educativos centrarse en los retos actuales"* (INCE, 1999, p. 23).

El objetivo es la evaluación del rendimiento pero pretende dar otro sentido al concepto, haciéndolo más independiente de los contenidos curriculares y definiéndolo en función de los conocimientos y destrezas necesarias para resolver problemas, razonar y aplicar ideas propias a situaciones de la vida cotidiana. Aunque el término competencia no era el aspecto central de la primera evaluación PISA, ha adquirido una gran relevancia y en evaluaciones posteriores también se incluyen las actitudes como otro elemento de la competencia, además de los conocimientos y las destrezas (Instituto de Evaluación, 2007b).

En la evaluación por competencias el contenido de los instrumentos de medida depende de los conocimientos que deben adquirir, las destrezas y actitudes necesarias (sustituyen a los procesos cognitivos) y la situación o contexto en el que se deben aplicar esos conocimientos, destrezas y actitudes.

El término competencia no aparece por primera vez en educación ligado a la evaluación PISA, tiene un inicios anteriores al proyecto *Design and Selection of Competencies* (DeSeCo) de la OCDE que originó la mencionada evaluación (De la Orden, 2011). Inicialmente estuvo ligado a la formación profesional americana y británica (Carabaña, 2011). Su traspaso a la educación general se inició con el proyecto DeSeCo, cuyo planteamiento era que la formación no puede basarse solo en la adquisición de conocimientos, también debe considerarse la utilización de esos conocimientos en situaciones nuevas y cotidianas de la vida adulta (Tiana, 2011).

Las competencias educativas tratan de diferenciarse, por tanto, de los contenidos curriculares y los procesos cognitivos utilizados hasta ahora. Se vinculan a las habilidades o destrezas que utilizan los sujetos para aplicar los conocimientos aprendidos en tareas relacionadas con aspectos cotidianos de la vida diaria. La enseñanza obligatoria española, desde la publicación de la LOE (2006), ha adoptado este sistema de competencias. Y las evaluaciones generales que se han llevado a cabo desde entonces también se orientan a la evaluación de competencias. En la marco de la evaluación general de diagnóstico las competencias se definen como:

> *"…capacidades de los sujetos para utilizar sus conocimientos, habilidades y actitudes en la comprensión de la realidad y en la resolución de problemas prácticos planteados en situaciones de la vida cotidiana; en resumen, la aplicación de los conocimientos en un contexto determinado para la resolución de un problema."* (Instituto de Evaluación, 2009, pág. 11).

Explorar el cambio que se ha producido en los objetivos y contenidos de la enseñanza no es objeto de este trabajo aunque, por supuesto, de ese cambio depende la elaboración final de las pruebas que van a ser utilizadas durante la evaluación, sobre todo en la construcción de ítems que permitan evaluar los procesos cognitivos o competencias definidas. La medida de las competencia "pasa por una prueba que incluya una demostración del dominio" (Castro, 2011, p. 111) y los instrumentos de medida se adaptan a esos requisitos. Incorporan un contexto, que simula una situación de la vida cotidiana, con el que se relacionan las

preguntas. Se incluyen otro tipo de preguntas además de ítems objetivos, principalmente, de respuesta construida.

El proceso de medida queda vinculado, por tanto, al tipo de resultados que van a ser evaluados. Una definición clara del constructo es fundamental y determinará tanto los instrumentos de medida, como la estimación de los resultados de logro. En esta fase deben tomarse decisiones respecto a los instrumentos de medida (utilizar test estandarizados o no, el formato de los ítems, la longitud de las pruebas, si conviene realizar diferentes formas paralelas de un mismo test, si se va a utilizar únicamente una medida de los resultados de los estudiantes o varias, etc.) y el análisis de las respuestas de los estudiantes a los ítems, principalmente decisiones sobre el modelo psicométrico.

La medida de logro académico trata de reflejar los resultados del proceso educativo. Con el proceso de medida se realiza una inferencia, a partir de datos observados (las respuestas de los estudiantes en el test), de un constructo que no se observa de forma directa. La medición cobra aquí especial relevancia porque permite conseguir las evidencias para realizar juicios de valor sobre un constructo no observado. La **Figura II.1** muestra ese proceso de medida.

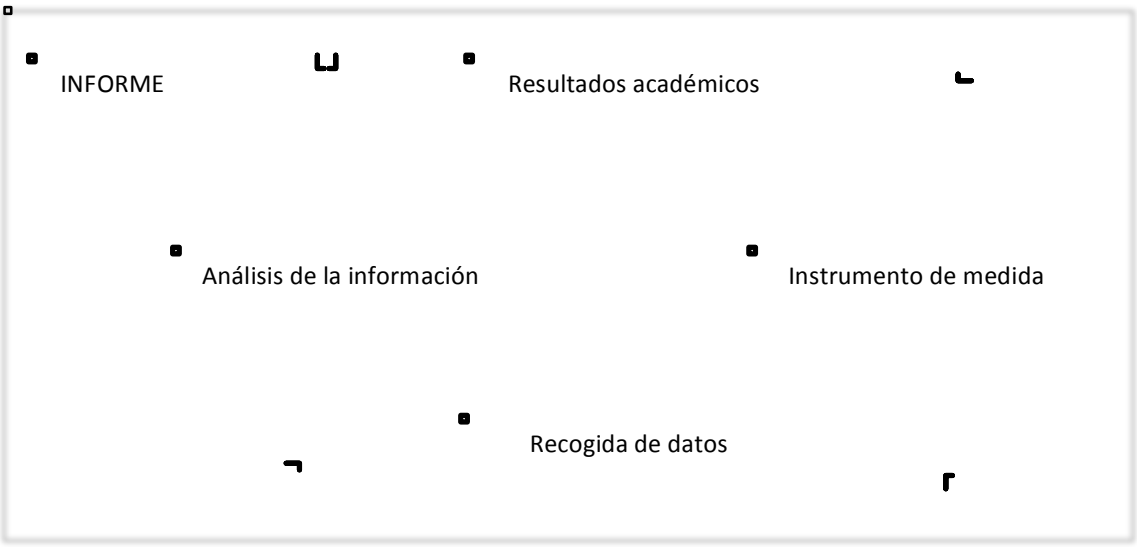


Figura II.1. Proceso de medida en evaluación educativa.
Fuente: Elaboración Propia

El proceso comienza con la definición del constructo que va a ser evaluado. Este constructo se traduce en ítems que componen el instrumento de medida, el test. Es la fase de operacionalización.

Una vez elaboradas las pruebas, comienza la fase de aplicación a la muestra de estudiantes, donde se recogen las respuestas de los estudiantes a los ítems que componen la información bruta de los resultados educativos. Mediante el tratamiento estadístico de esta información, utilizando un modelo psicométrico, se consigue la estimación de la puntuación concreta de un sujeto en el constructo evaluado. Los modelos basados en la Teoría Respuesta al Ítem son los más comunes en las evaluaciones generales nacionales (INECSE, 2003; Instituto de Evaluación, 2007; Instituto de Evaluación, 2009) e internacionales (INCE, 1999; INECSE, 2002; INECSE, 2004; INECSE, 2006; Instituto de Evaluación, 2007b)

Con la medida se logran evidencias suficientes que permiten llevar a cabo inferencias sobre ese objeto medido que se plasmen en resultados concretos que permitan alcanzar los objetivos planificados como, por ejemplo, el diagnóstico de la situación escolar en determinadas etapas educativa, saber qué proporción de alumnos alcanzan determinados niveles o llevar a cabo clasificaciones de escuelas en función de sus resultados.

El informe de resultados de evaluación puede elaborarse con los resultados extraídos de este proceso de medida. Sin embargo, los resultados académicos, medidos en forma de competencias o rendimiento, son un constructo, producto de factores del estudiante y su entorno, además de los procesos escolares. La cuestión es saber qué parte de esos resultados son producto de la acción de la escuela.

Las evaluaciones utilizan esas estimaciones del constructo como elemento clave informativo. Conocer qué parte de esos resultados se debe a factores escolares ha sido uno de los temas de investigación educativa más importante en las últimas décadas. El movimiento de eficacia escolar ha sido el encargado de llevarlo a cabo este tipo de estudios.

II.2 Eficacia Escolar

El estudio de los efectos de las escuelas sobre los resultados académicos comenzó con la publicación, en el año 1966, del trabajo de Coleman, Campbell, Hobson, McPartland, Mood, Weinfeld y York. Los autores afirmaba que las escuelas tienen poca influencia en el aprendizaje de los estudiantes. Como respuesta a estos estudios se generó toda una corriente de trabajos de investigación relacionada con la eficacia de las escuelas que alcanza hasta nuestros días.

Los estudios de eficacia escolar establecen un punto de partida en la importancia creciente que ha tenido la búsqueda de los efectos que la escuela, y los procesos educativos que se producen en ella, tiene sobre el rendimiento de sus estudiantes.

El denominado Informe Coleman, que estudió las relaciones entre factores escolares y del contexto familiar con el logro académico de los estudiantes (Coleman et al., 1966), desencadenó el inició de toda una corriente dirigida hacia el estudio de los efectos escolares. Para Coleman la importancia de los factores asociados al contexto es mucho mayor que las variables asociadas a la escuela.

Este informe fue criticado metodológicamente ya que utilizó para el análisis de los datos la técnica de regresión múltiple paso a paso, introduciendo cómo primeros predictores las variables del contexto socioeconómico, dejando poca varianza por explicar a las variables escolares, este efecto es producto de la colinealidad[30] de muchos de los predictores. Sin embargo, las investigaciones de Jencks y sus colaboradores (1971), que siguieron al estudio Coleman, confirman sus resultados y, además, el autor afirma que en el rendimiento académico lo más importante son las características de los propios estudiantes mientras que los factores escolares son poco relevantes. Sin embargo, multitud de estudios han mostrado cómo variables escolares, del aula y del docente influyen de forma significativa en el logro académico de los estudiantes (Raudenbush & Bryk, 1986; CIDE, 1990; Castejon, Navas & Sampascual, 1996; Goldstein, 1997; Gaviria, Martínez & Castro, 2004; Fernández & Blanco, 2004; Cervini, 2004; Ruiz & Castro,

[30]En el análisis de regresión múltiple, si los predictores que se incluyen en el modelo se encuentran relacionados entre sí, el orden de introducción puede afectar a los resultados. Esto es debido a que explican una parte común de la varianza de la variable criterio.

2006; Navarro & Redondo, 2007; López, Navarro, Ordoñez & Romero, 2009; Ruiz, 2009).

El interés por rebatir los resultados del mencionado informe Coleman dio lugar a numerosos estudios que formaron el movimiento de eficacia escolar. Estos estudios llevaron y llevan a cabo comparaciones cuantitativas de las escuelas, mostrando la importancia que determinados aspectos vinculados a la escuela (tiempo de aprendizaje, clima escolar, liderazgo, etc.) tienen sobre el rendimiento académico. Muchas revisiones se han hecho ya de las numerosas investigaciones de eficacia escolar[31] y no es pertinente volverlas a revisar en este trabajo.

Las investigaciones de eficacia escolar han ido evolucionando y aumentando su complejidad. Las técnicas estadísticas empleadas para el análisis de la información cambian desde los iniciales estudios correlacionales, que cuantificaban la relación existente entre las características de las escuelas y las aulas y los resultados de la enseñanza como el modelo de los 5 factores de Edmons (1979), hasta la utilización de técnicas estadísticas avanzadas como los modelos jerárquicos lineales.

Los modelos jerárquicos lineales, también conocidos como modelos multinivel, han tenido un doble desarrollo. Por un lado, en Inglaterra a cargo de Aitkin y Longford (1986) y Goldstein (1987; 1997) y, por otro, en Estados Unidos con las contribuciones de los profesores Bryk y Raudenbush (1986; 2002). Los modelos multinivel son una técnica estadística basada en la regresión y los modelos lineales de efectos mixtos. El aspecto destacable es que permite llevar a cabo análisis de regresión en diferentes niveles al mismo tiempo, es decir, ya que los estudiantes se encuentran agrupados en aulas y, éstas a su vez en escuelas que pertenecen a distintos distritos, comunidades autónomas, etc., el análisis multinivel descompone la varianza de cada uno de los niveles analizados y permite conocer qué proporción de esa varianza está explicando una determinada variable en cada uno de los niveles.

Creemers, Kyriakides y Sammons (2010) resumen en cuatro fases la evolución de la investigación sobre eficacia escolar:

[31]Para un mayor detalle revisar (Fernández & Gonzalez, 1997; Murillo, 2005; Creemers, Kyriakides & Sammons, 2010)

1. La primera fase, a comienzos de la década de los 80, se centró en el análisis del efecto diferencial que profesores o escuelas podían producir en el rendimiento, mostrando que la escuela importa.
2. Los estudios de la segunda fase, a finales de los 80 y principios de los 90, trató de buscar factores relacionados con la eficacia escolar, es decir, características asociadas a un mejor rendimiento académico. El modelo de 5 factores de Edmonds es uno de los ejemplos y destacaba el liderazgo, las expectativas, la potenciación de las destrezas básicas, un buen clima y una evaluación frecuente del progreso de los estudiantes como elementos clave relacionados con el rendimiento.
3. La tercera fase, a finales de los 90 y principios del nuevo siglo, se centra en la elaboración de modelos de eficacia educativa que apoyen teóricamente por qué determinados factores son importantes en la explicación de la varianza de los resultados de logro de los estudiantes.
4. En la última fase, a partir del año 2000 principalmente, los modelos son más complejos, centrados en el cambio a lo largo del tiempo y analizando aspectos como consistencia, estabilidad o eficacia diferencial. La educación es vista desde una perspectiva dinámica.

Este interés, iniciado por las investigaciones de eficacia escolar, en la comparación de escuelas de forma cuantitativa originó los primeros estudios que, durante la década de los 80 y principio de los 90, centran su atención en la elaboración de indicadores de rendimiento escolar utilizando, normalmente, las puntuaciones medias de los resultados de los estudiantes de un centro educativo. Es decir, para analizar los efectos escolares utilizaron las puntuaciones brutas de los test y clasificaron a las escuelas en función de ese rendimiento medio. Algunos ejemplos de esta metodología son los datos publicados por el gobierno de Inglaterra en sus iniciales Tablas de Rendimiento (*Performance Tables*) (Strand, 1998; Hibpshman, 2004; Ray, 2006). Estos trabajos recibieron críticas que apuntaban la necesidad de contextualizar las medidas del rendimiento escolar si se intenta buscar los efectos que tienen los centros, independientemente de factores ajenos a él (Fitz-Gibbon, 1992; Goldstein & Spiegelhalter, 1996; Goldstein, 1997).

Estas medidas evolucionaron hacia indicadores contextualizados, incluyendo las estimaciones de VA, que tratan de aislar los efectos escolares de factores ajenos a los procesos que se producen en las escuelas, como lo que un estudiante ya conoce o factores socioeconómicos del contexto familiar y escolar.

Esta corriente de investigación fue aumentando en número de estudios y relevancia pero también crecieron las críticas, sobre todo al hablar de los efectos escolares y su proceso de medida. Además, el tratamiento estadístico de la información para conseguir aislar estos efectos y evidenciar qué factores de los centros educativos producen un mayor aporte en el rendimiento conlleva un proceso analítico bastante complejo que debe estudiarse con cautela.

La cuestión es que si van a utilizarse las puntuaciones de rendimiento de los estudiantes para evaluar escuelas u otras unidades educativas, es necesario dotar al proceso con cierto rigor metodológico. No sería apropiado otorgar una nota numérica a un centro en función de, por ejemplo, la media de las puntuaciones que sus estudiantes han obtenido en un test de rendimiento. Si esa información se va a emplear para llevar a cabo comparaciones con centros debe considerarse que esas puntuaciones están afectadas por factores de contexto.

Si, en algún caso, se opta por la utilización de estas puntuaciones medias para conocer el estado de diferentes centros, solo se conseguiría una información sesgada. Por un lado, las poblaciones de estudiantes en los centros puede ser muy variada y, por tanto, la media obtenida por el centro dependerá del tipo de alumnado al que atiende y esto no demuestra realmente el trabajo de la escuela. Por otro lado, una única puntuación nos informa de la situación en un momento temporal concreto, sin tener en cuenta que sabían los estudiantes antes de ese momento de evaluación. Y ese conocimiento previo puede que no sea producto de la acción de la escuela a la que asiste.

II.2.1 Efectos escolares

Los estudios de eficacia escolar, que focalizan su atención en la búsqueda de indicadores del rendimiento de las escuelas, fueron el inicio de un creciente interés en la medición de los posibles efectos escolares. ¿Cómo averiguar la verdadera

aportación de la escuela al rendimiento de los estudiantes? es la cuestión principal que intenta resolver esta corriente.

La medida de esta eficacia implica que los resultados deben estar exentos de posibles variaciones debidas a otro tipo de factores ajenos a la escuela, es decir, que escapan a su control, como el nivel cultural y económico de la familia o la escuela. Por tanto, observar esa contribución escolar al aprendizaje conlleva señalar qué parte de los resultados depende del desempeño de las escuelas y cuál es el papel de otros factores que pueden estar relacionados con el resultado. De este modo es posible establecer una relación entre las escuelas y el logro de sus estudiantes.

Se han llevado a cabo estudios de los efectos escolares con los resultados de las evaluaciones generales realizadas en España. Coinciden en la utilización de los modelos jerárquicos lineales como técnica de análisis para tratar de forma separada la varianza que producen las escuelas en esos resultados educativos. Los datos de la evaluación PISA son los más utilizados en estos estudios que analizan los efectos escolares (Marchesi & Martínez, 2006; Navarro & Redondo, 2007; López, Navarro, Ordoñez & Romero, 2009; Ruiz, 2009) pero también se han llevado a cabo con los datos de la evaluación general de educación primaria realizada en 1995 (Murillo, 2005) o las últimas evaluaciones generales de diagnóstico (Instituto de Evaluación, 2010; 2011).

II.2.1.1 Tipos de efectos escolares

Los efectos escolares tienen ciertas connotaciones y se puede diferenciar entre, al menos, dos tipos (Raudenbush & Willms, 1995). En primer lugar, el llamado efecto Tipo A, es una combinación de factores relacionados con las características contextuales de las escuelas y las prácticas escolares. Este efecto está relacionado con la posibilidad de elección de centro por parte de los padres, en la que sus hijos podrían obtener mejores resultados. La diferencia entre la escuela a la que asisten sus hijos y otra escuela con alumnos de similares características determina los mejores resultados y, por tanto, se puede considerar un efecto de Tipo A.

Raudenbush y Willms (1995) definen el contexto escolar como aquellos factores sobre los que los educadores no tienen control, tales como la composición demográfica de la escuela y el ambiente de la comunidad donde se encuentra situada. Las prácticas escolares son, para los autores, la unión de estrategias de enseñanza, la estructura de la organización y las actividades de liderazgo de la escuela, es decir, factores, en principio, bajo el control escolar. Por tanto, el efecto Tipo A está compuesto por factores que dependen de la escuela (prácticas escolares) y otros que no (contexto escolar).

Otro efecto es el Tipo B, que diferencia la contribución sobre los resultados de los efectos del contexto escolar respecto de aquellos que producen las prácticas escolares. Este efecto es el objeto de estudio de los mencionados estudios de eficacia escolar.

Otros autores realizan una clasificación más concreta de los efectos escolares, añadiendo los de Tipo X y Z (Keeves, Hungi & Afrassa, 2005; Darmawan & Keeves, 2006). Los efectos de tipo X se estiman eliminando de los resultados, además de los efectos relacionados con las características de los estudiantes y el contexto escolar, los efectos de políticas educativas que se escapan al control escolar, por ejemplo, la titularidad. Finalmente, los de tipo Z también controlan el efecto que determinadas prácticas escolares tienen sobre los resultados. Los autores argumentan que los efectos tipo X son los más apropiados para estimar los efectos escolares sobre el aprendizaje, considerando el conocimiento previo de los estudiantes, el contexto de la escuela y algunos efectos de políticas educativas.

El modelo que plantean Raudenbush y Willms para analizar los efectos escolares utiliza los modelos jerárquicos lineales (1995). Este modelo estima ecuaciones de regresión para los diferentes niveles de agregación que se incluyen en el análisis, en este caso los estudiantes agrupados en escuelas. El primer nivel es, por tanto, el alumno y el segundo nivel la escuela. Cada ecuación tiene parámetros fijos (α, b y γ) que son los efectos de cada predictor sobre la puntuación de logro, y parámetros aleatorios (u y ε) que son los residuos de cada ecuación de regresión. El modelo planteado por los autores queda formulado en la ecuación Ec. II.1:

$$Y_{ij} = \alpha + b_w X_{ij} + \gamma_c \overline{X}_j + u_j + \varepsilon_{ij} \qquad \textbf{Ec. II.1}$$

Donde Y_{ij} es la puntuación actual del rendimiento de un estudiante *i* de la escuela *j*. X_{ij} es el rendimiento previo de los estudiantes, mientras que $\overline{X}_j$ es la media de la escuela *j* en ese rendimiento previo

El coeficiente α es la media general de rendimiento para todo el conjunto de estudiantes cuando el resto de predictores son iguales a cero. El coeficiente b_w hace referencia a la puntuación media general del rendimiento previo, el efecto intraescolar del rendimiento previo (la influencia del rendimiento previo de cada estudiante), es decir, el resultado de la regresión del rendimiento previo sobre la variable criterio (el rendimiento final evaluado); γ_c es el efecto contextual del rendimiento previo, en otras palabras, la influencia de asistir a una escuela con una media agregada de rendimiento previo $\overline{X}_j$, una vez controlado el efecto que ese conocimiento previo de cada estudiante tiene sobre la variable de resultados. Se controla tanto el rendimiento previo del estudiante como el de la escuela.

u_j es el efecto aleatorio de la escuela *j*, es decir, el efecto diferencial de la escuela *j* respecto a la media de todas las escuelas, es un residuo[32] de regresión. Y ε_{ij} el residuo aleatorio del nivel de los estudiantes, ambos se asumen como independientes y normalmente distribuidos con media cero y varianza σ_u^2 y σ_e^2, respectivamente.

El efecto tipo A, en este caso, sería el que índica la diferencia entre el rendimiento de un estudiante y el que obtendría si asistiera a una escuela típica. El concepto de escuela típica podría caracterizarse como aquella que atiende a estudiantes con características similares. Si pudiera distribuirse de forma aleatoria a los alumnos con contextos similares en diferentes escuelas, la puntuación de la escuela típica estaría producida por la media de los resultados de esos estudiantes.

Con los efectos tipo A se intenta identificar, además de la aportación producida por la práctica escolar, el producido por el contexto del centro educativo, es decir, no es una medida de eficacia escolar ya que se encuentra afectado por otros factores ajenos al control escolar.

[32]Es importante destacar esta vinculación de los efectos escolares con el residuo que llevan a cabo los autores. Se considera que el efecto escolar es el residuo obtenido una vez ajustados los resultados con diferentes variables. Las estimaciones de Valor Añadido están ligadas a ese residuo.

$$A_{ij} = \gamma_c \overline{X}_j + u_j \qquad \textbf{Ec. II.2}$$

Se asume que la escuela tiene un efecto en cada uno de los estudiantes y, por tanto, el efecto tipo A es el resultado de sumar el efecto diferencial de la escuela u_j (el efecto de las prácticas escolares, es el efecto escolar) más el efecto del contexto de la escuela. En este ejemplo, $\gamma_c \overline{X}_j$ es el efecto producido por escuelas con un determinado nivel de rendimiento previo, pero se pueden introducir en el modelo más predictores relacionados con el contexto del centro educativo. Y, por tanto, se han eliminado los efectos de las características de los estudiantes que en este modelo están representados por el rendimiento previo $b_w X_{ij}$

El efecto tipo B es, como se ha comentado, el efecto específico del centro ($B_{ij} = u_j$). Es la aportación de la escuela al rendimiento de sus estudiantes, una vez controlados los efectos del rendimiento previo de los estudiantes y de las escuelas. Esta puede ser una primera aproximación al concepto de Valor Añadido en educación pero el término es bastante amplio y alcanza mayor o menor complejidad dependiendo, principalmente, del tipo de metodología de análisis empleada para su estimación.

La consideración de los diferentes tipos de efectos escolares ayuda a diferenciar qué parte de la varianza de los resultados educativos es producto de los centros educativos y cuál no. Utilizar un tipo de efecto u otro dependerá de los objetivos que persiga la evaluación, por ejemplo, si es necesario comparar los resultados de los distintos tipos de centro o conocer el efecto diferencial de algunas prácticas escolares.

II.2.2 Estructura anidada de los datos del sistema educativo

Antes de comenzar con la conceptualización del VA en educación, es necesario describir por qué los modelos jerárquicos lineales son adecuados para analizar datos educativos.

Los datos que proceden de las ciencias sociales y del comportamiento y, esencialmente en educación, tienen una estructura anidada. Por ejemplo, las diferentes medidas repetidas anidadas dentro de cada individuo, a su vez, dichas personas se encuentran agrupadas dentro de organizaciones educativas y, además

estas organizaciones pueden estar anidadas dentro de distritos, comunidades e, incluso, países. Los modelos multinivel[33] representan cada uno de los niveles de agregación con un submodelo. Estos modelos expresan las relaciones que se producen entre variables dentro de un mismo nivel y especifican como las variables de un nivel influyen en otro.

La asociación jerárquica de los datos no es algo casual y, por tanto, debe tenerse en consideración al trabajar con datos educativos. Un ejemplo de esto es la posibilidad de que estudiantes con las mismas aptitudes sean agrupados en escuelas que utilizan criterios de selección o también por motivos socioeconómicos. Sin embargo, cuando el grupo está definido, todos sus miembros afectarán y serán afectados por el resto y tenderán a diferenciarse de otros grupos. Ignorar los efectos de los grupos puede invalidar las técnicas de análisis estadístico tradicionales que son utilizadas para el estudio de las relaciones entre datos de estas características. Estas técnicas estadísticas suelen incurrir en dos tipos de errores diferenciados (Hox, 1995):

- Asignar el mismo valor de las variables de las unidades macro, del contexto escolar o del grupo, a las unidades micro, es decir, a cada alumno, sin preocuparse por la posible variación de dichos factores entre los sujetos. Es lo que se conoce como falacia atomística, término acuñado por Alker en 1969.
- Realizar la media de cada variable del alumno para asignársela al grupo al que pertenece. Esto es factible para el estudio de las relaciones de nivel macro (centro), pero no para trasladar estas conclusiones al nivel del alumno. Este error se conoce como falacia ecológica, término que acuñó Robinson en 1950.

Los modelos multinivel ponen solución a este problema trabajando con los diferentes niveles al mismo tiempo. Con estos modelos es posible diferenciar la varianza explicada por cada predictor en los diferentes niveles de agregación

[33]Para un mayor detalle del funcionamiento de los modelos jerárquicos lineales, sus procesos de estimación y aplicaciones entre las que se encuentran el análisis de los efectos escolares y el Valor Añadido en Educación, pueden consultarse las siguientes referencias: Aitkin y Longford (1986), Raudenbush y Bryk (1986; 2002), Willms y Raudenbush (1989), Goldstein (1987; 1997; 1999) y Gaviria y Castro (2005).

seleccionados. Además, es posible realizar inferencias con variables que actúan a diferentes niveles, por ejemplo, la metodología didáctica del docente puede producir efectos diferenciales dependiendo del rendimiento de los alumnos, en algunas ocasiones, tienen mayor eficacia sobre alumnos con bajo rendimiento que con aquellos que poseen un nivel de logro alto.

Otro de los motivos por el que el análisis multinivel resulta útil al analizar este tipo de datos, es la falta de independencia de la información que proviene de observaciones individuales. Los alumnos de un mismo centro tienden a parecerse entre ellos (por ejemplo, algunas escuelas atraerán principalmente a alumnos con un nivel socioeconómico elevado, mientras que otras agruparán a alumnos con un estatus socioeconómico bajo). El grado de homogeneidad de los contextos viene dado por la autocorrelación o correlación intraclase. Las consecuencias de no tener en cuenta la autocorrelación son las siguientes (Gaviria & Castro, 2005):

- La información obtenida a nivel individual no es tanta como parece, debido a que los alumnos de los mismos centros educativos tienden a parecerse entre ellos. Por lo tanto, la información que proporcionan los estudiantes de una mismo escuela es menor que la que suministran los alumnos de distintos centros.
- Los errores típicos son demasiado pequeños debido a que los test estadísticos se basan en el supuesto de independencia de las observaciones. No obstante, en esta clase de estructuras poblacionales dicho supuesto no se cumple. Como consecuencia de ello es posible confirmar la existencia de resultados significativos cuando realmente son espurios.

Además de considerar la correlación intraclase, este tipo de análisis estadístico utiliza estimadores empírico bayesianos o BLUP[34] (*Best Linear Unbiased Predictor*) para llevar a las estimaciones de los residuos aleatorios de las escuelas, es decir, el efecto específico de los centros, los denominados como tipo A. Esta clase de estimadores tienen la característica denominada "*shrunk" o* "*shrunkage*", encogen o suavizan los resultados estimados hacia la media global de aquellas

[34]Más información sobre los estimadores BLUP en el apartado V.1.2.1

escuelas con poca fiabilidad en sus estimaciones. Por tanto, si la fiabilidad de una estimación para una determinada escuela es baja, tiende a no diferenciarse de la media global. Este tipo de estimación puede tener consecuencias positivas o negativas. El lado positivo es que si contamos con datos de escuelas poco fiables, normalmente aquellos centros educativos con una muestra muy pequeña, no tenderán a diferenciarse de la media general lo que nos asegura que no se tomarán decisiones erróneas en base a los resultados obtenidos. Y el lado negativo es que no se podrá saber si ese efecto escolar se debe a los procesos educativos que se dan en el centro porque aunque sus alumnos obtengan buenos resultados tenderán a parecerse a los de la media general.

Capítulo III: El Valor Añadido en Educación

En el contexto educativo y de forma más específica en el campo de las evaluaciones generales de logro escolar, ha aparecido recientemente una nueva metodología para medir los efectos que los centros educativos, los profesores o determinados programas educativos tienen sobre los resultados académico de los estudiantes. Esta metodología, denominada Valor Añadido (VA), tiene en cuenta los posibles efectos diferenciales producidos por el conocimiento previo de los estudiantes y, en ocasiones, también otras características del contexto, que no se encuentran bajo al control del centro educativo (p.ej.: las características socioeconómicas de las familias de los estudiantes) pero pueden influir en sus resultados.

A lo largo de este trabajo se utilizará el concepto de VA para hacer referencia tanto a esa metodología como a la información que se obtiene de ese proceso, es decir, las estimaciones de VA. De esta forma, se consideran los resultados como el VA obtenido por una determinada escuela o un determinado docente.

Del mismo modo, debido a que esta metodología es variada y se aplica de distintas formas, se va a utilizar el término Modelos de Valor Añadido (MVA) para hacer referencia a toda esa variedad de análisis estadísticos empleados para conseguir aislar los efectos escolares.

Los MVA utilizan análisis estadísticos para calcular la contribución de las escuelas o docentes a los resultados de los estudiantes, empleando dos o más

medidas de su logro académico obtenidas mediante test y, en ocasiones, utilizan otros datos del contexto económico y social del estudiante y de la propia escuela. Su aparición e interés creciente ha estado ligado a varios factores. Los estudios de eficacia escolar y el análisis de los efectos de las escuelas, mencionados en el capítulo anterior, pueden considerarse un antecedente. Pero también la aparición de los sistemas de evaluación basados en la rendición de cuentas (*accountability*), que tiene su principal representante en Estados Unidos desde la promulgación, en 2001, de la ley *No Child Left Behind* (NCLB). Y los análisis, desde una perspectiva económica, de la función de producción educativa.

III.1 Rendición de cuentas (Accountability)

En este tipo de sistemas de evaluación los centros son responsables de los resultados que obtienen sus estudiantes y deben responder ante aquellos que han invertido recursos para su financiación, sobre todo, si se utiliza dinero público. Tanto los padres que envían a sus hijos a una determinada escuela cómo los ciudadanos en general que pagan sus impuestos, tienen derecho a conocer qué se está haciendo con esa inversión. La idea que subyace en este tipo de sistemas es que los datos del logro académico de los alumnos son una información adecuada para averiguar cómo está funcionando una institución educativa. Y, por tanto, es posible tomar decisiones sobre las escuelas basadas en esos resultados.

En los sistemas de rendición de cuentas pueden distinguirse dos grandes grupos. Por un lado, los conocidos como sistemas de alto impacto (high-stakes), como los desarrollados en Estados Unidos, que utiliza los resultados de las evaluaciones con fines sancionadores, es decir, las escuelas que no consigan que sus estudiantes alcancen unos niveles anuales determinados de logro serán penalizadas. Este tipo de sistemas también puede tener una orientación positiva, estableciendo un sistema de premios en lugar de las sanciones. Por otro lado, los sistemas de bajo impacto (low-stakes) tienen un carácter más informativo y de diagnóstico de la situación escolar, los resultados de logro de las escuelas se emplean para conocer qué ocurre y hacia dónde se dirige el sistema educativo del que forman parte pero, en ningún caso, tiene una finalidad penalizadora.

La consideración de un sistema de evaluación, basado en la rendición de cuentas, como de alto o bajo impacto puede variar en función del país donde se esté desarrollando. Un mismo sistema puede ser considerado de alto impacto por algunos agentes y, en cambio, para otros será considerado una evaluación de bajo impacto. Por ejemplo, en países con poca trayectoria en el campo de la evaluación, el simple hecho de evaluar de forma externa a las escuelas ya puede ser considerado por algunos docentes o directores como un sistema de alto impacto, aunque la información solo sea utilizada para conocer la situación general del sistema educativo.

La mencionada ley estadounidense de 2001 NCLB, hace responsables a las escuelas de los niveles de rendimiento de sus alumnos y establece sanciones para aquellos centros educativos que no logren un progreso anual adecuado en concordancia con los objetivos de esta legislación. Por lo tanto, la educación en este país, teniendo en cuenta las pequeñas diferencias existentes entre los diferentes estados que lo componen, se ha convertido en un sistema de rendición de cuentas basado en estándares de rendimiento, dicho de otro modo, los centros y los profesores deben rendir cuentas y hacerse responsables del aprendizaje de sus estudiantes.

El gobierno federal adquiere el rol de árbitro en el establecimiento de los objetivos de rendimiento y del progreso que deben conseguir los estudiantes, dejando el establecimiento de las evidencias para realizar juicios de valor, incluyendo contenidos curriculares y la elección de los instrumentos de evaluación, a los diferentes estados. También hay varianza entre los estados en la forma de penalizar o premiar a las escuelas en función de los resultados. P.ej. el estado de Texas tiene uno de los sistemas con más alto impacto, penalizando a los centros con la retirada de parte de su financiación si no alcanzan los resultados establecidos. Esto produce efectos perversos en el sistema educativo y la propia evaluación porque las escuelas hacen lo posible por obtener buenos resultados en esas pruebas y, para ello, son capaces de cualquier cosa como sacar de las aulas, durante la aplicación de las pruebas de evaluación, a los estudiantes con peores resultados o dedicar las horas de clase a resolver este tipo de exámenes, incluso copiar (Vasquez & Darling-Hammond, 2008).

En definitiva, la financiación de la educación pública en Estados Unidos es un sistema de rendición de cuentas basado en incentivos. Los centros educativos consiguen estos incentivos (o penalizaciones) sí consiguen un progreso anual adecuado (*Adecuate Year Progress*) hacia los objetivos propuestos por la ley para el año 2014. Esta meta propuesta por la ley es que todos los estudiantes deben ser competentes en ese año, es decir, deben alcanzar unos niveles determinados de rendimiento. Y se penalizará o premiará a los centros en función de si avanzan de forma adecuada hacia este objetivo.

Hay una tendencia internacional creciente dirigida a establecer sistemas que midan el rendimiento del sector público en términos de eficacia y eficiencia, pero no todos utilizan sistema de premios y sanciones para las escuelas en función de sus resultados. Por consiguiente, si se pretende juzgar a las escuelas en función de los resultados de las evaluaciones, es conveniente contar con información real que refleje la aportación de las escuelas al aprendizaje de sus alumnos.

Los análisis del VA han sido una respuesta a la necesidad marcada por este tipo de sistemas de evaluación basados en la rendición de cuentas. No obstante, el término de VA no surge en el ámbito educativo sino que proviene de sectores más económicos o, más bien, industriales.

III.2 Función de Producción Educativa

En los sectores económicos e industriales el término VA se utiliza para cuantificar la diferencia entre el valor de un producto y los costes de los materiales empleados para su fabricación, el proceso entradas-producto (*input-output*). También es un concepto aplicado a los resultados de una empresa, donde puede definirse como el rendimiento neto de una empresa u organismo durante un periodo de tiempo determinado, es decir, los outputs o beneficios una vez que se han restado los inputs o costes.

En educación el término es un poco más ambiguo porque los análisis del VA centran sus esfuerzos en la medición de los efectos de las escuelas o docentes en el aprendizaje de sus estudiantes, aislándolos de otros factores ajenos al control escolar. Este aprendizaje se entiende como un cambio en los resultados escolares

pero el cambio no se operativiza como la simple de diferencia entre pretest y postest.

Si el término VA proviene del contexto económico, su vinculación con la educación también ha sido llevada a cabo por investigadores de ese mismo campo como Eric Hanushek (1972) o Richard Murnane (1975). Estos autores, en sus trabajos sobre los efectos de los profesores y las escuelas en los resultados educativos que utilizan análisis econométricos como la función de producción educativa, ya incluyen el término VA vinculado con el estudio de los efectos que determinados aspectos educativos tienen sobre el rendimiento.

Toda una corriente de investigación econométrica ha intentado trasladar los análisis input-output al sector educativo. Estos investigadores, de la misma forma que los que estudiaban la eficacia de las escuelas, citan el informe Coleman (1966) como el origen de los estudios de la función de producción educativa (Hanushek, 1971; 1979). Este citado informe parece haber provocado no solo la aparición de estudios desde el sector educativo (eficacia escolar) y desde el sector de la economía de la educación (función de producción educativa), también ha vinculado ambas metodologías para dar lugar a los estudios del VA en educación

Con los análisis de la función de producción educativa se relacionan los resultados educativos con diferentes variables controladas por las escuelas o las administraciones y, por tanto, susceptibles de ser cambiadas, es decir, centra su atención en la relación entre las entradas y los procesos del sistema educativo con los resultados que produce.

La función de producción, al relacionar estadísticamente los recursos y los resultados educativos, ayuda a valorar la productividad de una escuela o un sistema. La definición básica de la función de producción educativa es la siguiente (Ec. III.1) (Hanushek, 1972):

$$R = f(A, P, S, I) \qquad \textbf{Ec. III.1}$$

Donde:

- R = rendimiento académico del alumno, normalmente medido utilizando test.

- A = vector de variables de recursos y características familiares del alumno.
- P = vector de variables de la influencia de los compañeros de aula (*peers effect*)[35]
- E = vector de variables de recursos y características de la escuela.
- I= vector de características innatas del sujeto

Por tanto, los estudios de la función de producción educativa son análisis estadísticos que relacionan los resultados escolares observados en los estudiantes con las características de sus familias, de sus compañeros de aula, de las escuelas y las suyas propias. Hanushek (1971) argumenta que si el modelo para estimar la función de producción educativa incluye una o más puntuaciones del rendimiento previo de los estudiantes, también se estima el VA de las diferentes variables de entrada. Por su parte, Murnane (1975) especifica en su estudio que la información de resultados recogida sobre los estudiantes tiene un carácter longitudinal e informa sobre su progreso y, por este motivo, el análisis está focalizado en el VA producido por los diferentes recursos de la escuela.

Estos análisis de los economistas en educación, a través de la función de producción educativa, tratan de analizar el funcionamiento de ciertos elementos de entrada y proceso sobre los resultados de los estudiantes, una vez asilados ciertos factores contextuales, fundamentalmente lo que un alumno ya sabe. El VA, aunque tiene ciertas similitudes, centra su atención en los efectos producidos por las escuelas o los docentes en el progreso de sus estudiantes, intentando aislarlos de otros factores del contexto socioeconómico del estudiante y la escuela que pueden interferir en los resultados educativos, más bien, en el cambio que se produce en esos resultados.

Los estudiantes entran en las escuelas para aprender, crecer, desarrollarse y cambiar. Estos cambios son creados y mantenidos por las actividades y recursos de nuestras escuelas, por lo tanto, es la medida de estos cambios y la investigación de estas relaciones la que sostendrá ciertas actividades en las aulas y los recursos

[35]La influencia de los compañeros de clase es similar a la de las familias y se suele medir a través de variables agregadas del nivel socioeconómico individual en una determinada aula o escuela (Hanushek, 1971)

proporcionados a las escuelas. La noción de aprendizaje implica crecimiento y cambio y la medida de este cambio ofrece un pequeño avance para los que tratan de evaluar el aprendizaje individual.

Pero esta concepción inicial, tanto de la función de producción educativa como del VA mencionado por Hanushek y Murnane, tiene alguna deficiencia si el objetivo es estimar efectos escolares que se ajusten a la realidad educativa. El problema es que la variable de resultados se considera como una función lineal del rendimiento previo y otras covariables, o como una ganancia bruta (pretest menos postest). En ambos casos, los efectos escolares analizados se consideran fijos a lo largo de todos los grupos, es decir, todas las escuelas tendrían el mismo efecto sobre los resultados de aprendizaje. Y este fenómeno es muy poco probable en educación.

III.3 Valor Añadido en Educación

Los autores Anthony S. Bryk y Herbert I. Weisberf (1976) cambian esa concepción estática de los modelos econométricos de *input-output,* considerando los efectos de la intervención educativa como variantes o dinámicos. Introducen la idea, utilizando un modelo estadístico diferente a la regresión múltiple o el análisis de covarianza, de que un determinado programa educativo tendrá un efecto que producirá un cambio en la tasa de crecimiento en aprendizaje de los estudiantes. Es decir, diferentes programas pueden producir efectos distintos.

Lo que proponen es una aproximación alternativa a los análisis clásicos (análisis de regresión múltiple o análisis de covarianza) que utilizan técnicas estáticas basadas en la obtención de un resultado a partir de una función simple con varios predictores (como en el caso de la función de producción educativa).

Estos autores tratan de diferenciar la ganancia que se produce por la maduración natural de la que puede producir un determinado programa educativo. Para ello consideran dos mediciones del logro de los estudiantes, la primera antes del tratamiento (pretest) y la segunda al final (postest). Y, en lugar de establecer una función lineal sobre la ganancia bruta como ocurría en los modelos de Valor Añadido formulados por los economistas, utilizan ambas

mediciones como variables criterio, relacionándolas con la edad. De este modo estiman ese efecto de la maduración natural comparando los valores observados con los valores predichos tanto en el pretest como en el postest. Y consideran el Valor Añadido o el efecto del tratamiento como la diferencia media entre la puntuación observada en el postest, una vez aplicado un tratamiento que puede ser, por ejemplo, asistir a una escuela específica y la puntuación predicha en base a la maduración natural.

El modelo que plantean estos dos autores es un paso previo a los modelos jerárquicos lineales o modelos multinivel más actuales. La importancia de su modelo reside en la consideración del VA como una diferencia entre lo que se espera por maduración natural y lo que realmente se observa, es decir, entre lo esperado y lo observado. El modelo matemático que plantean es el siguiente:

$$Y_i = \beta a_i + \varepsilon_i \qquad \textbf{Ec. III.2}$$

Donde Y es la puntuación de rendimiento del individuo *i*; a_i representa la edad o la medida de tiempo que se utilice como referente, pueden ser meses como el caso del estudio que presentan, años, ocasiones de medida u otra unidad temporal. β es la tasa de crecimiento y se asume que es constante y conocida. Finalmente, ε_i es el componente aleatorio y es independiente de la edad, además puede variar entre ocasiones de medida.

Por tanto, con dos mediciones Y_1 e Y_2 y sin ningún tipo de intervención, el modelo queda formulado de la siguiente forma:

$$\begin{aligned} Y_{1i} &= \beta a_{1i} + \varepsilon_{1i} \\ Y_{2i} &= \beta a_{2i} + \varepsilon_{2i} \end{aligned} \qquad \textbf{Ec. III.3}$$

En la ecuación anterior (Ec. III.3), a_{1i} es la edad del sujeto en el pretest, a_{2i} es la edad en el postest y se puede averiguar el crecimiento debido a la maduración natural. Si se introduce algún tipo de tratamiento entre el pretest y el postest que suponemos incrementará el rendimiento al final, es decir, en la puntuación del postest, entonces:

$$Y_{2i} = \beta a_{2i} + \varepsilon_{2i} + v \qquad \textbf{Ec. III.4}$$

Los autores consideran el término v como el VA del programa (Bryk & Weisberg, 1976) y para estimarlo:

$$Y_{2i} - Y_{1i} - \beta(a_{2i} - a_{1i}) = v + \varepsilon_{2i} - \varepsilon_{1i} \quad \textbf{Ec. III.5}$$

Suponiendo que se cuenta con una muestra de sujetos de alguna población se puede definir entonces V como un estimador de v:

$$V = \overline{Y}_{2i} - \overline{Y}_{1i} - \beta(\bar{a}_{2i} - \bar{a}_{1i}) = v + \bar{\varepsilon}_{2i} - \bar{\varepsilon}_{1i} \quad \textbf{Ec. III.6}$$

En la ecuación Ec. III.6 los términos de la parte derecha son las medias muestrales. Por tanto V es un estimador insesgado de v como puede verse en Ec. III.7:

$$E(V) = E[\overline{Y}_2 - \overline{Y}_1 - \beta(\bar{a}_2 - \bar{a}_1)] = v \quad \textbf{Ec. III.7}$$

Y el crecimiento debido a la maduración natural en el sujeto *i* sería:

$$\Delta_i = \beta(a_{2i} - a_{1i}) \quad \textbf{Ec. III.8}$$

Y para el grupo se obtiene la media:

$$\bar{\Delta} = \beta(\bar{a}_2 - \bar{a}_1) \quad \textbf{Ec. III.9}$$

Además, la media esperada en el postest en base a la maduración natural sería:

$$\overline{Y}_1 + \bar{\Delta} \quad \textbf{Ec. III.10}$$

Y el VA:

$$V = \overline{Y}_2 - (\overline{Y}_1 + \bar{\Delta}) \quad \textbf{Ec. III.11}$$

Este parámetro V puede ser interpretado como la diferencia entre la media actual de rendimiento, el postest, y lo que se espera en base a la maduración natural. En el trabajo, los autores incluyen un modelo más complejo incluyendo otro tipo de covariables relacionadas con el contexto de los estudiantes como el sexo, la raza, los estudios de la madre y la interacción de ésta última con la raza. En su ejemplo elaboran 11 modelos, 10 tipos de tratamiento y un grupo de control de los que extraen las diferentes puntuaciones de VA.

Este trabajo es importante porque introduce una nueva concepción del VA, considerado como la diferencia entre una puntuación observada y una puntuación esperada, en este caso en base a la maduración natural. Llevando a cabo un análisis de regresión de la puntuación inicial del rendimiento sobre la edad se obtiene esa base de comparación, ese crecimiento debido a la maduración natural, es decir, si el sujeto no asiste a ningún tipo de centro. Otro trabajo que utiliza el mismo modelo de VA es el de Bryk y Woods (1980).

Esta definición del VA es la que utilizan algunos de los modelos más recientes, esa diferencia entre lo observado y lo esperado. Aunque en los MVA actuales la base de comparación es distinta, en lugar de considerar la maduración natural se consideran otros aspectos como la diferencia entre la aportación media de todos los centros que forman parte del estudio o una puntuación definida de antemano.

Otro paso más hacia la concepción más reciente del VA y su forma de medirlo se encuentra en el trabajo de Astin (1982). Este autor, aunque no propone un modelo de medida del VA, lleva a cabo una aproximación al término. El autor lo considera una herramienta útil para evaluar la calidad de las escuelas, llevando a cabo una concreción más adecuada a la educación que la simple importación de la definición acuñada para los sectores económicos e industriales.

Si se llevan a cabo evaluaciones educativas para conocer cómo está funcionando una escuela debe tenerse en cuenta el tipo de estudiantes con los que cuenta. Una única medida de sus resultados nos dirá en qué situación se encuentra pero nada acerca de los procesos que el centro educativo realiza para mejorar su aprendizaje. Si la escuela selecciona alumnos con niveles altos de rendimiento, lo más probable es que dichos estudiantes obtengan buenos resultados en un test y, por tanto, no se podrá diferenciar si ese resultado es producto del programa educativo de una escuela o de la propia capacidad del alumno. Este aspecto sugiere que ese test de rendimiento esta proporcionado poca información acerca de las escuelas al menos que se tenga en cuenta el rendimiento potencial de los estudiantes antes de ingresar en el centro (Astin, 1982).

El aspecto que Astin pone de relieve es fundamental para la comprensión del VA. Si se pretende evaluar una escuela a partir del rendimiento de sus

estudiantes lo lógico es poner el foco de atención en el cambio que producen antes y después de pasar por su programa educativo. Si se comparan escuelas en función de un único dato de rendimiento no se cuenta con una información real de sus resultados porque diferentes centros pueden contar con distintos grupos de estudiantes de muy diversa índole y con distintas capacidades.

El argumento básico que subyace a la aproximación del VA es que la eficacia de la escuela (o como Astin la denomina: la excelencia) reside en su habilidad para afectar a los estudiantes de forma favorable en su desarrollo intelectual. Por tanto, el interés son los cambios en el estudiantes desde el comienzo hasta el final de un determinado periodo o programa educativo y la mejor escuela es aquella que genera mayor crecimiento o aprendizaje.

Finalmente, los trabajos que han provocado el gran interés y desarrollo de MVA son los publicados por Sanders y Horn (1994) donde presentan el primer sistema de evaluación basado en medidas de VA, el modelo de Tennessee (TVAAS[36]: *Tennessee Value Added Assessment System*) y el de Sanders y Rivers (1996) donde prueban que los efectos de los profesores que han sido estimados usando las trayectorias de rendimiento de sus estudiantes, pueden predecir sus resultados de logro hasta al menos dos años en el futuro, es decir, que los profesores tienen efectos en el rendimiento de los estudiantes que influyen y permanecen en sus resultados hasta cuando ya no se encuentran bajo su enseñanza.

Otros ejemplos de modelos tempranos de VA que siguieron al desarrollado en Tennessee, fueron el de Chicago (Bryk, Thum, Easton & Luppescu, 1998) o el estudio desarrollado en Inglaterra con datos de educación primaria (Strand, 1998).

En definitiva, como muestra la **Figura III.2**, la concepción del VA en educación es producto de un proceso evolutivo que toma algunas premisas de los análisis de la economía de la educación, de aquí toma prestado el término y la base de los análisis de la función de producción educativa. También está basado en los estudios de eficacia escolar y la rama de investigación preocupada por analizar los efectos que las escuelas tienen sobre los resultados académicos. Y, como marco

[36]Actualmente se denomina EVAAS (*Evaluation Value Added Assessment System*)

global, el avance en las técnicas estadísticas de análisis de datos educativos y de las ciencias sociales que ha cambiado la forma de tratar esa información.

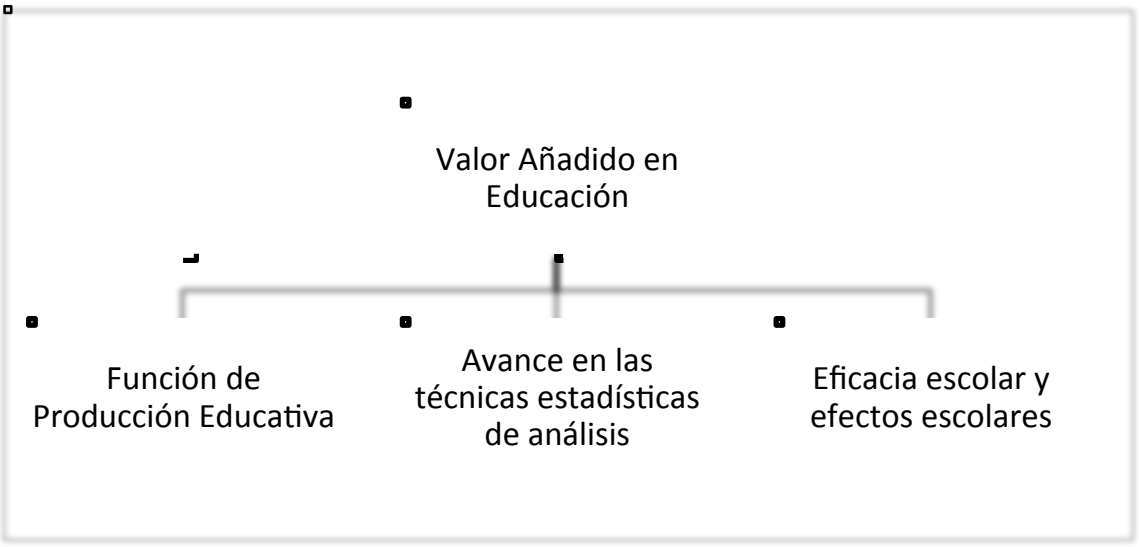


Figura III.2. Origen del Valor Añadido en Educación.

Fuente: Elaboración Propia

III.3.1 Definición de Valor Añadido

Considerando la diversidad de literatura existente en esta temática, es conveniente llevar a cabo una distinción entre diferentes términos que algunos autores utilizan para referirse al Valor Añadido.

El término VA puede analizarse desde dos aproximaciones fundamentales. La primera, más teórica, que trata de definir conceptualmente el Valor Añadido y, la segunda, relacionada con la forma de operativizarlo, de medirlo, directamente vinculada con su análisis y que hace referencia a la técnicas estadísticas utilizadas para calcularlo.

III.3.1.1 Valor Añadido como constructo teórico

El concepto de Valor Añadido en educación, aunque no existe una única definición, es una nueva forma de medir los efectos que las escuelas tienen en el rendimiento de sus estudiantes, entendiendo ese rendimiento como una ganancia o cambio en el aprendizaje. Esta metodología trata de proporcionar una buena estimación de la contribución de la escuela a ese aprendizaje, intentando aislarla de otros posibles factores que se escapan de su control, como el rendimiento previo y factores contextuales del estudiante. El término puede estar ligado a

contextos más económicos pero en educación adquiere unos matices particulares acordes con la realidad educativa que se evalúa.

Algunas definiciones de VA que encontramos en la literatura son las siguientes:

- La OCDE publica un informe en 2008 que recopila las mejores prácticas para evaluar el VA de las escuelas, en este trabajo se define el VA como la contribución de la escuela al progreso de los estudiantes hacia objetivos de la educación prescritos o indicados. La contribución es neta respecto de otros factores que contribuyen al progreso educativo de los estudiantes (OCDE, 2008).
- Otra definición del VA con carácter nacional es la que aportan Martínez-Arias, Gaviria y Castro (2009). Definiéndolo como la "contribución de la escuela al progreso neto de los estudiantes hacia objetivos de aprendizaje establecidos, una vez eliminada la influencia de otros factores ajenos a la escuela que pueden contribuir a dicho progreso" (pág. 17)
- Para Demie (2003), que utilizó los datos de las evaluaciones en Gran Bretaña, el VA en educación es el progreso relativo que los estudiantes hacen en las escuelas de una etapa educativa a otra, comparada con el progreso de otros estudiantes con rendimiento similar al inicio del periodo.
- Para Tekwe et al. (2004) el término VA se emplea para nombrar los diferentes métodos de evaluación del rendimiento de los profesores/escuelas que miden la ganancia en aprendizaje de los estudiantes de un año al siguiente y utilizan esta medida como base para el sistema de evaluación de rendimiento.
- Braun, Chudowsky y Koening (2010) mencionan que, en el contexto educativo, la metodología del VA hace referencia a los esfuerzos para medir los efectos que tienen en el rendimiento de los estudiantes sus actuales profesores, escuelas o programas educativos, teniendo en cuenta las diferencias en rendimiento previo y quizás otras medidas de características que los estudiantes llevan consigo a las escuelas.

Por tanto, el VA es una metodología de evaluación y, al mismo tiempo, es el dato, la información que este tipo de metodología proporciona. El VA está diseñado para la evaluación de diferentes aspectos del sistema educativo, puede centrarse principalmente en la evaluación de escuelas, profesores o algún programa educativo[37] específico como, por ejemplo, una determinada política educativa que reduzca el número de alumnos en las aulas.

Uno de los aspectos característicos de esta metodología es que se centra en el estudio del cambio en aprendizaje, es decir, trata de evaluar la ganancia o el crecimiento en el logro académico de los estudiantes que se produce dentro de un centro educativo. Willett (1989; 1994) lleva a cabo un estudio de los diferentes métodos para medir el cambio en el aprendizaje. Hace una distinción entre dos grandes grupos: por un lado, los métodos tradicionales, conocidos estrictamente como modelos de ganancia, que utilizan dos tomas de datos y son adecuados para medir el cambio intra-individual. Y, por otro, los métodos más actuales, conocidos como modelos de crecimiento, que amplían el número de mediciones (tres o más) y permiten analizar las ganancias entre diferentes grupos de sujetos y el estudio de variables que pueden afectar a ese crecimiento.

La definición de este cambio en aprendizaje puede realizarse de diversas formas, por ejemplo como cambios a lo largo de una escala de puntuaciones obtenidas mediante test de rendimiento, como la proporción de estudiantes que alcanzan o superan un determinado estándar, como la diferencia entre el crecimiento observado y el crecimiento esperado, etc. Por tanto, diferentes tipos de ganancia dan lugar a diferentes formas de medir el VA en educación.

Otro rasgo característico destacable es el concepto de aportación. Esta metodología trata de averiguar cuál es la verdadera aportación de una escuela al aprendizaje. Debe evaluarse el proceso que ocurre dentro de la escuela independientemente de otros elementos que pueden influir en los resultados pero que escapan al control de la institución educativa. Conviene diferenciar ese resultado de lo que los alumnos ya conocen, por eso se toman medidas del rendimiento previo de los estudiantes, y también considerar los efectos de otros

[37]El VA puede aplicarse en la evaluación de distintos aspectos educativos pero, para no repetir continuamente que también se aplica a docentes o programas educativos, se hará referencia únicamente a las escuelas.

elementos del contexto, tanto del estudiante como del centro, que pueden estar relacionados con el rendimiento.

Más de 40 años de investigación educativa sobre eficacia escolar, iniciados por la reacción ante el informe Coleman, se han encargado de demostrar que las simples puntuaciones medias de los resultados de los estudiantes de una escuela se encuentran parcialmente afectadas por otros aspectos del contexto familiar y las experiencias educativas previas, además de los efectos producidos por los procesos que se dan en la escuela en el momento concreto de la evaluación. Por tanto, si pretendemos evaluar los efectos de una escuela concreta deben considerarse estos aspectos porque los resultados brutos de rendimiento de un estudiante no solo están determinados por la acción escolar.

Si una única media bruta de rendimiento está afectada por otros factores ajenos a la escuela y, por tanto, no conviene utilizarla para llevar a cabo comparaciones entre escuelas. Por la misma razón, tampoco podemos averiguar si una escuela está aportando más al aprendizaje que otra comparando sus ganancias brutas entre las puntuaciones de sus estudiantes en dos cursos académicos distintos. Esa ganancia bruta también está afectada por otros factores que se escapan al control de la escuela.

Para averiguar esa aportación de las escuelas a la ganancia en aprendizaje se necesita una medida de comparación. Es decir, debe establecerse cuál va a ser el criterio o estándar que determine si una escuela está produciendo VA en sus estudiantes. Lo más usual es estimar el VA como la diferencia entre la ganancia observada de los estudiantes y la esperada, una vez que se han tenido en cuenta las diferencias iniciales entre estudiantes que pueden determinar sus resultados (Braun, Chudowsky & Koenig, 2010). Esa ganancia esperada puede ser la media de todas las escuelas de un distrito o la media de aquellas con las mismas características que la escuela que está siendo evaluada. La Figura III.3 representa el concepto de VA.

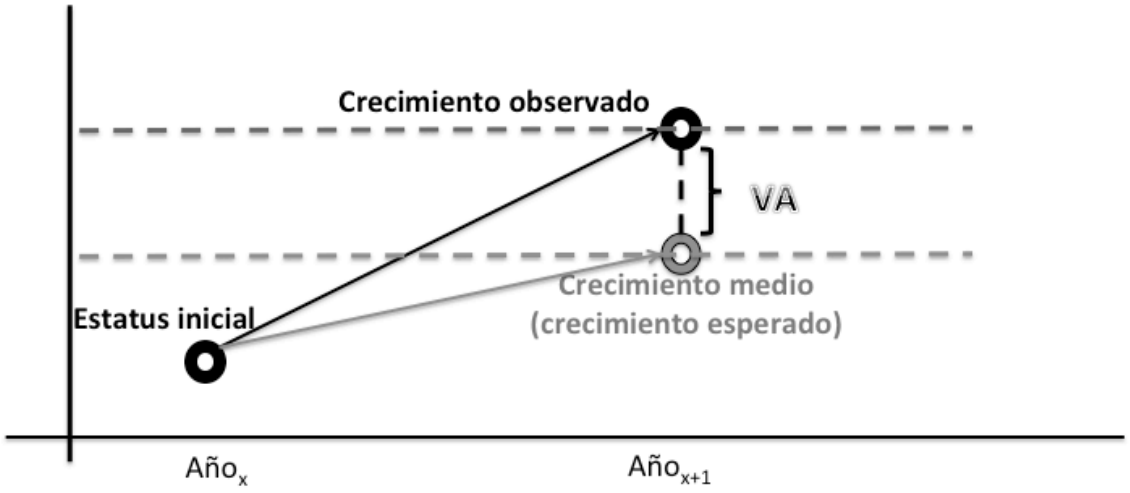


Figura III.3. Definición gráfica de Valor Añadido.
Fuente: Elaboración Propia a partir de Goldschmidt et al. (2005)

Se asume, por consiguiente, que con el aprendizaje escolar se produce un cambio en el estudiante y, por este motivo, el rendimiento debe medirse en forma de cambio e intentar superar las medidas transversales que no están representando realmente los procesos que ocurren dentro de la escuela. Este supuesto conlleva que el cambio en aprendizaje puede ser medido, y una de las formas de hacerlo es utilizar escalas de rendimiento construidas a partir de las respuestas de los estudiantes en diferentes test que aumentan su nivel de dificultad a lo largo de los diferentes cursos evaluados (Patz, 2007; Ballou, 2008; Briggs, Weeks & Wiley, 2008; McCaffrey, Lockwood, Doretz & Hamilton, 2003).

McCaffrey, Lockwood, Doretz y Hamilton (2003) señalan dos factores por los que las técnicas de análisis del VA tienen un interés creciente:

- Mantienen la promesa de separar los efectos de los profesores y las escuelas de los poderos efectos de factores no educativos como el contexto familiar.
- Pretenden mostrar diferencias en la efectividad entre profesores o escuelas. Si estas diferencias pueden estar justificadas y causalmente relacionadas con características específicas, el potencial para la mejora de la educación puede ser grande.

Al considerar que, una vez aislados factores contextuales y de rendimiento previo, el aprendizaje de los estudiantes es producto de lo que ocurre en la escuela y en el aula, se intenta dotar de cierta causalidad a los resultados de VA, responsabilizando a los centros de esos resultados (*accountability*). Sin embargo, la falta de asignación aleatoria de los estudiantes a las escuelas, al contrario de lo que ocurre en los estudios experimentales que establecen relaciones de causa-efecto, no puede paliarse del todo con controles estadísticos como los que utiliza el VA (McCaffrey, Lockwood, Doretz & Hamilton, 2003; Rubin, Stuart & Zanutto, 2004; Reardon & Raudenbush, 2008).

III.3.1.2 Modelos de Valor Añadido: la herramienta estadística

Los análisis del VA, a través de diferentes modelos, utilizan los resultados académicos de los estudiantes obtenidos mediante test para conseguir su objetivo. Existen otro tipo de modelos que también empelan este tipo de información y que conviene diferenciar de los MVA:

- **Modelos de Estatus**: conocidos también como modelos transversales, ofrecen una información sobre el rendimiento en un punto específico del tiempo, una fotografía de la situación. Para evaluar una escuela su información de rendimiento se compara con un objetivo establecido, por ejemplo, la media de un grupo específico de escuelas o un estándar prefijado. Se pueden extraer resultados sobre qué proporción de estudiantes se encuentran en unos niveles específicos de rendimiento o si rinden por encima o por debajo de los objetivos definidos para la comparación.
- **Modelos de Estatus Contextualizados**: utilizan una única toma de datos de rendimiento pero incluyen en sus modelos, mediante técnicas estadísticas de ajuste, variables contextuales que pueden afectar a ese logro académico. De esta manera se pretende tener en cuenta las diferencias existentes en los grupos de estudiantes que están matriculados en escuelas distintas. Estos modelos pueden ser considerados un primer paso hacia los MVA pero sin utilizar medidas del rendimiento previos de los estudiantes

- **Modelos de cambio cohorte a cohorte**: este grupo utiliza dos mediciones del logro académico en dos puntos temporales pero no de los mismos estudiantes. La aplicación de este modelo en la evaluación escolar permite conocer si ha habido cambios en la proporción de estudiantes que se situaban en los diferentes niveles de rendimiento entre ocasiones de medida.
- **Modelos de ganancia**: llevan a cabo un seguimiento del rendimiento de los mismos estudiantes en dos momentos temporales, por ejemplo, durante dos cursos consecutivos para analizar el cambio en aprendizaje. Tratan de analizar el progreso académico calculando puntuaciones de ganancia en el rendimiento. Centrando la atención en esta ganancia se pretende estudiar de forma más adecuada el proceso que ocurre dentro de las escuelas, ese proceso de aprendizaje. Los sistemas de evaluación que utilizan estos modelos quieren averiguar cuanta ganancia adquieren sus estudiantes, normalmente, de un curso al siguiente. No es usual que utilicen factores de contexto y, por tanto, no buscan saber que variables son responsable de ese crecimiento.
- **Modelos de Crecimiento o longitudinales**: también analizan el cambio en aprendizaje pero utilizan más de dos puntuaciones del rendimiento para estimar una o más pendientes de crecimiento en función del tiempo y un estatus o punto inicial de partida con la finalidad de analizar las trayectorias individuales de cambio. Las técnicas estadísticas para desarrollar estos modelos son variadas. La más común son los modelos jerárquicos lineales para estudiar la trayectoria de cambio (Singer & Willett, 2003), aunque también existen otras posibilidades como los modelos lineales mixtos o las ecuaciones estructurales (Rovine & Molenaar, 2002).

Los MVA se diferencian del resto porque son análisis estadísticos que tratan de conocer cuál es la contribución específica de una escuela, un docente o un programa educativo a la ganancia o crecimiento en aprendizaje, independientemente de otros factores ajenos a la escuela. Para aislar esa contribución se toman, al menos, dos mediciones del logro académico de los

estudiantes utilizando test y, algunos modelos, recopilan información de otras variables de los estudiantes o las escuelas (nivel socioeconómico, proporción de alumnos inmigrantes, nivel educativo de los padres, etc.). Principalmente tratan de responder a la pregunta ¿qué cantidad de cambio en el rendimiento de los estudiantes puede atribuirse a la escuela a la que se asisten? O de forma equivalente ¿Cuál es la contribución de una escuela concreta a ese cambio comparado con la contribución media de todas las escuelas? (Braun, Chudowsky & Koenig, 2010)

El debate sobre si deben emplearse medidas de ganancia o crecimiento para conseguir estimaciones más fiables de ese cambio en el aprendizaje es un debate con larga tradición (Willett, 1989a; Willett, 1994; Rogosa, 1995). Uno de los argumentos más contundentes en contra de las medidas de ganancia es que si se introducen predictores del cambio, los modelos con dos únicas tomas de datos pierden fuerza. No obstante, en países como Reino Unido o Polonia (Ray, 2006; Jakubowski, 2008) se desarrollan MVA que utilizan una de las mediciones del rendimiento como principal covariable en el análisis.

Como se ha comentado unos párrafos arriba, los modelos de VA estiman la diferencia entre una ganancia observada y una esperada. Esta diferencia se estima como un resultado numérico residual asociado con la intervención educativa específica (Ray, 2006). Por tanto, las estimaciones de VA son, principalmente, los residuos de regresión producidos por los diferentes análisis empleados para su obtención, son la parte de varianza que queda sin explicar. Estos residuos pueden obtenerse de varias formas, el tipo de puntuaciones y los análisis estadísticos desarrollados a través de un modelo dan lugar a esas estimaciones.

En el trabajo de Raudenbush y Willms (1995) denominan a estos residuos cómo efectos y, tomando como referencia sus definiciones, los análisis del VA o MVA intentan separar las contribuciones relativas de los distintos efectos para estimar, en la medida de lo posible, los efectos de tipo B atribuibles a la escuela. Para ello utilizan procedimientos estadísticos que intentan paliar la ausencia de aleatorización en la composición de las escuelas. Actuando de este modo, los modelos de rendición de cuentas y evaluación escolar que utilizan MVA

intervienen solo sobre la varianza que se encuentra bajo su control (Martínez-Arias, Gaviria & Castro, 2009), esa varianza residual.

Las MVA tratan de evaluar el efecto que tiene asistir a una escuela determinada en el aprendizaje de sus estudiantes, independientemente de otros factores asociados con el contexto que la escuela no puede controlar. Willms (2008) considera que el aprendizaje, de forma simple, puede definirse como una función de factores manejables por la escuela como, por ejemplo la calidad de la enseñanza, el clima del aula y la escuela y otros que no dependen del control escolar como el nivel socioeconómico de las familias o la propia habilidad de los estudiantes. El VA está interesado en los factores que influyen en el aprendizaje y sí pueden ser controlados por las escuelas pero su estimación se realiza a partir de los factores que no están bajo el control, es decir, esa concepción de residuo una vez que se controlan los factores de contexto:

> *VA = influencia de los factores bajo el control escolar = aprendizaje | factores familiares no controlables*

Willms destaca tres problemas con esta definición del VA:

a. La dificultad de especificar todas las variables de contexto que influyen en el aprendizaje y que escapan al control de las escuelas
b. No todos los factores que influyen en el aprendizaje de los estudiantes pueden clasificarse como controlables o no, pueden existir grados como, por ejemplo, la implicación de los padres en el aprendizaje que, inicialmente, puede ser considerado como un factor ajeno al control escolar, sin embargo, pueden existir programas escolares que traten de involucrar a los padres en el aprendizaje de sus hijos.
c. Las escuelas quieren y necesitan información de esos factores que sí están bajo su control y que influyen en el aprendizaje de los estudiantes. Sin ella, el VA únicamente informa de sí las escuelas están haciéndolo bien o no pero no ofrecen ninguna orientación de cómo mejorar sus situación.

Desde esta aproximación más técnica, los MVA son considerados una herramienta de análisis estadístico que proporciona medidas cuantitativas del

rendimiento de las escuelas. Los resultados pueden ser usados para evaluar aspectos del sistema educativo y de las propias escuelas. En este sentido, la implementación de un sistema de VA debe ser visto como un medio para conseguir un fin más que como un fin en sí mismo (OCDE, 2008).

En ocasiones, los MVA, dependiendo de factores como el número de predictores de contexto utilizados, de cómo se consideran esos efectos escolares, de qué tipo de medida de cambio se utiliza o el número de mediciones de rendimiento, necesitan de un proceso de análisis estadísticos altamente complejos para llevar a cabo la estimación final del VA de una escuela. Para aquellos que no son expertos en estadística este proceso es considerado un "caja negra" (OCDE, 2008, pág. 76). Elaborar una medida del VA implica la utilización de procesos de cálculo, normalmente estadístico, que pueden variar en su dificultad y entendimiento. La evolución en la aplicación de análisis estadísticos sobre datos educativos ha producido una mayor rigurosidad de los resultados y, a su vez, un aumento de su complejidad.

No obstante, si se utiliza este tipo de análisis como herramienta de evaluación del rendimiento de las escuelas, es inevitable el tratamiento estadístico de los datos pero realizar procesos demasiado complejos puede hacer menos comprensibles los resultados para las audiencias que no poseen conocimientos específicos sobre análisis estadísticos. Existe gran diversidad de publicaciones, incluso manuales que tratan diferentes aspectos metodológicos[38] de los procesos estadísticos de análisis del VA (McCaffrey, Lockwood, Koretz & Hamilton, 2003; McCaffrey, Koretz, Louis & Hamilton, 2004; Ballou, Sanders & Wright, 2004; Sean & Monczunski, 2007; Lockwood et al., 2007; Haegeland & Kirkeboen, 2008; Ferrão & Goldstein, 2009).

También hay una gran variedad de MVA[39] y en la literatura existen revisiones de esos trabajos (Sanders, 2006; Choi, Goldschmidt & Yamashiro, 2006; Sanders, 2006; Wright, Sanders & Rivers, 2006; Jakubowski, 2008). Estos modelos puede considerarse como la expresión matemática de lo que se pretende conseguir

[38]Las cuestiones metodológicas más importantes en la medida del VA se detallan en el capítulo IV

[39]Se dedica el Capítulo V de este trabajo a la descripción de diferentes modelos de Valor Añadido.

con la metodología, es decir, modelos estadísticos que proporcionan medidas cuantitativas del efecto de las escuelas sobre el crecimiento en aprendizaje y que sirven para llevar a cabo inferencias sobre su eficacia. En otras palabras, los MVA intentan dar respuesta a la pregunta ¿qué cantidad añade la escuela o el profesor al aprendizaje de los estudiantes? (Lissitz, Doran, Schafer & Willhoft, 2006).

Por tanto, los MVA son un tipo de modelos estadísticos que estiman la aportación de una escuela al progreso de los estudiantes, esta aportación es neta, libre de otros aspectos ajenos al control escolar. Además, esa contribución al progreso se analiza con respecto a un objetivo determinado previamente, normalmente estándares de rendimiento o diferencias respecto a la media global. La característica distintiva de los modelos de VA es la utilización de al menos dos puntuaciones de rendimiento de los estudiantes para poder estimar el cambio o el crecimiento en aprendizaje.

En este trabajo se considera el VA como la unión de la perspectiva teórica y técnica. Por un lado, es un constructo teórico que surge como consecuencia y necesidad derivada de la amplia difusión e interés alcanzado por las evaluaciones de rendimiento y el estudio de los efectos escolares. Un concepto ligado con la obtención de una medida precisa y justa del logro de las escuelas o docentes (o cualquier otra unidad posible de análisis como el distrito, la provincia, el país, etc.). Una puntuación que se encuentre libre de los efectos producidos por factores ajenos al control escolar y que sea capaz de distinguir entre los posibles efectos que pueden tener sobre los resultados de los estudiantes la escuela a la que asisten. Y, por otro lado, el conjunto de técnicas estadísticas utilizadas para la obtención de esta puntuación también se consideran VA, es el modelo. Es decir, no solo el resultado final es VA también el conjunto de técnicas empleadas para su estimación.

En resumen, el VA no se puede entender sin las técnicas de análisis que se emplean para conseguirlo. Además, reconoce que los estudiantes tienen diferentes niveles de capacidad y que provienen de diferentes contextos, y que estos factores afectarán a su progreso educativo y, por tanto, deben ser tenidos en cuenta. Analizan la ganancia, el cambio a lo largo de un curso, etapa o periodo escolar cuando, normalmente los estudios de rendimiento escolar se asocian a datos

transversales de logro académico. En consecuencia, los modelos de VA se ajustan más a la realidad intrínseca de la educación, es decir, hacer que sus estudiantes avancen y crezcan en aprendizaje. Se caracterizan por utilizar las puntuaciones de los estudiantes en test para estimar medidas de ganancia y los efectos que la escuela está provocando en ellas.

III.4 Finalidad del análisis del Valor Añadido

El objetivo principal del VA es atribuir cambios en el rendimiento de los estudiantes a los agentes responsables de esos cambios, principalmente, las escuelas. El resultado del VA es una estimación del efecto del centro educativo, una medida de su eficacia.

Si se pretende conocer cómo está afectando una determinada escuela al aprendizaje se debe estimar, en primer lugar, ese cambio o crecimiento en aprendizaje. En su forma más básica se calcula como un simple cambio en las puntuaciones de los test de los alumnos de un año a otro y modelos más complejos incorporan técnicas estadísticas que utilizan información longitudinal, no una simple relación pretest-postest, y pueden tener en cuenta diferentes factores del contexto del estudiante e incluso los efectos previos producidos por otras escuelas.

Para McCaffrey, Lockwood, Doretz y Hamilton (2003) el VA trata de dar respuesta a dos cuestiones generales:

- ¿Tienen las escuelas o los docentes un efecto diferencial en los resultados de sus estudiantes?
- ¿Cómo de efectiva es una escuela o profesor particular produciendo ganancia en el rendimiento? Y ¿qué escuelas o profesores son más o menos efectivos?

La primera cuestión plantea si los centros educativos están aportando o contribuyendo al logro académico de sus alumnos. Es decir, que un alumno asista a una escuela u otra producirá un efecto distinto en los resultados que obtenga durante el tiempo que se encuentre matriculado. Es posible responder a esta cuestión utilizando MVA que estimen la varianza del rendimiento que depende de

las escuelas. Si se utiliza un modelo adecuado es posible, además de estimar el crecimiento o la ganancia en aprendizaje, estimar su varianza y determinar qué proporción se puede atribuir a las escuelas.

La segunda pregunta requiere la estimación de los efectos individuales de cada escuela. Averiguar cuánto está influyendo o aportando un centro concreto al crecimiento de sus estudiantes. De esta manera es posible determinar qué escuelas son más eficaces y cuáles menos.

Este último aspecto puede ser problemático. Si los resultados de VA se utilizan en sistemas de evaluación basados en rendición de cuenta y penalizan a las escuelas en función de esta información (sistemas de alto impacto) se asume una relación causa-efecto[40] de los resultados. No obstante cualquier esfuerzo de aislar efectos causales a partir de datos que no han sido asignados de forma aleatoria es poco factible, por tanto, existen razones para cuestionar la capacidad de los métodos de VA para lograr este objetivo (Rubin, Stuart & Zanutto, 2004)

III.4.1 Utilidad de las estimaciones de Valor Añadido

La información que proporciona este tipo de metodología de evaluación puede ser utilizada por políticos, inspectores de educación, directores de los centros, profesores y padres con diferentes propósitos.

La utilización de modelos de VA para determinar el logro de las escuelas está justificada, principalmente, porque supone un gran avance y mejora respecto a otras técnicas existentes que persiguen el mismo objetivo, como la simple estimación de medias anuales de rendimiento de las escuela. No obstante, el uso que se haga de la información obtenida mediante esta potente herramienta puede convertirla en más o menos valiosa para la toma de decisiones en educación. Sobre todo debe tomarse mucha precaución si se utiliza como base o único aspecto que determine un sistema de sanciones o incentivos.

Los MVA proporcionan indicadores cuantitativos del rendimiento escolar que pueden facilitar la identificación de áreas para mejorar dentro de las escuelas

[40]Uno de los puntos críticos de las medidas del VA es la consideración de los resultados como un efecto causal de la situación escolar (Rubin, Stuart & Zanutto, 2004; Reardon & Raudenbush, 2008; Rothstein, 2009). En el capítulo IV (apartado IV.5) se amplía la información sobre la consideración causal o descriptiva de los MVA.

y los sistemas escolares y, además, permite la construcción de estándares de rendimiento hacia los que debe dirigirse un determinado centro educativo (OCDE, 2008). Por tanto, la mejora escolar puede ser uno de los usos principales del VA. También destaca su utilidad en los sistemas de evaluación de rendición de cuentas, aportando una información cuantitativa fiable que puede ser la base para la toma de decisiones en este tipo de evaluaciones. Finalmente, la utilización de este tipo de medidas en aquellos sistemas educativos que posibilitan la libre elección escolar por parte de los padres puede proporcionar una información precisa de la situación de las diferentes escuelas evaluadas que ayude a las familias en esta decisión.

En su informe, la OCDE (2008) también menciona que un sistema de evaluación que utilice un MVA como herramienta de análisis y cuyos resultados sirvan para disponer de una información detallada de cada centro educativo evaluado, con el objetivo de detectar posibles situaciones problemáticas o de éxito educativo se encontrará en el camino correcto. Aunque para tomar decisiones finales respecto a los centros evaluados debe tenerse en cuenta la situación real de cada escuela y saber interpretar la información acorde con dicha realidad.

Se recomienda que la información obtenida con los MVA se complemente con otras herramientas educativas encargadas de la revisión y el control de la vida del centro, como los inspectores de educación y los propios directores de las escuelas. La combinación de estos dos factores puede aumentar el poder del sistema de evaluación y conseguir una información más útil de la realidad educativa.

En función de si los resultados de VA tienen esa connotación causal que demandan los sistemas de rendición de cuentas de alto impacto o se interpretan como medidas descriptivas, pueden utilizarse con fines distintos. Los usos de la información obtenida se orienta, fundamentalmente, hacia cuatro ámbitos de actuación diferenciados:

- Una primera orientación es el diagnóstico de la situación de las escuelas, de la misma forma que otros sistemas de evaluación. Conocer la situación de los centros educativos a partir de los resultados que obtengan sus alumnos y poder establecer, en caso de

ser necesario, un plan de mejora adecuado. El VA puede proporcionar una información detallada y precisa de lo qué está ocurriendo en el centro en cuánto a rendimiento se refiere, sin interferencias de otros factores que pueden confundir los resultados. Además permite conocer el desarrollo de determinado grupos de alumnos y establecer proyecciones de su progreso en el tiempo. Por tanto, diagnóstico y mejora son dos funciones importantes en este proceso. El sistema de evaluación español sigue estos principios y son conocidos como sistemas de evaluación basados en la rendición de cuentas con bajo impacto (*low-stakes*).

- En segundo lugar, las evaluaciones llevadas a cabo sobre los estudiantes pueden formar parte de un sistema de rendición de cuentas, donde las escuelas son premiadas o penalizadas en función de sus resultados de rendimiento obtenidos. Los premios y castigos pueden ser de muy diversa índole, desde cambios en la financiación hasta cualquier tipo de dotación material. El VA puede ser la herramienta que proporcione una información más justa de los resultados escolares que sirva como base para la toma de decisiones respecto a los agentes evaluados en estos sistemas de rendición de cuentas con alto impacto (*high-stakes*)
- Una tercera línea de acción es la informativa. Los resultados pueden ser utilizados por las familias para determinar a qué centros deben acudir sus hijos, en el caso de que sea posible la libre elección de escuela. En Inglaterra, desde la utilización de las tablas de la liga, se ha estado implementando un sistema de clasificación de escuelas en función de los resultados. Este información se encuentra a disposición de los padres y puede influir e incluso dirigir su decisión sobre la elección de centro educativo para sus hijos.
- Finalmente, la investigación educativa es otro de los campos donde se utiliza la metodología del VA. Puede utilizarse, por ejemplo, para evaluar el efecto que tienen determinados programas educativos piloto aplicados en algunas escuelas ya que en el campo educativo es

muy difícil llevar a cabo una investigación completamente experimental.

El uso de los resultados obtenidos mediante el VA dependerá del tipo de política educativa implícita en el sistema educativo que se pretende evaluar y de los objetivos que persigue ese sistema con la evaluación. La información que proporcionan es amplia y permite dar respuesta a los diversos propósitos que pueden tener los sistemas de evaluación en educación.

No obstante, esta información es un dato derivado de los distintos análisis estadísticos que se llevan a cabo sobre puntuaciones en test aplicados a los estudiantes y puede estar sujeto a diferentes problemas metodológicos (Baker et al., 2010). Por tanto, deben tratarse los resultados con cautela y no utilizarla como única herramienta para la toma de decisiones sobre las escuelas, sobre todo en los sistemas de rendición de cuentas de alto impacto, ya que se estaría asumiendo una relación de causa-efecto entre los resultados de VA y lo que ocurre en los centros.

A continuación se expone con más detalle la cuatro formas de utilización de los resultados de VA que se han mencionado.

III.4.1.1 Para la mejora y el desarrollo de la escuela

Conocer qué ocurre en cada centro, cómo están evolucionando determinadas cohortes de alumnos e identificar aquellas que muestran algún tipo de problema, saber dónde se encuentran grupos de alumnos con alto o bajo rendimiento académico, realizar comparaciones de centros con características similares, elaborar estándares de crecimiento e identificar factores educativos que afectan a los resultados de logro son objetivos factibles que pueden lograrse con la utilización de modelos de VA como herramienta de evaluación educativa. Es posible realizar diagnósticos concretos de la situación educativa que guíen los procesos mejora escolar.

Cualquier tipo de actividad, sea educativa o no, que pretende ser mejorada necesita saber cómo está funcionando en la actualidad, es decir, requiere una evaluación de su situación actual y, a su vez, necesita una medida precisa de su rendimiento, necesita conocer su eficacia. La información obtenida mediante la

utilización de MVA puede emplearse en la identificación de aquellos sectores de los sistemas educativos (profesores, escuelas, programas, etc.) que están logrando o incidiendo en mayor medida en el aprendizaje de sus estudiantes y, además diagnosticar cuáles son los aspectos que necesitan mejorar. Los MVA crean medidas precisas del rendimiento de las escuelas que fortalecen la toma de decisiones educativa, pudiendo llevar a cabo actuaciones sobre una base más sólida e informada que permita la mejora de la situación escolar.

También es posible evaluar los efectos de iniciativas o programas específicos de mejora (cambios de ratio, nuevos recursos como la pizarra electrónica, etc.). En este sentido, se está hablando de detectar aspectos que influyan de una determinada manera en el rendimiento del alumnado, pero esta situación no es susceptible de ser evaluada con una única toma de datos de logro académico ya que la implantación de un determinado programa o recurso no tiene efectos inmediatos. Por tanto, la información que proporciona el VA tiene una mayor utilidad porque hace referencia a la evolución de la escuela y no solo indica la situación actual de los centros educativos. Así, con este tipo de resultados, se podrá apoyar y legitimar una determinada actuación educativa.

La información que proviene de los MVA puede ser utilizada para varios propósitos de mejora de la escuela pero únicamente si es utilizada por los actores que pueden influir en el proceso y/o los resultados (OCDE, 2008). Si los inspectores de educación y, sobre todo, profesores y directores llegan a ser capaces de comprender y utilizar la información que proporcionan estos modelos tendrán una ayuda extra para organizar los recursos escolares.

Una de las utilidades del VA, desde la perspectiva de mejora, es su capacidad para realizar perfiles detallados del progreso y la evolución de los estudiantes. Esta información permite conocer qué sectores del alumnado están evolucionando de forma positiva e identificar a aquellos que no se encuentran en esa situación (Thum, 2002; 2003; Doran & Izumi, 2004). De esta forma es posible administrar recursos y programas específicos de mejora hacia esos sectores.

Elaborar perfiles de las escuelas que incluyan, además de los datos de VA, la relación con otras variables de contexto y factores vinculados a los procesos que ocurran en las escuelas puede ayudar a que sean los propios centros educativos los

que utilicen esta información para su auto-evaluación y mejora (Demie, 2003). Este aspecto implica que los responsables en los centros sean capaces de interpretar la información de VA, por lo que el modo de presentar los datos adquiere especial relevancia.

Combinando las trayectorias de crecimiento en el rendimiento de los estudiantes con las estimaciones de VA de las escuelas es posible realizar proyecciones del logro académico futuro de un estudiante (OCDE, 2008). Con esta información las escuelas y sus administradores pueden detectar, en función de unos estándares especificados previamente o comparando los resultados de crecimientos con los de otros grupos de similares características, que proporción de ellos logrará alcanzar esos objetivos en un periodo de tiempo determinado.

El análisis del VA incluyendo las relaciones entre diferentes elementos educativos de entrada y los resultados de rendimiento puede informar sobre qué estrategias o recursos educativos están funcionando y cuáles no y, de esta manera, guiar las decisiones políticas y la distribución de los recursos educativos.

Otro de los usos de la información de VA es la elaboración de proyecciones sobre la evolución del rendimiento de las escuelas que ayuden a la planificación, distribución de recursos y toma de decisiones educativas. Dichas proyecciones pueden utilizarse para identificar posibles resultados futuros, señalando si las trayectorias de rendimiento obtenidas se prolongarán en el tiempo y, también si se pueden alcanzar determinados objetivos de rendimiento fijados de antemano. Con este tipo de datos es posible actuar sobre los problemas con antelación.

Además, un sistema educativo puede desarrollarse a través de sus resultados de valor añadido si, por ejemplo, la información es útil para identificar escuelas que necesitan una evaluación más específica y detectar centros con bajo rendimiento o grupos de estudiantes considerados en riesgo (Young, 1999). Y los resultados aumentarían su utilidad con un incremento de los mecanismos de intercambio de información, de modo que estas escuelas puedan disponer de las mejores prácticas que se están llevando a cabo en centros con alto VA.

En resumen, la información que procede de modelos de VA puede utilizarse para variedad de propósitos en la mejora de la escuela pero es muy importante que los principales responsables de la enseñanza en las escuelas dispongan de esta

información y, además sepan interpretarla. El director y su equipo, junto con los docentes son los agentes educativos que tienen una vinculación directa en el proceso de enseñanza y, por tanto, son ellos los que tienen la capacidad de dotar de utilidad a las estimaciones obtenidas con los MVA. Por tanto, si los resultados se utilizan con propósitos informativos y de mejora estos agentes pueden considerar este tipo de evaluaciones como un elemento positivo y de ayuda. En cambio, si la información se utiliza en sistemas con fuertes medidas de rendición de cuentas con sanciones severas para las escuelas, puede distorsionarse ya que se corre el riesgo de que las escuelas intenten evitar los malos resultados en las pruebas que sirven para construir estas medidas.

III.4.1.2 Para la rendición de cuentas en educación (Accountability)

Los sistemas que responsabilizan a las escuelas de sus resultados, es decir, los sistemas de rendición de cuentas pueden beneficiarse de la utilización de los MVA. Las administraciones que implementan este tipo de sistemas pretenden conseguir información sobre el funcionamiento de sus escuelas y docentes en términos, principalmente, de eficacia y eficiencia[41]. La escuela es tratada como un servicio que debe rendir cuentas de los recursos públicos de los que dispone. Debido a que el dinero empleado proviene de las arcas públicas y son los contribuyentes los que corren con los gastos tienen el derecho a estar informados sobre su utilización.

La cuestión clave en este tipo de sistemas es su base de comparación. Normalmente los centros educativos se comparan con un estándar[42] establecido o con los resultados de otras escuelas que forman parte del conjunto de una misma administración. Se penaliza a los centros porque no consiguen alcanzar ese estándar o porque están rindiendo por debajo de lo esperado en comparación con

[41]La eficacia está relacionada con el logro de los objetivos que persigue una determinada escuela (ver apartado II.2), mientras que la eficiencia focaliza su atención en los recursos y su utilización (más información sobre la medida de la eficiencia en López, E. (2012)).

[42]Los acercamientos a la utilización de estándares van desde la definición de objetivos educativos generales (competencias mínimas en España) hasta la formulación de expectativas precisas de rendimiento en determinadas áreas de conocimiento. Algunos países han establecido estándares educativos como puntos de referencia e introducido marcas de logro que los estudiantes en una edad concreta deben alcanzar, como en la ley estadounidense NCLB. La aplicación de modelos de VA requiere que el rendimiento de las escuelas se mida en comparación con otras o con una estándar predeterminado.

escuelas de similares características. Este aspecto genera la necesidad de contar con procesos de evaluación de los logros educativos de las escuelas de forma precisa y justa. Las puntuaciones de VA pueden ser las medidas precisas y justas del rendimiento de escuelas que se utilicen en este tipo de sistemas, ya que informan de hasta qué punto las escuelas han mejorado el rendimiento de sus estudiantes.

Los análisis del VA han adquirido gran relevancia en Estados Unidos para ayudar con la tarea que encomienda la ley NCLB. Ser una herramienta que analiza la eficacia de las escuelas y poseer gran versatilidad para adaptarse a diversas situaciones de evaluación son las características clave que han hecho del VA uno de las técnicas más utilizadas en el país. Además, ha habido un aumento de las opiniones en contra de la utilización de datos transversales de rendimiento, y no analizar el cambio, en los sistemas de rendición de cuentas de alto impacto (Raudenbush, 2004; Choi, Goldschmidt & Yamashiro, 2006; Thum, 2009)

Este tipo de sistemas utiliza las estimaciones de VA como herramienta para la detección de esa eficacia en los docentes o centros (Dossett & Muñoz, 2003) y, por tanto, para la toma de decisiones respecto a la asignación de los incentivos. Las medidas de VA están destinadas a ayudar en la identificación de qué escuelas y profesores son más o menos efectivos, así como las áreas en las cuáles tienen una efectividad diferente.

Los sistemas de rendición de cuentas utilizan los incentivos con la finalidad de mejorar los resultados académicos de sus estudiantes. La utilización del VA mejora las decisiones basadas en puntuaciones brutas de los test que podrían estar distorsionadas por otro tipo de factores ajenos a la escuela. A pesar de esto, el uso de determinados incentivos puede afectar a la situación de las escuelas y el aprendizaje que se produce en ellas pero sin conseguir su mejora. Por ejemplo, si se premia el aumento en los niveles de logro, los centros podrían preparar específicamente a sus alumnos para la superación de ciertos tipos de pruebas de rendimiento, obviando muchos contenidos educativos. Esto se debe a que se crea una motivación especial hacia el aumento específico del rendimiento en esas evaluaciones olvidando el resto de las áreas escolares (OCDE, 2008; Baker et al., 2010).

Otra mejora que se consigue con el VA es un mayor ajuste a la realidad educativa. Al no centrar el foco de la evaluación en lo que un alumno sabe sino en lo que cambia dentro de la escuela, se pretende que los centros no seleccionen alumnos con alto rendimiento y se olviden de otros alumnos menos capaces. De esta forma, los incentivos se basan en medidas precisas del progreso en el aprendizaje del alumnado, evitando que se premie en mayor medida a la escuelas que cuentan solo con estudiantes de alta capacidad.

Con las estimaciones de VA y utilizando la información individual de cada uno de los centros educativos evaluados, se podrían premiar situaciones concretas de las escuelas como, por ejemplo, aumentar los niveles de rendimiento de los estudiantes inmigrantes en comprensión lectora o mejorar el rendimiento en matemáticas de los grupos con bajo nivel de logro.

III.4.1.3 Para la elección escolar

Una tercera vía de utilización de los resultados obtenidos mediante evaluaciones educativas que emplean la metodología del VA, es proporcionar información a las familias sobre los centros educativos para ayudarles en su proceso de decisión sobre la elección de la escuela a la que asistirán sus hijos. Conviene mencionar que en muchos países los padres no tienen la posibilidad de elegir la escuela en la que sus hijos van a estudiar, simplemente son dirigidos directamente a la escuela local sin tener en cuenta los deseos de la familia. Por otro lado, los altos requisitos de determinadas escuelas en ocasiones dificultan la libertad total de elección escolar.

La elección de centros educativos por parte de los padres es posible y se basa en numerosas razones: proximidad geográfica, programas ofrecidos en la escuela, los compañeros con los que su hijo va a integrarse u orientación religiosa. Estos son algunos ejemplos de factores sobre los cuales las familias pueden basar sus decisiones relativas a la elección escolar. Las puntuaciones de VA de las escuelas puede llegar a ser otro factor importante por el que las familias y estudiantes elijan la escuela a la que desean asistir (OCDE, 2006). Por ejemplo, en Inglaterra, en 1992, antes de introducir su sistema de análisis del VA, ya se empleaban las denominadas "*Performance Tables*", que actualmente tienen el

nombre de "*School and College Achievement and Attainment Tables*", con el objetivo de proporcionar información sobre el rendimiento que ayude a los padres en la elección de la escuela y proporcionando incentivos a los centros que aumentaran sus niveles. A partir de 2002 se introdujeron puntuaciones de VA con el propósito de proporcionar una información más precisa y consistente del rendimiento de las escuelas (Ray, 2006; Ray, McCormack & Evans, 2009).

La gran precisión de las medidas obtenidas por MVA es fundamental para desarrollar sistemas efectivos de elección escolar, permitiendo a los padres contar con información adecuada del rendimiento de los diferentes centros educativos que sirva como base para su toma de decisiones escolares. La utilización de puntuaciones de VA, en este tipo de sistemas de elección escolar, mejora la información que se proporciona a los padres mediante las puntuaciones brutas de rendimiento. De esta forma, los familiares reciben una información adicional sobre la cual basar sus decisiones para la elección de la escuela más adecuada.

III.4.1.4 Para la investigación

En investigación educativa es muy difícil, o más bien imposible, llevar a cabo estudios totalmente experimentales que permitan la distribución aleatoria de los estudiantes en las escuelas donde se va a probar, por ejemplo, un nuevo programa educativo y aquellas que continúan con la enseñanza habitual. Esa falta de aleatoriedad puede ser paliada, en parte, utilizando la metodología de análisis del VA.

Las características de los MVA que controlan el rendimiento previo de los estudiantes y estudian el crecimiento en aprendizaje son adecuados para llevar a cabo este tipo de investigación exploratoria. Además es posible introducir otras características de los estudiantes y las escuelas que ayuden a controlar otros factores que puedan estar relacionados con el logro académico (Hanushek, 2003).

También es posible llevar a cabo investigaciones exploratorias que relacionen diferentes variables del profesorado o de los centros con las puntuaciones de VA, de forma que se averigüe la importancia de cada uno de los factores que se incluyan en los estudios.

Además existe la posibilidad de estudiar diferentes tipos de análisis estadístico para la estimación del VA y comprobar las posibles diferencias que producen en las clasificaciones de las escuelas (Betebenner, 2004; Kane & Staiger, 2008).

III.5 Beneficios del Valor Añadido

Los MVA, como herramienta para llevar a cabo la evaluación de los efectos escolares, han conseguido superar a otras técnicas de análisis. Tal vez se deba a la flexibilidad de este tipo de modelos para adaptarse a diferentes contextos de evaluación, puede ser que se deba a que estos modelos tratan de aislar la contribución de los centros educativos al logro de sus alumnos de la forma más justa posible, independientemente de factores contextuales del estudiantes que no se encuentran bajo el control de dichos centros o, quizás sea porque se aproxima en mayor medida al análisis de la realidad educativa midiendo el cambio en aprendizaje.

En consecuencia, ha habido un aumento de la financiación y apoyo que recibe el desarrollo de este tipo de modelos en diferentes sistemas educativos. Esto se ve reflejado en la implantación de esta metodología en diferentes sistemas de evaluación, principalmente en Estados Unidos e Inglaterra (Sanders & Horn, 1994; Doran & Izumi, 2004; McCaffrey & Hamilton, 2007; Ray, 2006; Ray, McCormack & Evans, 2009). También existe gran diversidad de publicaciones relacionadas con la investigación y desarrollo de los MVA y sus aspectos metodológicos.

Las mejoras y beneficios que proporcionan los MVA no se encuentran únicamente ligadas al aumento del rigor y la precisión metodológica en el tratamiento de los datos educativos, sino que mejora todo el proceso de evaluación. La OCDE menciona algunos de los beneficios que fueron destacados por los profesores durante un programa de entrenamiento en este tipo de metodología que se llevó a cabo en Polonia (2008) :

- La objetividad de los resultados del VA que destaca el buen trabajo de escuelas con alumnos desaventajados y supera las comparaciones basadas en puntuaciones brutas de los test.

- La precisión de las evaluaciones cuantitativas y los métodos estadísticos empleados.
- La gran transparencia y comparabilidad resultantes de los métodos de VA para evaluar escuelas.
- El potencial para mejorar la evaluación escolar interna del progreso de los estudiantes, especialmente mediante análisis adicionales del nivel escolar como, por ejemplo, el análisis de las puntuaciones de VA para grupos específicos de estudiantes (con alto o bajo rendimiento, inmigrantes, etc.).
- Los beneficios derivados del extenso entrenamiento y las consultas públicas previas a la implementación de un modelo de VA.

Otras ventajas asociadas al VA:

- Al tener en cuenta datos de crecimiento y ganancia y no solo un punto de referencia de rendimiento, es posible identificar centros que realmente hacen progresar a sus estudiantes eliminando posibles elementos de sesgo como la selección de estudiantes de altas capacidades en determinadas escuelas.
- Son el complemento perfecto de los sistemas de evaluación, ya sean para la rendición de cuentas o para cualquier otro propósito, los MVA producen medidas precisas del rendimiento escolar que pueden ser utilizadas como base para la toma de decisiones sobre política educativa.
- Son útiles para estudiar determinadas poblaciones de alumnos dentro de las escuelas o los distritos. De este modo, observando la evolución de estos estudiantes, es posible detectar posibles problemas y poner en marcha estrategias para solucionarlos.
- Las medidas de VA son útiles para comparar grupos de escuelas y probar si un determinado programa educativo tiene mejores resultados académicos que otro o incluso determinar la eficacia diferencial de distintos docentes.
- Una vez obtenida una medida precisa de VA se abre un abanico de posibilidades para relacionar determinados factores educativos del

centro y del estudiantes, determinando así los posibles efectos que estos pueden tener en el progreso académico.

Pero los MVA no son solo ventajas. El aumento de la precisión metodológica también implica un aumento de la complejidad del cálculo y la estimación de las puntuaciones finales. Se utilizan técnicas estadísticas complejas que pueden ser muy difíciles de comprender e interpretar si no se poseen conocimientos estadísticos. El aumento en el grado de complejidad de los análisis, en ocasiones, no permite conocer con claridad qué quiere decir realmente esa puntuación final de VA, ese residuo asociado a cada centro educativo evaluado. Por tanto, una información precisa y en profundidad sobre el proceso estadístico de análisis es fundamental en el desarrollo de este tipo de medidas.

III.6 Problemas vinculados al Valor Añadido

Las estimaciones finales de VA asociadas a las escuelas se consiguen después de varios procesos de análisis estadístico. Estos procesos pueden hacer poco entendibles los resultados para aquellos que no son expertos en análisis estadísticos avanzados. Este aspecto conlleva, por un lado, complicaciones vinculadas a la metodología de análisis de la información y, por otro, a cuestiones relacionadas con el trabajo que requiere la planificación, diseño y puesta en marcha del operativo necesario para llevar a cabo este tipo de evaluaciones basadas en el VA.

Algunos de los problemas metodológicos relacionados con el análisis del VA que más destacan en la literatura son, por ejemplo, el tipo de modelo estadístico adecuado para llevar a cabo el análisis de la información, el tipo de medida del rendimiento, el número de ocasiones de medida y la elección de la trayectoria de crecimiento.

Rogosa y Willet (1983) consideran que la medida de ganancia es una medida fiable para estimar el cambio individual, pero estos modelos no son tan efectivos para proporcionar información que determinados predictores individuales o de las escuelas pueden tener sobre la tasa de ganancia. No obstante, existen VAM que utilizan dos mediciones del rendimiento del estudiante, es decir,

comparando el rendimiento actual con el rendimiento previo (Meyer, 1997; Demie, 2003; Ray, 2006).

Algunos de los MVA, principalmente aquellos utilizan modelos multinivel longitudinales para llevar a cabo el análisis y estimar una pendiente de crecimiento a lo largo del tiempo, necesita contar con escalas de rendimiento capaces de comprobar esa evolución, es decir, que puedan cuantificar la cantidad de cambio en el constructo evaluado a lo largo de un continuo definido por la propias características métricas de la escala. Las escalas verticales son una opción recomendada cuando se desarrollan este tipo de modelos (Singer & Willett, 2003; Goldschmidt, Choi & Martinez, 2004). Conocer que características tienen y como se construyen es un elemento de análisis fundamental en la elaboración de estimaciones de VA porque la métrica de la escala puede condicionar los resultados de crecimiento (Yen, 1986; Ballou, 2009).

La introducción de predictores de contexto en los MVA es otro aspecto discutido y que puede afectar a las puntuaciones finales (Ballou, Sanders & Wright, 2004; Hibpshman, 2004; Tekwe et al., 2004; Choi, Goldschmidt & Yamashiro, 2006; Lockwood et al., 2007; Haegeland & Kirkeboen, 2008; Ferrão, 2009). Esta cuestión está vinculada directamente con la utilización de modelos de crecimiento, para algunos investigadores, el sujeto ejerce su propio control al focalizar la atención en el cambio y, por tanto, no es necesaria la inclusión de predictores de contexto (Sanders & Horn, 1994; Stevens, 2005; Stevens & Zvoch, 2006). Pero no todos los autores están de acuerdo con esta afirmación y opinan que incorporarlos sí es importante (Bryk & Raudenbush, 2002; Hibpshman, 2004; Ferrão , 2009).

La conceptualización del VA como un residuo acarrea ciertos problemas como la ambigüedad en su definición, que dependerá de las variables que se incluyan en el MVA. El residuo es la variación que no se ha modelado, es decir, una vez que se controlan los factores de contexto, se asume que esa variación residual es el resultado de los procesos que ocurren en la escuela, el efecto escolar. No obstante, esta consideración del VA como un residuo ha sido objeto de crítica, sobre todo si se utiliza dicha puntuación como única fuente de información en sistemas de rendición de cuentas que premian o castigan a los agentes evaluados,

es decir, la consideración puramente causal del VA (Rubin, Stuart & Zanutto, 2004; Ray, 2006; Reardon & Raudenbush, 2008; Rothstein, 2009).

Los aspectos metodológicos problemáticos mencionados y otras cuestiones como el tratamiento de los datos perdidos o los errores y precisión en las estimaciones se detallan en el siguiente capítulo[43].

Las cuestiones de planificación y puesta en marcha de este tipo de evaluaciones son otro aspecto que requieren un gran esfuerzo y trabajo. El VA no requiere únicamente tomar decisiones respecto al análisis de la información de rendimiento académico de los estudiantes, sino que para su desarrollo es necesario tomar decisiones durante la planificación de la evaluación que permitan conseguir este tipo de puntuaciones.

Durante la planificación inicial de la evaluación es necesario:

- La colaboración de los encargados de la administración educativa. Sin un compromiso de los políticos encargados de la toma de decisiones, educativas, principalmente respecto a la evaluación en educación, es imposible el desarrollo de un MVA. Desde el inicio del proceso se debe contar con el apoyo, tanto de los dirigentes como de los responsables directos en los centros educativos (inspectores, directores y profesores). Actuando de esta forma, se facilita el desarrollo del proceso de evaluación y el tratamiento posterior de los resultados alcanzados. También deben concretarse las tareas de cada uno de los participantes en el proceso.
- Conocer de antemano qué objetivos persigue la evaluación y qué utilización se hará de los resultados obtenidos, son aspectos de suma importancia y deben considerase durante el diseño. Ya se ha comentado la variedad de usos de las puntuaciones de VA, por ejemplo la rendición de cuentas, el diagnóstico y conocimiento de la situación escolar y la selección de escuelas por parte de los padres. Además, es posible combinar estos usos, por ejemplo, utilizando la información para detectar e incentivar el progreso de grupos de estudiantes con bajo rendimiento dentro de los centros y, además, detectar centros

[43]Ver *Capítulo IV* para más información

que se encuentran significativamente por encima de la media para llevar a cabo estudios pormenorizados que identifiquen buenas prácticas en estas escuelas y poder hacer fluir la información obtenida.

- Estimar el coste de la ejecución de un modelo de VA. El precio final dependerá básicamente del tamaño de la muestra a evaluar y del formato de recogida de la información de resultados y de contexto (número de ocasiones de medida, ítems objetivos o de respuesta construida, pruebas en lápiz y papel o evaluación telemática, etc.). El coste de la puesta en marcha suele ser más elevado que su posterior mantenimiento y algunas factores pueden abaratar costes. Por ejemplo, la evaluación asistida por ordenador y la utilización de pruebas adaptativas. De este modo, será posible contar con una base de datos de ítems equiparada que permita su replicación anual, ya que los estudiantes de un mismo grupo no tienen por qué contestar a las mismas preguntas aunque finalmente todos hayan contestado a una prueba con propiedades psicométricas equivalentes.

Durante el diseño del MVA es necesario:

- Decidir cómo se va a llevar a cabo la evaluación del logro académico de los estudiantes. Se disponen de datos o es necesaria la elaboración de pruebas *ad hoc*. Deben tomarse decisiones respecto a qué contenidos van a ser evaluados y la forma de hacerlo. Una decisión importante es el tipo de ítems que van a incluir los instrumentos de medida. En el caso de ítems de respuesta construida es necesaria la formación de correctores y, por tanto, habrá un incremento en los costes.
- Contar con un sistema de identificación de estudiantes que permita su seguimiento a lo largo de las distintas ocasiones de medida. Uno de los requisitos de algunos MVA es contar con varias puntuaciones de logro de un mismo estudiantes y, por tanto, es necesario identificar a los alumnos durante las distintas evaluaciones. Si el número de cursos a evaluar es amplio una de las cuestiones clave a resolver es cómo seguir al alumnado en su paso de una etapa educativa a otra, aunque

cambien de centro educativo, por ejemplo, cuando finalizan la educación primaria y comienzan la secundaria.

- Decidir como se va a llevar a cabo el escalamiento de las puntuaciones para lograr la comparabilidad de los datos, es decir, situar en una escala común las diferentes puntuaciones obtenidas mediante los test para poder medir el cambio o crecimiento en aprendizaje. Este aspecto influye también en el diseño de los propios instrumentos de medida y la forma de llevarlo a cabo puede producir diferencias en las estimaciones de VA.
- Tomar decisiones respecto a la contextualización de los modelos, ¿es necesaria la introducción de covariables en el modelo? y si es así, ¿cuáles van a incluirse y cómo van medirse?
- Determinar el MVA más adecuado. Elegir entre un determinado modelo u otro no es una decisión sencilla y un estudio piloto puede ayudar a elegir un modelo apropiado a la situación. El testeo de diferentes modelos y la inclusión de parámetros que nos ayuden a conocer la fiabilidad de los resultados obtenidos ayuda a mejorar su transparencia. Se deben tomar decisiones respecto a qué técnicas de análisis estadísticos utilizar o que procesos de estimación van a emplearse para el cálculo de las puntuaciones finales de VA.

En la presentación de los resultados de VA debe tenerse en cuenta que:

- Los resultados se deben vincular con la práctica educativa y utilizarse para la mejora escolar. Es necesario que los encargados de aplicar las políticas educativas en el centro escolar sepan interpretar y utilizar las estimaciones de VA y, por tanto, precisan de formación en estas cuestiones.
- Los agentes que tienen capacidad para tomar decisiones en las escuelas (directores, profesores y otros miembros de la organización) y que, en definitiva, son el objeto de la evaluación (recordemos que el VA evalúa, principalmente, escuelas o docentes a partir de los resultados de sus estudiantes) han de estar informados sobre el funcionamiento de este tipo de metodología si se pretende que los

resultados tengan una buena acogida. Conviene diseñar programas de información dirigidos a aquellos que son objeto directo de los de las evaluaciones, con la finalidad de que conozcan el proceso de análisis de la información y cómo se estiman las puntuaciones de VA de las escuelas.

- La transparencia de los resultados es fundamental, una complejidad excesiva en el modelo y los procesos de estimación puede hacer menos entendible la información para aquellos que no posean altos conocimientos de estadística. Si, por ejemplo, los inspectores de educación, los directores de los centros y los propios profesores deben comprender y utilizar la información, ésta debe ser clara y sin ningún tipo de sesgo que provoque confusión en la interpretación de los resultados.

El número de aproximaciones que tratan de analizar el VA en educación es bastante alto. Las distintas metodologías, es decir, los MVA difieren en aspectos relacionados con la concepción del cambio en aprendizaje, la elaboración de escalas de rendimiento, el tipo de análisis estadístico, etc. En consecuencia, los dos próximos capítulos (*Capítulo IV* y *Capítulo V*) están dedicados a detallar los aspectos metodológicos más relevantes en la elaboración de MVA y a la descripción de diferentes aproximaciones que tratan de estimar ese VA en educación, respectivamente.

Capítulo IV: Aspectos metodológicos en el análisis del Valor Añadido

Las cuestiones metodológicas que se desarrollan en este apartado están ligadas directamente con la medida de cambio necesaria en las evaluaciones que utilizan la metodología de VA. Para conseguir una variable de resultados que refleje el cambio en aprendizaje es fundamental que las puntuaciones de la variable de resultados estén en una escala común. La construcción de este tipo de escalas es uno de los temas principales del capítulo. También se incluye la descripción de las distintas medidas de cambio, diferenciando entre ganancia o crecimiento. Además se tratan otros aspectos metodológicos que destacan en la investigación sobre VA, como las consideración de causalidad de las estimaciones de VA asociadas a las escuelas, si los MVA deben contextualizarse con la inclusión de predictores del contexto del estudiante y/o la escuela, el posible efecto de los casos perdidos en las estimaciones, la incertidumbre producida por el error muestral o la volatilidad de las puntuaciones de cambio.

Las estimaciones de VA que se utilizan en los sistemas de evaluación son producto de los análisis estadísticos empleados para su obtención, es decir, de los MVA[44] que se desarrollan. Estos análisis pueden variar en complejidad y algunos de los que se utilizan en la actualidad utilizan técnicas con un nivel de dificultad

[44]Ya se ha mencionado en el capítulo anterior (apartado III.3.1) que el término VA puede analizarse desde dos aproximaciones fundamentales. La primera, más teórica, que trata de definirlo conceptualmente y, la segunda, directamente vinculada con la forma de operativizarlo, de medirlo, y que hace referencia a la técnicas estadísticas de análisis utilizadas para calcularlo, los Modelos de Valor Añadido (MVA)

tan alto que su comprensión resulta difícil para los que no poseen unos conocimientos estadísticos suficientes. Este aspecto es el que ha provocado la denominación de "caja negra" a algunos análisis del VA (OCDE, 2008, pág. 76).

La elaboración y estimación de puntuaciones de VA conlleva tomar decisiones metodológicas en la formulación del modelo estadístico, es decir, el MVA. Si estas cuestiones no son tratadas adecuadamente, los análisis del VA probablemente valorarán de forma errónea la eficacia de muchos profesores y escuelas y podrían obstaculizar los esfuerzos sistemáticos para mejorar la educación (McCaffrey, Koretz, Louis & Hamilton, 2004). No obstante, todos los análisis del VA comparten ciertos aspectos que pueden amenazar a la validez de las estimaciones de los efectos de las escuelas o docentes para los que han sido diseñados.

Algunas publicaciones llevan a cabo una revisión de algunos de las cuestiones metodológicas que pueden sesgar las estimaciones de VA (McCaffrey, Lockwood, Koretz & Hamilton, 2003; Doran, 2003; Lockwood, Louis & McCaffrey, 2003; McCaffrey, Koretz, Louis & Hamilton, 2004; Lockwood et al., 2007; Armein-Beardsley, 2008; Braun, Chudowsky & Koenig, 2010).

La elección de un determinado MVA conlleva asumir ciertos riesgos y cualquiera de ellos puede ser imperfecto (Wiley, 2006). Un descripción de estas cuestiones puede ayudar a mejorar la comprensión de estos modelos estadísticos empleados para evaluar la eficacia escolar.

La primera cuestión que trata este capítulo es la utilización, por parte de las evaluaciones basadas en el VA, de test estandarizados para obtener las diferentes puntuaciones de resultados y la atención que ponen dichos modelos en el cambio en aprendizaje, en lugar de analizar una única medida del logro académico. Otro de los aspectos a destacar, en relación con la utilización de diferentes medidas de logro académico y si se quiere contar con datos de rendimiento comparables, es la construcción de una escala que permita conocer cómo está cambiando el constructo evaluado, las escalas verticales son una de las opciones más utilizadas (Sanders, Saxton & Horn, 1997; Zvoch & Stevens, 2003; Ballou, Sanders & Wright, 2004; McCaffrey, Koretz, Louis & Hamilton, 2004; Stevens & Zvoch, 2006) para estimar el crecimiento en rendimiento del estudiante.

En segundo lugar, otro aspecto clave en los MVA, es la modelización del crecimiento en el aprendizaje. El análisis de la ganancia ha sufrido una evolución desde su consideración inicial como diferencia entre una puntuación de pretest y otra de postest, hasta modelos estadísticos altamente complejos que utilizan datos longitudinales de los resultados de los estudiantes. Es necesario, por tanto, una descripción de este proceso.

Además de analizar el cambio en aprendizaje, el VA trata de estimar puntuaciones que reflejen la aportación de las escuelas a ese cambio, independientemente de otros factores que no se encuentran bajo el control escolar. El debate sobre la contextualización de los modelos y los posibles efectos de incluir predictores del entorno familiar y socioeconómico del estudiante en los MVA es otro de los aspectos que se trata en este capítulo (Ballou, Sanders & Wright, 2004; Haegeland & Kirkeboen, 2008; Ferrão, 2009).

Contextualizar un MVA implica la inclusión de determinados predictores, tanto del contexto de los estudiantes como el las escuelas, que pueden estar relacionados con los resultados académicos pero que no son producto de los procesos que se producen en los centros educativos. Se incluyen con la finalidad de llevar a cabo un control estadístico de los mismos y evitar su influencia en las estimaciones finales de VA. Esta contextualización se vincula directamente con la consideración de las puntuaciones de VA cómo efectos causales, que es otra de las cuestiones que ha generado cierto debate (Rubin, Stuart & Zanutto, 2004; Ballou, Sanders & Wright, 2004; Reardon & Raudenbush, 2008; Kane & Staiger, 2008; Rothstein, 2009; Koedel & Betts, 2009). El problema aparece sobre todo en los sistemas de evaluación que utilizan las puntuaciones de VA como única fuente de información para tomar decisiones respecto a las escuelas o los docentes (principalmente en los sistemas de rendición de cuentas *high-stakes*), si se utilizan con esta finalidad se asume que dichas puntuaciones tienen una relación causa-efecto respecto a los procesos que se producen en la escuela. Este aspecto debe tratarse con cautela y, por tanto, es conveniente realizar una aclaración en este capítulo.

IV.1 Variable de resultados en los análisis del Valor Añadido

La utilización de test estandarizados es la forma de contar con herramientas que permitan extraer información de los resultados de los estudiantes de forma objetiva. Si se utilizan, por ejemplo, las calificaciones que los docentes dan a sus alumnos durante el curso académico no sería posible realizar comparaciones de las puntuaciones otorgadas por diferentes profesores a grupos distintos de estudiantes. Los docentes utilizan criterios de evaluación distintos y los resultados no dependen únicamente de la nota en un examen.

Las puntuaciones obtenidas mediante los test estandarizados son la materia prima que utilizan los MVA en sus análisis (Sanders, Saxton & Horn, 1997; Bryk, Thum, Easton & Luppescu, 1998; Demie, 2003; Ballou, Sanders & Wright, 2004; Doran & Izumi, 2004; McCaffrey & Hamilton, 2007; Choi, Seltzer, Herman & Yamashiro, 2007; Jakubowski, 2008). En consecuencia, la estimación final de los efectos de las escuelas estará influenciada por la calidad de esos resultados que proporcionan las pruebas de rendimiento. Por esta razón, un buen diseño y elaboración de los instrumentos de medida es fundamental para conseguir estimaciones fiables del VA de las escuelas.

Para poder medir el VA es necesario medir el cambio en el aprendizaje de los estudiantes y esto no sería posible si las puntuaciones de rendimiento no cumplen dos requisitos indispensables (Thum, 2003):

- La noción de cambio tiene poco sentido si el constructo que se mide es diferente entre ocasiones de medida. Asumiendo que dicho constructo es cualitativamente constante, medir el cambio es detectar las variaciones en grado o cantidad en ese constructo.
- Medir la cantidad de cambio no es posible si el instrumento de medida o la propia escala cambia de forma desconocida.

Los análisis del VA, al utilizar puntuaciones de cambio, necesitan contar con medidas de resultados que puedan compararse a lo largo de los diferentes cursos evaluados, es decir, que los diferentes resultados de los test derivados de las

distintas tomas de datos llevadas a cabo permitan conocer la evolución del logro académico del estudiante.

Medir el aprendizaje como un proceso de cambio implica tomar diferentes mediciones del constructo que va a ser evaluado. Pero no solo basta con eso, es necesario que los instrumentos de medida empleados sean capaces de medir la evolución de ese constructo a lo largo del tiempo e identificar cuál es el cambio que se produce entre las distintas ocasiones de medida. Un requisito necesario en las evaluaciones longitudinales que tratan de seguir el progreso de los estudiantes a lo largo de un periodo de escolarización, entre los que se encuentran aquellas basadas en el VA o en modelos de crecimiento, es contar con una escala vertical válida (Singer & Willett, 2003; Chin, Kim & Nering, 2006).

IV.2 Escalas verticales de rendimiento

La utilización de una escala vertical de rendimiento, también denominadas escala de desarrollo, que sea capaz de identificar los cambios en el logro de los estudiantes a lo largo de varios cursos académicos, es una característica de algunos MVA (Bryk, Thum, Easton & Luppescu, 1998; Ponisciak & Bryk, 2005; Zvoch & Stevens, 2006; Castro, Ruíz & López, 2009). Su elaboración es un aspecto clave en la evaluación porque la estimación final de los efectos de los centros educativos dependerá de como haya sido construida la escala (Yen, 1986; Chin, Kim & Nering, 2006; Jungnam, 2007; Briggs, Weeks & Wiley, 2008; Briggs & Weeks, 2009; Briggs & Betebenner, 2009). No obstante, no todos los MVA emplean este tipo de escalas, aquellos que no lo hacen es fundamentalmente por dos motivos: porque únicamente utilizan dos puntuaciones del logro académico para desarrollar un modelo de regresión multinivel utilizando el rendimiento previo como covariable (Demie, 2003; Ray, McCormack & Evans, 2009), o porque no son un requisito necesario para el modelo, como ocurre en el de Tennessee (Sanders & Horn, 1994; Sanders, Saxton & Horn, 1997) o en el de percentiles de crecimiento desarrollado por Betebenner (2009).

En cambio, cuando se utilizan modelos de ganancia sí es necesario contar con este tipo de escala. El cálculo de la diferencia entre las puntuaciones requiere,

además de medir el mismo constructo, tener las mismas unidades de medida (Reckase, 2008). También es algo imprescindible si se utiliza un modelo longitudinal para estimar una pendiente de crecimiento vinculada al tiempo y el residuo asociado a cada escuela a través de un análisis de regresión multinivel (Bryk, Thum, Easton & Luppescu, 1998; Ponisciak & Bryk, 2005; Zvoch & Stevens, 2006; Briggs, Weeks & Wiley, 2008)

Se denomina escalamiento vertical (*vertical scaling*) al proceso empleado para situar en una escala común las puntuaciones de test que difieren en dificultad pero que están diseñados para medir el mismo constructo (Kolen & Brennan, 2004). Las escalas verticales se elaboran a partir de distintos tests, cada uno de los cuales está desarrollado para que se adecúe a los estudiantes de un determinado grado o edad. Este proceso pone en una misma escala las puntuaciones de varios test que miden un constructo similar pero en niveles educativos distintos, es decir, cambian en dificultad y contenido pero miden contenidos afines.

Este proceso es denominado anclaje (*linking*) por los autores Kolen y Brennan (2004) y calibración (*calibration*) por Myslevy (1992) y Linn (1993) pero también es conocido como equiparación para la comparabilidad.

Además del escalamiento vertical, dependiendo del diseño de los instrumentos de medida, puede ser necesario otro tipo de anclaje. Por ejemplo, si además del test que se aplica en cada uno de los cursos evaluados, se ha utilizado más de una forma para un mismo grado o nivel académico, el proceso de equiparación horizontal es otro requisito. Este proceso trata de ajustar las diferencias en dificultad que existen en diferentes test que han sido diseñados para tener contenido y dificultad similar. Por ejemplo, cuando se diseñan dos formas paralelas para un mismo curso.

La equiparación horizontal es denominada estrictamente equiparación (*equating*) por Kolen y Brennan (2004), Myslevy (1992) y Linn (1993). Es el tipo de equiparación más potente y produce puntuaciones totalmente intercambiables.

La utilización de escalas de rendimiento equiparadas verticalmente que permiten comparar cuantitativamente el cambio que se produce en el logro académico en los diferentes grados evaluados, y elaboradas empleando Teoría Respuesta al Ítem (TRI) demandan la propiedad de intervalo. Sin embargo, existe

un cierto debate entre los expertos en psicometría sobre si este tipo de escalas realmente tienen esa propiedad (Reckase, 2008; Ballou, 2009; Briggs & Betebenner, 2009). Este es uno de los puntos problemáticos asociados a la elaboración de estas escalas.

Otro aspecto de controversia es que el constructo que trata de medir un determinado test, por ejemplo matemáticas, puede ser muy distinto a lo largo de los diferentes grados, es decir, los contenidos pueden ser cambiantes y también la importancia que cada uno de ellos tiene, por lo que resulta más difícil la consideración unidimensional de dicho constructo cuando se evalúan cursos o grados muy separados. Además, otro aspecto problemático relacionado con los contenido que incorporan los instrumentos de medida es la utilización de una determinada medida de resultados u otra o la variación de los contenidos, que puede afectar a las estimaciones finales de VA (Lockwood et al., 2007).

Los cambios de etapa o ciclo educativo también pueden determinar la estructura de la escala de puntuaciones de rendimiento y, por tanto, afectar al diseño de los MVA (Gaviria, Biencinto & Navarro, 2009). Por tanto, la diferencia en los contenidos del currículum de los cursos evaluados es una de las mayores amenazas que puede afectar a la elaboración de este tipo de escalas. El problema se amplia en los cursos superiores porque los contenidos son menos uniformes (McCaffrey, Lockwood, Koretz & Hamilton, 2003). Martineau (2006) analiza los posibles efectos de la violación de este supuesto de unidimensionalidad del constructo de medida en la construcción de escalas verticales.

Un último aspecto problemático es el señalado por Patz (2007) y Ballou (2009). Ambos indican que las medias de las ganancias en los grados superiores de una escala vertical y también la varianza, se ve reducida empleando TRI. Esta cuestión evidencia que, con este tipo de escala, la ganancia puede no tener el mismo significado en diferentes partes de la escala.

Estos tres aspectos problemáticos que se han mencionado (propiedad de intervalo, dimensionalidad del constructo y varianza en el crecimiento) se detallan en el siguiente apartado de características de las escalas verticales (IV.2.1).

La elaboración de una escala vertical es, por tanto, un punto crítico en aquellos MVA que las utilizan como variable de resultados. Las estimaciones

finales de VA dependerán de cómo haya sido construida la escala (Ray, 2006; Tong & Kolen, 2007; Jungnam, 2007; Lockwood et al., 2007; Ito, Sykes & Yao, 2008; Briggs & Betebenner, 2009).

IV.2.1 Características de las escalas verticales

Una escala vertical indica que si un estudiante incrementa su puntuación en esa escala, en consecuencia, está aumentando su conocimiento y habilidad en el constructo evaluado (Lissitz, Doran, Schafer & Willhoft, 2006). Por consiguiente, si las puntuaciones obtenidas a través de los test quieren compararse entre las diferentes ocasiones de medida, deben estar equiparadas verticalmente en una escala común.

La forma de construir este tipo de escalas puede derivar en propiedades distintas. Este apartado se dedica a describir los rasgos principales de estas escalas verticales.

IV.2.1.1 Propiedad de intervalo

La propiedad de intervalo en una escala supone, principalmente, asumir que la distancia entre los diferentes puntos es la misma, por ejemplo, entre la puntuación 5 y 10 hay la misma distancia que entre la 10 y la 15. Se ha generado debate sobre si las escalas verticales elaboradas bajos los supuestos de la TRI pueden asumir esta propiedad (Briggs & Betebenner, 2009; Ballou, 2009), es decir, si las diferencias entre los distintos puntos de la escala tienen un mismo significado.

Esta característica sería cierta si el modelo psicométrico utilizado para construir las escala es una representación matemática verdadera de la relación entre el logro del estudiante y sus respuestas a los ítems (Martineau, 2009). Los modelos psicométricos representan la probabilidad que tiene un alumno, con un determinado nivel de rendimiento, de responder a un ítem correctamente. En consecuencia, en los ítems con buen funcionamiento los estudiantes con alto rendimiento tendrán una alta probabilidad de responder de forma correcta a los ítems. Sin embargo, la forma exacta de la relación entre estos dos aspectos, la probabilidad de responder correctamente y el rendimiento, es el centro del debate.

Algunos autores destacan que no hay una forma satisfactoria de validar las condiciones necesarias para dotar a este tipo de escalas con la propiedad de intervalo (Briggs, Weeks & Wiley, 2008) y que únicamente es posible tratarlas como escalas ordinales. No obstante, si se utilizan escalas ordinales, en lugar de las de intervalo, se cuenta con una información menos precisa y tampoco son una solución para contar con medidas de resultados en el análisis del VA (Thum, 2006).

Una opinión opuesta tiene Reckase (2008) que sí considera que las escalas verticales de rendimiento construidas bajo los supuestos de la TRI poseen la propiedad de intervalo, por dos motivos:

- Si el modelo TRI empleado ajusta con los datos, sí se cuenta con una escala de intervalo porque la forma de la función dentro de esta teoría no está definida al menos que la escala de rendimiento tenga dicha propiedad;
- Si se considera la forma específica de la distribución de las puntuaciones verdaderas en el test y la distribución observada coincide con la distribución asumida, entonces puede confirmarse que los resultados tienen esa propiedad. Normalmente se asume una distribución normal del rendimiento de los estudiantes, por tanto, si el número de respuestas correctas sigue este tipo de distribución puede considerarse que la escala tiene la propiedad de intervalo.

IV.2.1.2 Dimensionalidad del constructo evaluado

Las escalas verticales tratan de medir una habilidad a lo largo de diferentes cursos. Esta evolución a lo largo del tiempo puede ser problemática si el constructo medido cambia significativamente con los avances que se producen entre los cursos académicos. Un determinada habilidad, por ejemplo matemáticas, puede cambiar en contenido de primero a cuarto curso de educación secundaria o variar la importancia que se da a esos contenidos en cada curso. Este factor puede provocar una variación de la unidimensionalidad del constructo y por tanto afectar a la escala que asume este supuesto. Las escalas elaboradas bajo los supuestos de la TRI asumen la unidimensionalidad del constructo medido.

La utilización de una determinada medida de rendimiento u otra o la variación de contenidos puede afectar a las estimaciones finales de VA. Lockwood et al., (2007) llevan a cabo un estudio empírico donde prueban la sensibilidad de los MVA a la utilización de dos subescalas de rendimiento en matemáticas, comprobando que las estimaciones se ven más afectadas por la variación de la escala que por cambios en los modelos estadísticos como la introducción de predictores de contexto o la utilización de distintos análisis (ganancia, ajuste de covariables, modelos con persistencia de los efectos o con posibilidad de variación de los efectos de un año al siguiente).

Los cambios de etapa o ciclo educativo pueden determinar la estructura de la escala de puntuaciones de rendimiento y, por tanto, afectar al diseño de los modelos de VA (Ray, 2006; Gaviria, Biencinto & Navarro, 2009). Al considerar este aspecto, se presenta la siguiente cuestión:

¿es mejor abarcar un amplio abanico de cursos, incluso etapas completas, o centrar la atención en ciclos educativos concretos?

La diferencia en los contenidos del currículo es una de las mayores amenazas que puede afectar a la elaboración de las escalas verticales de logro. Esta variación en los contenidos puede cambiar la dimensionalidad del constructo evaluado y son menos uniformes a medida que los cursos analizados son superiores. Por tanto, parece más adecuado elaborar modelos diferenciados cuando se evalúan etapas educativas distintas para evitar problemas vinculados a los cambios sustanciales que pueden producirse en los contenidos.

Las escalas verticales elaboradas bajos los supuestos de la TRI tratan el constructo evaluado como unidimensional a lo largo de las diferentes cursos. Schafer (2006) menciona las deficiencias derivadas de asumir este supuesto cuando realmente no es una característica del constructo y qué requisitos deben tener los test para poder llevar a cabo la construcción de este tipo de escalas:

- Los contenidos que se enseñan en los cursos inferiores pueden ser bastante diferentes a los que se enseñan en los grados superiores, aunque tengan la misma denominación.

- Los test, en algunos diseños, incluyen ítems que los estudiantes más jóvenes pueden no haber estudiado nunca y también ítems que los estudiantes de mayor edad pueden no haber estudiado recientemente.
- Los estudiantes de cursos inferiores pueden recibir una puntuación mayor de la que se merece porque los ítems de los cursos superiores son tratados esencialmente como datos perdidos.
- El crecimiento entre diferentes regiones de una escala vertical desarrollada a lo largo de un gran número de cursos no es comparable porque la construcción se lleva a cabo tomando como base la posición de los ítems, más que utilizar la información sobre crecimiento.
- La descripción de los niveles de rendimiento sobre lo que los estudiantes saben o pueden hacer para una misma puntuación es diferente cuando los cursos estudiados son distintos.

Los posibles efectos de la violación de este supuesto de invarianza en la medida necesario en la construcción de las escalas verticales es estudiado por Marnineau (2006). Este autor trata de responder a la siguiente cuestión:

¿si los cambios que se producen en el constructo varían en grado y tipo, se distorsionan los resultados de los MVA basados en el crecimiento?

Martineau encuentra que el cambio en la mezcla de constructos a lo largo de los diferentes cursos puede tener consecuencias cuando se utilizan escalas verticales en el análisis del VA. Concluye que no hay escala vertical que pueda ser utilizada de forma válida en evaluaciones de alto impacto para la estimación del VA sobre el crecimiento del estudiante. La utilización de puntuaciones de escalas que son empíricamente multidimensionales puede no tener una utilidad práctica en la medida del VA de unidades concretas (escuelas o docentes). El autor recomienda la medida de un único constructo considerando solo dos grados, el curso actual y el curso previo, para obtener una estimación del VA. Denomina a esta aproximación como análisis del VA en una pareja de grados empíricamente unidimensionales. Otra opción recomendada para solucionar el problema de cambio en el constructo es la utilización de métodos de escalamiento vertical multidimensionales basados en TRI (Martineau, 2009).

Como consecuencia de lo anterior, la validez de las inferencias basadas en una escala vertical disminuirán, de forma general, con la distancia entre los cursos evaluados.

IV.2.1.3 Varianza del crecimiento a lo largo de la escala

Las escalas verticales tienden a reducir los cambios que se producen en el crecimiento en aprendizaje de los grados superiores evaluados. Ballou (2009) señala que las medias de las ganancias en los grados superiores de una escala vertical y también la varianza del rendimiento disminuye cuando se utiliza TRI. Se plantea, por tanto, la siguiente cuestión:

¿los estudiantes de los cursos superiores aprenden menos que los que se encuentran en los cursos iniciales?

Este aspecto es una evidencia de que, con este tipo de escala, la ganancia no tiene el mismo significado en diferentes partes de la escala. Por ejemplo, los estudiantes que crecen cinco puntos en la parte baja de la escala no consiguen la misma ganancia que aquellos que ganan los mismos cinco puntos en la parte alta de la escala. Patz (2007) también destaca este problema.

En consecuencia, será menos problemático comparar los datos de dos cursos adyacentes dentro de una misma cohorte de estudiantes. Y, de forma opuesta, la comparación de datos que proceden de estudiantes con una gran distancia entre los cursos a los que asisten, por ejemplo comparar 3º de Educación Primaria con 4º de Educación Secundaria Obligatoria, puede acarrear mayores problemas. Por ejemplo, los estudiantes de los cursos que se encuentran en la parte alta de la escala sufrirán el denominado *efecto techo* (Betebenner & Linn, 2009), es decir, el crecimiento disminuye porque se aproximan a las puntuaciones máximas de la escala. Un fenómeno similar, pero con consecuencias distintas, se produce en los cursos de la parte baja de la escala. Es conocido como el *efecto suelo* que produce un mayor crecimiento de los estudiantes que se sitúan al comienzo de la escala.

Los efectos de suelo y techo son problemáticos cuando se analiza el crecimiento entre cursos pero no para el que se produce dentro de un mismo curso. Este fenómeno está vinculado con la propiedad de intervalo que reclaman

las escalas verticales de rendimiento. Si se producen estos efectos, los intervalos entre puntuaciones de la escala no serán iguales, es decir, crecer 10 puntos de rendimiento al principio y al final de la escala no tiene el mismo significado.

Reckase (2010) afirma que las escalas elaboradas con TRI, al no tener límites, reducen estos efectos en comparación con el porcentaje de respuestas correctas u otros métodos de calificación similares. Por su parte, Braun, Chudowsky y Koenig (2010) comentan que las escalas de intervalo son solo una forma de medir que puede no coincidir con el valor que socialmente se da a las diferencias entre esos intervalos. Un ejemplo claro que ponen los autores sobre este fenómeno es lo que ocurre con la temperatura, es una escala de intervalo en el sentido de que se necesita la misma cantidad de energía para aumentar un grado la temperatura de un objeto. Sin embargo, no es de intervalo cuando se atiende a la comodidad de la gente, es decir, si la temperatura pasa de 20 a 25 grados no produce la misma sensación que si cambia de 30 a 35. Ocurrre algo similar con las escalas elaboradas con TRI que son de intervalo si se tiene en cuenta su definición, pero es improbable que lo sea si se tiene en cuenta el valor que la sociedad pone a los cambios que se producen en diferentes puntos de la escala.

En resumen, si se utiliza este tipo de escala es importante reconocer que los resultados del análisis del VA pueden ser sensibles a las características particulares de la misma. Utilizar la TRI es la opción más común para llevar a cabo este proceso de escalamiento vertical y conseguir unos resultados comparables. Y, aunque los análisis TRI tienen probada fiabilidad y validez para el estudio de constructos latentes y la construcción de escalas, deben tomarse ciertas precauciones cuando las escalas verticales son el aspecto crucial para la toma de decisiones en sistemas de evaluación basados en la rendición de cuentas.

La utilización de las escalas verticales será la opción adecuada cuando el número de cursos que va a ser anclado no es muy alto, alcanzando su mayor valor cuando se comparan cursos adyacentes. La escala comienza a producir algunos problemas cuando aumentan las distancias entre los grados evaluados y, por tanto, también en el contenido enseñado.

IV.2.2 Elaboración de una escala vertical

El proceso de elaboración de la escala vertical es un aspecto que puede estar presente cuando se diseña una evaluación basada en el VA. Es indispensable en aquellos que utilizan modelos longitudinales con pendiente de crecimiento o los que emplean una puntuación de ganancia como variable de resultados. Para analizar el cambio que se produce durante un periodo de tiempo es necesario contar con información de logro que se encuentre operativizada en una escala común. De esta forma podrá ser comparada.

La toma de decisiones en el proceso de construcción de una escala vertical puede producir variaciones en las estimaciones finales de VA que se logran con la información proporcionada por los test. Una correcta elaboración de la escala de rendimiento conducirá hacia el logro de estimaciones de VA más fiables y útiles para la evaluación de centros escolares.

Hay gran variedad de trabajos que tratan de evaluar el proceso de anclaje para la construcción de una escala vertical, a continuación se mencionan algunos. El tema central de estos estudios es comprobar el efecto que puede tener sobre las puntuaciones de la escala vertical las decisiones que se toman respecto a los distintos aspectos metodológicos del proceso de anclaje.

El primero de esos aspectos es el tipo de diseño adecuado para la recogida de la información como parte del proceso de equiparación. El trabajo de Tong y Kolen (2007) compara los resultados producidos por un diseño de ítems comunes frente a un test de anclaje. Otros trabajos analizan las variaciones que se producen al cambiar el número de ítems comunes que incluyen los test (Chin, Kim & Nering, 2006) o las consecuencias de cambiar la longitud de un test de anclaje (Lee & Ban, 2010)

El segundo aspecto investigado es la elección de un determinado modelo psicométrico. Lo usual es utilizar modelos logísticos TRI y hay trabajos que comparan modelos de uno y tres parámetros (Willett, 1997; Chin, Kim & Nering, 2006; Briggs, Weeks & Wiley, 2008; Briggs & Weeks, 2009). También hay trabajos que comparan los modelos logísticos con una escala elaborada con el método Thurstone (Yen, 1986; Kolen & Brennan, 2004; Tong & Kolen, 2007).

En tercer lugar, el proceso de escalamiento y calibración de los ítems también es objeto de estudio. Cuando se utilizan modelos TRI, situar las puntuaciones de distintas formas de un test, ya sea para la equiparación horizontal o el anclaje vertical, requiere transformar también los parámetros de los ítems a través de la calibración. Para ello existen diferentes opciones de llevarla a cabo, como la calibración conjunta, fija y por separado:

- En la Calibración Conjunta (CC) la estimación de los parámetros de los ítems de las distintas formas del test se realiza al mismo tiempo. Esto se consigue considerando los ítems de las diferentes pruebas como si fuera un único test y tratando aquellos ítems que no deben ser contestados por algunos estudiantes como perdidos por diseño.
- En la Calibración Fija (CF) los parámetros de los ítems comunes estimados inicialmente se fijan en el momento de llevar a cabo la estimación de los de la otra prueba. De esta manera se sitúan en una escala común.
- En la Calibración por Separado (CS) los parámetros de los ítems de las distintas formas se estiman por separado. Una vez realizado el proceso se lleva a cabo la trasformación lineal de los mismos. Esta transformación puede realizarse de diferentes modos: por un lado los métodos momento media/media (Loyd & Hoover, 1980) y media/sigma (Marco, 1977) y, por otro, los métodos de curva característica del ítem de Haebara (1980) y Stocking y Lord (1983).

Algunos trabajos se han encargado de comparar los tres métodos en el contexto de anclaje vertical (Jungnam, 2007; Ito, Sykes & Yao, 2008; Kang & Petersen, 2009); una calibración híbrida que mezcla por separado y conjunta es la que utilizan Briggs, Weeks, y Wiley (2008); otros solo comparan los resultados producidos por la calibración conjunta y por separado (Chin, Kim & Nering, 2006; Tong & Kolen, 2007); y Lee y Ban (2010) comparan, además de la calibración conjunta y la calibración por separado, otra metodología denominada transformación de la habilidad[45].

[45]Más información sobre esta metodología en Kolen y Brennan (2004, pág. 164)

Finalmente, el cuarto aspecto está relacionado con el proceso de calificación (*scoring*) de los sujetos, la estimación de la habilidad. El proceso de calificación puede llevarse a cabo utilizando tres tipos de estimación principalmente: Máxima verosimilitud (ML), Bayes o esperada a posteriori (EAP) y Bayes Modal o máxima a posteriori (MAP). Estos métodos utilizan todo el patrón de respuestas de los sujetos para llevar a cabo el proceso. Por este motivo, sujetos con la misma proporción de respuestas correctas pero con diferente patrón pueden obtener puntuaciones distintas. A estos métodos de estimación se añade el de cuadratura de la distribución (QD) (*quadrature distribution*) (Tong & Kolen, 2007; Jungnam, 2007) o una variación de los métodos ML y EAP basados en el total de puntuaciones correctas y no en el patrón completo (Jungnam, 2007).

IV.2.2.1 Diseño de recogida de información

Siguiendo a Kolen y Brennan (2004), hay tres tipos de diseño que son más comunes en el proceso de equiparación horizontal: Diseño de grupos aleatorios, diseño de grupo único y diseño de grupo único con contrabalanceo. Para el anclaje vertical los más comunes son el diseño de ítems comunes, diseño de grupos equivalentes y diseño de test de anclaje

El tipo de diseño de recogida de información es uno de los factores que determina el proceso de equiparación. La inclusión de ítems de anclaje entre los test de los diferentes cursos evaluados para llevar a cabo este proceso es el diseño más utilizado (Chin, Kim & Nering, 2006; Ito, Sykes & Yao, 2008; Tong & Kolen, 2007; Jungnam, 2007; Briggs & Weeks, 2009). También existe la posibilidad de utilizar un test de anclaje como en el estudio de Lee y Ban (2010). En su trabajo, en la primera aplicación se administra una única forma (Forma A) que será la base de la escala, en la siguiente los estudiantes deben responder a una nueva forma (Forma B) además de la anterior Forma A, que se reparte con un diseño en espiral. En el proceso en espiral se alternan las diferentes formas de los test en los paquetes que se envían a las aulas y se asigna, por ejemplo, la forma A al primero del listado, la B al siguiente y así sucesivamente. En otras aulas se comienza distribuyendo por la B. De esta manera, con el reparto aleatorio de los cuadernillos, se pretende conseguir dos grupos equivalentes. La forma que se administra en ambas aplicaciones es el elemento de anclaje.

Para Kolen y Brennan (2004), en términos estadísticos, un conjunto mayor de ítems comunes produciría un menor error de equiparación. Los autores sugieren que al menos el 20% de los ítems de un test sean comunes en un diseño de equiparación horizontal. En este tipo de equiparación, la horizontal, un mayor número de ítems comunes puede ser preferible (Kang & Petersen, 2009). No obstante, existe poca investigación respecto a cuáles son los efectos del número de ítems comunes en un proceso de equiparación vertical. Patz (2007) recomienda utilizar al menos 15 ítems representativos del dominio para proporcionar un anclaje sólido. Chin, Kim y Nering (2006) recomiendan que el rango de los parámetros de dificultad del conjunto de ítems comunes sea variado, es decir, que cubra el continuo de la habilidad estimada para contener el sesgo y los errores. Estos autores simulan datos para estudiar, entre otros factores, como afecta el número de ítems comunes y la variación de sus parámetros de dificultad a las estimaciones de la habilidad. Sus resultados muestran, que independientemente del número de ítems de anclaje, si el rango del parámetro de dificultad de estos reactivos varía hasta dos desviaciones típicas en el rango de la habilidad estimada, es decir, si cubren la mayor parte del continuo de puntuaciones en el rasgo, entonces las estimaciones mejoran, conteniendo el sesgo y los errores de estimación en niveles bajos.

Tong y Kolen (2007) comparan los resultados producidos por un diseño de ítems de anclaje con los de un test de anclaje y señalan que el primer diseño tiende a mostrar mayor crecimiento que el segundo. No obstante, ambos tienden a disminuir el crecimiento a medida que el curso evaluado es superior, es decir, en ambos casos los cursos inferiores muestran un mayor crecimiento que los superiores. Para contrastarlo calculan tamaños del efecto utilizando las medias y desviaciones típicas de cada par de cursos adyacentes. También encuentran que esa distancia en los cursos inferiores es mayor en el caso del diseño de ítems comunes.

IV.2.2.2 Modelo psicométrico

Los procesos de equiparación y de anclaje están directamente vinculados con el de escalamiento (*scaling*), es decir, el proceso que convierte las respuestas de los sujetos en el test, las puntuaciones brutas, en un nuevo conjunto de números

con determinadas características. Por tanto, el proceso de escalamiento consiste en asociar números u otros indicadores con el rendimiento de los estudiantes evaluados (Kolen & Brennan, 2004) y su producto es una escala de puntuaciones.

Normalmente, la escala se construye a partir de las respuestas a un solo test pero cuando se emplean varias formas es necesario transformar las puntuaciones para lograr la comparabilidad. Existen varias formas de llevar a cabo esa transformación pero este trabajo se centra en aquellos basados en la Teoría de Respuesta al Ítem (TRI). Este tipo de metodología es la más empleada para la construcción de las escalas verticales que se utilizan en los análisis de VA (Briggs, Weeks & Wiley, 2008). No obstante, antes de comenzar con la descripción de los modelos TRI, conviene llevar a cabo una pequeña descripción del método Thurstone que también se utiliza en diferentes estudios de comparación de metodologías de equiparación y anclaje (Yen, 1986; Kolen & Brennan, 2004; Tong & Kolen, 2007).

El método de escalamiento Thurstone asume la normalidad de la distribución de la habilidad de los estudiantes dentro de un mismo curso y utiliza el número de respuestas correctas para llevar a cabo el proceso. Con esta metodología el número de respuestas correctas se transforma a puntuaciones normalizadas y se establece una escala de desarrollo a lo largo de esas puntuaciones. Para llevar a cabo el anclaje y establecer las relaciones entre cursos se utiliza la media y desviación típica[46].

En cambio, los modelos TRI tratan de dar una fundamentación probabilística a la estimación de constructos no observables (rasgo latente). Las respuestas de los sujetos a los ítems dependen de determinadas características de los reactivos y del nivel de rasgo que el sujeto posee. Para ello estiman una función de probabilidad de respuesta a cada uno de los ítems que forman parte de un test en función de los parámetros de cada ítem y la puntuación obtenida por el sujeto. Utilizar modelos TRI que ajusten bien con un conjunto de datos tiene determinadas ventajas (Roberts & Ma, 2006):

[46]Más detalle en Kolen y Brennan (2004, pág. 393)

- La interpretación de los parámetros de los ítems será invariante con respecto a la distribución del rasgo latente de los estudiantes que responden al test. Esta propiedad permite la construcción de grandes bancos de ítems en los cuales las características de diferentes conjuntos de reactivos están determinadas por diferentes muestras de estudiantes.
- La interpretación del parámetro del sujeto (el rasgo) será invariante con respecto a la distribución de los ítems del test. Esto permite elaborar medidas a partir de diferentes ítems administrados a los sujetos, es decir, aunque los alumnos respondan a diferentes ítems es posibles estimar una puntuación en el rasgo que sea comparable.
- Puede calcularse la precisión de cada modelo de estimación del parámetro. Esto permite aproximar la medición del error asociado con la estimación del rasgo latente de cada individuo.

Kolen y Brennan (2004) denominan calibración (*calibration*) al proceso de cálculo de los parámetros de los ítems de un test a través de la transformación y la estimación.

Para el tratamiento de ítems dicotómicos, como los utilizados en este trabajo, existen modelos psicométricos de uno, dos y tres parámetros que estiman la probabilidad que tiene un sujeto de responder correctamente un ítem. En un modelo de tres parámetros la probabilidad de acertar un ítem depende de la relación entre los diferentes parámetros del ítems (dificultad, discriminación y pseudoazar) y la habilidad del sujeto:

- Parámetro a (discriminación del ítem): permite diferenciar entre los sujetos con aptitud inferior a la posición del ítem (dificultad) y los que la tienen superior. Es proporcional a la pendiente de la recta tangente en el punto de máxima pendiente de la Curva Característica del Ítem (CCI).
- Parámetro b (dificultad del ítem): cantidad de rasgo necesaria para responder correctamente al ítem. Está en la misma escala que el rasgo. Es el punto de máxima pendiente de la CCI.

- Parámetro c (pseudoazar): es la probabilidad de acertar el ítem al azar cuando no se sabe nada. Es la asíntota inferior de la CCI (probabilidad de acierto en el menor nivel del rasgo.

En cambio, en un modelo de un parámetro, solo depende de la dificultad del ítem, todos los ítems discriminan igual y no hay pseudoazar. El modelo de Rasch es un caso especial del anterior donde la discriminación de los ítems se fija a 1.

Con el modelo logístico de tres parámetros, la probabilidad de respuesta correcta de un sujeto a un determinado ítem en función de su nivel en el rasgo se calcula a través de la siguiente formula:

$$P(\theta|a,b,c) = c + (1-c)\frac{e^{Da(\theta-b)}}{1+e^{Da(\theta-b)}} \qquad \textbf{Ec. IV.1}$$

Donde a, b y c son los parámetros de los ítems; θ es la puntuación obtenida en el constructo evaluado; *e* es la base de los logaritmos neperianos; y *D* es una constante para ajustar a la normalidad los resultados (-1,7). Se utiliza esta constante porque la escala en los modelos TRI es indeterminada, es decir, el punto de partida y la unidad se fijan de forma arbitraria. Con la constante se utiliza la distribución normal como referencia.

Tomando como base esta ecuación, es posible llevar cabo una transformación lineal de la puntuación en el constructo (θ) de una escala X, obtenida con el test X, en una nueva escala Y de la siguiente manera:

$$\theta_Y = A\theta_X + B \qquad \textbf{Ec. IV.2}$$

Donde A y B son las constantes en esa transformación (la pendiente y el intercepto respectivamente) y θ_Y y θ_X son las puntuaciones de un sujeto en las dos escalas. Transformar la escala implica también transforma los parámetros de los ítems utilizando esas constantes. De esta forma, se mantienen las mismas probabilidades de respuesta estimadas:

$$a_Y = \frac{a_X}{A}$$

$$b_Y = Ab_X + B \qquad \textbf{Ec. IV.3}$$

$$c_Y = c_X$$

a_X, b_X y c_X son los parámetros de un ítem en la escala X; a_Y, b_Y y c_Y son los de un ítem en la escala Y. Es conveniente mencionar que el parámetro c (azar) es independiente de la transformación. Por tanto, si un modelo TRI ajusta con los datos, llevando a cabo esta transformación lineal de la escala se obtendrá el mismo ajuste, siempre que se transformen los parámetros de los ítems (Kolen & Brennan, 2004).

Elegir el modelo psicométrico para el análisis de las respuestas de los sujetos es otro de los factores sobre los que debe tomarse una decisión en el proceso de equiparación, es decir, transformar las respuestas a los ítems en una escala que refleje el constructo evaluado. Los trabajos consultados estudian las diferencias principalmente entre dos metodologías de escalamiento de las puntuaciones. Por un lado, los modelo de TRI que consideran la respuesta del estudiante a un ítem del test como una función probabilística que incluye información sobre la habilidad latente del sujeto y las características de los ítems. Y, por otro, el método Thurstone que utiliza el número de respuestas correctas en el test para construir una escala normalizada.

Briggs, Weeks y Wiley (2008) y Briggs y Weeks (1997) estudian el efecto que tienen sobre la elaboración de una escala vertical y la interpretación del crecimiento la utilización de modelos TRI de uno y tres parámetros. Los resultados muestran que el modelo de un parámetro produce puntuaciones medias y desviaciones típicas menores que las calculadas con el modelo de tres parámetros. Respecto al crecimiento entre cursos, la tendencia se repite, los tamaños del efecto de ambos modelos psicométricos indican un mayor crecimiento en los grados inferiores y menor distancia en los superiores. Yen (1986), Kolen y Brennan (2004) y Tong y Kolen (2007) comparan los resultados de un modelo TRI de tres parámetros y el método Thurstone, y destacan que con este último método la dispersión de las puntuaciones es mayor en los cursos superiores, de forma opuesta a lo que ocurre con el modelo de tres parámetros.

Los tamaños del efecto son mayores con el método Thurstone, por tanto, muestra un mayor crecimiento entre grados consecutivos. Estas distancias son mayores en los cursos inferiores y se reducen cuando se analizan los cursos superiores, tanto para el modelo de tres parámetros como para el método Thurstone los tamaños del efecto se reducen. Con datos simulados los dos modelos psicométricos tienden a producir resultados similares, esta diferencia con las estimaciones llevadas a cabo con datos reales pueden deberse a la posible violación de los supuestos de los modelos (Tong & Kolen, 2007).

IV.2.2.3 Métodos de calibración

Se incluye, a continuación, una breve descripción de los tres principales procesos de calibración que se pueden llevar a cabo desde TRI. Estos procesos de calibración tienen el objetivo principal de situar las estimaciones del constructo y los parámetros de ítems de diferentes test una escala común. Las tres metodologías descritas (calibración por separado, conjunta y fija) se llevan a cabo de forma similar en el proceso de equiparación horizontal y en el anclaje vertical.

IV.2.2.3.1 Calibración por Separado (CS)

Empleando esta metodología los parámetros de los ítems de las distintos test se estiman por separado. Una vez realizado el proceso se lleva a cabo la trasformación lineal descrita anteriormente en el apartado IV.2.2.2. Para la obtención de las constantes A y B existen diferentes aproximaciones. Por un lado, están los métodos media/media (Loyd & Hoover, 1980) y media/sigma (Marco, 1977) y, por otro, los métodos de curva característica del ítem de Haebara (1980) y Stocking y Lord (1983).

El primero, utiliza la media de los parámetros *a* y *b* estimados en los ítems comunes de dos test X e Y para obtener las constantes de transformación.

$$A = \frac{\mu(a_X)}{\mu(a_Y)} \qquad \textbf{Ec. IV.4}$$

$$B = \mu(b_Y) - A\mu(b_X)$$

Donde μ son las medias de los parámetros de discriminación y dificultad (a y b) de ambas escalas. Situando los valores de Ec. IV.4 en Ec. IV.5:

$$\theta_Y = A\theta_X + B \qquad \textbf{Ec. IV.5}$$

Se obtiene el nuevo rasgo:

$$\theta_Y = \left[\frac{\mu(a_X)}{\mu(a_Y)}\right]\theta_X - \left[\frac{\mu(a_X)}{\mu(a_Y)}\right]\mu(b_X) + \mu(b_Y) \qquad \textbf{Ec. IV.6}$$

El Segundo utiliza las medias y desviaciones típicas de los parámetros b estimados en los ítems comunes de los dos test a equiparar X e Y.

$$A = \frac{\sigma(b_Y)}{\sigma(b_X)}$$
$$B = \mu(b_Y) - A\mu(b_X) \qquad \textbf{Ec. IV.7}$$

Donde σ son las desviaciones típicas de los parámetros de dificultad de los ítems comunes en los test *X* e *Y*, y μ son las medias de esos mismos parámetros. Con esta metodología, el rasgo en la nueva escala es:

$$\theta_Y = \left[\frac{\sigma(b_Y)}{\sigma(b_X)}\right]\theta_X - \left[\frac{\sigma(b_Y)}{\sigma(b_X)}\right]\mu(b_X) + \mu(b_Y) \qquad \textbf{Ec. IV.8}$$

Un problema de estas dos formas de transformación se produce cuando dos curvas características del ítem similares se han estimado por una combinación distinta de los valores de *a*, *b* y *c*. Los métodos basados en estas curvas consideran todos los parámetros de forma simultánea. Hay métodos que minimizan las diferencias entre las curvas características del test de dos grupos como el método Stocking y Lord o entre las curvas características de los ítems como el método Haerbara.

Ambas metodologías utilizan la relación lineal establecida en la igualdad vinculada a la estimación de la probabilidad de responder correctamente a un ítem, en función de sus parámetros, cuando se lleva a cabo la transformación formulada en la siguiente ecuación (Ec. IV.9):

$$P_{ij}\left(\theta_{Yi}\middle|a_{Yj}, b_{Yj}, c_{Yj}\right) = P_{ij}\left(A\theta_{Xi} + B; \middle|\frac{a_{Xj}}{A}, Ab_{Xj} + B, c_{Xj}\right) \quad \textbf{Ec. IV.9}$$

Los subíndices *i* y *j* hacen referencia al sujeto y el ítem respectivamente. El método de Stocking-Lord calcula las diferencias cuadradas entre las curvas características de un test para un valor determinado del rasgo θ_i, que es el sumatorio de las curvas características de los ítems comunes entre formas (ver Ec. IV.10).

$$SLdif(\theta_i) = \left[\sum_{j=1}^{V} P_{ij}\left(\theta_{Yi}\middle|\hat{a}_{Yj}, \hat{b}_{Yj}, \hat{c}_{Yj}\right) - \sum_{j=1}^{V} P_{ij}\left(A\theta_{Xi} + B; \middle|\frac{\hat{a}_{Xj}}{A}, A\hat{b}_{Xj} + B, \hat{c}_{Xj}\right)\right]^2 \quad \textbf{Ec. IV.10}$$

Una vez obtenida la diferencia, se acumula utilizando todas las puntuaciones de los sujetos y los parámetros *A* y *B* se obtienen cuando el proceso iterativo encuentra los valores que minimizan el siguiente criterio:

$$SLCrit = \sum_{i} SLdif(\theta_i) \quad \textbf{Ec. IV.11}$$

El método Haerbara utiliza la suma de las diferencias cuadradas entre las curvas características de cada ítem común para un sujeto con un determinado nivel de rasgo (θ_i), la suma de esas diferencias es:

$$Hdif(\theta_i) = \sum_{j=1}^{V}\left[P_{ij}\left(\theta_{Yi}\middle|\hat{a}_{Yj}, \hat{b}_{Yj}, \hat{c}_{Yj}\right) - P_{ij}\left(A\theta_{Xi} + B; \middle|\frac{\hat{a}_{Xj}}{A}, A\hat{b}_{Xj} + B, \hat{c}_{Xj}\right)\right]^2 \quad \textbf{Ec. IV.12}$$

Estas diferencias se suman con los resultados de todos los casos y la obtención de las constantes *A* y *B* es similar al método Stocking y Lord: la iteración que minimiza ese valor.

IV.2.2.3.2 Calibración Conjunta (CC)

La estimación de los parámetros de los ítems de las distintas formas del test se realiza al mismo tiempo con un modelo TRI multigrupo. Esto se consigue tratando las respuestas de los sujetos en cada aplicación como un grupo distinto y debido a que no todos los ítems deben ser contestados por los estudiantes al

contar con diferentes instrumentos de medida, se tratan como reactivos perdidos por diseño.

La CC es la manera para establecer una escala común en todos los grupos de sujetos que responden a diferentes test, en el momento de llevar a cabo la estimación de los parámetros de esos ítems y de la habilidad (Kolen & Brennan, 2004). Esto quiere decir que las respuestas de dos o más grupos que contestan a formas distintas del test (con ítems comunes) pueden calibrarse de forma simultanea como si hubieran respondido a todos al mismo conjunto de ítems. Con este procedimiento, una única ejecución del software es suficiente para situar en una escala común los parámetros de los ítems y la habilidad del sujeto, sin llevar a cabo más transformaciones a posteriori.

El software utilizado para la calibración debe permitir la opción multigrupo (por ejemplo BILOG-MG) y al contar con ese tipo de ítems perdidos por diseño, se necesita una codificación específica para diferenciarlos de los omitidos.

La calibración conjunta parece generalmente menos afectada por el diseño del bloque de ítems comunes (número de ítems comunes o la variación en sus parámetros de dificultad). Sin embargo, cuando el número de grupos que debe ser anclado es amplio, este tipo de calibración puede producir resultados menos estables e incluso no alcanzar la convergencia. Cuando el continuo en el que se distribuye el rasgo estimado se amplía demasiado, es decir, cuando las diferencias en el rasgo entre los grupos evaluados son grandes, los resultados tampoco son satisfactorios (Chin, Kim & Nering, 2006).

IV.2.2.3.3 Calibración Fija (CF)

El principio básico de la metodología de Calibración Fija es mantener invariantes los parámetros de los ítems comunes entre aplicaciones. En un modelo TRI de tres parámetros se fijan tanto los de discriminación, como los de dificultad y pseudoazar (a, b y c).

Los parámetros de los ítems del test que va a ser utilizado como base de la escala se estiman en primer lugar, de forma separada. Para situar en esa escala el test administrado en la aplicación siguiente, se utilizan los valores estimados para los ítems comunes como referencia en el proceso de anclaje. Estos parámetros

permanecen fijos mientras que los ítems específicos de la nueva forma son calibrados, así los valores estimados se sitúan en la escala del instrumento de medida utilizado como base de la escala.

Esta metodología comparte algunas propiedades de la CC y puede producir resultados de anclaje más estables que las transformaciones en la escala, por ejemplo, eliminando parte del error de equiparación producido por la falta de precisión de las funciones de transformación y, además, tiene en cuenta el parámetro de adivinación para llevar a cabo el cambio de escala (Jungnam, 2007)

El tipo de calibración es otro factor clave tanto en el proceso de anclaje vertical como en el de equiparación horizontal. Kolen y Brennan (2004) señalan que la CC puede ser preferible, en teoría, para llevar a cabo el anclaje vertical ya que utiliza toda la información disponible de los parámetros de los ítems al realizar una única estimación, mientras que la CS necesita varias estimaciones, además de la transformación de los parámetros para situarlos en una escala común. En cambio, en la práctica, la CS puede ser preferible ya que se pueden comparar las estimaciones de los parámetros aplicación a aplicación y, de esta forma, identificar ítems que tienen un comportamiento diferente entre cursos o aplicaciones. Por tanto, con la CS el no cumplimiento del supuesto de unidimensionalidad de la habilidad medida puede causar menos problemas que con la CC. Además la CC puede tener problemas de convergencia al estimar un gran número de ítems en una sola ejecución, incluyendo aquellos que se consideran perdidos por diseño.

El trabajo de Chin, Kim y Nering (2006) con datos simulados encuentran que la CC se encuentra generalmente menos afectada por el número de ítems comunes y su rango de dificultad. En cambio, este método de calibración puede presentar dificultades cuando el número de grupos a anclar es alto y/o las diferencias entre los resultados de las distintas aplicaciones son grandes. Los resultados también muestran que la CC produce estimaciones medias del rasgo más altas que la CS.

Briggs, Weeks y WIley (2008), trabajando con datos reales y un modelo de TRI de tres parámetros, señalan que la CS produce medias y desviaciones típicas mayores que una calibración híbrida que mezcla conjunta y por separado. En

cambio, con un modelo de un parámetro los resultados entre los dos métodos de calibración son similares. Si estudiamos los tamaños del efecto, en general, tienden a disminuir a medida que se avanza en los cursos estudiados. De forma concreta, la CS produce tamaños del efecto mayores que la híbrida, excepto entre los dos últimos cursos evaluados. Estos mismos autores, analizando las trayectorias de crecimiento utilizando un modelo jerárquico lineal para el estudio del VA, encuentran diferencias en las estimaciones producidas por la CS y la híbrida en el caso del modelo de tres parámetros. No existe tanta distancia en las estimaciones de ambos métodos con un modelo de un parámetro.

Jungnam (2007) compara los tres métodos de calibración y concluye que la CC muestra menos crecimiento además de una disminución suave de ese crecimiento grado a grado comparado con la CS y CF. Si se analizan los resultados de los test de resolución de problemas matemáticos, no existe casi variación en la diferencia de medias entre grados consecutivos al comparar los tres métodos de calibración. Únicamente en los cursos inferiores la CC produce diferencias de medias menores que CS y CF, cuyas diferencias de medias son similares. Además, si se comparan las desviaciones típicas producidas por los diferentes métodos de calibración las diferencias son casi inexistentes. Lo mismo ocurre al comparar los tamaños del efecto y las distancias horizontales de los diferentes tipos de calibración.

Kang y Petersen (2009) también comparan los tres métodos de calibración pero con datos simulados y en un contexto de equiparación horizontal. Una característica de este estudio es que incluyen dos tipos de software distinto para llevar a cabo la calibración fija (BILOG_MG y PARSCALE) debido a que existe un funcionamiento diferencial de ambos programas durante el proceso de CF. Durante la calibración PARSCALE actualiza la distribución a priori de la habilidad en cada ciclo del algoritmo EM (esperanza-maximización) empleado para la estimación. Los resultados muestran que la CF con BILOG produce las estimaciones más pobres, con mayor error y sesgo. Otra de las características de esta calibración es el menor tamaño de sus desviaciones típicas, independientemente del número de ítems comunes empleados o del tamaño muestral.

IV.2.2.4 Estimación de la habilidad

Una vez que los parámetros de los ítems se encuentran en una misma escala, ya sea en un solo test o para realizar la equiparación de varios test, se lleva a cabo la calificación (*scoring*) de los sujetos. Es el paso final en el proceso de anclaje vertical, se debe optar por el método adecuado para estimar la puntuación en el constructo evaluado.

El método de estimación más común es Máxima Verosimilitud (MV) y lleva a cabo el proceso utilizando una función de verosimilitud que tiene en cuenta todo el patrón de respuestas del sujeto. El problema con este tipo de estimación es que no produce valores para sujetos que no han respondido correctamente a ningún ítem o que han contestado bien todos los reactivos de la prueba. Las otras dos aproximaciones, Esperada a Posteriori (EAP) y Máxima a Posteriori (MAP), utilizan estadística bayesiana para llevar a cabo la estimación y difieren de la anterior en los supuestos sobre la distribución de la habilidad estimada.

Los procedimientos bayesianos combinan la información que proporciona la función de verosimilitud con supuestos sobre la distribución de la habilidad en la población. Esta distribución asumida es la distribución previa. Combinando esta distribución previa con la información de verosimilitud se construye una distribución ajustada denominada distribución a posteriori. Las dos perspectivas bayesianas difieren en el parámetro que utilizan de la distribución a posteriori. Mientras EAP utiliza la media de la distribución, MAP emplea la moda.

La utilización de estimadores bayesianos, también se denominan BLUP[47], como el EAP o el MAP tienden a contraer la escala ya que las estimaciones se encuentran contraídas en torno a la media de la población. De esta manera introducen menor varianza en las estimaciones y tienden a producir errores medios cuadráticos más pequeños (Tong & Kolen, 2007) que con los estimadores de MV.

Según Jungnam (2007) los métodos bayesianos introducen un mayor sesgo en las estimaciones del rasgo, sobre todo en los extremos de la distribución. En cambio, las estimaciones MV no están segadas en test largos, aunque como se ha

[47]Ver apartado II.2.2 y V.1.2.1 para más información.

mencionado en este mismo apartado, no calculan valores para los estudiantes con ninguna o todas las respuestas correctas.

Tong y Kolen (2007) comparan un total de cinco formas de estimar las puntuaciones del rasgo incluyendo MV, QD, MAP y dos versiones de EAP empleando todo el patrón de respuestas correctas y el total de correctas. Los resultados indican que las estimaciones medias y las distancias horizontales son muy similares entre todos los métodos de estimación pero son MV y QD los que cuentan con mayor variabilidad en sus estimaciones. Respecto a los tamaños del efecto, son los estimadores bayesianos los que muestran más distancia entre grados consecutivos, aunque la tendencia con todas las metodologías es la disminución de esas diferencias a medida que se avanza hacia los grados superiores. Los autores señalan que con datos simulados y cumpliendo los supuestos TRI, los estimadores EAP y QD, producen las estimaciones más precisas.

Jungnam (2007) compara cinco metodologías distintas para la estimación del rasgo de los sujetos. Incorpora en su estudio el método de QD, MV y EAP basados en todo el patrón de respuestas del sujeto y los que denomina pseudo-MV y pseudo-EAP que se basan en el total de respuestas correctas para llevar a cabo la estimación final de la habilidad. El autor no encuentra prácticamente distinción entre las diferencias de medias producidas por los distintos métodos de estimación del rasgo. Únicamente destaca que el método MV empleando una calibración fija produce diferencias de medias mayores entre los cursos 3 y 4 que con las otras metodologías. Respecto a la variabilidad de las estimaciones, de manera general, los métodos EAP y pseudo-EAP producen desviaciones típicas más altas que los métodos MV y pseudo-MV. Con el método QD la dispersión es menor que los métodos MV pero mayor que los EAP. Esta tendencia es más clara utilizando la calibración por separado.

Con respecto a los tamaños del efecto, Jungnam (2007) encuentra que los obtenidos con EAP y pseudo-EAP son similares entre ellos y mayores que los producidos por los métodos MV y pseudo-MV, cuyos valores son similares también. Los estimadores QD producen tamaños del efecto ligeramente inferiores a los métodos EAP pero mayores que los MV. De forma concreta, en los test de matemáticas, utilizando calibración por separado, los tamaños del efecto sufren un

decrecimiento a medida que se el curso incrementa, de forma similar en todas las metodologías de estimación de la habilidad. Finalmente, las distancias horizontales calculadas para grados consecutivos con los distintos tipos de estimación producen resultados similares y siguen esa tendencia decreciente, de la misma manera que los tamaños del efecto, a medida que se incrementa al grado evaluado.

Lee y Ban (2010) estudian las diferencias entre los dos estimadores bayesianos (EAP y MAP) con el de máxima verosimilitud (MV). En su estudio con datos simulados encuentran, como era de suponer, que los resultados de los métodos bayesianos tienden a parecerse entre ellos. Las varianzas estimadas a través de EAP y MAP son menores que aquellas calculadas con MV. Sin embargo el sesgo es mayor utilizando los estimadores bayesianos, como también apuntan Tong y Kolen (2007), llegando a valores inaceptables con formas cortas de test (25 ítems), debido al efecto de encogimiento (*shrunkage*) de las puntuaciones. En cambio, los estimadores MV producen menos error que los bayesianos. Los autores recomiendan la utilización de estimadores MV en la práctica, debido a la cantidad de sesgo producida por los estimadores bayesianos cuando las poblaciones a equiparar difieren considerablemente en su habilidad, especialmente con test cortos.

Briggs y Weeks (1997) y Briggs, Weeks y Wiley (2008) comparan únicamente dos métodos de estimación MV y EAP y, de la misma forma que la tendencia encontrada en el resto de estudios mencionados, las estimaciones bayesianas producen datos con una menor variabilidad que la MV. Sus resultados también muestran que el estimador EAP cuenta con tamaños del efecto mayores.

IV.3 Ganancia y crecimiento: datos longitudinales

Los modelos de análisis que utilizan puntuaciones en un único punto temporal y que empelan algunas evaluaciones, son deficientes para observar diferencias entre escuelas porque pueden confundir los aspectos específicos de los estudiantes con los procesos de instrucción y calidad de la enseñanza de las escuelas. Es preferible la utilización de métodos alternativos que estudian el

cambio que se produce en el nivel de conocimiento o habilidad (Tekwe et al., 2004).

Una de las principales características de los MVA es el uso de al menos dos puntuaciones del logro académico de los estudiantes para estimar los efectos de las escuelas o los docentes, ya que busca determinar cuál es la aportación que hacen al cambio que se produce en el rendimiento de los estudiantes durante un periodo determinado en el que se encuentran bajo la influencia del centro educativo. Esa aportación debe estar libre de otros factores contextuales del estudiante y de la propia escuela, que pueden estar relacionados con el logro pero que no son controlables por parte de los agentes educativos evaluados.

El cambio puede medirse de dos maneras distintas, en forma de ganancia o de crecimiento. La ganancia se calcula utilizando únicamente dos puntuaciones del rendimiento de los estudiantes (pretest-postest), mientras que para estimar el crecimiento se necesita más de dos medidas del logro educativo. Los MVA pueden utilizar tanto medidas de ganancia como de crecimiento para conseguir su propósito.

Las medidas de ganancia[48] pueden utilizarse de tres formas en los VA:

- Calcular el incremento entre las dos mediciones de rendimiento y utilizar esa puntuación como variable criterio en los análisis del VA (ganancia bruta).
- Utilizar la puntuación del postest como variable criterio y el pretest como principal covariable en el modelo (ganancia residual).
- Considerar ambas puntuaciones, pretest y postest, como variables de resultado (ganancia estimada).

Tanto la ganancia bruta como la estimada necesitan que las puntuaciones de las dos medidas de rendimiento se sitúen en una escala común para poder calcular o estimar el cambio. No es necesario en el caso de la ganancia residual ya que el pretest es una covariable que se utiliza para ajustar los resultados del postest.

[48]Una descripción más detallada de los modelos se lleva a cabo en el capítulo V, concretamente en el Apartado V.2.2.

Utilizar más de dos mediciones de rendimiento (datos longitudinales) para analizar el cambio es una alternativa a los análisis de la ganancia. Los MVA analizan los datos longitudinales desde dos perspectivas distintas:

- Mediante modelos multinivel longitudinales que estiman el estatus inicial y una pendiente de crecimiento a lo largo del tiempo. El VA se estima como un residuo asociado a las escuelas que también depende de esa función temporal (Bryk & Raudenbush, 2002).
- Mediante modelos lineales mixtos que estiman tantos coeficientes y residuos como aplicaciones de medida. El VA se estima en términos de ganancia entre dos aplicaciones consecutivas (Sanders & Horn, 1994).

Rogosa y Willet (1983) consideran que la medida de ganancia es una medida fiable para estimar el cambio individual, pero estos modelos no son tan efectivos para proporcionar información sobre los efectos que determinados predictores individuales o de las escuelas pueden tener sobre la tasa de ganancia. Se necesitan más de dos mediciones para poder estimar con precisión la influencia de predictores sobre esas tasas de cambio. No obstante, existen VAM que utilizan dos mediciones del rendimiento del estudiante, es decir, comparando el rendimiento actual con el rendimiento previo, principalmente los modelos de ganancia residual que se desarrollan en Reino Unido (Ray, 2006; Meyer, 1997; Demie, 2003).

Los modelos de ganancia tienen propiedades estadísticas problemáticas porque los ajustes hechos para la variación entre escuelas con los estudiantes es débil (OCDE, 2008). Además, si se introducen predictores del cambio los modelos con dos únicas tomas de datos pierden fuerza (Willett, 1989a; Willett, 1994). Los diseños longitudinales, al permitir evaluar la trayectoria del crecimiento de los alumnos durante un periodo de tiempo, son considerados por algunos autores como los más adecuados para evaluar el progreso de los alumnos y la eficacia de las escuelas (Singer & Willett, 2003; Stevens & Zvoch, 2006; Thum, 2009). Este tipo de modelos están ligados a la construcción de escalas longitudinales de rendimiento capaces de medir el crecimiento de los estudiantes a lo largo de un periodo de tiempo y requiere el escalamiento de las puntuaciones de logro, como las escalas verticales mencionadas en este capítulo.

Los modelos de crecimiento se asemejan a los MVA pero, normalmente, asocian esa trayectoria de cambio al nivel de los estudiantes y no al de las escuelas o profesores. Los MVA son una variación de los de crecimiento, es decir, cuando el crecimiento se asocia al nivel de la escuela o del docente y se analizan la aportación que realizan a este cambio en el logro académico de sus estudiantes, independientemente de factores ajenos al control escolar. El crecimiento se estudia considerando un nivel de logro determinado o tomando una referencia normativa.

IV.3.1 ¿Por qué utilizar una medida de crecimiento?

El análisis del cambio es un fenómeno complejo que puede separarse en dos fases (Coleman, 1975). La primera es la etapa en la que se analiza el cambio individual, es decir, a través de diferentes tomas de datos a lo largo de un periodo de tiempo se estudia el crecimiento en el aprendizaje de un individuo. Pero existe una segunda fase, que Coleman considera de mayor relevancia para la investigación del cambio, consiste en relacionar las diferencias en el crecimiento individual con diferentes características de los alumnos y su entorno y de las escuelas.

Los estudios de cambio no comenzaron con el análisis longitudinal del crecimiento sino que, en un primer momento, se empleaban dos únicos puntos temporales para la medida del rendimiento, un pretest y un postest, es un estudio de la ganancia (Willett, 1989a; 1994). En estudios más recientes, la utilización de más de dos mediciones de logro para una mejor estimación y modelización del cambio se ha impuesto a las anteriores (Thum, Easton & Luppescu, 1998; Zvoch & Stevens, 2003; Bryk, Singer & Willett, 2003; Stevens & Zvoch, 2006). A pesar de esta mejor consideración de los modelos de crecimiento, la utilización de estas técnicas de análisis de la ganancia o crecimiento dependerá del tipo de datos del que se disponga y contar con dos o más mediciones del rendimiento está ligada a las características y el diseño de la evaluación.

Willett (1989a) considera las distintas metodologías para la estimación de la ganancia (ganancia bruta, residual y estimada) métodos tradicionales para medir el cambio. El autor también afirma que estos tres tipos de medida de la ganancia pueden llegar a ser buenos estimadores del cambio intra-individual. Pero

si se pretende asociar la ganancia con predictores del cambio o vincularla a otro nivel de estudio como las escuelas, estas medidas del cambio pierden su fuerza. Dos únicas tomas del rendimiento académico no son suficientes para explicar la trayectoria de aprendizaje de un alumno.

Los diseños de dos mediciones están lejos de ser una estrategia óptima para la recogida de datos que pretendan estudiar la evolución del aprendizaje individual, esto es debido a que proporcionan solo una mínima información sobre el crecimiento. Los modelos que utilizan dos únicas ocasiones de medida son mejores que aquellos que utilizan una pero no mucho mejores (Rogosa, 1995). Los modelos con más de dos tomas de datos tienen dos ventajas fundamentales (Willett, 1997):

A. Un mayor número de ocasiones de medida aumenta la fiabilidad de la estimación del crecimiento individual.
B. Una mayor flexibilidad para estimar el crecimiento que permite al investigador probar diferentes modelos de crecimiento distintos al lineal.

Una de las aproximaciones al análisis longitudinal son las medidas de crecimiento. En ella cada trayectoria de crecimiento individual se representa matemáticamente por un modelo de crecimiento que describe el estatus verdadero como una función en el tiempo. Esta perspectiva inspecciona las trayectorias individuales de crecimiento permitiendo, entre otros aspectos metodológicos, identificar el modelo matemático de crecimiento, imposible de averiguar con solo dos mediciones donde únicamente puede ser lineal, es decir, no es posible estimar la forma de la función de crecimiento (Stevens & Zvoch, 2006).

Otra de las cuestiones relacionadas con estos modelos de medidas múltiples es el número de ocasiones en las que debe evaluarse al alumno. Tres o cuatro mediciones, con suficiente espacio temporal entre ellas, pueden capturar la forma y dirección del cambio en el aprendizaje. No obstante, si la trayectoria individual de cambio es muy compleja, serán necesarias más ocasiones de medida con menor tiempo entre ocasiones de medida (Willett, 1994).

La selección del punto de partida es un aspecto importante en los modelos que utilizan medidas múltiples del logro académico. La utilización de un determinado momento u otro como referente inicial puede tener efectos en la evaluación y la interpretación de los modelos, así como una influencia en la estimación del crecimiento inversamente proporcional al número de ocasiones de medidas con las que se cuente (Stevens & Zvoch, 2006), es decir, está decisión sobre el punto de partida tendrá una menor influencia sobre las estimaciones a medida que el número de tomas de datos sea mayor.

Rogosa (1995) muestra como la variación en la elección de este punto inicial puede cambiar el sentido y el valor de la correlación entre el estatus inicial y el crecimiento, que puede cambiar de positiva a negativa.

IV.3.2 Importancia de la relación entre estatus inicial y crecimiento

La posible relación entre el estatus inicial y la puntuación de cambio ha sido un aspecto de controversia metodológica en el campo de los modelos que tratan de estimar la ganancia o trayectoria de crecimiento (Rogosa, 1995; Willett, 1997; Seltzer, Choi & Thum, 2002). El sentido de esa relación puede dar lugar a dos fenómenos distintos. Si la correlación es positiva los sujetos con puntuaciones iniciales altas tenderán a crecer más y se denomina efecto mateo. En cambio, si es negativa, los sujetos con resultados más bajos tendrán tasas de cambio más altas que aquellos con puntuaciones elevadas, este fenómeno es conocido como efecto de regresión hacia la media (ERM en adelante).

Rogosa (1995) demuestra que el ERM solo es posible cuando existe una correlación negativa entre estatus inicial y cambio. El mismo autor también estudia como el valor de la correlación está influido por la elección de un determinado punto de partida y, por tanto, puede ser un artefacto del diseño. Lo mismo opina Willet (1994; 1997) y se pregunta por qué se debería esperar un valor concreto para la relación entre estos dos parámetros si la manera en que la gente cambia puede producirse de maneras distintas, es decir, esa relación es consecuencia de la propia historia de crecimiento, es un hecho inevitable de la vida.

El término regresión hacia la media *(regression to the mean)* fue detectado en primera instancia por Francis Galton (1886) mientras estudiaba la relación

entre las alturas de padres e hijos. Cuando examinó a los padres con altura superior a la media, encontró que sus hijos tendían también a ser más altos que el promedio pero se acercaban más a la altura media que lo que lo estaban sus padres. Observó el mismo fenómeno con padres con una altura por debajo de la media, sus hijos también tenían una altura por debajo de la media pero más cercana a la altura media de la población.

Este efecto se ha estudiado principalmente con la utilización de puntuaciones de ganancia. El ERM se produce con la simple regresión de un pretest sobre una puntuación de postest y ocurre cuando dos variables están correlacionadas de forma imperfecta (Healy & Goldstein, 1978; Doran, 2003).

Este efecto se define como la distancia de la línea regresión hasta la línea de correlación perfecta. Y como la línea de regresión no puede nunca ser igual que la de la correlación perfecta es inevitable encontrarse con este efecto (Campbell & Kenny, 1999). Por tanto, Los análisis pretest-postest o de dos mediciones son susceptibles al ERM que puede afectar a la validez de las inferencias sobre el rendimiento de las escuelas (Healy & Goldstein, 1978; Rocconi & Ethington, 2006). Sin embargo, los modelos de crecimiento pueden paliar este tipo de efecto (Stevens & Zvoch, 2006).

Nesselroade, Stigler y Baltes (1980) destacan que los diseños con dos únicas mediciones pueden maximizar los problemas asociados al ERM. En los modelos con más de dos ocasiones de medida esos efectos dependerán de los patrones de correlación que se produzcan entre mediciones. Si el patrón de correlaciones es constante no se debería esperar ese efecto más allá de la segunda ocasión de medida. En cambio, un patrón decreciente en las correlaciones puede ser indicador de ERM a lo largo de todas las mediciones. Finalmente, si los valores de la correlación aumentan a lo largo de las aplicaciones se produce un fenómeno que los autores denominan egresión desde la media. También destacan la existencia de factores que pueden afectar al estatus inicial y, por tanto, al ERM, por ejemplo, los procesos de selección en los centros[49], una menor fiabilidad de la puntuación

[49]La distribución de los estudiantes en los centros no suele ocurrir de forma aleatoria. Determinados factores como la cercanía del centro, nivel socioeconómico familiar, etc. pueden determinar la población de una escuela concreta. Por tanto, pude ocurrir que algún centro educativo acumule estudiantes con un determinado nivel de rendimiento, si esos niveles se sitúan

inicial o, sobre todo en modelos longitudinales, el rasgo medido puede ser inestable o cambiante a lo largo del tiempo. En estos casos, se recomienda utilizar la primera ocasión de medida como elemento de control y la segunda medición como punto de partida.

Para algunos autores el principal causante de este tipo de artefacto estadístico es el error de medida (Healy & Goldstein, 1978; Rogosa, 1995; Willett, 1997; Ladd & Walsh, 2002) tanto del pretest como del postest, que puede determinar esa tendencia hacia la media de las puntuaciones más extremas. Los estudiantes con una puntuación alta en el pretest pueden experimentar un descenso en su puntuación del postest simplemente porque este postest puede tener asociado un error aleatorio positivo y alto. La situación opuesta puede darse en aquellos estudiantes con una baja puntuación en el pretest, que pueden experimentar un cambio por encima de la media.

Rogosa (1995) también destaca que si las puntuaciones observadas no son estimadores fiables para calcular la ganancia verdadera, la correlación entre los valores observados de estatus inicial y cambio también son una estimación pobre de la correlación verdadera, que se encuentra sesgada negativamente.

Existen diferentes trabajos que tienen el objetivo de paliar los efectos de esa relación entre estatus inicial y cambio, sobre todo cuando el efecto de la relación se asocia al descrito ERM. La ganancia residual o regresión pretest-postest ha sido una de las formas utilizadas para obtener estimaciones del cambio no relacionadas con el estatus inicial, tratando de describir el cambio verdadero que obtendría un sujeto si todos parten del mismo punto inicial. Otro método para tratar el fenómeno de ERM en los modelos pretest-postest es el propuesto por Rocconi y Ethington (2006) siguiendo las indicaciones de Roberts y Ma (2006). Los autores recomiendan, después de encontrar una correlación negativa entre cambio y estatus inicial, llevar a cabo un ajuste en la puntuación inicial como muestra la siguiente ecuación (Ec. IV.13)

$$x' = x + (1 - r_{xx})(\mu - x) \qquad \textbf{Ec. IV.13}$$

en los extremos el centro puede experimentar un cambio en la siguiente ocasión de medida debido situándose más cerca de la media.

Donde x′ la puntuación inicial ajustada; x es la puntuación inicial observada; r_{xx} es la fiabilidad test-retest; y μ es la media global en ese estatus inicial.

Este efecto, por tanto, puede ser un problema en aquellos diseños que tratan de compensar las diferencias iniciales entre grupos no equivalentes o en los estudios que tratan de comparar las diferencias en la ganancia de grupos que tienen valores iniciales muy diferentes, como podría darse en los datos de un grupo de escuelas. En estos casos es posible que se estimen efectos de las escuelas como distintos al estándar establecido cuando realmente no existen o incluso cambiar la dirección de los efectos.

Cuando se utilizan más de dos ocasiones de medida el ERM es más complicado de identificar. La técnica de análisis de regresión multinivel con la que normalmente se elaboran modelos longitudinales de crecimiento, trata de minimizar los efectos que puede tener la falta de aleatorización de los sujetos en las escuelas. Un modelo multinivel con coeficientes aleatorios estima los residuos de las escuelas utilizando los mencionados estimadores bayesianos BLUP[50], que tienden a suavizar las estimaciones de los grupos con valores extremos hacia la media empleando la fiabilidad de los estimadores como elemento de ponderación. De esta forma, el ERM puede estar provocado por este tipo de estimación de los efectos de las escuelas (Armein-Beardsley, 2008, Sanders & Wright, 2008). En los centros con poca muestra de estudiantes, si han obtenido puntuaciones extremas en la primera ocasión de medida, tenderán a situarse en los valores medios por el efecto de encogimiento.

Otro factor que puede confundirse con el ERM en los análisis de crecimiento es la propia escala de medida. En las escalas verticales se pueden producir los denominados efectos suelo y techo, es decir, estudiantes con puntuaciones muy bajas tenderán a crecer en las siguientes mediciones y, en el lado opuesto, los estudiantes que obtienen una puntuación muy alta al comienzo de la evaluación tienen menos posibilidades de crecimiento.

Para lidiar con este posible artefacto del diseño en los modelos de crecimiento los autores Marsh y Hau (2002) prueban, bajo condiciones simuladas, modelos multinivel longitudinales con cuatro mediciones del rendimiento. Estos

[50]Ver apartado II.2.2 y V.1.2.1 para más información.

análisis estiman un estatus inicial y una pendiente de crecimiento como parámetros principales y consideran los efectos de las escuelas como aleatorios. Construyen tres modelos distintos de crecimiento: un modelo de dos niveles (estudiante y escuela) utilizando las diferentes puntuaciones pretest como covariables, es un modelo de ajuste de covariables jerárquico; un modelo estándar de curva de crecimiento con tres niveles (tiempo, estudiante y escuela); y un tercer modelo similar al anterior pero empleando la primera medición como covariable, por lo que se cuenta con una ocasión de medida menos en la función de tiempo. El segundo modelo tiene cuatro puntos temporales (tiempo=0,1,2,3) y el tercero tres (tiempo=0,1,2). Con los datos simulados no se espera encontrar cambios en la varianza y las puntuaciones entre ocasiones de medida pero los autores detectan que el modelo estándar de crecimiento (modelo 2) introduce cierta varianza que atribuyen al ERM. Concluyen que el modelo que utiliza el rendimiento previo como covariable principal no produce este efecto y resulta más adecuado para estimar los efectos reales de las escuelas.

Castro, Ruiz y López (2009) analizan este efecto vinculado al análisis del VA con modelos multinivel de crecimiento. Las autoras proponen introducir el rendimiento inicial de forma ajustada, definido como la distancia entre la puntuación bruta del estudiante en la primera ocasión de medida y la media global en esa misma medición en función del tiempo (ver Ec. IV.14), como principal predictor de la pendiente de crecimiento. Un coeficiente negativo y significativo de este parámetro señalaría la presencia del ERM en los datos.

$$Y'_{1i} = (Y_{1i} - \overline{Y}_1)(t - t_0) \qquad \textbf{Ec. IV.14}$$

Donde Y_{1i} es la puntuación inicial del estudiante *i*, $\overline{Y}_1$ es la media global de todos los sujetos en la primera ocasión de medida; y $t - t_0$[51] es la función vinculada al tiempo. Con este parámetro se introduce un "término de interacción entre niveles (estudiante y tiempo) que cuantifica la relación entre el nivel inicial del alumno y el paso del tiempo. Además, muestra la tasa de crecimiento del

[51]La notación $t\text{-}t_0$ se incorpora cuando se considera la misma distancia entre ocasiones de medida. Por ejemplo, si se realizaron cuatro mediciones de rendimiento desde un hipotético 5º curso hasta 8º, el primer valor de la función de tiempo es 5-5=0; Así se establece el punto inicial, el siguiente 6-5=1, etc.

alumno cuando se incrementa su nivel inicial un punto con respecto al conjunto de centros" (pág. 117).

El mismo término de ajuste es utilizado en otros trabajos pero con una finalidad distinta: comprobar la relación entre estatus inicial y crecimiento dentro de las escuelas y verificar si determinadas poblaciones de estudiantes dentro de los centros poseen diferentes ritmos de crecimiento y que características pueden estar determinando esos ritmos distintos (Seltzer, Choi & Thum, 2002; Choi, Seltzer, Herman & Yamashiro, 2007).

Por tanto, el tipo de relación existente entre el estatus inicial y cambio puede influir en las estimaciones finales del VA de las escuelas. El tipo de modelo empleado para esa estimación puede hacer variar esa relación. Elegir un punto inicial determinado, incluir más de dos tomas de datos en los análisis o introducir el rendimiento previo como covariable son algunas de las alternativas para tratar de paliar ese efecto.

IV.4 ¿Efecto causal o medida descriptiva?

Los modelos de evaluación basados en la rendición de cuentas, que toman decisiones sobre las escuelas en función de los resultados de VA, tratan de dotar con carácter causal a las estimaciones. Asumir esta relación causa-efecto entre los resultados de VA y el trabajo en los centros es un aspecto discutido. El debate gira en torno a si el VA tiene un significado causal o simplemente representa una medida descriptiva (Rubin, Stuart & Zanutto, 2004; Ballou, Sanders & Wright, 2004; Reardon & Raudenbush, 2008; Kane & Staiger, 2008; Briggs, 2008; Koedel & Betts, 2009; Rothstein, 2009).

Estimar los efectos que un profesor o una determinada escuela tienen sobre el aprendizaje a partir de los resultados de los estudiantes y utilizarlos para tomar decisiones sobre los agentes evaluados, sobre todo, para aplicar sanciones o premios, conlleva la dotación de sentido causal a los resultados. Briggs (2008) afirma que esa interpretación causal depende de dos factores:

- La naturaleza de la intervención educativa subyacente que está siendo parametrizada en el modelo e evaluación.

- El uso destinado a los resultados estimados.

El mayor inconveniente en este aspecto es que las relaciones causa-efecto normalmente están asociadas con estudios experimentales y generalmente requieren que los sujetos, en este caso estudiantes, estén distribuidos aleatoriamente en los diferentes tratamientos.

En el caso de los sistemas educativos, los tratamientos son las diferentes escuelas, docentes o ambos, dependiendo de la unidad que interese analizar. Sin embargo, en los sistemas educativos actuales es poco probable que los estudiantes se distribuyan de forma aleatoria. Factores geográficos y de coste son los dos grandes determinantes de la elección de una determinada escuela por parte de los padres.

Las escuelas pueden atender a poblaciones de estudiantes con diferencias sustanciales en capacidad y contextos familiares. Con esta situación es muy difícil saber si los efectos escolares estimados con un modelo determinado son producto de los centros educativos o se deben a factores de contexto. Por esta razón, la simple comparación de escuelas en términos de puntuaciones medias o de ganancias medias en test puede ser errónea.

En consecuencia, los resultados no podrán atribuirse directamente a las escuelas si existen otros factores que están enmascarando los resultados reales. Por ejemplo, puede ser que una determinada escuela solo atienda a estudiantes con un nivel socioeconómico familiar alto y, por tanto, unos buenos resultados pueden estar motivados por ese efecto y no por la calidad de la enseñanza que se produce en ella.

Debido a esta carencia de aleatorización, en algunos casos, los resultados de VA se obtienen llevando a cabo ajustes estadísticos utilizando variables cuantificables e identificables (sexo, condición de inmigrante, nivel socioeconómico, etc.) con el objetivo de nivelar las posibles diferencias iniciales entre los grupos de estudiantes o entre los grupos de escuelas (titularidad, porcentaje de alumnado inmigrante, etc.). Procediendo de esta forma, se pretende evitar el posible efecto de variables que perjudican a la precisión de las estimaciones de los resultados de las escuelas o los docentes. Los análisis del VA

"*tratan de capturar las ventajas de los experimentos aleatorios cuando no se han llevado a cabo*" (OCDE, 2008, pág. 108). Sin embargo, incluso cuando se realizan los ajustes estadísticos no es suficiente para apoyar las reclamaciones causa-efecto del rendimiento de las escuelas o los docentes (Wiley, 2006).

Para poder llevar a cabo relaciones causa-efecto, los estudios deben cumplir ciertos supuestos. Condiciones que solo es posible asumir completamente en estudios experimentales puros:

- Los estudiantes asignados a los diferentes tratamientos, en el caso de los MVA, profesores o escuelas, deben tener la misma probabilidad de asistir a cada uno de ellos y, en consecuencia, de obtener los resultados de rendimiento en cualquiera de las diferentes escuelas. Por tanto, los estudiantes deberían estar aleatoriamente asignados en las diferentes escuelas para que cada una de ellas cuente con una mezcla de alumnos con características similares (Rubin, Stuart & Zanutto, 2004; Reardon & Raudenbush, 2008; Martineau, 2009)
- El valor estable en la unidad de tratamiento (STUVA) es otra de las condiciones. Todos los estudiantes asignados a una escuela deben recibir los mismos estímulos o tratamiento y que no haya interferencia entre los estudiantes, es decir, el resultado de cada estudiante debe ser independiente de los resultados obtenidos por el resto de sus compañeros o por otros estudiantes del sistema (Rubin, Stuart & Zanutto, 2004; Reardon & Raudenbush, 2008)
- Los datos perdidos pueden influir en las estimaciones. En los estudios longitudinales los datos perdidos pueden llegar a ser un problema y afectar a la estimación del VA. Estudiantes que cambian de centro, sobre todo si se pretende observar la transición de primaria a secundaria, alumnos que no asisten el día de la aplicación de la prueba, etc. son factores que pueden provocar la pérdida de datos.
- La comparación causal de escuelas se lleva a cabo estudiando las medias globales de diferentes unidades de análisis y buscando diferencias en sus resultados. Y para hacerlo es necesario cuantificar esas diferencias. En consecuencia, se necesita que las distribuciones

de las puntuaciones que provienen de los test se encuentren en una escala de intervalo (Reardon & Raudenbush, 2008; McCaffrey, Lockwood, Koretz & Hamilton, 2003; Martineau, 2009)

- Los análisis han de incluir todos los factores que pueden influir en el constructo evaluado para poder aislar los efectos de las escuelas o los decentes. Sin embargo, no es posible demostrar de forma concluyente que todos los factores importantes han sido tenidos en consideración (Martineau, 2009). Por un lado, es prácticamente imposible controlar todas aquellas variables de contexto que influyen en la estimación de los efectos de las escuelas o los profesores (McCaffrey, Koretz, Louis & Hamilton, 2004). Y por otro, si los efectos estimados para una escuela están relacionados con alguna variable, es decir, son heterogéneos, y esta variable no se incluye en el modelo, las estimaciones pierden fiabilidad (Reardon & Raudenbush, 2008)

Rubin, Stuart y Zanutto (2004) señalan que los análisis de los efectos escolares no están estimando cantidades causales, excepto bajo supuestos extremos e irrealistas. Estos modelos no deberían ser vistos como la estimación de efectos causales de las escuelas o los profesores, más bien proporcionan medidas descriptivas. Martineau (2009) también señala los problemas que conllevan las atribuciones causales de los resultados de VA y señala que es más válido interpretarlos de forma descriptiva.

Este problema se acentúa más cuando las estimaciones de VA se utilizan para medir la eficacia de los docentes y tomar decisiones de alto impacto (*high stakes)*. Dentro de un determinado centro educativo los diferentes grupos, es decir, las aulas, pueden organizarse en función de determinados factores como la habilidad de los propios estudiantes. Si el objetivo del VA es dar incentivos a los profesores más eficaces, o sea, que producen un mayor VA, este factor podría sesgar totalmente los resultados.

Trabajos recientes tratan de probar la influencia de esta falta de asignación aleatoria en los resultados de VA, observando el posible sesgo que pueden producir. Desde una perspectiva más económica del análisis del VA, Rothstein (2009) afirma que es peligroso interpretar las estimaciones de VA como efectos

causales. La falta de aleatoriedad, es la que produce esta situación. El autor afirma que, bajo el supuesto de que los modelos estadísticos pueden absorber los efectos de esta distribución no aleatoria de estudiantes, es posible paliar el sesgo que produce. Si se incluyen en los modelos ciertos elementos de control de los procesos de asignación de estudiantes en las aulas puede evitarse parte de ese sesgo, aunque depende de dos factores:

- si es posible observar los factores que determinan la asignación (rendimiento previo, por ejemplo). Rothstein utiliza modelos de ganancia para llevar a cabo sus pruebas
- si estas variables correlacionan con los términos de error y si pueden variar con el tiempo.

Koedel y Betts (2009) analizan este problema de falta de aleatorización en modelos de crecimiento y observan una reducción del sesgo en las estimaciones de VA obtenidas por algunos modelos. Según los autores, un modelo de efectos fijos para los estudiantes y que evalúa docentes durante tres años consecutivos reduce el sesgo producido por la asignación en las estimaciones de los efectos de los docentes.

Por su parte, Kane y Staiger (2008) utilizan los efectos ajustados de los docentes estimados en estudios no experimentales previos para compararlos con los obtenidos en un experimento donde hubo una asignación aleatoria de los profesores en diferentes grupos de estudiantes. Concluyen que los efectos de modelos que controlan las puntuaciones previas y características de los estudiantes son un buen predictor de los resultados posteriores, explicando la mitad de la variación de los efectos de los docentes en los experimentos. Pero disminuye o, como dicen los autores, se desvanece en las siguientes evaluaciones.

En definitiva, en algunas evaluaciones, lo que se intenta es dotar a las puntuaciones de VA con carácter causal, en otras palabras, la diferencia entre las aportaciones estimadas de dos escuelas se interpreta como un reflejo de las diferencias en su eficacia para hacer progresar el aprendizaje de sus estudiantes. Si realmente fuera posible aislar esta contribución se dispondría de una base sólida sobre la que poder tomar decisiones sobre los centros. No obstante, esto solo sería posible con diseños experimentales puros, donde los sujetos se distribuyen de

forma aleatoria en las escuelas y aunque los análisis del VA intenta paliar esa falta de aleatoriedad, llevar a cabo dicha inferencia causal puede ser un aspecto problemático. Las estimaciones de VA han de ser interpretadas como medidas descriptivas y, como señala Linn (2008), estos resultados siguen teniendo valor sin hacer inferencias causales respecto a la calidad de las escuelas y pueden utilizarse como indicadores para identificar escuelas que requieren una investigación con mayor profundidad.

IV.5 Contextualización

Los resultados de VA llevan implícito que los efectos estimados para las escuelas o los docentes están libres de las posibles variaciones que pueden producir factores ajenos al control escolar, como el contexto socioeconómico familiar de los estudiantes o los que los propios alumnos ya saben antes de comenzar con la evaluación. Es decir, se intenta aislar el efecto de estos elementos educativos (escuelas o profesores) del resto de posibles variables que determinan el aprendizaje de un estudiante.

La introducción de predictores de contexto en los modelos de VA es un aspecto discutido y que puede afectar a las estimaciones finales (Ballou, Sanders & Wright, 2004; Hibpshman, 2004; Tekwe et al., 2004; Keeves, Hungi & Afrassa, 2005; Choi, Goldschmidt & Yamashiro, 2006; Lockwood et al., 2007; Haegeland & Kirkeboen, 2008; Ferrão, 2009).

Sanders, Saxton y Horn (1997) argumentan que una de las ventajas de un sistema basado en el análisis de las ganancias de los estudiantes es que no necesita incorporar covariables en el modelo porque cada estudiante, por decirlo así, ejerce un control sobre sí mismo. Stevens y Zvoch (2006) también comparten este argumento, cuando el modelo de crecimiento está basado en el seguimiento del rendimiento individual del estudiante, las características del alumno son estables y permanecen constantes a lo largo del tiempo y no pueden confundir la estimación del crecimiento. Por otro lado, Raudenbush y Bryk (2002) indican que la introducción de ajustes con variables del contexto de los estudiantes son importantes por dos motivos:

- Porque los estudiantes normalmente no están asignados de forma aleatoria en las organizaciones, fallar en el control del contexto puede sesgar las estimaciones de los efectos de las organizaciones.
- Porque si las covariables de los estudiantes tienen alta relación con el constructo estudiado, controlar esos factores incrementará la precisión de cualquier estimación del efecto de las escuelas y la fuerza de los contrastes de hipótesis debido a la reducción de la varianza sin explicar en el nivel de los estudiantes.

Por tanto, la introducción o no de covariables también es un aspecto clave en la construcción de modelos de VA y no existe un acuerdo respecto a si es necesario incluir o no estos predictores en los análisis, ni sobre cuáles son los más adecuados.

De acuerdo con la OCDE (2008), el uso de características socioeconómicas en modelos contextualizados de VA puede tener un impacto negativo en la equidad y eficiencia de la toma de decisiones, sin embargo mucho de esto depende de cómo se utilice la información que proporciona el VA. Por ejemplo, modelos de VA que no utilizan factores contextuales pueden identificar centros con un bajo rendimiento académico y, por tanto, un bajo VA y, dichos centros, podrían estar compuestos de estudiantes con bajo nivel socioeconómico. Sin embargo, en los modelos contextualizados de VA estos mismos centros podrían tener un bajo rendimiento pero un VA cercano a la media, ya que se han extraídos los efectos que las variables de contexto tienen sobre el rendimiento y, en consecuencia, su posición ha cambiado. Esto podría favorecerles o perjudicarles dependiendo de con qué objetivos se usen los resultados. Si se diera financiación extra a aquellos centros con un bajo VA para aumentar su rendimiento, si se utilizan modelos contextualizados, les perjudicaría porque los resultados no mostrarían nada problemático en estos centros al no diferenciarse significativamente de la media. Si, al contrario, se utilizaran modelos sin contextualizar, les favorecería.

En los modelos de ganancia los factores contextuales tienen una mayor influencia en las estimaciones que en los modelos de crecimiento. Tekwe et al. (2004) compararon cuatro modelos distintos de ganancia, tres sin contextualizar y un modelo contextualizado. Si se analizan las correlaciones entre las diferentes

estimaciones de VA que se obtienen con los distintos modelos no se observa demasiada diferencia entre los modelos sin ajustar, con valores por encima de 0,95. Sin embargo, el modelo ajustado es el que produce estimaciones de VA diferenciadas del resto, con correlaciones entre 0,50 y 0,90 dependiendo de la materia y el grado evaluado.

Estos mismos autores argumentan que si los resultados de VA tienen el objetivo de dirigir los recursos a las escuelas de bajo rendimiento, no se deberían incluir dichas covariables porque distorsionarían los resultados de aquellos centros educativos con mayor proporción de estudiantes con un nivel socioeconómico más bajo. No obstante, si los resultados se orientan a premiar a las escuelas con mayor rendimiento y no se ajustan los modelos, les perjudicaría.

La inclusión o no de covariables del estudiante no es tema sencillo de resolver. Controlar las covariables no es una simple inclusión de predictores en el modelo de regresión. McCaffrey, Koretz, Louis y Hamilton (2004) argumentan que si las covariables están correlacionadas con los efectos de la escuela pueden provocar errores sistemáticos en su estimación. El sesgo se produce porque, con la introducción de las covariables, el modelo atribuye el efecto verdadero de las covariables, y una parte del efecto de la escuela que correlaciona con dicha covariable, al efecto estimado por las covariables e incorrectamente elimina, del efecto de la escuela, la parte que correlaciona con la covariable. Por tanto, el efecto estimado para la escuela solo recibe la parte residual del efecto verdadero que no está correlacionado con la covariable. Por ejemplo, si una determinada escuela con un VA verdadero muy alto tiene a una gran proporción de alumnos que proceden de familias con un alto nivel socioeconómico y utilizamos ese predictor para ajustar el modelo, los efectos de la eficacia de esa escuela se estimarán por debajo de lo que realmente ocurre. Al contrario, excluyendo la covariable del modelo se produce el efecto opuesto. Por este motivo, parece que controlar las covariables del estudiante no es suficiente para eliminar los efectos de las características del contexto.

Lockwood et al. (2007) encuentran una alta estabilidad de las estimaciones cuando comparan modelos que incluyen o no covariables, incluso cuando existe gran variedad de estudiantes con características diferentes en las unidades de

análisis. Afirman que los métodos de análisis del VA mantienen su promesa de aislar los efectos de las características del contexto del estudiante ajenas al control de los profesores o las escuelas. No obstante, no se obtienen resultados similares cuando se utilizan variables de centro a partir de la agregación de los resultados de los estudiantes en los modelos.

Hibpshman (2004) afirma que los VAM que no incorporan covariables de los estudiantes o de las escuelas producen estimaciones que pueden ser fiables para identificar profesores o escuelas en los extremos de una distribución de eficacia, es decir, diferentes de la media, pero podrían estar sesgados a favor de aquellos que trabajan con poblaciones de estudiantes más aventajados. Utilizando datos de varios años solo se puede controlar parcialmente ese problema.

En un estudio empírico Choi, Goldschmidt y Yamashiro (1994) analizan el efecto de la inclusión del rendimiento previo o el nivel socioeconómico (medido a través de la variable ayuda con el comedor escolar) sobre los rankings de escuelas. Prueban tres modelos distintos, uno con cada predictor por separado y otro que incluye ambos. Al analizar los rankings encuentran altas correlaciones entre el modelo que incluye el estatus inicial como predictor y el que incorpora el estatus socioeconómico como principal covariables, con valores de 0,97. Una correlación perfecta entre el modelo que incluye únicamente el nivel inicial como predictor y el que incorpora tanto el nivel inicial como el estatus socioeconómico, indica que el estatus inicial del estudiante captura muchos de los efectos que el nivel socioeconómico intenta medir. Por tanto, controlando el estatus inicial, el modelo recoge los efectos que el nivel socioeconómico puede tener en los resultados de los estudiantes. Sin embargo, el alumnado puede tener otras oportunidades de aprendizajes más allá de las que ofrece la escuela y que no son tenidas en cuenta a través de la vía del rendimiento inicial (Doran, 2003)

Alguno de los MVA incluye un vector de características de los estudiantes entre los que destacan, principalmente, como indicadores del estatus socioeconómico familiar, si se acogen a la opción de comida gratis o a precio reducido (Webster & Mendro, 1997; Ballou, Sanders & Wright, 2004) o la construcción de un índice de estatus socioeconómico (SES) (Sanders & Horn, 1994; Raudenbush, 2004). También la introducción del rendimiento previo de los

estudiantes, incluso en otra materia, es otra de las opciones más utilizadas (Sanders & Horn, 1994; Lockwood et al., 2007). También sexo y raza son variables que se incluyen para modelar la posible heterogeneidad de los resultados. La inclusión de predictores del contexto de las escuelas, además de los factores del estudiante, es otra de las prácticas utilizadas en los análisis del VA (Thum, 2003; Keeves, Hungi & Afrassa, 2005)

No es adecuado asumir que la introducción de covariables en el modelo va a producir resultados sin sesgo. Si el efecto de la escuela no estuviera correlacionado con la covariable, entonces introducirla no sesgaría los resultados. La cantidad de sesgo dependerá de varios factores, como el tamaño de la correlación entre los efectos y las covariables, y entre las covariables y las puntuaciones de rendimiento (McCaffrey, Lockwood, Koretz & Hamilton, 2003). Debe tenerse en cuenta que el principal propósito de la inclusión de predictores de contexto en el modelo es reducir el sesgo de la medidas de rendimiento de las escuelas y poder establecer una relación causal entre los efectos estimados y la ganancia en rendimiento de los estudiantes. Y, en ocasiones, la inclusión de predictores producirá el efecto contrario.

Una posible solución es estudiar la relación entre las covariables y los efectos aleatorios estimados. Si las correlaciones son significativas entonces debemos plantearnos eliminar esas covariables del modelo pero si, de forma opuesta, no se encuentra correlación significativa, su introducción en el modelo no producirá sesgo. Otra opción es utilizar un modelo de efectos fijos para introducir las covariables de los estudiantes y después utilizar las puntuaciones ajustadas para elaborar un modelo de efectos aleatorios (Ballou, Sanders & Wright, 2004).

IV.6 Otras cuestiones metodológicas

En este apartado se mencionan otros de los aspectos que son objeto de estudio cuando se desarrollan MVA en educación y pueden resultar problemáticos. Cuando se desarrolla uno de estos modelos es conveniente llevar a cabo pruebas empíricas para conocer los efectos provocados por la toma de decisiones en diferentes aspectos metodológicos. Además de los mencionados en este capítulo, la

investigación sobre VA también destaca el estudio de los casos perdidos, el sesgo y la estabilidad de las estimaciones de VA y el tipo de efecto de las escuelas (fijos o aleatorios)

A. Datos perdidos

En primer lugar, en los estudios longitudinales es inevitable contar con datos incompletos del rendimiento de los estudiantes. Los datos perdidos pueden llegar a ser un problema y afectar a la estimación de las puntuaciones de VA.

Los MVA son sensibles a la naturaleza de esos datos y el proceso de análisis implementado (McCaffrey, Lockwood, Koretz & Hamilton, 2003; McCaffrey, Koretz, Louis & Hamilton, 2004; Zaidman-Zait & Zumbo, 2005). Los factores que provocan la perdida de datos en este tipo de evaluaciones son los siguientes:

- Estudiantes que cambian de centro, sobre todo si se pretende observar la transición de primaria a secundaria,
- Alumnos que no asisten el día de la aplicación de la prueba por diferentes motivos

En el caso de modelos de evaluación basados en la rendición de cuentas de alto impacto y que penalizan a los centros menos eficaces, los equipos directivos pueden llegar a recomendar la no asistencia de alumnos con un nivel de rendimiento más bajo. Todos estos factores afectan a los datos con los que contará un determinado modelo de VA para llevar a cabo sus estimaciones finales de los efectos de las escuelas o los docentes.

Una posible solución es la utilización de únicamente los datos completos pero esto solo es adecuado si los datos perdidos tienen una distribución aleatoria. Y puede que no ocurra en los estudios longitudinales con datos educativos (Rubin, Stuart & Zanutto, 2004; Martineau, 2009) porque los estudiantes que no realizan las prueban suelen situarse en la parte baja de la distribución del rasgo evaluado, son los estudiantes con peores resultados.

B. Sesgo de las estimaciones de Valor Añadido

En segundo lugar, existen factores que afectan directamente a la precisión de las estimaciones. La utilidad de los resultados obtenidos con un MVA dependerá de la cantidad error que producen las estimaciones estadísticas.

El sesgo es una medida de imprecisión y, por tanto, un estimador estará sesgado si la media de los valores obtenidos a lo largo de muchas repeticiones del estudio no tiende hacia el valor verdadero. Las principales fuentes de error que pueden afectar a las estimaciones finales de los efectos de las escuelas o los docentes son las siguientes:

- El sobreajuste de un modelo con demasiadas variables que se vea afectado por la colinealidad o, de forma inversa, un modelo con predictores que no representan el contexto socioeconómico del alumno (OCDE, 2008).
- El error de medida es otra fuente posible de sesgo. Los supuestos de la teoría clásica de regresión asumen que las variables explicativas están medidas sin error. Por tanto, los modelos de VA en los que se utiliza el rendimiento previo o el nivel socioeconómico como predictor principal, como ocurre en los modelos contextualizados, pueden tener fuentes de error. Sumados al error de estimación que acompaña a la puntuación de rendimiento utilizada como variable criterio. Tener en cuenta el error de medida en los modelos puede reducir el sesgo y el error de las estimaciones y, de esta forma, proporcionar estimaciones más fiables. La consideración de los errores de medida de las covariables en los análisis de VA puede cambiar tanto la magnitud de los coeficientes estimados como sus desviaciones estándar (Hutchison, 2004; Goldstein, Kounali & Robinson, 2008; Ferrão & Goldstein, 2009).
- El error muestral. Está relacionado con la varianza de los parámetros estimados e influye en el grado de dispersión que puede tener la estimación realizada (McCaffrey, Lockwood, Koretz & Hamilton, 2003). La varianza se utiliza para construir intervalos de confianza en torno a la estimación de los efectos de cada escuela que permita

comprobar las diferencias entre el VA de distintas escuelas y con respecto a la media global. Obviamente, si los intervalos de confianza son pequeños se identificarán más fácilmente las escuelas con un VA significativamente por encima o debajo de la media.

C. Efectos escolares fijos o aleatorios

En tercer lugar, el tratamiento de los efectos escolares puede afectar a los resultados (Tekwe et al., 2004; Sanders & Wright, 2008). Asumir los efectos de interés para el VA, es decir, los efectos que producen las escuelas o los docentes en el crecimiento, como fijos o aleatorios[52] conlleva una variación en la forma de estimarlos.

Los modelos de efectos aleatorios asumen que las unidades estudiadas son una muestra de una amplia población de unidades similares pero no observadas y, por tanto, la variabilidad entre las unidades observadas describe la variabilidad en la población. En cambio si los efectos son tratados como parámetros fijos esas unidades observadas son las únicas que interesan y, por tanto, son la población (McCaffrey, Lockwood, Koretz & Hamilton, 2003).

La estimación mediante modelos de efectos fijos utiliza únicamente los datos de los estudiantes de cada escuela para realizar el proceso. De otro modo, cuando los efectos son tratados como aleatorios, la estimación se lleva a cabo a través de los mencionados estimadores bayesianos[53] o BLUP. La característica principal de este tipo de estimación es que utilizan los datos de otras escuelas para estimar cada efecto de una escuela particular, produciendo el mencionado efecto de encogimiento. Y provoca que las escuelas con tamaños muestrales pequeños no se diferencien de la media global.

D. Estabilidad de las estimaciones de Valor Añadido

En cuarto lugar, la volatilidad o falta de estabilidad de los resultados. Este aspecto se encuentra implícito en los de MVA que suponen cambios en el rendimiento de los estudiantes producidos por los efectos de las escuelas.

[52]Más información sobre este aspecto en el apartado V.1.2

[53]Se analiza con mayor detalle en el apartado V.1.2.1 y en II.2.2.

En cierta medida se espera que los centros puedan cambiar sus puntuaciones a lo largo del tiempo. No obstante, si el cambio se produce de forma errática puede estar motivado por algún tipo de sesgo, por ejemplo, pasar de tener VA positivo a tenerlo negativo. Otros factores que pueden afectar a la estabilidad de los resultados son:

- El cambio en la medida utilizada para evaluar el rendimiento de los estudiantes. Diferentes medidas de rendimiento de los estudiantes producen mayor variación en las estimaciones de VA que la utilización de diferentes modelos con una medida de rendimiento determinada (Lockwood et al., 2007; Sean & Monczunski, 2007)
- El tamaño de los conglomerados evaluados (centros o aulas). Aquellos con pocos estudiantes (menos de 20) pueden sufrir los efectos de una gran dispersión en los resultados y, por tanto, la falta de precisión a la hora de estimar diferencias significativas respecto a la media global de VA (Lockwood, Louis & McCaffrey, 2003).
- La introducción de datos de contexto. Es posible que las variables de contexto cambien con el tiempo y, en consecuencia, pueden provocar la inestabilidad de los resultados de VA. Por ejemplo, una decisión política que decide dotar de ordenadores las aulas de los centros puede afectar a las estimaciones de VA durante ese año (Choi, Goldschmidt & Yamashiro, 2006).

Capítulo V: Modelos estadísticos para el análisis del Valor Añadido en Educación

Cuando se habla de Modelos de Valor Añadido (MVA en adelante) se hace referencia a diferentes análisis estadísticos utilizados para estimar las contribuciones de las escuelas o los docentes al crecimiento en aprendizaje de sus estudiantes. Son una familia específica de modelos estadísticos que se emplean para realizar inferencias sobre la eficacia de unidades educativas, habitualmente escuelas y/o profesores. Se caracterizan por poner el foco de atención en los patrones de ganancia de las puntuaciones de los estudiantes a lo largo del tiempo, en lugar de en el estatus en un momento concreto. En particular, tratan de extraer de los datos brutos las contribuciones de las escuelas o los profesores a las trayectorias estimadas (Braun & Wainer, 2006).

Para Lissitz, Doran, Schafer y Willhoft (2006) los MVA son un tipo muy especial de modelos de crecimiento que añaden significado al cambio que se produce en el aprendizaje del estudiante si se compara con una cantidad de cambio esperado o con los cambios que se han producido en otros estudiantes con otras experiencias educativas (p.ej. diferentes profesores, escuelas o distritos) y otros contextos (p.ej. nivel socioeconómico, género o etnia, o rendimiento inicial). Es decir, el MVA es una herramienta estadística que facilita el aislamiento de los efectos de los docentes o las escuelas para poder probar su eficacia, independientemente de otros factores ajenos al centro, y comparando los resultados con un valor de referencia. Las distintas perspectivas se diferencian,

principalmente, en cuestiones concretas relacionadas con el desarrollo de los modelos estadísticos utilizados para conseguir las estimaciones.

La revisión en profundidad de los aspectos metodológicos de cada uno de estos modelos estadísticos ha llevado a muchos autores a cuestionarse la validez de los mismos y a formular problemas vinculados con esta metodología de análisis de los efectos escolares: ¿realmente las estimaciones del VA se pueden interpretar como efectos causales?, ¿deben ajustarse los modelos con la utilización de covariables? o ¿las diferentes medidas de los propios sujetos son suficientes para establecer un control? y ¿es más adecuado utilizar modelos de crecimiento o de ganancia?. Estos son algunos de los principales interrogantes que han surgido con el desarrollo de MVA y se han tratado en el capítulo anterior.

Otros estudios tratan de llevar a cabo un análisis comparativo de los distintos MVA que actualmente se utilizan en evaluaciones generales. Estos trabajos comparan empíricamente los resultados que producen varios de los modelos utilizados con mayor frecuencia en el análisis del VA (Tekwe et al., 2004; McCaffrey, Koretz, Louis & Hamilton, 2004; Choi, Goldschmidt & Yamashiro, 2006; Sanders, 2006; Lockwood et al., 2007), principalmente aquellos que se desarrollan en Estados Unidos para su sistema de evaluación basado en la rendición de cuentas. Otros trabajos realizan comparaciones descriptivas de las características de los distintos modelos (Hibpshman, 2004; Goldschmidt et al., 2005; Wiley, 2006).

Uno de los modelos que mayor número de revisiones ha acumulado es el ya mencionado modelo de Tennessee, que inicialmente tenía el nombre de *Value Added Assessment System* (TVAAS) (Sanders & Horn, 1994; Sanders, Saxton & Horn, 1997) y, en la actualidad se denomina *Evaluation Value Added Assessment System* (EVAAS). También es conocido como modelo estratificado o de capas (*Layered Model*) por considerar los efectos previos de los docentes como permanentes en los cursos posteriores. Es uno de los modelos pioneros en la utilización del VA para referirse a los análisis empleados para averiguar la eficacia de los docentes y las escuelas a partir de las puntuaciones de rendimiento de los estudiantes en varias ocasiones de medida.

El EVAAS es una de las aproximaciones más complejas que se emplea para el análisis del VA, utiliza modelos lineales mixtos multivariados para intentar aislar

los posibles efectos de los profesores y las escuelas en el aprendizaje de los estudiantes. No obstante, no todas las técnicas de análisis del VA alcanzan el mismo grado de dificultad en su análisis estadístico. También existen modelos más sencillos que pretenden lograr el mismo objetivo, como los modelos de ajuste de covariables o la ganancia residual.

V.1. Descripción de los modelos

Existe una gran variedad de aproximaciones al análisis del VA en educación. Los modelos varían en el tratamiento estadístico de la información y la definición del análisis del cambio en aprendizaje, diferenciándose en el grado de complejidad y en los supuestos que subyacen en cada uno de ellos. A pesar de las diferencias, todos comparten la misma necesaria condición: relacionar los cambios en el rendimiento individual del estudiante con los profesores o las escuelas a las que asisten (Wiley, 2006). Ya se han detallado las características principales de los MVA[54], en resumen son las siguientes:

- Son análisis estadísticos que proporcionan medidas cuantitativas del rendimiento de las escuelas, también pueden asociarse a los docentes, los distritos, etc. y sus resultados tienen por objetivo evaluar aspectos del sistema educativo en general o de las escuelas en particular. Por tanto, el VA de las escuelas es una estimación producida por los diferentes modelos estadísticos y se considera la aportación que realizan dichas escuelas sobre el cambio en aprendizaje de sus estudiantes.
- Ponen la atención en el cambio en rendimiento, es decir, utilizan al menos dos puntuaciones de logro de los estudiantes para elaborar los modelos. Ese cambio puede considerarse como una ganancia (dos tomas de datos) o crecimiento (más de dos mediciones).
- Tratan de separar los efectos escolares de otros factores, relacionados con el rendimiento escolar pero que son ajenos al proceso que se produce escuela (principalmente efectos de la familia, compañeros y

[54]Más información sobre los Modelos de Valor Añadido en el Apartado III.3.1.

capacidad individual). De esta forma, el aprendizaje que se produce en los estudiantes puede ser atribuido de manera apropiada a estos agentes educativos. Existen dos grandes perspectivas, ya mencionadas, sobre la contextualización[55] de los MVA, los que consideran que no es necesario (Sanders, Saxton & Horn, 1997; Stevens & Zvoch, 2006) y los que opinan que es un requisito de los MVA (Bryk & Raudenbush, 2002)

Cada una de las diferentes aproximaciones que tratan de analizar el VA en educación difiere en ciertos aspectos relacionados con el desarrollo de los modelos estadísticos para llevar a cabo las estimaciones. Por ejemplo, el tratamiento de la medida del cambio (ganancia o crecimiento), la consideración de los efectos de las escuelas (fijos o aleatorios y anidados o cruzados), la introducción de covariables de contexto (modelos contextualizados o sin contextualizar) o cómo consideran la variable de resultados (univariante o multivariante). También coinciden en otros puntos, normalmente en la utilización de Modelos Lineales Mixtos (MLM en adelante) para conseguir su objetivo. Los modelos de regresión multinivel[56] pueden considerarse una variación de estos.

Un MLM es un modelo lineal paramétrico para datos anidados, longitudinales o para medidas repetidas que cuantifican la relación entre una variable dependiente continua y varias variables predictoras (West, Welch & Gallecki, 2007). Estos modelos pueden incluir parámetros de efectos fijos asociados con una o más covariables y efectos aleatorios relacionados con uno o más niveles de varianza (estudiantes, docentes, escuelas, etc.). Mientras que los parámetros de los efectos fijos describen la relación entre las covariables y la variable dependiente para toda la población, los efectos aleatorios son específicos de grupos o sujetos dentro de una población. En consecuencia, los efectos aleatorios son utilizados directamente en la modelización de la varianza aleatoria de la variable dependiente en los distintos niveles de agrupación de los datos. La mezcla de efectos fijos y aleatorios da lugar al nombre de MLM

[55]Más información en el apartado IV.5.

[56]Los modelos jerárquicos lineales pueden incluirse dentro de este grupo de técnicas de análisis, son una variante. Aitkin y Longford (1986) llevan a cabo una revisión de estos modelos de efectos mixtos para la evaluación de la eficacia de las escuelas, incluyendo los modelos multinivel.

Siguiendo el modelo general de ecuaciones mixtas de Henderson (1975), quedaría formulado de la siguiente manera:

$$Y = X\beta + Zr + \mathrm{e} \qquad \textbf{Ec. V.1}$$

Donde Y es todo el vector de puntuaciones de rendimiento que se incluyen en el modelo con n x 1 puntuaciones observadas para cada sujeto; *X* es una matriz conocida con diseño n x p; β es un vector con p dimensiones que representa los distintos efectos fijos y *r* es el vector de efectos aleatorios con q dimensiones; *r* es una matriz n x q dimensiones asociada a los efectos aleatorios; y *e* es un vector aleatorio no observable n x 1 que representa la variación sin considerar, es decir, el error. Ambos, *r* y *e*, tienen media cero y varianza:

$$Var\begin{bmatrix} r \\ e \end{bmatrix} = \begin{bmatrix} R & 0 \\ 0 & T \end{bmatrix} \qquad \textbf{Ec. V.2}$$

El término residual, en caso de emplearse datos longitudinales, puede asumir una matriz de varianzas covarianzas T que refleje la correlación entre los residuales de las puntuaciones de los estudiantes. Un ejemplo sencillo de MLM, sin efectos aleatorios, sería un ANOVA de medidas repetidas con cuatro tomas de datos, como el de la siguiente ecuación (Ec. V.3)

$$Y = X\beta + \mathrm{e}$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + [e_t] \qquad \textbf{Ec. V.3}$$

Con esta estructura el coeficiente β_0 es la puntuación en la cuarta toma de datos y el resto las diferencias de las puntuaciones de otras aplicaciones con la última. Un modelo de regresión jerárquica con dos niveles (tiempo y estudiante) en el que se comparan, por ejemplo, dos estudiantes, desde la perspectiva de los MLM, quedaría formulado como en la ecuación siguiente (Ec. V.4):

$$\text{Estudiante 1: } \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 1 & 3 & 0 & 0 \\ 1 & 4 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \qquad \textbf{Ec. V.4}$$

$$\text{Estudiante 2: } \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 1 & 2 \\ 1 & 3 & 1 & 3 \\ 1 & 4 & 1 & 4 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

Este modelo estima dos parámetros, el estatus inicial y la pendiente de crecimiento, además de los efectos individuales. En este caso, los dos primeros coeficientes hacen referencia al estatus inicial y a la pendiente de crecimiento (β_0 y β_1). Los otros dos coeficientes representan las diferencias entre los estudiantes en estos dos parámetros. Para el caso del estudiante 1 son iguales a 0, por tanto, los valores β_0 y β_1 son sus puntuaciones, y las del estudiante 2 serían ($\beta_0 + \beta_2$) para el intercepto y ($\beta_1 + \beta_3$) para la pendiente.

Si estos dos estudiantes fueran los únicos de la población objeto de estudio, como en el ejemplo anterior, sus puntuaciones se incluirían como parámetros fijos en el modelo, sin embargo, en educación lo más usual es contar con muestras de una población mucho mayor y se consideran efectos aleatorio, como en la ecuación Ec. V.5:

$$Zr + \mathrm{e} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} r_1 \\ r_2 \end{bmatrix} + [e_t] \qquad \textbf{Ec. V.5}$$

Al introducir parte aleatoria en el modelo se cuenta con otra matriz de datos. r_1 y r_2 representan las diferencias en el intercepto y la pendiente de cada estudiante y los coeficientes fijos β_2 y β_3 no formarían parte de la ecuación, sino que a cada estudiante de la muestra se le estima un residuo que refleja su distancia respecto a las medias globales en estatus inicial y pendiente de crecimiento. Los MVA son extensiones más complejas que pueden añadir un nivel más de agregación para estimar los efectos de los centros o los docentes o incluir otras covariables de contexto. A continuación se detallan esos aspectos que pueden diferenciar a las distintas perspectivas de análisis del VA.

V.1.1 Medida del cambio: ganancia vs. crecimiento

Las puntuaciones de ganancia[57] son más sensibles a posibles artefactos del diseño que los modelos de crecimiento. Por ejemplo, el mencionado efecto de regresión hacia la media (ERM) que puede producirse cuando la relación entre el estatus y el cambio en aprendizaje es negativa (Rogosa, 1995). No obstante, son fiables para medir el cambio en rendimiento del estudiante, sobre todo al trabajar con los errores de medida de las puntuaciones de logro y tratar de estimar la ganancia verdadera (Rogosa & Willett, 1983). A pesar de esta probada fiabilidad no son capaces de trazar la trayectoria de crecimiento en aprendizaje porque únicamente cuentan con dos mediciones. Un modelo con más tomas de datos permite probar diferentes formas de crecimiento de los estudiantes.

Otro factor a considerar en contra de las puntuaciones de ganancia es que no son una herramienta muy potente para el análisis de los predictores de ese cambio (Willett, 1989a; 1994). Los modelos longitudinales superan algunas de las debilidades mencionadas y parecen más adecuados para el análisis de la eficacia de las escuelas (Singer & Willett, 2003; Stevens & Zvoch, 2006; Choi, Goldschmidt & Yamashiro, 2006; Thum, 2009). No obstante, no todos los modelos longitudinales son capaces de medir el cambio, Singer y Willet (2003) señalan tres características en las que coinciden los estudios longitudinales destinados a medir el cambio:

- Más de dos ocasiones de medida.
- Los valores de la variable dependiente deben cambiar sistemáticamente con el tiempo. Las cuestiones relacionadas con la construcción de escalas verticales de rendimiento deben considerarse para cumplir con este aspecto.
- Contar con una medida de tiempo razonable. Son los objetivos de la evaluación los que determinan esa medida de tiempo. Por ejemplo, si se desea evaluar los cambios en función del curso o se incluyen varias mediciones dentro del mismo grado. El espacio transcurrido entre aplicaciones juega un papel importante en la toma de decisiones sobre este aspecto.

[57]En el Capítulo IV (Apartado IV.3) también se habla de este aspecto.

Existen modelos con más de dos ocasiones de medida que no pueden considerarse modelos de crecimiento. Por ejemplo, los que ajustan la variable de resultados con varios predictores del rendimiento previo como ocurre en MVA de Dallas (Webster & Mendro, 1997; Webster, 2005). Y los MLM que utilizan más de dos puntuaciones, como el EVAAS (Sanders & Horn, 1994; Sanders, Saxton & Horn, 1997), pero estiman el cambio como ganancia entre aplicaciones. Los modelos multinivel que estiman un parámetro para la pendiente de crecimiento como una función del tiempo, son los que se ajustan a esta consideración de modelos de crecimiento.

El desarrollo de estos modelos de crecimiento se realiza normalmente con la técnica de análisis de regresión jerárquica o multinivel, aunque existen otras técnicas para el análisis de múltiples tomas de datos, es el caso de los MLM y los modelos de ecuaciones estructurales (Bryk & Raudenbush, 2002).

En el caso de las ecuaciones estructurales (Rovine & Molenaar, 2002), el primer nivel de los modelos multinivel correspondería al modelo de medida de los modelos de ecuaciones estructurales, los parámetros de crecimiento individual se corresponden con las variables latentes y, finalmente, el modelo estructural con el segundo nivel de los modelos jerárquicos. Esta aproximación al estudio del cambio requiere que la estructura temporal de los datos sea equilibrada, sin embargo, permite el tratamiento aleatorio de los datos perdidos, por los que el número de tomas puede variar entre sujetos. Además, con esta técnica es posible estimar varias estructuras de covarianza alternativas.

Los modelos multivariados de medidas repetidas o MLM tienen su caso más representativo en el MVA de Tennessee (EVAAS) (Sanders & Horn, 1994). Analizan la trayectoria individual de cada estudiante observando su avance a lo largo de múltiples ocasiones de medida con un diseño que se complica a medida que aumentan las ocasiones de medida y los grupos a analizar. Y, aunque los modelos multinivel son un caso específico de los modelos lineales mixtos, se diferencian en varios aspectos (Rogosa & Willett, 1983):

- El primer nivel, en los modelos jerárquicos lineales, representa la trayectoria individual de cambio como una función de los parámetros de estatus inicial y crecimiento de un individuo más el error aleatorio

con distribución normal, media cero y varianza común entre mediciones, es un escalar. El nivel dos describe la variación de ese punto de partida y pendiente de crecimiento a través de toda la muestra de individuos y, un tercer nivel o incluso más, en el caso de considerar otras agrupaciones de nivel superior (aulas, escuelas, distritos, etc) que refleje esa varianza aleatoria. Sin embargo, los modelos lineales mixtos requieren, por un lado, la especificación de los principales efectos e interacciones que describen la trayectoria para los diferentes subgrupos (efectos cruzados) y, por otro lado, la descripción de la varianza y covarianza de las medidas repetidas a través del tiempo. Por tanto, equivaldría a dos únicos niveles, el primero incluye los términos residuales de las distintas puntuaciones de cada estudiante con una estructura de los residuos que permita la correlación. Y, normalmente, los docentes como segundo nivel donde se incluyen los efectos cruzados de los estudiantes con los docentes ya que permiten los cambios de profesor en cada toma de datos.

- Una segunda distinción entre estos modelos está relacionada con el tipo de datos apropiados para su realización. Los MLM requieren que las mediciones se lleven a cabo con la misma distancia temporal ya que se basa en cálculos de las ganancias. En cambio, los modelos jerárquicos son más flexibles respecto a las características de los datos, ya que las observaciones repetidas se consideran anidadas dentro de una persona, pudiendo variar tanto en número como en la separación en el tiempo.

- Los modelos del cambio individual desde su aproximación jerárquica, permiten estudiar el efecto de los factores de los centros educativos y las comunidades sobre el desarrollo individual a través del tiempo, incluyendo un tercer nivel de análisis. Esto es muy difícil de estimar en los lineales mixtos que se centran en el estudio de los efectos cruzados de los docentes, las aulas o las escuelas principalmente. Añadir covariables es muy costoso en términos de procesos de cálculo y estimación.

Los modelos lineales mixtos aplicados al análisis del VA han recibido algunas críticas:

- La gran complejidad del modelo necesita contar con herramientas de cálculo de gran potencia para poder lograr la convergencia y estimación (Wiley, 2006).
- No tiene en cuenta la interacción entre donde comienzan las escuelas y cuanto crecen (Goldschmidt et al., 2005). Esto no ocurre con la aproximación jerárquica a este tipo de análisis del VA.
- Normalmente no incorporan covariables del contexto del estudiante, sus autores argumentan que no es necesario porque es el propio estudiante es el que actúa como elemento de control (Sanders, Saxton & Horn, 1997). Y aunque este aspecto ha sido criticado (Kupermintz, Shepard & Linn, 2001; Armein-Beardsley, 2008) una modificación del modelo con la inclusión de estos predictores no provocó variaciones sustanciales en sus estimaciones (Ballou, Sanders & Wright, 2004).
- Problemas de identificación de los estudiantes con los diferentes profesores pueden incrementar los valores perdidos y afectar a las estimaciones (McCaffrey, Lockwood, Doretz & Hamilton, 2003; Armein-Beardsley, 2008).

V.1.2 Efectos de las escuelas: fijos vs. aleatorios

Los modelos de efectos aleatorios asumen que las unidades estudiadas son una muestra de una amplia población de unidades similares pero no observadas y, en consecuencia, la variabilidad entre las unidades observadas describe la varianza en la población. En cambio, si los efectos son tratados como parámetros fijos esas unidades observadas son las únicas que interesan y, por tanto, son la población (McCaffrey, Lockwood, Koretz & Hamilton, 2003). Los modelos para estimar el VA de las escuelas normalmente dotan a esos efectos de carácter aleatorio (Sanders, Saxton & Horn, 1997; Webster, 2005; Ponisciak & Bryk, 2005)

Los análisis que tratan los efectos estimados para las escuelas como coeficientes fijos en el modelo asumen que las escuelas incluidas en la muestra son

toda la población de interés y no hay varianza aleatoria entre esos centros educativos. El modelo más sencillo de efectos fijos[58], sin considerar el cambio en aprendizaje, es un modelo de regresión simple:

$$Y_i = \beta_0 + e_i$$
$$e_i \sim N(0, \sigma_e^2) \qquad \textbf{Ec. V.6}$$

Donde Y_i es la puntuación en logro escolar de un estudiante *i*, β_0 es la puntuación promedio de toda la población y se calcula añadiendo un vector de unos como predictor; y e_i es el residuo de cada estudiante, es decir, la diferencia entre la puntuación predicha de un estudiante y su puntuación observada. Se asumen como independientes y normalmente distribuidos, con media cero y varianza común para todos los estudiantes. Si $\widehat{Y}_i$ es la puntuación predicha sobre la recta de regresión de un estudiante e Y_i su puntuación observada, entonces el residuo para ese sujeto es el siguiente:

$$e_i = Y_i - \widehat{Y}_i \qquad \textbf{Ec. V.7}$$

El valor predicho para el estudiante *i* de una escuela se basa en la estimación de la ecuación con todos los datos disponibles de la muestra. Entonces, una supuesta estimación del VA se calcula como la media de los residuos estimados de cada uno de los estudiantes de una escuela. Por tanto, si los estudiantes de una determinada escuela obtienen un rendimiento alto en sus puntuaciones finales (en comparación con estudiantes de otras escuelas con características similares) sus residuos correspondientes tienden a ser positivos, produciendo una estimación positiva del VA para la escuela.

El modelo de regresión simple no incluye ninguna medida de cambio en aprendizaje y, por tanto, no es un MVA real. Se utiliza como ejemplo básico de análisis de efectos fijos. La importancia reside en la forma de calcular las estimaciones de las escuelas a partir de los residuos de los estudiantes.

Una posible variación del modelo es incluir los efectos de cada una de las escuelas como predictores en la ecuación. Por tanto, habrá tantos coeficientes

[58]Para más información sobre los modelos de efectos fijos ver Aitkin y Longford (1986)

como escuelas evaluadas y se debería añadir el término de sumatorio de efectos de las escuelas en la ecuación.

En los modelos de efectos aleatorios, los parámetros estimados para medir las contribuciones de las escuelas al rendimiento de sus estudiantes se tratan como coeficientes aleatorios. Considera que los coeficientes de las escuelas también pueden variar en torno a la media global, y no únicamente los de los estudiantes. En este tipo de modelos los efectos estimados para una escuela particular están determinados por los datos de todas las otras escuelas y de esa misma escuela. Y asumen que los centros de la muestra son una representación de la población objeto de estudio.

Estos análisis, en su aproximación más básica, están formados por dos ecuaciones de regresión que funcionan a distintos niveles. En primer lugar la regresión en el nivel de los estudiantes y, en segundo, el nivel de los centros educativos donde se modela la variación debida a las escuelas. Las ecuaciones de este modelo de efectos aleatorios son la Ec. V.8 para el nivel de estudiantes y la Ec. V.9 para las escuelas:

Nivel 1 (estudiantes) $$Y_{ij} = \beta_{0j} + e_{ij}$$ **Ec. V.8**

Nivel 2 (escuelas) $$\beta_{0j} = \beta_{00} + r_{0j}$$ **Ec. V.9**

Los coeficientes de primer nivel (Ec. V.8) hacen referencia al estudiante *i* de la escuela *j*; β_{0j}, es el efecto fijo, la media en rendimiento de los estudiantes de la escuela *j;* y e_{ij} es el efecto aleatorio, el residuo asociado a cada estudiante.

El coeficiente fijo de primer nivel pasa formar una ecuación de segundo nivel (Ec. V.9). En el nivel de centros β_{00} es la media global en rendimiento para todas las escuelas y r_{0j} es la desviación de una determinada escuela *j* respecto a la media global, el residuo de la escuela. Por tanto, el efecto de asistir a una determinada escuela *j* es $\beta_{00} + r_{0j}$

Los residuos de ambas ecuaciones se asumen como independientes y con distribución normal como se muestra en Ec. V.10:

$$e_{ij} \sim N(0, \sigma_e^2)$$
$$r_{0j} \sim N(0, \sigma_r^2) \qquad \textbf{Ec. V.10}$$

El tratamiento de los efectos de los centros como fijos o aleatorios puede tener consecuencias en los resultados de VA. La estimación mediante modelos de efectos fijos utiliza únicamente los datos de los estudiantes de cada escuela para realizar el proceso. De otro modo, cuando los efectos son tratados como aleatorios, la estimación se lleva a cabo a través de estimadores bayesianos o BLUP (*Best Linear Unbiased Predictor*).

V.1.2.1 Estimadores Bayesianos de los efectos (BLUP)

La característica principal de este tipo de estimación es que utilizan los datos de otras escuelas para estimar cada efecto de una escuela particular. El mencionado efecto de encogimiento (*shrunk o shrunkage*):

$$r_j = \lambda_j(\bar{e}_j) \qquad \textbf{Ec. V.11}$$

Donde r_j es la estimación del efecto de una escuela *j*, es decir, el VA; $\bar{e}_j$ es la media de los residuos de los estudiantes de esa escuela. El parámetro λ_i, normalmente la fiabilidad del estimador, toma valores entre 0 a 1 y se utiliza para ponderar la estimación.

El efecto de esta ponderación sobre las estimaciones de las escuelas está determinado por la ecuación de Kelly[59] y también se conoce como estimador empírico bayesiano. La nueva estimación está ajustada por el factor de la ecuación Ec. V.12:

$$\lambda_j Y_j + (1 - \lambda_j)\overline{Y} \qquad \textbf{Ec. V.12}$$

Donde λ_j es la fiabilidad de la estimación del rendimiento observado de una escuela (Y_j), es decir, la media de los resultados de sus estudiantes; $\overline{Y}$ es la media global. La puntuación observada de la escuela tendrá mayor peso de forma proporcional a la fiabilidad de la medida. Cuando la fiabilidad de la puntuación

[59]En Gaviria y Castro (2005, pág. 72) se detalla esa relación.

observada es alta, se añade un mayor valor a esa puntuación. En cambio, si es baja se le da más importancia a la media global.

Este tipo de estimadores protegen contra posibles errores de clasificación si un determinado estudiante obtiene puntuaciones muy alejadas de la media, tanto por encima como por debajo. Tienen especial importancia en aquellas unidades de análisis con poca muestra, es decir, pocos estudiantes en un aula o escuela. Sin la utilización de estimadores BLUP los profesores o las escuelas podrían injustamente ser clasificados alejados de la media debido a la poca cantidad de datos disponibles.

Las estimaciones BLUP reducen la varianza de la estimación de los efectos de las escuelas en comparación con los modelos de efectos fijos. Fuerzan las estimaciones de los efectos a desviarse hacia la media global y, por tanto, si una escuela cuanta con poca muestra, en consecuencia, la precisión de las sus estimaciones será baja y pueden verse afectadas. En ese caso, los efectos verdaderos de escuelas que pueden ser eficaces tenderán a no diferenciarse de la media y, en el lado opuesto, escuelas con unos efectos reales bajos tenderán a parecerse a los resultados medios. La cantidad de ese encogimiento dependerá de la precisión de la media del centro y de la varianza entre escuelas (McCaffrey, Lockwood, Koretz & Hamilton, 2003)

En consecuencia, este tipo de estimación puede afectar a la precisión para identificar escuelas significativamente diferentes de la media. La imprecisión resultante de la variabilidad de las puntuaciones de los estudiantes dependerá del tamaño global de la muestra, del número de alumnos por clase y la relación de la varianza de las escuelas y los profesores con las varianzas de los errores residuales. Un estudio de esta cuestión es la llevada a cabo por McCaffrey, Koretz, Louis y Hamilton (2004). Los autores prueban que un modelo con efectos aleatorios de los docentes es capaz de distinguir efectos estadísticamente diferentes de la media global. Encuentran que, con una muestra moderada (20 alumnos por aula y un total de 10 aulas), tanto el modelo general para datos longitudinales, como el TVAAS identifican entre un 25% y 33% de docentes distintos de la media.

Tekwe et al. (2004) llevan a cabo un estudio empírico que compara los resultados de distintos modelos que utilizan únicamente dos puntuaciones de rendimiento para estimar el VA de las escuelas, incluyen modelos jerárquicos lineales de efectos fijos y aleatorios con y sin covariables, y una variación del modelo EVAAS con dos únicas medidas. Los autores no encontraron diferencias importantes entre los modelos que utilizan estimaciones BLUP, es decir, los modelos de efectos aleatorios con o sin predictores. Las correlaciones de Pearson entre las estimaciones de los distintos modelos oscilaron entre 0,97 y 1. Esos valores de correlación al comparar las estimaciones BLUP con el modelo de efectos fijos sin covariables oscilan entre 0,50 y 0,90, dependiendo del modelo de efectos aleatorios.

Para Sanders y Wright (2008) las estimaciones BLUP son el método adecuado para calcular el impacto de los centros en el progreso académico de los estudiantes por dos motivos:

- Se ha probado teóricamente que las estimaciones bayesianas proporcionan la máxima correlación con la estimación el efecto verdadero.
- Proporcionan cierta protección contra estimaciones espurias producidas por muestras pequeñas.

La elección de una opción u otra dependerá de los datos utilizados, del modelo de análisis empleado y del objetivo que persiguen las estimaciones de VA. Por ejemplo, si el estudio de los componentes de varianza son la unidad principal de interés como ocurre con el análisis del VA, que pretenden aislar la varianza que producen las escuelas en el aprendizaje, lo normal es estimar los efectos como aleatorios. En cambio, si la intención es llevar a cabo inferencias sobre un grupo particular de docentes o escuelas, los modelos de efectos fijos pueden ser más adecuados porque las estimaciones BLUP pueden estar segando los efectos de los que se sitúan en los extremos de la distribución. Además, las estimaciones producidas por los modelos de efectos fijos pueden estar acompañadas de elevados errores (McCaffrey, Lockwood, Koretz & Hamilton, 2003)

V.1.3 Efectos de las escuelas: anidados vs. cruzados

En ocasiones, las estructuras de los datos que provienen de la educación pueden alcanzar una mayor complejidad que la agrupación totalmente jerárquica o anidada. En función de si la evaluación está diseñada para seguir la trayectoria de los estudiantes cuando cambian de aula o centro educativo o no, es posible considerar los efectos de las escuelas de dos formas distintas: anidados[60] o cruzados.

Lo usual en un sistema educativo es que los estudiantes cambien de profesor cada curso. Por tanto, al medir longitudinalmente un constructo, debe tenerse en cuenta que el resultado puede ser producto de la influencia de docentes distintos. Y Las estimaciones finales de VA pueden estar sesgadas si no se considera este tipo de estructura (Meyers & Beretvas, 2006; Grenn, 2010; Daniel, 2012).

Cuando los datos educativos son considerados completamente anidados en escuelas, si algún sujeto cambia de centro deja de formar parte de la evaluación. Cuando las evaluaciones abarcan un amplio rango de cursos la probabilidad de que los estudiantes cambien de escuela aumenta y los datos perdidos también. En cambio, los efectos cruzados al considerar el cambio de escuela o docente disminuyen la cantidad de valores perdidos (Hill & Goldstein, 1998)

Una estructura de datos completamente anidada sería la siguiente:

Sujeto/escuela	A	B	C	D
1	XXXX			
2	XXX			
3		XXXX		
4		XX		
5			XXXX	
6			XXXX	
7				XXX
8				XXXX

Tabla V.1. Estructura anidada de datos longitudinales

En la tabla anterior (Tabla V.1) cada fila representa un sujeto y las columnas representan cuatro escuelas distintas. Las distintas *X* son cada una de las

[60]La importancia de esta estructura anidada en los datos que proceden de Ciencias Sociales y, sobre todo, de educación ha sido tratada en el Capítulo II (Apartado II.2.2).

mediciones que se realizan sobre los estudiantes. Las diferentes mediciones (nivel 1) están anidadas en cada estudiantes (nivel2) y estos, a su vez, se agrupan en escuelas (nivel3). Como puede observarse todas las mediciones de un mismo estudiantes se llevan a cabo en la misma escuela. Por ejemplo, los estudiantes 1 y 2 pertenecen a la escuela A, el 3 y 4 a la B, etc. No hay cambios de estudiantes entre escuelas. En cambio, cuando las alumnos pueden cambiar de escuela o de profesor a lo largo de la evaluación:

Sujeto/escuela	A	B	C	D
1	XX	X	X	
2	X		XX	
3		X	XX	X
4		XX		
5	XX		XX	
6		XX	XX	
7			X	XX
8		XX	X	X

Tabla V.2. Estructura cruzada de datos longitudinales

En este caso, las diferentes mediciones también se anidan en cada sujeto pero los estudiantes y las escuelas se encuentran cruzados en un único nivel y no en dos como en el caso de los modelos completamente anidados. Por ejemplo, el estudiante 1 es evaluado en el centro A en dos ocasiones, otra en el B y la cuarta en el C.

Cuando los modelos analizan los efectos aleatorios de los estudiantes que varían entre aulas o escuelas, se denominan modelos de efectos cruzados. En estos análisis no se asume una estructura completamente anidada (Sanders & Horn, 1994; Sanders, Saxton & Horn, 1997; McCaffrey, Lockwood, Doretz & Hamilton, 2003; Ballou, Sanders & Wright, 2004), los estudiantes pueden cambiar de aula o de escuela de un año a otro y el diseño debe estar preparado para poder seguirlos.

En los modelos que consideran los efectos aleatorios completamente anidados, los niveles de análisis están formados por las agrupaciones naturales que se dan en educación. En los modelos de crecimiento se distinguen principalmente tres niveles, las diferentes puntuaciones de rendimiento anidadas en cada estudiante y estos agrupados en escuelas. El cambio se estima como una pendiente, lineal o no, que depende del tiempo (Bryk & Raudenbush, 2002; Zvoch & Stevens, 2003; 2006) y se estima la aportación de la escuela a ese cambio a través del

residuo asociado a esa pendiente. No se tiene en cuenta los cambios de aula o profesor durante la evaluación y si un estudiante cambia de centro durante el proceso se considera un caso perdido.

Cuando la evaluación tiene por objetivo el análisis de los efectos de los docentes el diseño se complica. Estimar un único residuo asociado al crecimiento de los docente no sería adecuado porque los docentes pueden atender a alumnos distintos en cada medición. Una forma posible de considerar estos efectos cruzados la ofrece la perspectiva de los MLM, por ejemplo, el modelo EVAAS esta diseñado para seguir a los estudiantes cuando cambian de docente, incluso identificar los efectos cuando un estudiante cambia de profesor en un mismo curso académico. Desde la perspectiva de los modelos multinivel también es posible analizar este tipo de efectos con los modelos de clasificación cruzada.

V.1.4 Efectos de las escuelas: persistentes vs. cambiantes

Los análisis del MVA, mencionados en el epígrafe anterior, que incluyen los efectos cruzados como el EVVAS y el modelo de clasificación cruzada, consideran que estos efectos permanecen en el estudiante cuando cambia de docente. Los resultados de un estudiantes están en función de los efectos del docente actual pero también de los que ha tenido en las mediciones anteriores. Estos modelos tienen esta concepción de persistencia de los efectos porque consideran que el aprendizaje de los estudiantes se acumula y los docentes provocan incrementos en el conocimiento de sus alumnos. Incluyen los efectos pasados añadiéndolos a los efectos actuales en el modelo.

El modelo EVVAS recibe el nombre también de "*layered model*" o modelo estratificado por esta característica de acumulación de los efectos de los docentes anteriores en los resultados actuales. Estos efectos se suman a los anteriores y permanecen constantes. En consecuencia, la contribución total del profesor a la varianza de las puntuaciones de resultados aumenta con el tiempo, incluso cuando la varianza total no lo hace (McCaffrey, Koretz, Louis & Hamilton, 2004).

Otra perspectiva del análisis del VA, con un modelo similar al EVAAS, sí permite una disminución de los efectos de los docentes en las mediciones sucesivas (McCaffrey, Lockwood, Doretz & Hamilton, 2003). No asumen que los

efectos previos de los docentes permanecen constantes en los resultados de los años posteriores una vez que los estudiantes han cambiado de aulas o incluso de centro. El modelo de persistencia asume que ese efecto puede suavizarse o incluso desaparecer con el tiempo, aunque también permite la posibilidad de que permanezcan constantes. Para ello, incluye unos parámetros que describen el cambio en esos efectos a lo largo del tiempo. Son los parámetros de persistencia que dan nombre al modelo.

El modelo de efectos completamente persistentes vincula los resultados actuales del estudiantes con el docente actual y los previos en una ocasión de medida determinada. Sin persistencia, únicamente se vincula con el docente de el año en el que se realiza la medición. Esta cuestión puede verse de forma más clara con un ejemplo de la matriz de datos:

			Persistencia			No-Persistencia			Efecto parcial		
Estudiante	Año	Docente	A	B	C	A	B	C	A	B	C
1	1	A	1			1			0,5	0,5	
1	2	B	1	1			1		0,5	1	
1	3	C	1	1	1			1	0,5	1	1
2	1	B		1			1			1	
2	2	C		1	1			1		1	1
2	3	A	1	1	1	1			1	1	1

Tabla V.3 Matriz de datos con y sin persistencia

Por ejemplo, el estudiante 1 recibe la docencia del profesor A el primer curso, la del B en el segundo y la del C en el tercero. Con la consideración de la persistencia de los efectos, la aportación del docente A permanece constante durante las otras mediciones situando el valor 1 también en las celdas correspondientes a ese docente. Si los efectos de los docentes desaparecen por completo entre aplicaciones, la matriz solo incluye el valor 1 cuando el estudiante ha recibido su docencia en ese curso, de lo contrario el valor es cero. Una tercera opción es que un estudiante reciba la enseñanza de un profesor durante medio curso y la otra mitad con otro docente distinto. En este caso, el efecto del docente en el estudiante no es del 100% al no haber completado todo el curso, en lugar de 1 se incluye un valor inferior que disminuye el peso del efecto previo del docente, en función del tiempo que haya estado bajo su enseñanza.

La consideración o no de este supuesto puede afectar a las estimaciones finales de los efectos de los docentes o las escuelas. McCaffrey, Koretz, Louis & Hamilton (2004) comprueban empíricamente esta afirmación. Los autores para llevar a cabo este estudio formulan un modelo general[61] que puede incluir los diferentes modelos que analizan en su estudio y que permite que los efectos de los docentes disminuyan con el tiempo. Comparan los resultados obtenidos por este modelo con los producidos por el EVAAS que no permite esa disminución de los efectos. Concluyen que asumir efectos del profesor como parámetros que no van suavizándose con el tiempo no está teórica o empíricamente justificada y parece no ser del todo plausible. Según los autores, que los efectos disminuyan es lo normal en muchas de las investigaciones en Ciencias Sociales.

Otro estudio compara los resultados del modelo que considera la persistencia total de los efectos previos de los docentes en los años posteriores y otro en el que desaparecen (Daniel, 2012), bajo diferentes condiciones simuladas con datos longitudinales como el número de mediciones, el número de docentes evaluados o las tasas de valores perdidos. La autora confirma que el modelo que acumula los efectos de los docentes obtiene mejores resultados de estimación y ajuste.

V.1.5 Ajuste de los modelos: contextualizados vs. no contextualizados

La contextualización ha sido uno de los aspectos analizados en este trabajo (ver apartado IV.5) y no conviene volver a profundizar en el tema. A modo de resumen, el ajuste de los MVA es visto desde dos perspectivas distintas: Por un lado, Sanders, Saxton y Horn (1997) argumentan que no es necesario incorporar covariables de los estudiantes en los modelos debido a que cada estudiante ejerce un control sobre sí mismo. Stevens & Zvoch (2006) también comparten este argumento. Por otro lado, Raudenbush y Bryk (2002) indican que la introducción de ajustes con variables del contexto de los estudiantes si es importante.

[61]Más información sobre el modelo general formulado por McCaffrey, Koretz, Louis y Hamilton en el apartado V.2.3.2.2

V.1.6 Variable dependiente: univariante vs. multivariante

Se consideran modelos univariantes aquellos que utilizan una única medición de resultados educativos como variable dependiente o una puntuación de ganancia. En cambio, los multivariantes utilizan más de una medida como variable criterio, normalmente la matriz completa de puntuaciones de logro en los diferentes momentos temporales.

Los modelos univariantes están vinculados normalmente a los análisis de ganancia. La ganancia bruta emplea la diferencia entre las puntuaciones de pretest y postest como variable criterio y la residual solo utiliza la puntuación del postest. En este análisis de la ganancia residual, el pretest actúa como covariable. Dentro del grupo de medidas de ganancia, la estimada es una excepción porque las dos tomas de datos actúan como variable de resultados.

Los modelos multivariados tienen su principal exponente en los modelos de crecimiento desde su aproximación multinivel. Estos análisis utilizan las diferentes puntuaciones de logro como variable de resultados en el modelo. Lo mismo ocurre con el MLM empleado en el modelo EVAAS pero con una pequeña diferencia respecto al análisis jerárquico: no estiman una pendiente de crecimiento en función del tiempo, sino ganancias entre aplicaciones consecutivas calculadas con toda la matriz de datos a partir del rendimiento estimado en cada aplicación de medida.

V.2. Clasificación de Modelos de Valor Añadido

Antes de comenzar con la descripción de los distintos MVA es necesario mencionar que no son la única técnica utilizada para la evaluación de la eficacia de las escuelas que analiza el cambio en el logro académico, existen otros modelos que tratan de conseguir este objetivo por otros caminos. Principalmente se pueden distinguir tres grupos: modelos de cambio cohorte a cohorte, modelos de cambio basados en la ganancia y los basados en el crecimiento. Estos dos últimos grupos, cuando dirigen sus resultados a la estimación de los efectos de las escuelas libres de otros elementos ajenos su control, son denominados específicamente MVA.

Es conveniente llevar a cabo una descripción de todas las posibles opciones de análisis de ese cambio, desde los modelos transversales que analizan el cambio que se produce cohorte a cohorte o la simple ganancia bruta, hasta los análisis más complejos como el modelo EVAAS y los modelos multinivel en su perspectiva longitudinal.

Todos los modelos descritos pueden considerar los efectos de los centros educativos como fijos o aleatorios pero para evitar la redundancia únicamente se incluye únicamente la aproximación aleatoria. La única variación es que la varianza entre las escuelas desaparece y en la mayor parte de las ocasiones los modelos que consideran esos efectos fijos son análisis de regresión simple o múltiple.

V.2.1 Modelos de cambio cohorte a cohorte

Estos modelos analizan los resultados obtenidos por una escuela en un curso con los que obtienen los estudiantes de ese mismo curso el año siguiente, no son más que modelos de estatus comparados. Por ejemplo, comparando los resultados que han obtenido los estudiantes de cuarto curso con los que obtuvieron los alumnos de ese mismo curso el año anterior.

Desde esta perspectiva se llevan a cabo mediciones anuales de un curso específico dentro de una escuela y analizan las puntuaciones medias con respecto a un estándar establecido. El avance o progreso se define como el porcentaje de estudiantes que alcanzan ese nivel establecido de antemano en un año determinado (Goldschmidt et al., 2005) y se comparan los resultados anuales obtenidos por los centros de enseñanza para saber si mejoran o empeoran.

En Estados Unidos esta clase de modelos se denominan de Progreso Anual Adecuado (AYP) de las escuelas e indican el porcentaje de estudiantes que consigue alcanzar o superar una puntuación determinada a la que denominan "*proficiency*" (competente). Y tienen el objetivo final de alcanzar el 100% de alumnos competentes en una fecha establecida.

Estos modelos aunque son sencillos de calcular y fácilmente entendibles no reflejan realmente la situación escolar y son uno de los indicadores de rendimiento más débiles por varios motivos (Choi, Goldschmidt & Yamashiro, 2006):

- Los resultados pueden estar contaminados por el rendimiento previo y otros factores que no permiten diferenciar si se deben realmente a los procesos de enseñanza y aprendizaje que se producen en una escuela determinada en una año concreto.
- Al examinar solo el porcentaje de alumnos que alcanzan el nivel competente se obvia el movimiento que se produce entre los grupos de estudiantes que no se encuentran en ese nivel. Por tanto, la información de la situación escolar no es completa.
- No siguen la trayectoria de cambio en aprendizaje de una misma cohorte de estudiantes. Por tanto, no se evalúa el crecimiento sino la situación en un momento temporal concreto.

Los modelos de estatus que analizan los cambios que se producen de una cohorte de estudiantes a otra son una opción para llevar a cabo la evaluación de la eficacia de las escuelas, pero los modelos que utilizan dos o más puntuaciones de rendimiento de los estudiantes a lo largo del tiempo reflejan mejor su trayectoria de aprendizaje (Willett, 1994) y, por consiguiente, la situación educativa que se produce en los centros educativos que intentan producir un cambio en sus estudiantes.

V.2.2 Modelos de ganancia

Estos modelos analizan la misma cohorte de estudiantes en dos cursos consecutivos con el objetivo de determinar si los alumnos han hecho o no, en términos medios, progreso. Es posible diferenciar tres formas distintas de medir la ganancia: los análisis que utilizan la puntuación de ganancia como variable dependiente (ganancia bruta), los que ajustan la variable criterio (rendimiento en el año evaluado) utilizando como principal covariable el rendimiento previo (ganancia residual o modelo de ajuste de covariables) y los que consideran las puntuaciones del pretest y postest como variables de resultados en el modelo (ganancia estimada). Todos han sido considerados en su conjunto como modelos de ganancia (Hibpshman, 2004; McCaffrey, Lockwood, Doretz & Hamilton, 2003) o técnicas tradicionales para medir el cambio (Willett, 1989a; 1997).

Los modelos de ganancia asumen que el rendimiento, tanto de los estudiantes como de las escuelas, no es una cuestión sencilla de analizar y, por tanto, no es suficiente con una única medida de logró académico en un determinado momento temporal. Es necesario observar el progreso, considerado como el cambio entre un pretest y un postest, si se quiere contar con medidas más precisas de la eficacia de las escuelas.

Definen ese cambio como un incremento que se mide comparando el estatus individual o grupal antes de un periodo de aprendizaje y su nivel después de este periodo. Este tipo de análisis trata el cambio individual no como un proceso de desarrollo continuo a través del tiempo, sino como la cantidad de aprendizaje que se produce en el periodo de tiempo que trascurre entre el pre-test y el pos-test.

Los MVA que emplean el análisis de la ganancia relacionan el cambio que se produce en el rendimiento de un estudiante con el profesor que les ha impartido clase o la escuela a la que han asistido durante ese periodo. Por ejemplo, para cada estudiante el efecto del rendimiento previo es extraído de la puntuación actual para calcular una medida del cambio en el logro académico de los estudiantes en una determina materia y en un aula particular. La media de todas las ganancias de los estudiantes de una misma aula se calcula y se comparan con la de otros grupos de estudiantes. Es posible comparar diferentes profesores dentro de un mismo centro o buscar diferencias con la media global de todos los centros. El VA de la escuela será la diferencia entre la ganancia de un determinado centro y la del grupo con el que se compara, por ejemplo, media del distrito, media de centros de sus características (públicos o privados) o la media global.

Los modelos de ganancia pueden incorporar covariables para controlar los efectos del contexto del estudiante en caso de ser necesario. Estos análisis evolucionaron desde la simple diferencia de dos puntuaciones conocida como ganancia bruta, pasando por la regresión de una puntuación pretest sobre la puntuación postest (ganancia residual o ajuste de covariables), hasta alcanzar la última variación con las ganancias estimadas donde tanto el pretest como el postest se consideran medidas de resultados.

V.2.2.1 Ganancia bruta

El principal objetivo de la enseñanza es producir aprendizaje y ¿qué es el aprendizaje? Es un cambio, por lo tanto, la cantidad y tipo de aprendizaje que se produce en un individuo puede medirse comprobando el estatus individual o grupal antes de dicho periodo, con el nivel alcanzado una vez producido el periodo de aprendizaje. Esta sería la forma básica de medir el cambio en el aprendizaje de un individuo y se denomina ganancia bruta o absoluta. Es decir, mide la distancia existente entre el nivel actual de rendimiento y el nivel inicial. Si tomamos dos medidas, la primera al comienzo del curso (Y_0) y otra al final (Y_1), el producto resultante de la diferencia ($Y_1 - Y_0$) es la ganancia bruta.

Esta ganancia bruta puede estar sesgada por el error de medida asociado a las puntuaciones de logro utilizadas. Si se pretende hallar la ganancia verdadera debe aludirse a la clásica distinción entre puntuación observada, puntuación verdadera y errores de medida. Siguiendo la Teoría Clásica de los Test, el estatus de un individuo, medido por algún test de rendimiento o cualquier otro instrumento, es una combinación lineal de dos componentes independientes: un componente sistemático o fijo (puntuación verdadera) y uno aleatorio (error). El modelo básico de medida será:

$$Y_{ti} = V_{ti} + e_{ti} \qquad \textbf{Ec. V.13}$$

El subíndice *t* es la ocasión de medida, igual a cero para el pretest e igual a uno para el postest; el subíndice *i* indica qué individuo de la población está siendo representado. El símbolo V_{ti} índica el estatus verdadero. La variable *Y* es una medida falible de *V* y ha sido obtenida con la presencia de un error de medida (e_{ti}). Dichos errores son independientes y se distribuyen de forma normal con media cero y varianza constante.

Una simple medida de ganancia, definida como el cambio entre el estatus inicial y final, puede obtenerse muy fácilmente como la resta de las dos puntuaciones de logro observadas:

$$G_i = Y_{1i} - Y_{0i} \qquad \textbf{Ec. V.14}$$

En la ecuación Ec. V.14 G_i es la diferencia observada para un individuo *i*. Sustituyendo en la ecuación las puntuaciones observadas por las verdaderas el resultado sería el siguiente:

$$\varepsilon_i = e_{1i} - e_{0i} \qquad \textbf{Ec. V.15}$$

Por lo tanto:

$$G_i = \Delta_i + \varepsilon_i \qquad \textbf{Ec. V.16}$$

Cuando un test o instrumento de medida es administrado a un individuo, la medida observada combina la puntuación verdadera con un error aleatorio que le acompaña. Si se tiene en cuenta este aspecto, la diferencia bruta es fácil y cómoda de estimar y es una medida del cambio verdadero del sujeto que no está sesgada (Rogosa & Willett, 1983; Willett, 1989a). Algunos autores han criticado esta medida de ganancia por su alta correlación con el estatus inicial, con el pretest. Sin embargo, Willet (1994) asegura que incluso si la ganancia bruta careciera siempre de fiabilidad, esto no sería necesariamente un problema para medir el cambio intra-individual, es decir, la baja fiabilidad de la diferencia bruta no implica exclusivamente que el cambio haya sido medido de forma imprecisa. Además, la correlación entre cambio y estatus inicial no señala directamente falta de fiabilidad, la relación entre el estatus inicial de rendimiento y el crecimiento puede ser producto de la historia de cambio del individuo, es decir, el nivel actual de logro puede determinar el estatus futuro y, por tanto, la correlación entre cambio y nivel inicial es un hecho habitual en el aprendizaje.

Una vez calculada la ganancia bruta, puede utilizarse como variable dependiente en modelos de efectos fijos o aleatorios y no conformarse únicamente con la media de las ganancias brutas de los estudiantes de un centro como medida de su eficacia. Si los análisis de la ganancia bruta se llevan a cabo utilizando variables del contexto de los estudiantes y las escuelas como predictores y se estiman los efectos asociados a las escuelas pueden ser considerados MVA.

A continuación se formula un modelo multinivel que emplea las puntuaciones de ganancia como variable dependiente y considera los efectos de los centros como variables aleatorias. Estos análisis incluyen ecuaciones de regresión

en dos niveles. Por un lado, los estudiantes y, por otro, la regresión a nivel de escuelas, donde se modela la variación de los interceptos ajustados de las escuelas obtenidos de la regresión del primer nivel. Un modelo multinivel completamente aleatorio desde esta perspectiva es el siguiente:

Nivel 1 (estudiantes): $$G_{ij} = \beta_{0j} + \beta_{qj}X_{qij} + e_{ij}$$ **Ec. V.17**

Nivel 2 (escuelas): $$\beta_{0j} = \beta_{00} + \beta_{q}W_{sj} + r_{0j}$$ $$\beta_{qj} = \beta_{q0} + r_{qj}$$ **Ec. V.18**

Si los coeficientes de la ecuación de primer nivel (Ec. V.17) se sustituyen por las ecuaciones de segundo nivel (Ec. V.18), el resultado es el siguiente:

$$G_{ij} = \beta_{00} + \beta_{q}W_{sj} + (\beta_{q0} + r_{qj})X_{qij} + r_{0j} + e_{ij}$$ **Ec. V.19**

La Ec. V.19 es un modelo de regresión con dos niveles de análisis. En primer lugar, la ganancia bruta de cada estudiante (G_{ij}) se ajusta con predictores del contexto del estudiante (X_{qij}) (género, nivel socioeconómico, si es repetidor, etc.). Por tanto, β_{0j} es la ganancia media de todos los estudiantes de la escuela *j* y β_{qj} es la variación sobre la ganancia de la escuela *j* producida por cada variable del contexto del estudiante.

Cada coeficiente de primer nivel es una ecuación en el nivel dos. Donde β_{00} es la ganancia media global para todas las escuelas; β_{q} es la variación producida por cada predictor del contexto escolar (W_{sj}) (titularidad de los centros, proporción de estudiantes inmigrantes, selección de estudiantes, etc.) en la ganancia media; y β_{q0} es la variación sobre la ganancia global de cada predictor de primer nivel.

Los coeficientes aleatorios, es decir, los residuos de ambas ecuaciones, se asumen como independientes y normalmente distribuidos (ver Ec. V.20). El residuo de segundo nivel tiene una matriz de varianzas-covarianzas R que depende de los predictores de la ganancia introducidos en el modelo:

$$e_{ij} \sim N(0, \sigma_e^2)$$
$$r_{qj} \sim N(0, R) \qquad \textbf{Ec. V.20}$$

r_{0j} es la desviación de una determinada escuela *j* respecto a la media global. Por tanto, el efecto que produce sobre la ganancia asistir una determinada escuela *j*, una vez aislados determinados factores contextuales, es $\beta_{00} + r_{0j}$, el VA de la escuela.

V.2.2.2 Ganancia residual o ajuste de covariables

En un intento por avanzar en la construcción de una medida de ganancia fiable y con la finalidad de evitar la correlación entre el cambio o la ganancia estimada y el pretest, se opta por la utilización de la ganancia residual (Willett, 1994). Este análisis considera el nivel inicial (pretest) como la principal covariable del nivel actual (postest). La regresión de Y_1 sobre Y_0 produce $E(Y_1|Y_0)$, por lo tanto, la ganancia quedaría definida de la siguiente forma:

$$Y_1 - E(Y_1|Y_0) \qquad \textbf{Ec. V.21}$$

Es decir:

$$Y_{i1} = \beta_0 + \beta_1 Y_{i0} + e_i$$
$$e_i \sim N(0, \sigma_e^2) \qquad \textbf{Ec. V.22}$$

En la ecuación Ec. V.22 el término e_i es la ganancia residual una vez que se ha introducido como principal covariable del rendimiento la puntuación previa obtenida por el estudiante (Y_{i0}). Esta ganancia resultante es el residuo producido por el análisis de regresión, es decir, la puntuación observada menos la puntuación estimada utilizando el rendimiento previo como principal covariable en el análisis.

Tiene algunos problemas metodológicos, por ejemplo, las covariables deben estar medidas sin error, tampoco informa de cuánto ha cambiado un individuo en un atributo determinado, sino de cuanto rinde un estudiante eliminando los efectos de su rendimiento previo. No obstante, una ventaja es que no necesita que las puntuaciones de rendimiento se encuentren situadas en una escala común.

La ganancia residual, es decir, el residuo de regresión, se estima normalmente a través de mínimos cuadrados ordinarios y es difícil asumir que no están sesgados para la estimación de los efectos de los docentes o las escuelas por dos motivos (Doran, 2003):

- La estimación OLS asume que la correlación intraclase es igual cero. Este supuesto es insostenibles cuando se considera la naturaleza de la organización de las escuelas.
- La desviación respecto a la recta de regresión en el análisis pretest-postest puede deberse al efecto de un artefacto del diseño, la regresión hacia la media, y llevar a cabo inferencias incorrectas respecto a las escuelas o los docentes, considerándolas como causantes de esa desviación.

Los modelos de efectos aleatorios pueden superar algunos de los inconvenientes mencionados. El análisis residual de la ganancia también puede considerar los efectos fijos o aleatorios. Los últimos asumen que no solo existe varianza entre estudiantes, también entre las escuelas y modifican sus análisis para poder adecuarlos a esta situación. Son los modelos que se emplean normalmente en los análisis de la función de producción en educación (Hanushek, 1979). Si se incluyen predictores de contexto, además del pretest y se extraen esos efectos escolares también puede ser considerado un MVA.

Su formulación es similar al modelo de efectos aleatorios especificado en el apartado de ganancia bruta (ver Ec. V.17 y Ec. V.18) pero utilizando solo el postest como variable dependiente, en lugar de la puntuación de ganancia bruta, e incluyendo el pretest y "q" variables de contexto como predictores, como puede verse en Ec. V.23:

$$\text{Nivel 1 (estudiantes): } Y_{1ij} = \beta_{0j} + \beta_1 Y_{0ij} + \beta_q W_{qij} + e_{ij}$$

$$\text{Nivel 2 (escuelas): } \beta_{0j} = \beta_{00} + r_{0j} \qquad \textbf{Ec. V.23}$$

$$e_{ij} \sim N(0, \sigma_e^2)$$

$$r_{0j} \sim N(0, \sigma_r^2)$$

Los subíndices *i* y *j* hacen referencia a los estudiantes dentro de las escuelas. β_{0j} es el punto de corte, el rendimiento medio de los estudiantes de la escuela *j* cuando el resto de predictores toman un valor igual a cero. En consecuencia, su interpretación dependerá de las covariables introducidas y del significado del valor cero del pretest, por ejemplo, si ha sido centrado respecto a la media global[62], de la siguiente forma:

$$\beta_1(Y_{0ij} - \overline{Y}_0) \qquad \textbf{Ec. V.24}$$

En ese caso, β_{0j} es la media en el postest de los estudiantes de la escuela *j* con un nivel medio de rendimiento previo; β_1 y β_q son los coeficientes de regresión que indican los efectos del pretest y las diferentes covariables de contexto (W)

Los residuos de ambas ecuaciones (e_{ij} y r_{0j}) se asumen como independientes y normalmente distribuidos. En la ecuación de segundo nivel, los interceptos ajustados de las escuelas obtenidos en la ecuación de primer nivel se distribuyen aleatoriamente en torno a la media de todas las escuelas (β_{00}) y, por tanto, las desviaciones respecto a esta media, los residuos de segundo nivel (r_{0j}), se consideran estimaciones del VA de las escuelas.

En Inglaterra suele utilizarse este tipo de análisis para estimar sus puntuaciones de VA (Ray, 2006). En un estudio comparativo, también en Inglaterra, de las estimaciones obtenidas usando regresión múltiple y modelos multinivel, es decir, utilizando efectos fijos o aleatorios para el cálculo del VA de las escuelas, los resultados mostraron que ambos conjuntos de datos estaban correlacionados aproximadamente al 0.99 (Fitz-Gibbon, 2001).

Otro ejemplo de ganancia residual más complejo es el MVA estadounidense de Dallas en el estado de Texas. Este modelo combina regresión lineal y regresión multinivel.

[62]Para más detalle ver: Koeing y Lissitz (2001) y Bryk y Raudenbush (2002)

Modelo de Valor Añadido de Dallas

Este análisis del VA implementado en Dallas desde principios de los 90 (Webster & Mendro, 1997; Webster, 2005) se denomina también modelo de dos etapas y combina regresión múltiple con modelos jerárquicos lineales. El modelo de Dallas se puede considerar un ejemplo de ganancia residual con ajustes de covariables que tiene connotaciones particulares. Una característica peculiar es que lleva a cabo dos ajustes diferenciados tanto del rendimiento actual como del rendimiento previo.

En la primera fase de análisis (ver Ec. V.25) se ajustan las puntuaciones de rendimiento de los estudiantes, tanto la puntuación actual, como las de rendimiento previo (puede ser más de una) que entran en juego en la segunda etapa. El ajuste se lleva a cabo mediante un número de características relevantes de los estudiantes como raza, sexo, estatus socioeconómico y competencia lingüística.

En la segunda fase (Ec. V.26 y Ec. V.27), se lleva a cabo una regresión de la puntuación ajustada del postest sobre las puntuaciones ajustadas de rendimiento previo y un conjunto de covariables de las escuelas, como el porcentaje de minorías y estatus socioeconómico, emplenando un modelo jerárquico lineal de dos niveles.

Por tanto, en la primera fase:

$$Y_{ti} = \beta_0 + \beta_1 W_i + \cdots + \beta_q W_{qi} + e_i \qquad \textbf{Ec. V.25}$$

Y_{ti} es el rendimiento actualo el rendimiento previo del estudiante i. El subíndice t indica la ocasión de medida; W son un conjunto de características de los estudiantes (raza, género, nivel de pobreza y también indicadores del contexto socioeconómico); β son los diferentes coeficientes de regresión; Finalmente, el residuo e_i es independiente, con distribución normal y varianza común para todos los estudiantes. En un modelo de regresión lineal múltiple.

Los residuos estimados en cada regresión se estandarizan y se utilizan como variables dependientes en la segunda fase del proceso, realizando un análisis de regresión jerárquica con dos niveles: los estudiantes y los centros educativos.

Suponiendo que disponemos de una puntuación actual del rendimiento y dos puntuaciones de rendimiento previo de los estudiantes, en el nivel 1:

$$Y_{tij} = \beta_{0j} + \beta_{1j}Y_{t-1ij} + \beta_{2j}Y_{t-2ij} + e_{ij} \qquad \textbf{Ec. V.26}$$

Y_{tij} es la nueva variable dependiente una vez ajustado en la primera fase, es el residuo estandarizado del nivel de rendimiento actual del alumno *i* en la escuela *j*.

Y_{t-1ij} y Y_{t-2ij} son las dos puntuaciones de rendimiento previo que han sido ajustadas en la primera fase.

β_{0j}, β_{1j} y β_{2j} son los diferentes coeficientes de regresión, es decir, la media de los estudiantes de la escuela *j* y los efectos diferenciales producidos por las dos puntuaciones de rendimiento previo en la escuela *j*. En el caso de contar con más variables de rendimiento previo aumentaría el número de coeficientes.

e_{ij} son las desviaciones de los estudiantes respecto a su escuela específica, independientes y normalmente distribuidas, es decir, el residuo asociado a los alumnos.

Cada coeficiente de regresión de nivel 1 pasa a ser una ecuación en el nivel 2, por tanto quedaría formulado de la siguiente forma:

$$\beta_{0j} = \beta_{00} + \sum_{k=1}^{m} \beta_{01}W_j + r_{0j}$$

$$\beta_{1j} = \beta_{10} + \sum_{k=1}^{m} \beta_{11}W_j + r_{1j} \qquad \textbf{Ec. V.27}$$

$$\beta_{2j} = \beta_{20} + \sum_{k=1}^{m} \beta_{21}W_j + r_{2j}$$

W son el conjunto de características de las escuelas (indicadores demográficos de la composición del centro educativo, indicadores socioeconómicos del estatus de la comunidad escolar, movilidad escolar y agrupación) que se utilizan para ajustar las variables dependientes.

β son los diferentes coeficientes de regresión. β_{00} es la media global en rendimiento β_{10} y β_{20} son los efectos medios globales de las dos puntuaciones de rendimiento previo.

r_{0j} es la desviación específica de la escuela *j* respecto a la media global de rendimiento, una vez controladas las características escolares, el rendimiento previo y ajustando previamente las variables de rendimiento con indicadores del contexto de los estudiantes. Este residuo es el VA en el modelo. El resto de residuos de segundo nivel también asumen una distribución normal:

$$r_{kj} \sim N(0, \sigma^2_{rk}) \qquad \textbf{Ec. V.28}$$

Los índice de VA globales para una escuela concreta en el MVA de Dallas se construyen como una media ponderada de la estimación de los efectos escolares (r_{0j}) en diferentes cursos y grados.

V.2.2.3 Ganancia estimada

Por último, otro salto en la estimación de medidas de ganancia se produce al considerar el nivel inicial (pre-test) y el actual (pos-test) como medidas de resultados. De esta forma la ganancia quedaría expresada de la siguiente manera:

$$E(Y_1) - E(Y_0) \qquad \textbf{Ec. V.29}$$

Entonces,

$$Y_{0i} = \beta_{00} + e_{0i}$$
$$Y_{1i} = \beta_{10} + e_{1i} \qquad \textbf{Ec. V.30}$$

Donde β_{00} y β_{10} son las medias de todos los estudiantes en el pretest y en el postest; e_i es el residuo en cada una de las dos mediciones, es decir, lo que se desvía cada estudiante de la media estimada, son independientes y se distribuyen con de forma normal con media cero y varianza común para todos los estudiantes

Por tanto la ganancia estimada es:

$$G_i = (\beta_{10} + e_{1i}) - (\beta_{00} + e_{0i}) \qquad \textbf{Ec. V.31}$$

Si estas ganancias se asocian con escuelas o docentes también pueden considerarse análisis del VA, es decir, miden la aportación de estas al cambio en rendimiento de los estudiantes. Para llevar a cabo este análisis es necesario un modelo de dos niveles[63] (estudiante y escuela):

$$\text{Nivel 1 (estudiantes): } Y_{tij} = \beta_{0ij} + \beta_{1ij} + r_{tij} \qquad \textbf{Ec. V.32}$$

De la misma forma, en notación matricial:

$$\begin{bmatrix} Y_{1ij} \\ Y_{2ij} \end{bmatrix} = \begin{bmatrix} 10 \\ 01 \end{bmatrix} \begin{bmatrix} \beta_{0ij} \\ \beta_{1ij} \end{bmatrix} + \begin{bmatrix} r_{0ij} \\ r_{1ij} \end{bmatrix} \qquad \textbf{Ec. V.33}$$

$$r_{ij} \sim N(0, R) \qquad \textbf{Ec. V.34}$$

En el primer nivel se estiman dos coeficientes fijos: la puntuación media de los estudiantes de la escuela *j* en la primera medición, es decir, el pretest (β_{0ij}) y la puntuación media en el postest (β_{1ij}). Y también estima los coeficientes aleatorios (r_{tij}) que son las diferencias entre la puntuación de un estudiante y la de su escuela.

Ambos residuos se distribuyen de forma normal con medias cero y matriz de varianzas-covarianzas R (Ec. V.35). Con varianza σ^2_{r00} y σ^2_{r11}, respectivamente y covarianza σ^2_{r10}.

$$R = \begin{bmatrix} \sigma^2_{r00} \\ \sigma^2_{r10} \sigma^2_{r11} \end{bmatrix} \qquad \textbf{Ec. V.35}$$

En el nivel 2 (escuelas) cada coeficiente de nivel 1 pasa a ser una ecuación en el segundo nivel (Ec. V.36):

$$\begin{aligned} \beta_{0j} &= \beta_0 + u_{0j} \\ \beta_{1j} &= \beta_1 + u_{1j} \end{aligned} \qquad \textbf{Ec. V.36}$$

[63]Aunque es un modelo de dos niveles para su desarrollo es necesario un nivel que defina la estructura multivariante de la variable dependiente pero sin varianza aleatoria (Rasbash, Steele, Browne y Goldstein, 2009)

β_0 es la media general en el pretest β_1 en el postest; u_{0j} es la aportación de una escuela *j* al estatus inicial; y, finalmente, u_{1j} es el efecto diferencial de la escuela *j* sobre el postest. Por tanto, la ganancia estimada para una escuela:

$$G_j = (\beta_1 + u_{1j}) - (\beta_0 + u_{0j}) \qquad \textbf{Ec. V.37}$$

En resumen, las medidas de ganancia han evolucionado desde la simple diferencia de medias para conseguir análisis fiables que reflejen el cambio en aprendizaje de los estudiantes. Sin embargo, los modelos que utilizan más de dos puntos temporales parecen más adecuados para analizar ese cambio.

Un problema específico de los modelos de ganancia bruta y ganancia estimada es que no consideran los efectos de los cursos previos en el análisis de la ganancia. Es decir, en estos análisis lo ha sucedido en los cursos anteriores no influye en los cambios medidos entre los cursos posteriores.

Los modelos de ganancia residual o ajuste de covariables tienen otro inconveniente y es que pueden producir unos resultados sesgados al introducir el rendimiento previo como covariable. Los modelos de regresión asumen que los predictores están medidos sin error y las puntuaciones de los test son medidas falibles del rendimiento, es decir, se encuentran acompañadas de un error de medida. Introducir el rendimiento previo sin considerar este error de medida puede ser problemático (McCaffrey, Lockwood, Doretz & Hamilton, 2003; Wiley, 2006; Choi, Goldschmidt & Yamashiro, 2006).

Tekwe y otros autores (2004) llevan a cabo un estudio comparativo de modelos que utilizan las puntuaciones de ganancia como variable de resultados y una versión del modelo EVAAS con únicamente dos tomas de datos y no encuentran diferencias sustanciales en las estimaciones producidas por cada uno de ellos. Excepto con respecto al modelo ajustado, el que introduce covariables de contexto de los estudiantes en el análisis de regresión multinivel. Esta cuestión es un indicador de que los modelos con dos únicas tomas de datos pueden verse afectados, en mayor medida, por las variables contextuales. Añadir más tomas de datos en los análisis, es decir, utilizar datos longitudinales, puede ser una manera de suavizar estos efectos.

V2.3 Modelos de crecimiento

Estos modelos estadísticos de análisis utilizan más de dos puntuaciones del rendimiento de los estudiantes como variable dependiente. No utilizan una única puntuación de resultados o la ganancia como variable dependiente, sino que incluyen el vector entero de puntuaciones de logro de los estudiantes a lo largo de diferentes momentos temporales.

Dentro de esta categoría es posible distinguir entre, por un lado, aquellos que consideran los datos completamente anidados durante todo el proceso, normalmente los niveles de anidamiento son las diferentes ocasiones de medida, los estudiantes y las escuelas, y se desarrollan a través de los modelos de regresión multinivel o modelos jerárquicos. Y, por otro lado, los que permiten variaciones de los estudiantes entre las diferentes unidades de análisis, por ejemplo que los estudiantes cambien de aula o de escuela de un curso al siguiente, de un momento de medida a otro. El aspecto del anidamiento puede abordarse desde dos aproximaciones distintas: los modelos lineales mixtos o la perspectiva multinivel.

Los modelos que permiten el cruce entre los diferentes niveles consideran, normalmente, esos efectos estimados de los docentes o las escuelas como persistentes y constantes en el tiempo. Pero es posible flexibilizar ese supuesto modelando la persistencia de esos efectos. El modelo de persistencia es un ejemplo de análisis que considera los efectos de cursos anteriores como cambiantes a lo largo del tiempo, pueden disminuir entre aplicaciones (McCaffrey, Lockwood, Doretz & Hamilton, 2003; McCaffrey, Koretz, Louis & Hamilton, 2004).

Las características fundamentales que comparten los modelos de VA que utilizan medidas de crecimiento de los estudiantes son las siguientes:

- Siguen la trayectoria de crecimiento en rendimiento de un mismo estudiante a lo largo de un periodo de tiempo determinado. Para ello utilizan más de dos mediciones de su logro académico. Lo importante no es saber cuánto rinde una escuela de media sino conocer cuánto cambio produce durante un periodo de tiempo concreto.
- La perspectiva multinivel estudia el crecimiento durante un periodo de tiempo determinado con dos parámetros principales: el estatus

inicial y una tasa de cambio en el rendimiento, una pendiente de crecimiento. Las perspectiva de los modelos lineales mixtos no incorpora esta tasa de crecimiento, en su lugar incluye un coeficiente para cada ocasión de medida y el efecto de cada escuela o docente asociado a esos coeficientes.

- Las puntuaciones deben estar en una escala común para poder ser comparadas, es decir, si se quiere observar el crecimiento a lo largo de varios cursos académicos las medidas que se utilicen para evaluar el rendimiento debe situarse en la misma escala para poder calcular la cantidad de crecimiento logrado por los estudiantes. Normalmente se utilizan escalas verticales de rendimiento.
- Se analizan los efectos de las escuelas o los docentes sobre la ganancia en aprendizaje de sus estudiantes. El foco de interés no es el estudiante, es el profesor, la escuela o incluso el distrito. Se busca determinar qué cantidad de cambio, producido en el logro escolar del estudiante, puede ser atribuida a una escuela o profesor particular.
- Para conocer si las escuelas rinden por encima o por debajo de lo esperado se debe conocer qué cantidad es la esperada. Es posible utilizar la media de crecimiento de escuelas con similares características o algún estándar especificado previamente.

Esta clase de modelos pueden incorporar, o no, variables del contexto del estudiante o de las escuelas. Por ejemplo, el modelo EVAAS no incluye porque asumen que la utilización de medidas múltiples de rendimiento de cada estudiante elimina ese efecto del contexto porque es el propio estudiante el que sirve como elemento de control (Sanders, Saxton & Horn, 1997; Sanders & Wright, 2008).

V.2.3.1 Modelo de efectos anidados

El crecimiento se define en estos modelos como una función a lo largo del tiempo. Estos modelos de crecimiento completamente anidados asumen la existencia de una pendiente de crecimiento entre las diferentes ocasiones de medida de un estudiante. En otras palabras, las diferentes puntuaciones de resultados se anidan en cada individuo, por tanto, la trayectoria de cambio se

representa como una función específica de cada individuo más un error aleatorio (Willett, 1989a; Bryk & Raudenbush, 2002; Goldstein & Woodhouse, 2001). A su vez, los estudiantes pueden estar agrupados en aulas y/o escuelas, incluso niveles superiores. Sin embargo, los modelos anidados no siguen a los alumnos que cambian de unidad durante la evaluación.

El modelo de crecimiento lineal de un individuo, con una tasa constante de cambio, es una función lineal que tiene en cuenta la variable tiempo. La representación algebraica del modelo de crecimiento contiene dos parámetros: el estatus inicial de un alumno i en la primera medición (β_{0i}) y la tasa de ganancia o pendiente (β_{1i}), de la siguiente manera:

$$Y_{ti} = \beta_{0i} + \beta_i(T) + e_{ti} \qquad \textbf{Ec. V.38}$$

Y_{ti} es la puntuación de rendimiento de un individuo i en un momento temporal t y depende de su estatus inicial (β_{0i}), es decir, el nivel de logro en el momento la primera ocasión de medida, y la tasa de crecimiento (β_i) que está en función del tiempo T, que depende de las ocasiones de medida y de la variable utilizada para la identificación de ese cambio temporal (edad, curso evaluado, número de ocasiones de medida, meses, etc.).

Para cada sujeto se estima una curva de crecimiento individual que representa su desarrollo con el paso del tiempo. Los parámetros principales estimados, difieren entre estudiantes, cada uno con su nivel inicial y una tasa específica de cambio, como muestra la Figura V.1:

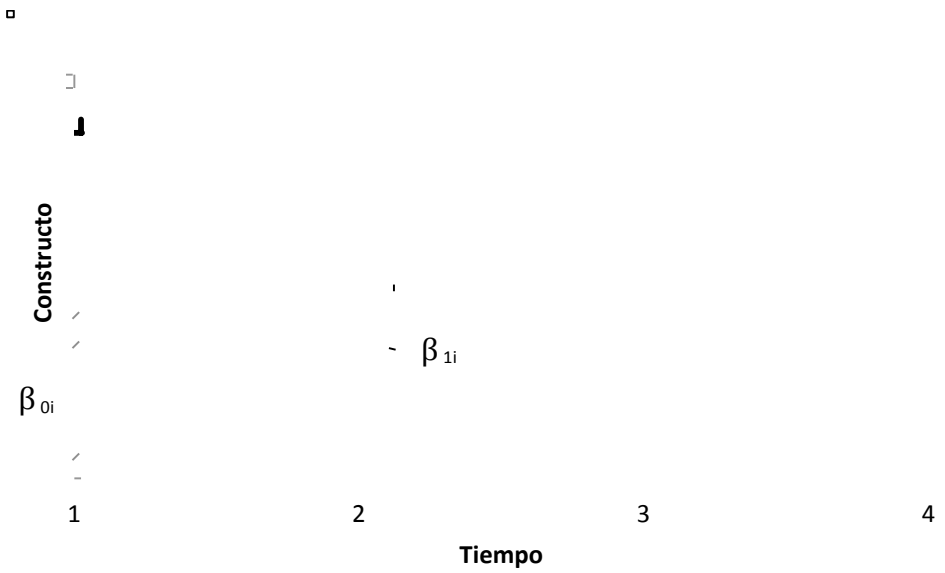


Figura V.1. Estatus inicial y crecimiento en rendimiento en función del tiempo.
Fuente: Elaboración Propia

La utilización de un modelo lineal de crecimiento no es la única opción posible para modelar la trayectoria de cambio de un estudiante. Modelos cuadráticos, cúbicos, exponenciales o funciones spline son algunas de las alternativas. Una de las tareas más importantes del investigador es seleccionar la función de tiempo adecuada a la realidad evaluada, por tanto, un análisis empírico del comportamiento de distintas medidas temporales es importante y se lleva a cabo en este trabajo.

Debido a que los estudiantes pueden diferir tanto en sus puntos de partida como en sus tasas de crecimiento y, además, estos alumnos se encuentran agrupados en aulas o escuelas que también pueden incorporar varianza entre sus puntos de partida y sus pendientes. Por tanto, conviene ampliar el modelo de crecimiento longitudinal añadiendo al menos dos niveles más (estudiante y escuelas pero también es posible añadir varianza entre aulas, distritos, comunidades autónomas, etc.).

Un ejemplo de modelo de crecimiento que considera el anidamiento completo de los datos para llevar a cabo el análisis del VA es el que se desarrolla desde la aproximación multinivel, los denominados modelos de curva de crecimiento.

Modelo multinivel de crecimiento

El modelo que se describe a continuación es completamente aleatorio porque sus parámetros (estatus inicial y pendiente de crecimiento) varían de forma aleatoria en los niveles superiores (estudiante y escuelas). Por tanto, los estudiantes y las escuelas pueden tener puntos de partida y trayectorias de crecimiento distintas que se capturan con los residuos aleatorios de ambos niveles. El residuo de cada escuela asociado al crecimiento es la estimación del VA.

El residuo asociado a la pendiente de crecimiento cada escuela cuantifica la distancia respecto a la tas de crecimiento global. Es la aportación diferencial de un centro educativo, su VA, y tiene una concepción distinta a los modelos de efectos cruzados que estima un residuo escolar en cada una de las aplicaciones. Esto se debe a que los estudiantes pueden cambiar de escuela (o docente) entre aplicaciones y, por tanto, no enseñan a los mismos grupos de estudiantes.

También es posible considerar los efectos de las escuelas como fijos en el modelo y, en consecuencia, no sería necesario añadir el tercer nivel de agregación. De esta forma, la media de los residuos de la pendiente de crecimiento de los estudiantes de cada centro sería suficiente para calcular el VA de una escuela.

A modo de ejemplo, se desarrolla el modelo multinivel de crecimiento para una cohorte de estudiantes a lo largo de cuatro ocasiones de medida, considerando tres niveles de anidamiento (tiempo, estudiantes y aulas). En este caso, la variable criterio consta de cuatro puntuaciones de resultados escaladas verticalmente. La variable tiempo está definida por los meses transcurridos entre aplicaciones y se considera la primera medición como el punto de partida $T(0, 8, 13, 20)$. Se utilizan meses porque la distancia entre aplicaciones es distinta. En el caso de recoger la información siempre en el mismo momento temporal la variable que define el tiempo podría cambiarse por el número de aplicación (0,1,2,3,…)

Los diferentes niveles de anidamiento en este modelo son los siguientes:

- Nivel 1: El tiempo, las ocasiones de medida, es decir, la trayectoria de crecimiento.
- Nivel 2: El estudiante, las ocasiones de medida se encuentran anidadas en cada sujeto

- Nivel 3: El centro, los estudiantes se agrupan en escuelas. El modelo puede especificarse mucho más identificando aulas o profesores.

Nivel 1 (modelo de crecimiento individual)

$$Y_{tij} = \beta_{0ij} + \beta_{1ij}(T) + e_{tij} \qquad \textbf{Ec. V.39}$$

En la ecuación Ec. V.39 Y_{tij} Es el logro en una ocasión de medida determinada t (1,2,3,4) de un alumno i de una escuela j.

β_{0ij} es el estatus inicial de partida del estudiante i de la escuela j y β_{1ij}su tasa de crecimiento. Esta tasa de crecimiento debe multiplicarse por la variable tiempo definida previamente T. Por ejemplo, el rendimiento estimado para un estudiante en la primera ocasión de medida es:

$$Y_{1ij} = \beta_{0ij} + \beta_{1ij}(0) + r_{1ij} = \beta_{0ij} + e_{1ij} \qquad \textbf{Ec. V.40}$$

Es igual al estatus inicial más el residuo. En la segunda medición es igual al estatus inicial más la pendiente estimada por los meses transcurridos entre aplicaciones, como muestra la siguiente ecuación:

$$Y_{2ij} = \beta_{0ij} + \beta_{1ij}(8) + e_{2ij} \qquad \textbf{Ec. V.41}$$

Este tipo de modelos asumen que el error residual de primer nivel tiene distribución normal con media cero y varianza común entre aplicaciones:

$$e_{tij} \sim N(0, \sigma^2) \qquad \textbf{Ec. V.42}$$

Nivel 2 (Alumnos).

$$\beta_{0ij} = \beta_{0j} + r_{0ij}$$

$$\beta_{1ij} = \beta_{1j} + r_{1ij} \qquad \textbf{Ec. V.43}$$

$$r_{ij} \sim N(0, R)$$

Cada coeficiente de nivel 1 pasa a ser una ecuación en el segundo nivel. En la ecuación Ec. V.43 β_{0j}es la media del estatus inicial de los estudiantes de la escuela *j*; β_{1j}es la media de la tasa de crecimiento dentro de esa escuela *j*. Los residuos de

los estudiantes asociados al estatus y la pendiente son r_{0ij} y r_{1ij} respectivamente. Ambos residuos se distribuyen de forma normal con medias cero y matriz de varianzas-covarianzas R:

$$R = \begin{bmatrix} \sigma^2_{r00} \sigma^2_{r01} \\ \sigma^2_{r10} \sigma^2_{r11} \end{bmatrix} \qquad \textbf{Ec. V.44}$$

Estos modelos asumen la correlación entre los residuos del estatus inicial y el crecimiento pero no entre las diferentes puntuaciones de logro de los estudiantes anidadas dentro de los propios alumnos. En este nivel es posible introducir características del contexto de los estudiantes para ajustar los resultados si fuera necesario. Deben estar libres de error de medida y no correlacionar con los residuos para evitar un sesgo en los resultados.

Nivel 3 (escuelas):

$$\beta_{0j} = \beta_{00} + u_{0j}$$
$$\beta_{1j} = \beta_{10} + u_{1j} \qquad \textbf{Ec. V.45}$$

Este nivel representa las medias globales, de toda la muestra, en el estatus inicial (β_{00}) y la tasa de crecimiento (β_{10}). Las medias tienen residuos aleatorios asociados a cada escuela que indican la aportación diferencial respecto a esa media global en estatus inicial (u_{0j}) y respecto a la pendiente de cambio en aprendizaje asociada al tiempo(u_{1j}).Este último residuo puede considerarse el VA de una escuela j y se encuentra vinculado también al tiempo.

Si se sustituye para formular una única ecuación se observa qué residuos del estudiante y la escuela se encuentran vinculados a esa función de tiempo:

$$Y_{tij} = \beta_{00} + \beta_{10}(T) + r_{0ij} + r_{1ij}(T) + u_{0j} + u_{1j}(T) + e_{tij} \qquad \textbf{Ec. V.46}$$

Este tipo de análisis es capaz de modelar patrones de crecimiento no lineales con la inclusión de más términos en la ecuación. El modelo descrito es el denominado condicional o nulo, es decir, sin la inclusión de ningún predictor en los diferentes niveles estudiados.

Las estimaciones de los efectos aleatorios de los docentes o las escuelas son tipo BLUP o estimaciones empírico bayesiano y tienen la peculiaridad ya descrita de encoger los resultados estimados de las escuelas con poca muestra hacia la media global.

El MVA desarrollado en Chicago (Bryk, Thum, Easton & Luppescu, 1998; Ponisciak & Bryk, 2005) es una ejemplo de este tipo de análisis. Alguno de los modelos realizados por el CRESST[64] (Choi, Seltzer, Herman & Yamashiro, 2007) en California siguen las mismas premisas. Zvoch y Stevens también emplean este tipo de modelos estadísticos para analizar los datos longitudinales de rendimiento (2003; 2006)

V.2.3.2 Modelo de efectos cruzados

Las estructuras de medición longitudinales pueden alcanzar una dimensión más compleja que supera el anidamiento de los datos, sobre todo cuando el número de ocasiones de medida se incrementa o el objetivo de la evaluación son los profesores. Los estudiantes pueden tardar más tiempo en cambiar de escuela pero si el foco de análisis son los efectos del docente lo más probable es que cambien de profesor de un curso a otro. Existe un grupo de modelos que incorpora efectos aleatorios cruzados que permiten que un estudiante reciba, por ejemplo, el efecto de dos profesores distintos en un mismo curso académico o en dos ocasiones de medida diferentes.

Los modelos de análisis estadístico utilizados comúnmente para tratar este tipo de estructura son los modelos lineales mixtos (Sanders & Horn, 1994; Sanders, Saxton & Horn, 1997; McCaffrey, Lockwood, Koretz & Hamilton, 2003; McCaffrey, Koretz, Louis & Hamilton, 2004). Pero es posible considerar esta estructura cruzada desde la aproximación de los modelos multinivel con los modelos de clasificación cruzada (*cross-classified*) (Bryk & Raudenbush, 2002).

Es posible hacer una distinción dentro de los modelos que permiten el cambio entre unidades de análisis. Los que consideran los efectos estimados como persistentes de un año al siguiente, es decir, el efecto del docente permanece

[64]Center for Research on Evaluation, Standards & Student Testing es un centro vinculado a la Universidad de UCLA. Web: http://www.cse.ucla.edu/

constante en los años posteriores, cuyo principal representante es el modelo EVAAS, también denominado modelo estratificado porque añade los efectos previos de los docentes en los resultados actuales (Sanders & Horn, 1994; Sanders, Saxton & Horn, 1997); y aquellos que permiten una disminución de estos efectos previos cuando se incluyen en los resultados. El modelo de persistencia es el principal ejemplo (McCaffrey, Lockwood, Doretz & Hamilton, 2003).

V2.3.2.1 Modelo de efectos persistentes

En esta sección se describen los dos principales modelos que consideran la persistencia de los efectos de las escuelas o los docentes y que permiten la estimación de efectos cruzados, es decir, que los estudiantes puedan cambiar de profesor, aula o escuela entre las ocasiones de medida. El modelo EVAAS desde la perspectiva de los modelos lineales mixtos y el de efectos cruzados desde la perspectiva multinivel.

El Modelo de Valor Añadido de Tennessee (EVAAS)

El *Tennessee Value-Added Assessment System* (TVAAS) (Sanders & Horn, 1994; Sanders, Saxton, & Horn, 1997) también conocido como modelo estratificado (*Layered Model*) o en capas, ha sido uno de los pioneros en llevar a cabo el análisis del VA y se desarrolló en el estado de Tennessee con el objetivo de aprovechar los diferentes datos longitudinales de rendimiento del programa "*Comprehensive Assessement Program*". Actualmente se denomina *Education Value-Added Assessment System* (EVAAS) y lo desarrolla la empresa de software estadístico SAS[65]. Utiliza para el análisis los mencionados MLM de Henderson, adecuados para estudiar muestras complejas como datos longitudinales y estructuras de datos anidadas o cruzadas. El EVAAS en una de las aproximaciones más complejas.

Tennessee fue el primer estado que adoptó formalmente los análisis de VA como parte de una iniciativa para el mejora escolar. En relación con el trabajo desarrollado por William Sanders en la Universidad de Tennessee, en 1993 una ley estatal requirió que las escuelas y distritos recogieran e hicieran llegar datos de los estudiantes al profesor Sanders. Esta legislación explícitamente prohibía la

[65]Más información sobre el modelo EVVAS en su página web: http://www.sas.com/govedu/edu/k12/evaas/index.html

utilización de los resultados de VA como herramienta para la rendición de cuentas de escuelas o profesores, estos debían servir exclusivamente para el desarrollo escolar y dejó en manos de cada distrito la decisión de utilizar o no los resultados del modelo EVAAS.

El modelo EVAAS es longitudinal y recopila datos de los sujetos en Matemáticas, Ciencias, Estudios Sociales y destrezas lectoras y de lenguaje, desde el tercer al octavo grado. Los análisis son dirigidos por cada distrito escolar y los informes de las escuelas proporcionan resultados del curso en el que se lleva a cabo la evaluación, los dos años previos y la media de crecimiento de los tres años.

La variable input es toda la matriz de datos de los estudiantes en los diferentes grados, materias, docentes y escuelas. Por este motivo la complejidad de la estimación aumenta en gran medida. De forma diferente a los modelos multinivel de crecimiento, no estima una pendiente de la ganancia sino que se encuentra implícita en el modelo. Estima una puntuación en cada ocasión de medida y la varianza asociada a esas medias entre estudiantes, docentes y escuela de forma cruzada. El cambio se estima como ganancia entre aplicaciones considerando esa varianza diferencial.

El modelo EVAAS no incluye covariables para controlar el contexto del estudiante o de la escuela, sino que ajusta los resultados añadiendo las puntuaciones previas de los efectos de los docentes en los análisis de los siguientes cursos, de ahí proviene la denominación de modelo estratificado o por capas. Los autores consideran que con datos longitudinales multivariados cada estudiante se controla con su propia trayectoria y se evita así la necesidad de incorporar variables de contexto del estudiante en el modelo (Sanders & Horn, 1994).

Una modificación del modelo EVAAS se probó con la finalidad de conocer las variaciones producidas por la inclusión de variables demográficas y de estatus socioeconómico del estudiante (Ballou, Sanders & Wright, 2004) y analizar sus posibles efectos en las estimaciones de VA. Los autores encontraron correlaciones altas (0,9) entre el modelo con y sin predictores. También señalan que ambos tienen una potencia similar para identificar efectos de los docentes como estadísticamente diferentes de la media.

El modelo EVAAS tiene en cuenta dos supuestos clave sobre las puntuaciones de los test (Wiley, 2006):

- Las puntuaciones de rendimiento de un estudiante reflejan los efectos del profesor actual así como los de los docentes que ha tenido anteriormente y asume que dichos efectos permanecen constantes y estables a lo largo del tiempo.
- Las puntuaciones de rendimiento también capturan las características personales de cada estudiante.

El modelo EVAAS produce estimaciones de VA de los docentes teniendo en cuenta los posibles efectos de las correlaciones entre las diferentes puntuaciones del estudiante, para ello utiliza una matriz de varianzas-covarianzas sin estructura que permita la covarianza entre los residuos de las puntuaciones del mismo estudiantes pero no entre los residuos de distintos alumnos (McCaffrey, Koretz, Louis & Hamilton 2004).

El modelo EVAAS causó un gran interés entre los educadores y ha provocado la creciente utilización de modelos de VA en diferentes países y para diferentes fines. Algunas cuestiones que hicieron este modelo tan interesante fueron las siguientes:

- El proceso enfatiza el crecimiento y, por tanto, es consistente con el propósito escolar de mejorar a sus estudiantes.
- La metodología de análisis empleada permite tratar con grandes series de datos, aunque no se cuente con toda la información de un estudiante, es decir, haya datos perdidos en algún punto temporal.
- Permite estimar los efectos educativos sobre las ganancias en rendimiento sin la necesidad de contar con medidas directas de ganancia de cada estudiante. En su lugar, las estimaciones de la ganancia pueden obtenerse con alta precisión de la solución del vector de ecuaciones del modelo mixto, porque las covarianzas entre todas las puntuaciones de cada estudiante son incluidas en el modelo (Sanders, Saxton & Horn, 1997). Incluye una estructura del error

residual de las puntuaciones de los estudiantes que permita la correlación.

- Centrándose en esa medida del cambio, el propio sujeto actúa como control y no son necesarias otras variables de contexto.
- Puede utilizarse para evaluar programas y personas y agregaciones en centros, distritos, etc.
- Utiliza estimaciones BLUP para calcular los efectos de los docentes.
- El modelo trata de aislar causas, extrayendo los efectos el rendimiento previo de la puntuación de logro actual. No obstante, asumir que estos modelos están estimando efectos causales es uno de los aspectos más criticados.

También se han hecho objeciones a este modelo de análisis del VA. Se ha calificado esta aproximación como demasiado técnica. Es decir, cómo puede explicarse a los que toman las decisiones en los centros educativos o al público en general, las implicaciones que tienen para la educación la interacción de los efectos del rendimiento previo, los efectos de los datos perdidos, las influencias del contexto, la multidimensionalidad de un dominio latente, etc. También se han criticado aspectos relacionados con las características metodológicas del modelo como no incluir predictores de contexto, utilizar estimadores bayesianos de los residuos, etc. (Kupermintz, Shepard & Linn, 2001; Kupermintz, 2002; Hibpshman, 2004; McCaffrey, Koretz, Louis & Hamilton, 2004; Rubin, Stuart & Zanutto, 2004; Tekwe et al., 2004; Goldschmidt et al., 2005; Wiley, 2006)

Con este modelo existe la posibilidad de no utilizar una escala vertical del logro académico porque estima un efecto diferencial para cada docente o escuela en cada aplicación de medida y considera los efectos previos en los resultados actuales. Pero si decide utilizar esa medida de resultados no podría calcularse la ganancia entre aplicaciones, únicamente la aportación diferencial del docente en cada aplicación.

Las puntuaciones de resultados de un estudiante con el modelo EVAAS serían las siguientes:

$$Y_{1i} = \beta_1 + Pu_{1j} + r_{1i}$$

$$Y_{2i} = \beta_2 + Pu_{1j} + Pu_{2j} + r_{2i}$$

$$Y_{3i} = \beta_3 + Pu_{1j} + Pu_{2j} + Pu_{3j} + u_{3i}$$

$$Y_{4i} = \beta_4 + Pu_{1j} + Pu_{2j} + Pu_{3j} + Pu_{4j} + r_{4i}$$

Ec. V.47

El modelo añade un coeficiente fijo para cada ocasión de medida y también efectos aleatorios cruzados mediante un sistema complejo de matrices que permite el cambio de estudiantes entre los distintos docentes e incluye los efectos previos en los resultados actuales.

Y_{ti} es la puntuación en el test en un determinado año *t* del alumno *i*; β_t es la media global para ese mismo año *t*; u_{tj} es la contribución del profesor *j* en el año *t*, es decir, el VA del docente. Este VA son desviaciones respecto a la media global y se distribuyen de forma normal con media cero y varianzas específicas para cada año, además de ser independientes dentro y entre las diferentes ocasiones de medida; P es la medida de proporción de escolarización de un estudiante con un determinado docente, si no ha sido atendido por ese profesor el valor será 0; r_t es el componente residual del estudiante en el año *t*, con distribución normal y con una matriz de varianzas-covarianzas intra-estudiantes no restringida que permite diferentes patrones de correlación entre mediciones pero no entre estudiantes.

El análisis EVAAS elabora un único modelo para los resultados de diferentes materias educativas por grado, de diferentes cohortes de estudiantes que pertenecen a escuelas distintas y que son enseñados por profesores distintos. Este modelo se ha simplificado para incluir solo cuatro puntuaciones de logro de un estudiante escaladas verticalmente que han cambiado de profesor en cada aplicación pero dentro de una misma escuela:

$$Y_{tij} = \beta_t + \sum_{t=1}^{T} Pu_{t(ij)} + r_{tij}$$

Ec. V.48

En este modelo el cambio se encuentra implícito. Realmente se estiman las medias globales en cada una de las ocasiones de medida como parámetros fijos en el modelo y el residuo asociado aleatorio del estudiante asociado a cada uno de

ellos. La estimación de los efectos cruzados de los docentes se lleva a cabo estimando un residuo para cada uno de ellos en cada ocasión de medida, utilizando una matriz de coeficientes que permita el cambio entre docentes en un mismo curso y también los cambios entre cursos. Además, los errores residuales de los estudiantes (r_{tij}) asociados a cada puntuación de rendimiento pueden estar correlacionados para un mismo sujeto y, de esta forma, captar la estructura longitudinal de la variable dependiente.

La Tabla V.4 muestra un ejemplo de los coeficientes utilizados para estimar los efectos aleatorios de los docentes. Son dos sujetos de una misma escuela que asisten a clases con profesores distintos en cada medición. El estudiantes uno es un caso especial porque asistió, en un mismo año, al 50% de las clases del profesor A y al 50% con el B y por eso muestra un coeficiente igual a 0,5.

			Disminución			
Estudiante	Año	Docente	A	B	C	D
1	1	A	0,5	0,5		
1	2	B	0,5	1		
1	3	C	0,5	1	1	
1	4	D	0,5	1	1	1
2	1	B		1		
2	2	C		1	1	
2	3	A	1	1	1	
2	4	D	1	1	1	1

Tabla V.4. Ejemplo de efectos aleatorios cruzados

El estudiante dos recibe la docencia del profesor B en la primera medición y esos efectos permanecen constantes en las otras aplicaciones por eso también tiene un valor 1 en el resto de medidas.

Los coeficientes fijos β_t se incluyen en el modelo llevando a cabo una codificación de contraste. Por ejemplo, de la siguiente manera:

$$\begin{bmatrix}\beta_1\\ \beta_2\\ \beta_3\\ \beta_4\end{bmatrix}\begin{bmatrix}1&0&0&0\\ 0&1&0&0\\ 0&0&1&0\\ 0&0&0&1\end{bmatrix} \qquad \textbf{Ec. V.49}$$

Cuando se incluyen los cuatro predictores en el modelo, cada uno de ellos refleja la puntuación media estimada en cada aplicación. En este caso β_1 es la puntuación media en la primera ocasión de medida. Si se pretende analizar las

diferencias entre dos grados adyacentes e identificar el efecto de cada docente se deben resolver las ecuaciones. Por ejemplo, la diferencia entre las puntuaciones obtenidas por un estudiante los dos primeros años es:

$$Y_2 - Y_1 = (\beta_2 - \beta_1) + Pu_2 + (r_2 - r_1) \qquad \textbf{Ec. V.50}$$

Se asume que el efecto previo del profesor permanece con el estudiante cuando progresa. Por tanto, el efecto del profesor en la ganancia del estudiante es lo que queda una vez eliminado el efecto de la ganancia del estudiante, la ganancia en la media global y la contribución de factores característicos del estudiante si los incluyera (Ballou, Sanders & Wright, 2004):

$$Pu_2 = (Y_2 - Y_1) - (\beta_2 - \beta_1) - (r_2 - r_1) \qquad \textbf{Ec. V.51}$$

Modelo de clasificación cruzada (cross-classified)

Desde la perspectiva de los modelos jerárquicos lineales también es posible identificar los cambios de los estudiantes entre profesores o escuelas a lo largo de las diferentes mediciones mediante el modelo de clasificación cruzada (Bryk & Raudenbush, 2002). La principal diferencia con el modelo EVAAS es que asume que el crecimiento en rendimiento es predecible y con una tasa a largo del tiempo. La tasa de crecimiento se analiza de la misma forma que en la aproximación completamente anidada de los VAM. Por tanto, para cada ocasión de medida la ecuación quedaría formulada de la siguiente forma:

$$\begin{aligned}
Y_{1i} &= \beta_0 + T_0(\beta_1 + r_{1i}) + r_{0i} + u_{1j} + e_{1i} \\
Y_{2i} &= \beta_0 + T_1(\beta_1 + r_{1i}) + r_{0i} + u_{1j} + u_{2j} + e_{2ij} \\
Y_{3i} &= \beta_0 + T_2(\beta_1 + r_{1i}) + r_{0i} + u_{1j} + u_{2j} + u_{3j} + e_{3ij} \\
Y_{4i} &= \beta_0 + T_3(\beta_1 + r_{1i}) + r_{0i} + u_{1j} + u_{2j} + u_{3j} + u_{4j} + e_{4ij}
\end{aligned} \qquad \textbf{Ec. V.52}$$

Y_{ti} es la puntuación en el momento temporal *t* para el alumno *i* pero en cada aplicación las escuelas pueden estar compuestas por estudiantes distintos; β_0 es la media global del estatus inicial y β_1 es la tasa de crecimiento vinculada al término lineal de crecimiento T, que en este ejemplo toma los valores 0, 1, 2 y 3 para las cuatro ocasiones de medida. De esta forma, la primera aplicación es el estatus

inicial porque el término de crecimiento es igual a cero. La ecuación general queda formulada en Ec. V.53.

$$Y_{tij} = \beta_{00} + r_{0i} + T_{ti}(\beta_{10} + r_{1i}) + \sum_{j=1}^{J}\sum_{t=0}^{T} Du_{t(ij)} + e_{tij} \qquad \textbf{Ec. V.53}$$

Los coeficientes *r* son los efectos aleatorios de los estudiantes asociados al estatus inicial (r_{0i}) y al crecimiento (r_{1i}) y se distribuyen de forma normal con una matriz de varianzas-covarianzas:

$$\begin{pmatrix} r_{0i} \\ r_{1i} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{r0} & \\ \sigma_{r1,r0} & \sigma^2_{r1} \end{pmatrix}\right] \qquad \textbf{Ec. V.54}$$

$u_{t(ij)}$ es el efecto aleatorio de la escuela que se encuentra cruzado con los estudiantes debido a esa posibilidad de cambio de docente entre aplicaciones. Se define como desviaciones esperadas en el crecimientos que se producen cuando el estudiante se encuentra con el profesor *j* en la ocasión de medida *t*. Este efecto se suma a lo largo de todos los profesores y aplicaciones en un mismo estudiante, es persistente y se acumula. La puntuación del estudiante se atribuye al efecto del profesor actual y también de los docentes previos. Tiene una distribución normal con media cero y matriz de varianzas-covarianzas constante entre aplicaciones, también es independiente de otros efectos del modelo ($N(0, \tau_{u00})$).

La acumulación de efectos del docente depende de la variable D que toma un valor igual a 1 cuando el estudiante recibe la enseñanza del profesor *j* en la aplicación *t*, y el valor cero en caso contrario.

e_{tij} es el término de error residual intra-sujetos $N(0, \sigma^2)$, la varianza de este término residual es constante a lo largo de las mediciones. Es un escalar porque se asume que el de crecimiento captura los efectos que las características del estudiante tiene sobre las puntuaciones de logro (McCaffrey, Lockwood, Doretz & Hamilton, 2003).

Si separamos los niveles, en este caso encontramos únicamente dos. En el nivel 1 se define el crecimiento a lo largo del tiempo, de la misma forma que en el modelo de crecimiento multinivel completamente anidado:

$$Y_{tij} = \beta_{0ij} + \beta_{1ij}(T) + e_{tij} \qquad \textbf{Ec. V.55}$$

En el nivel 2 se incluye tanto la variación entre estudiantes en el estatus inicial y el crecimiento (r) como los efectos cruzados de estudiantes y escuelas ($u_{t(ij)}$) y ambos aparecen en el mismo nivel del modelo jerárquico como muestra la ecuación Ec. V.56:

$$\beta_{0ij} = \beta_0 + r_{0i} + \sum_{j=1}^{J}\sum_{t=0}^{T} Du_{t(ij)} \qquad \textbf{Ec. V.56}$$

$$\beta_{1ij} = \beta_1 + r_{1i}$$

En la tasa de crecimiento, en lugar de utilizar el crecimiento medio de los estudiantes de un docente *j*, como en los modelos completamente anidados, en el modelo de clasificación cruzada se emplea la media de todos los estudiantes de la muestra. La tasa de crecimiento incluye el residuo aleatorio de los estudiantes, la diferencia respecto a la media global (r_{1i}), mientras que el modelo multinivel incorpora la diferencia respecto a los estudiantes de un determinado docente o escuela.

El modelo de clasificación cruzada considera la persistencia de esos efectos pero también es posible asumir que desaparezcan. En el supuesto caso de que un estudiante cambie de docente en cada aplicación pueden darse dos posibilidades: que los efectos de los docentes desaparezcan de un año a otro o que se acumulen. La puntuación predicha también variará en función de esta condición.

En el caso que los efectos desaparezcan:

$$\begin{aligned}
\hat{Y}_{1i1} &= \beta_0 + r_{001} + u_{001} \\
\hat{Y}_{2i2} &= \beta_0 + r_{00i} + u_{002} + \beta_1 + r_{10i} \\
\hat{Y}_{3i3} &= \beta_0 + r_{00i} + u_{003} + 2(\beta_1 + r_{10i}) \\
\hat{Y}_{4i4} &= \beta_0 + r_{00i} + u_{004} + 3(\beta_1 + r_{10i})
\end{aligned} \qquad \textbf{Ec. V.57}$$

La ganancia del año 1 al siguiente sería:

$$Y_{2i2} - Y_{1i1} = \beta_1 + r_{10i} + u_{002} - u_{001} \qquad \textbf{Ec. V.58}$$

En cambio, si los efectos de la escuela se acumulan los valores predichos serían los siguientes:

$$\widehat{Y}_{1i1} = \beta_0 + r_{00i} + u_{001}$$

$$\widehat{Y}_{2i2} = \beta_0 + r_{00i} + u_{001} + u_{002} + \beta_1 + r_{10i}$$

$$\widehat{Y}_{3i3} = \beta_0 + r_{00i} + u_{001} + u_{002} + u_{003} + 2(\beta_1 + r_{10i}) \qquad \textbf{Ec. V.59}$$

$$\widehat{Y}_{4i4} = \beta_0 + r_{00i} + u_{001} + u_{002} + u_{003} + u_{004} + 3(\beta_1 + r_{10i})$$

Y la ganancia del año 1 al 2 es:

$$Y_{2i2} - Y_{1i1} = \beta_1 + r_{10i} + u_{002} \qquad \textbf{Ec. V.60}$$

V.2.3.2.2 Modelo de efectos no permanentes

La principal diferencia con los modelos anteriores es que estos permiten una disminución de los efectos previos de los docentes en las estimaciones del año estudiado. La metodología de análisis es similar a la utilizada en el modelo EVAAS pero incluye un parámetro, que se estima cada año, y permite la variación de los efectos aleatorios previos de los docentes o las escuelas. Este modelo, denominado modelo de persistencia, fue desarrollado por McCaffrey y sus colaboradores (2003).

Modelo de persistencia

El modelo de persistencia es una aproximación al análisis del VA desde los modelos lineales mixtos, igual que el modelo propuesto por Sanders y Horn en Tennessee (EVAAS). La única diferencia entre abmos es la inclusión del parámetro que identifica la persistencia de los efectos de los docentes de años previos, por ese motivo es conocido como modelo de persistencia. En este caso, no trata los efectos de los docentes como constantes en el tiempo, sino que se vuelven a estimar en los grados sucesivos.

Un modelo con tres ocasiones de medida desde esta perspectiva se formula en la ecuación Ec. V.61:

$$Y_1 = \beta_1 + Pu_1 + r_1 \qquad \textbf{Ec. V.61}$$

$$Y_2 = \beta_2 + a_{21}Pu_1 + Pu_2 + r_2$$

$$Y_3 = \beta_3 + a_{31}Pu_1 + a_{32}Pu_2 + Pu_3 + r_3$$

Los parámetros a_{21}, a_{31} y a_{32} determinan la persistencia de los efectos previos de los profesores en las puntuaciones de rendimiento del año actual. La ganancia, en este caso, si depende de los efectos en los años previos:

$$Y_3 - Y_2 = (\beta_3 - \beta_2) + (a_{31} - a_{21})r_1 + (a_{32} - 1)u_2 + u_3 + r_3 - r_2 \qquad \textbf{Ec. V.62}$$

Por tanto, la ganancia entre la segunda y tercera aplicación depende del efecto del docente en el curso a través del término $(a_{32} - 1)u_2$ (resta el parámetro de persistencia en el curso tres menos el del curso dos, como el parámetro u_2 es el efecto del docente que ha enseñado al estudiante en el segundo curso y no tiene ese parámetro, por eso es igual a 1) y del profesor del grado 1 incluyendo $(a_{31} - a_{21})u_1$.

Modelo general

Los mismos autores del modelo de persistencia formulan un modelo general para resultados longitudinales del estudiante (McCaffrey, Koretz, Louis & Hamilton, 2004) que incluye los efectos del docente, de la escuela y posibles covariables. Este tipo de análisis engloba tanto al modelo de persistencia como al EVAAS y su formulación para la primera toma de datos se incluye en la ecuación Ec. V.63:

$$Y_{i0} = \mu_0 + \beta_0 x_i + \gamma_{00} z_{i0} + \lambda_{i0k}\eta_{0k} + \phi_{i0j}\theta_{0j} + \epsilon_{i0} \qquad \textbf{Ec. V.63}$$

La puntuación Y_{ig} es para cada estudiante i en el grado g, por tanto, g=0 es el primer punto de recogida de datos.

μ_0 es la media de todos los estudiantes en ese grado.

x_i y z_{i0} son covariables fijas y covariables que pueden variar en el tiempo, respectivamente. Por ejemplo, género, raza (fijas) o circunstancias especiales durante la prueba (variantes). También pueden incluir variables de las escuelas como porcentaje de estudiantes inmigrantes, titularidad, etc. Por tanto, β y γ son los coeficientes de regresión asociados a los predictores.

η_{0k} es el efecto de la escuela k en el grado 0. Por ejemplo, las desviaciones de la media de la escuela respecto a la media general, aunque también puede ser con respecto a algún estándar establecido previamente.

λ_{i0k} es una medida de la cantidad de escolarización de un alumno en la escuela k durante ese grado evaluado. Si el estudiante no ha sido atendido por la escuela el valor será cero.

De la misma forma θ_{0j} son los efectos del profesor y ϕ_{i0j} es la medida de ponderación que refleja el tiempo que un estudiante *i* recibe la enseñanza del profesor *j*. ϵ_{i0} es el término de error residual y se distribuye de forma normal $N(0, \sigma^2_{\epsilon 0})$

El modelo aumenta su complejidad con la inclusión de varias variables dependientes de rendimiento o factores que afectan a las puntuaciones de logro académico. En la segunda toma de datos sería:

$$Y_{i1} = \mu_1 + \beta_1 x_i + \gamma_{11} z_{i1} + (\omega\lambda_{i0k}\eta_{0k} + \lambda_{i1k}\eta_{1k}) + (\alpha\phi_{i0j}\theta_{0j} + \phi_{i1j}\theta_{1j}) + \epsilon_{i1} \quad \textbf{Ec. V.64}$$

La principal diferencia con los modelos que no permiten la variación de los efectos en el tiempo son los parámetros ω y α, que capturan el efecto de los años previos tanto del profesor como del centro. Si tienen un valor igual a cero no hay contribuciones con respecto al rendimiento actual. Si al contrario tienen un valor igual a uno el efecto permanece de forma perpetua y contribuyen en la misma medida que el año anterior. También permite variaciones intermedias.

La ecuación se modificaría para los años posteriores añadiendo más términos de efectos del profesor y de la escuela, y el VA se estima en términos de ganancia anual de la misma forma que el resto de modelos de efectos cruzados.

V.2.4 Modelo de percentiles de crecimiento

Mención aparte merece el MVA desarrollado por Betebenner (2009). El autor, con la finalidad de eliminar los problemas asociados a las puntuaciones de VA operativizadas como un residuo de regresión y diseñadas para analizar la progresión hacia un estándar determinado, propone otra metodología de estimación para unir crecimiento, estándares y rendición de cuentas. Este modelo,

al contrario de lo que hacen la mayor parte de técnicas de análisis longitudinales, no busca explicar la variabilidad de las puntuaciones de los estudiantes relacionándola con la variabilidad de los centros o profesores sino que trata de describir esta variabilidad.

Mediante los rangos de percentiles de crecimiento de los estudiantes y a través de la regresión cuantílica (Koener y Bassett, 1978) trata de informar sobre cómo es el crecimiento de los estudiantes en diferentes rangos de valores de la variable dependiente. Este tipo de modelos no necesitan de la construcción de escalas verticales para medir el crecimiento en un determinado constructo porque examinan otro tipo de cambio que está relacionado con el avance o retroceso en la posición dentro de una distribución de estudiantes con características similares, es decir, con un rendimiento previo equivalente.

Se analiza, en lugar de las escalas longitudinales de rendimiento, una cuantificación normativa del crecimiento, el percentil de crecimiento. Y se lleva a cabo transformando la distribución condicional a términos de probabilidad:

$$\text{Percentil de crecimiento} \equiv \Pr(\text{Rendimiento Acutal}|\text{Rendimiento Anterior}) \times 100 \qquad \textbf{Ec. V.65}$$

El cálculo de los percentiles de crecimiento de los estudiantes se basa en la estimación de la densidad condicional asociada con la puntación de los alumnos en el tiempo *t* utilizando el rendimiento previo en las ocasiones de medida como variables condicionantes. Dada esta densidad condicional, el percentil de crecimiento de los estudiantes se define como el percentil de la puntuación en la densidad condicional en el tiempo *t*. El percentil refleja la verosimilitud de ese resultado considerando el rendimiento previo y, por tanto, convirtiendo los resultados a términos de probabilidad. La estimación de la densidad condicional se realiza usando regresión cuantílica.

Este modelo es opuesto a aquellos que buscan una causa de los sucesos, lo que intenta es proporcionar un marco normativo para el progreso de los estudiantes y, de esta forma, promover el cambio hacia aproximaciones más descriptivas y no de relaciones causa-efecto (Linn , 2008).

Con esta descripción de las opciones más relevantes para el análisis del VA en educación finaliza la aproximación teórica de este trabajo. La parte empírica de esta tesis engloba los próximos tres capítulos. El siguiente capítulo se encarga de describir los datos de la evaluación longitudinal que se utilizan para llevar a cabo las diferentes pruebas empíricas. Y los dos restantes describen los estudios empíricos que se han desarrollado en esta tesis. El primero de ellos está relacionado con la elaboración de la escala vertical de los resultados y el segundo con la elaboración del MVA adecuado a los datos de la evaluación.

Parte empírica:

Diseño, muestra, equiparación vertical y modelos de valor añadido

Capítulo VI: Instrumentos de medida, muestra y características de los datos

Los datos utilizados en este trabajo no son simulados. Han sido recogidas en la investigación empírica realizada bajo el amparo del proyecto I+D "El valor añadido en educación y la función de producción educativa: un estudio longitudinal", cuyo investigador principal es el profesor José Luís Gaviria Soto. Este proyecto fue financiado por el Ministerio de Ciencia y Tecnología (Ref. SEC20,003-09742).

Este estudio piloto de evaluación se incluyó en los planes generales de actuación de la inspección educativa en los cursos 2005-2006[66], 2006-20,007[67] y 2007-2008[68]. Y su principal característica es que recoge información de los mismos estudiantes al inicio y final de dos cursos académicos. Para ello se utilizaron test de rendimiento, elaboradas adhoc, de matemáticas y compresión lectora y se estudiaron tres cohortes diferentes de estudiantes: 5º y 6º de Educación Primaria (EP), 1º y 2º de ESO y 3º y 4º también de ESO. No obstante, para la realización de esta tesis se ha seleccionado únicamente la información de matemáticas de la cohorte de primer ciclo de educación secundaria obligatoria

[66]RESOLUCIÓN de 7 de septiembre de 2005, de la Viceconsejera de Educación, por la que se aprueba el Plan General de Actuación de la Inspección Educativa para el curso 2005-2006.

[67]RESOLUCIÓN de 2 de octubre de 2006, de la Viceconsejera de Educación, por la que se aprueba el Plan General de Actuación de la Inspección Educativa para el curso 2006-2007.

[68]RESOLUCIÓN de 21 de septiembre de 2007, del Viceconsejero de Organización Educativa, por la que se aprueba el Plan General de Actuación de la Inspección Educativa para el curso 2007-2008.

debido a la falta de crecimiento encontrada entre las dos últimas aplicaciones en análisis previos (Castro, Ruíz & López, 2009).

Esta evaluación tiene por objetivo el estudio del Valor Añadido (VA en adelante) de las escuelas que formaron parte de la muestra. Con estos datos y previamente al desarrollo de esta tesis, se han llevado a cabo estudios de varios aspectos: el estudio la dimensionalidad de las puntuaciones utilizadas como medidas de resultados y estimadas con Teoría Respuesta al Ítem (TRI) (Lizasoain & Joaristi, 2009), los patrones de relación entre esas puntuaciones y las diferencias que se producen entre cohortes (Gaviria, Biencinto & Navarro, 2009) y el estudio de la forma del crecimiento en los modelos multinivel longitudinales (Castro, Ruíz & López, 2009).

Además de los estudios realizados, es necesario realizar comprobaciones empíricas de otros aspectos metodológicos que resultan imprescindibles si se pretenden obtener estimaciones del VA de las escuelas con los resultados de la evaluación. Probar diferentes metodologías para lograr la comparación de las diferentes mediciones de resultados realizadas sobre el mismo estudiantes o una comparación de las formas de estimar el VA se hacen indispensables. Este trabajo es, por tanto, un análisis secundario de los datos recogidos mediante la mencionada evaluación piloto.

El diseño de la tesis es no experimental (expostfacto), se analiza la información una vez que los hechos ya se han producido. No se aplica ningún programa experimental que busque el cambio de los niveles de rendimiento, sino que su finalidad es fundamentalmente investigadora e informativa. No obstante, los análisis del VA tratan de captar esa relación causa-efecto[69] que se establece en los estudios experimentales al vincular los resultados de los estudiantes a los efectos que producen escuelas o docentes. En este caso, esas escuelas o docentes juegan el papel de tratamientos diferentes que reciben los estudiantes y las condiciones experimentales se buscan mediante el análisis estadístico de los resultados académicos.

A continuación se detallan las características de los instrumentos utilizados para la recogida de información y de la muestra que participó en el estudio.

[69]Más información sobre la relación causa-efecto en los análisis del VA en el apartado IV.4

VI.1 Diseño de recogida de información

El diseño de los instrumentos de medida para la recogida de información de rendimiento en matemáticas que se empleó en la investigación puede considerarse un diseño mixto. Por un lado, se elaboraron dos formas para ser aplicadas en cada una de las cuatro mediciones, como muestra la Figura VI.1. En cada aplicación las formas (A y B) comparten una proporción de ítem comunes que sirven para llevar a cabo la equiparación horizontal, también incorporan otro grupo de ítems con las formas de la siguiente aplicación utilizando un diseño cruzado, es decir, los ítems que en una aplicación se sitúan en la forma A en la aplicación siguiente pasan a la forma B. Este último grupo de ítems se utiliza para llevar a cabo el anclaje vertical.

La finalidad de este diseño es logar dos puntuaciones del rendimiento en cada uno de los grados evaluados, una al principio y la otra al final de cada uno de ellos. Otra de las características distintivas de este diseño es la utilización de dos formas paralelas de cada instrumento de medida en cada aplicación.

Las formas A y B, con el objetivo de conseguir grupos equivalentes, se repartieron de forma aleatoria entre los estudiantes que participaron en el estudio utilizando para ella la técnica en espiral. Por tanto, en teoría, se cuenta con grupos equivalentes. Por tanto, este diseño mixto permite emplear distintas metodologías de calibración para la equiparación horizontal al emplear ítems de anclaje y grupos equivalentes.

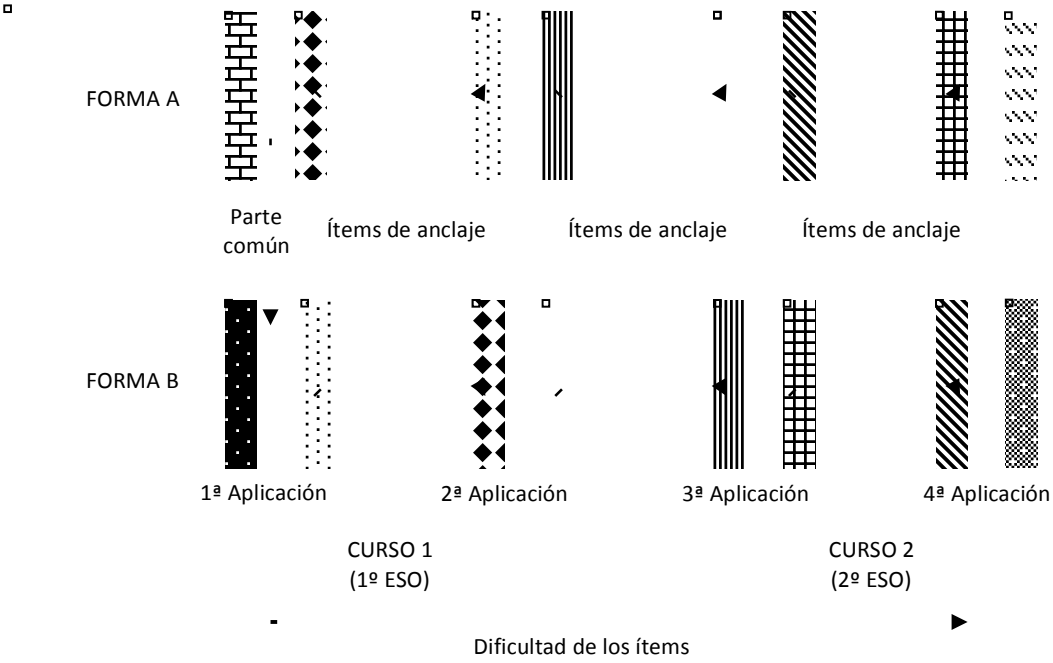


Figura VI.1. Diseño de test con ítems de anclaje en aplicaciones longitudinales.
Fuente: Elaboración propia

Los instrumentos de evaluación incluyen un total de 40 ítems en cada una de las formas en las cuatro aplicaciones. No obstante, se decidió eliminar ciertos ítems del análisis debido a erratas como incluir dos respuestas correctas o no tener respuesta correcta. Además, durante los procesos de análisis y estimación TRI la correlación biserial puntual negativa de algunos ítems impedía la convergencia del modelo y se eliminaron del análisis para lograr el ajuste

La Tabla VI.5 muestra el número de ítems que incorporan cada una de las formas de cada aplicación una vez depuradas. Para facilitar la interpretación se denomina ítems comunes a aquellos que son iguales entre las formas de una misma aplicación e ítems de anclaje a los que comparten los dos instrumentos de aplicaciones consecutivas, es decir, los reactivos necesarios para llevar a cabo el anclaje vertical.

		Aplicación			
		1	**2**	**3**	**4**
Forma A	Comunes	19	18	19	30**
	Anclaje	8 (aplicación 2)	10 (aplicación 1) + 10 (aplicación 3)	10 (aplicación 2) + 7* (aplicación 4)	10 (aplicación 3)
	Específicos	10		3	
	TOTAL	37	38	39	40
Forma B	Comunes	19	18	19	30**
	Anclaje	10 (aplicación 2)	8 (aplicación 1) + 10 (aplicación 3)	10 (aplicación 2) + 10 aplicación 4)	8 (aplicación 3)***
	Específicos	9			2
	TOTAL	38	36	39	40

*Uno de los ítems comunes se utiliza de anclaje con la aplicación 4 Forma B, por este motivo los 7 ítems de anclaje de la forma A de la 3º aplicación pasan a ser ocho en la forma B de la 4º. Además se eliminó un ítem común en el proceso de calibración con TRI debido a una correlación biserial puntual negativa. Por tanto, el total es de 18 en ambas formas.

**En la cuarta aplicación se eliminó un ítem común debido a una correlación biserial puntual negativa, por tanto el total es de 29.

***Se eliminó un ítem de anclaje con la aplicación 3 debido a su correlación biserial puntual negativa por lo que únicamente habrá 7 ítems de anclaje en la forma B.

Tabla VI.5. Distribución de los ítems entre los distintos instrumentos de medida elaborados

Teniendo en cuenta ítems específicos, comunes y de anclaje, el total en las 8 formas diseñadas para la medición longitudinal del rendimiento se cuenta con un total de 170 ítems. Sin embargo, los procesos de depuración dejan un total de 162 ítems para analizar, se ha eliminado por tanto un 5% de los reactivos.

Los ítems comunes entre formas son los utilizados para llevar a cabo la equiparación horizontal. Es la cuarta aplicación la que cuenta con mayor número de este tipo de ítems, un total de 29. El número se reduce en las tres primeras aplicaciones a 18 o 19.

Los ítems de anclaje se emplean en el anclaje vertical y se cuenta con aproximadamente 19 en cada aplicación, repartidos entre las dos formas. Los reactivos de anclaje cambian de una forma a otra cuando se avanza en las aplicaciones, es decir, los 10 ítems de anclaje de la forma B en la primera aplicación pasan a la forma A en la segunda aplicación. El anclaje de la segunda con la tercera aplicación es el que cuenta con una mayor número de ítems comunes, los diez en cada forma. El número desciende a 17 para anclar la tercera y cuarta aplicación.

En la evaluación de matemáticas desarrollada por el mencionado I+D se optó por un diseño longitudinal de medida del rendimiento con cuatro tomas de

datos a lo largo de dos cursos académicos completos. Ya a se ha detallado en el capítulo anterior[70] alguno de los modelos para el análisis del VA más importantes, además existen varios trabajos que los analizan y clasifican (Thum, 2002; Tekwe et al., 2004; Ballou, Sanders & Wright, 2004; Darmawan & Keeves, 2006; Lockwood et al., 2007). No todos ellos utilizan un diseño longitudinal del rendimiento. Existen análisis que emplean únicamente dos puntuaciones del rendimiento de los estudiantes. Por tanto, no es necesario utilizar más de dos mediciones del logro para conseguir medir el VA de la escuela, no obstante con esas dos únicas tomas de datos es muy difícil medir más allá de un cambio, es decir, no se puede observar el crecimiento. Además, cuando se incluyen predictores de ese cambio este tipo de modelos no son tan potentes, en términos estadísticos (Willett, 1994; Choi, Goldschmidt & Yamashiro, 2006; Martínez-Arias, Gaviria & Castro, 2009)

VI.2. Población y muestra

La evaluación realizada en el proyecto I+D recogió información de matemáticas y comprensión lectora de tres cohortes distintas de estudiantes de la comunidad de Madrid. Para este trabajo, como se ha mencionado más arriba, se ha seleccionado una de ellas. La población de referencia es, por tanto, el conjunto de centros educativos (tanto públicos, como privados y privados concertados) de la Comunidad de Madrid que imparten el primer ciclo de Educación Secundaria Obligatoria (ESO) en el año académico 2005-2006. La población está constituida por un total de 64.137 alumnos agrupados en 749 centros de enseñanza que imparten esta etapa educativa.

Los centros pueden ser de dos tipos: Por un lado aquellos que ofertan, junto la etapa de Educación Primaria (EP), también el primer ciclo o la etapa completa de secundaria. La mayoría son centros concertados o privados. Por otro lado, los centros que no ofertan la etapa de EP junto con la secundaria. La mayor parte son centros públicos y también incorporan la secundaria no obligatoria, es decir, los cursos de bachillerato.

[70]Más información sobre los modelos de valor añadido en el Apartado V.2.

	Centros que imparten ESO	Estudiantes matriculados en 1ºESO
Públicos	304	35633
Concertados	324	28504
Privados	121	
Total	749	64137

Tabla VI.6 Población de estudiantes y escuelas por titularidad

VI.2.1 Procedimiento de muestreo

Se utilizó un muestreo polietápico en dos fases. En la primera fase se calculó tamaño inicial necesario en cada uno de los cursos, para ello se utilizó un muestreo aleatorio simple utilizando el tamaño de la población de referencia, un nivel de confianza del 95% y un error muestral máximo de 0,1σ. No obstante, dado que las unidades primarias de muestreo son los centros, es necesario llevar a cabo una segunda fase donde el tamaño se corrige teniendo en cuenta el efecto de este tipo de diseños. Los factores que se utilizan para realizar la corrección son el tamaño de las escuelas y la correlación intraclase[71]. El muestreo por conglomerados tiene un efecto negativo, es decir, el tamaño de la muestra aumentará con respecto al muestreo aleatorio simple en función de esos dos factores mencionados.

El número medio de alumnos matriculados en 1º de la ESO en los diferentes centros educativos de la Comunidad de Madrid es de 86 aproximadamente, según los datos y cifras del curso 2005-2006[72]. Este número de estudiantes será el tamaño del conglomerado.

Para el cálculo del tamaño muestral aplicaremos el muestreo aleatorio simple y, seguidamente, se introducirá la modificación necesaria por el efecto del diseño.

$$M = \frac{N\frac{k^2}{e^2}}{N + \frac{k^2}{e^2}} \qquad \textbf{Ec. VI.1}$$

[71]La autocorrelación (ρ) representa una medida de homogeneidad interna de los grupos en el análisis de regresión multinivel, estableciendo la similitud entre las unidades de nivel individual (Gaviria & Castro, Modelos jerárquicos lineales, 2005), es decir, el grado de homogeneidad de los estudiantes dentro de las escuelas. Este parámetro también es un indicador de la proporción de varianza de los estudiantes se debe a las escuelas

[72]Enlace:
http://www.educacion.gob.es/dctm/ministerio/horizontales/estadisticas/indicadores-publicaciones/cifras/2008/cifrasd3-08.xls?documentId=0901e72b80858a58

Donde N es la población de referencia, en este caso 64.137; K es la puntuación típica correspondiente al nivel de confianza seleccionado, es decir, con un nivel de confianza del 95%, k sería igual a 1,96; y e es el error muestral máximo que será de 0,1 desviaciones típicas.

Con estos datos la muestra sería inicialmente 382 estudiantes seleccionados de forma aleatoria entre toda la población. Sin embargo, es necesario aplicar el mencionado efecto de diseño (F) que considere las unidades a seleccionar (escuelas) y la homogeneidad de los estudiantes dentro de esas unidades (correlación intraclase):

$$F = 1 + (B - 1)\,\rho \qquad \textbf{Ec. VI.2}$$

Donde B es el tamaño del conglomerado, es decir, el tamaño medio de los grupos[73] (86 estudiantes), es decir, el número medio de estudiantes matriculados en 1º de ESO en las distintas escuelas; y ρ es la autocorrelación. En estudios con datos de edades similares en España (Ruiz & Castro, 2006; Navarro & Redondo, 2006; Ruíz, 2009; López, Navarro, Ordoñez & Romero, 2009) su valor es aproximadamente 0,2 por lo que utilizaremos es número para el cálculo del efecto producido por el diseño. Con estos datos F es igual a 18.

El tamaño muestral definitivo será:

$$M' = F * M \qquad \textbf{Ec. VI.3}$$

Por lo tanto, el total de alumnos necesarios asciende a 2.292. Es necesario recordar que los estudiantes se encuentran agrupados en conglomerados de tamaño 86, por lo que debemos realizar el muestreo sobre los grupos (M'/B). Finalmente, el total de grupos de nuestra muestra es de 80 aproximadamente.

VI.3 Características de los datos

A continuación se lleva a cabo un estudio de los requisitos previos necesarios para la elaboración de una escala de rendimiento en matemáticas. Para

[73]En la muestra inicial se seleccionaron todos los grupos de estudiantes de 1º de ESO de cada centro educativo pero una reducción muestral por motivos de financiación en la tercera aplicación provocó la eliminación de algunos grupos completos.

evitar repeticiones en la numeración de las distintas aplicaciones va a utilizarse A1 para hacer referencia a la primera toma de datos, A2 para la segunda y A3 y A4 para las dos últimas.

En primer lugar, y antes de utilizar modelos TRI para la equiparación y estimación de las puntuaciones de rendimiento en matemáticas, se realiza una revisión de la dimensionalidad de las distintos test. Este supuesto ya ha sido analizado con los datos que se utilizan en esta tesis en un trabajo de los profesores Lizasoain y Joaristi (2009).

En segundo lugar se lleva a cabo un análisis descriptivo básico poniendo énfasis en el estudio de los casos perdidos. Este aspecto es importante porque entre la segunda y tercera aplicación (A2 y A3) la muestra tuvo que reducirse por motivos de presupuesto. No se eliminaron escuelas, sino que fueron grupos de estudiantes, es decir, aulas las que sufrieron ese proceso. De esta forma, no se pierden unidades de referencia para la estimación del VA pero sí tamaño muestral de los conglomerados y esto puede aumentar el sesgo de las estimaciones. Por tanto, conviene llevar a cabo un análisis de este grupo de sujetos eliminado y decidir si deben eliminarse también en las primeras aplicaciones (A1 y A2) cuando se desarrollen los distintos Modelos de Valor Añadido.

VI.3.1 Dimensionalidad e independencia de campo

Para la construcción de una escala común entre los distintos instrumentos de medida se han desarrollado modelos bajo los supuestos de la Teoría Respuesta al Ítem. Con la finalidad de conseguir una escala con garantía los datos analizados deben cumplir los supuestos de unidimensionalidad del constructo evaluado (rendimiento en matemáticas) y de independencia local, es decir, que la probabilidad de responder correctamente a un ítem no depende de la respuesta a otros ítems.

El análisis en profundidad de la dimensionalidad de las pruebas utilizadas en este trabajo se detalla en el trabajo de Lizasoain y Joaristi (2009). Los autores concluyen que para la cohorte de estudiantes de 1º y 2º de Educación Secundaria Obligatoria las pruebas mantienen la unidimensionalidad esencial.

En la práctica solo se pone a prueba la dimensionalidad porque implica la independencia local: "Si el supuesto de unidimensionalidad exige que la respuesta del examinado al ítem esté determinada solo por su nivel de rasgo latente, es evidente que dicha respuesta no podrá estar influenciada por cómo haya contestado los anteriores ítems" (Muñiz & Fidalgo, 2005, pág. 82). Por lo tanto, se pueden asumir los supuestos necesarios para llevar a cabo el ajuste de modelos TRI.

VI.3.2 Análisis de valores perdidos

Los datos de la Tabla VI.7 se han calculado teniendo en cuenta a los sujetos con puntuación de rendimiento en la ocasión de medida analizada, es decir, descartando los valores perdidos. La media inicial de rendimiento en matemáticas es aproximadamente cero, con una desviación típica cercana a uno, valores propios del carácter normal de la distribución.

La mayor pérdida de estudiantes se produce entre A2 y A3 debido, en gran parte, a esa reducción muestral intencionada como muestra la Tabla VI.7.

Descriptivos		A1	A2	A3	A4
N		5106	4882	2870	2939
Media		0,051	0,482	1,299	1,659
DT		0,871	0,816	0,679	0,726
Varianza		0,758	0,666	0,461	0,528
Percentiles	5	-1,380	-0,863	0,113	0,280
	10	-1,065	-0,608	0,407	0,710
	25	-0,550	-0,105	0,853	1,213
	50	0,037	0,483	1,318	1,715
	75	0,654	1,050	1,783	2,150
	90	1,191	1,546	2,155	2,552
	95	1,502	1,848	2,380	2,759

Tabla VI.7 Estadísticos descriptivos de las puntuaciones de rendimiento en las 4 aplicaciones.

Las medias de rendimiento en las cuatro aplicaciones reflejan la trayectoria creciente de los estudiantes. Se observa un gran cambio entre A2 y A3, aproximadamente 0,8 puntos en la escala del rasgo, el doble que entre el resto de mediciones. Estos resultados ponen de manifiesto la necesidad de analizar los

resultados de esos estudiantes eliminados en la tercera aplicación y conocer qué nivel de rendimiento alcanzan en la primera toma de datos.

En el anexo I[74] puede consultarse la perdida muestral de cada uno de los centros educativos que forman parte de la muestra, además de la media y desviación típica obtenida en cada una de las aplicaciones. La reducción de la muestra que se produce por esa eliminación es del 36%. Pero si se analizan la pérdida en cada escuela, en algunos casos como en los centros educativos 21 y 31, que pierden casi el 70%. A pesar de la reducción, la mayor parte de ellas (el 80%) conserva el tamaño mínimo recomendado para realizar el análisis multinivel en condiciones óptimas, al menos 30 casos como recomiendan Kreft y De Leeuw (1998). Por su parte, Lockwood, Louis y McCaffrey (2003) mencionan que los centros educativos con tamaños inferiores a 20 sujetos pueden ser un problema, sobre todo si se utilizan estimadores bayesianos BLUP[75] porque no tenderán a diferenciarse de la media global.

El total de centros educativos evaluados es de 65 pero no de todos ellos se posee información sobre los resultados en las cuatro aplicaciones. Las escuelas con información completa es del 94% aproximadamente. No se dispone información, en una ocasión de medida, de los centros 25, 62 y 64. Concretamente del centro 25 en A2, del 64 en A3 y del 62 en A4. Otro dato destacable es que solo se posee información de un estudiante en la A3 de la escuela 34.

El tamaño de estos centros en A1 es variado, oscila desde más de 200 sujetos hasta un mínimo de 20 aproximadamente. En A3 y A4 lógicamente esta dispersión de los tamaños se reduce. Los centros con más número de estudiantes han sido los encargados de aportar mayor proporción de estudiantes a esa reducción muestral. El número máximo de estudiantes en A4 no llega 90, sin embargo, esa disminución no es tan notable en los centros con baja muestra. Esto es debido a la eliminación de grupos completos de estudiantes en las escuelas que tenían un mayor número de ellos participando en esta evaluación.

Los casos perdidos entre aplicaciones dependen entonces de dos tipos de situaciones: En primer lugar, la mortalidad experimental, típica en estudios

[74]Información detallada en la Tabla AI.1
[75]Más información en el aparado V.1.2.1

longitudinales, que incluye a los sujetos que no realizaron alguna de las aplicaciones por ausencia o porque dejó toda la prueba en blanco. En este caso, los estudiantes pueden volver a ser medidos en otra aplicación o, en cambio, perderlos definitivamente. Y, en segundo lugar, la mencionada reducción muestral llevada a cabo entre A2 y A3 por motivos de presupuesto, es decir, una disminución intencionada. La Tabla VI.8 describe las distintas situaciones:

Rendimiento A1		P_A1-A2	E_A2-A3	P_A2_A3	P_A3-A4
N	**Válidos**	658	1802	269	219
	Perdidos	0	238	17	0
Media		-0,355	-0,105	0,230	0,115
DT		0,883	0,855	0,843	0,816
Varianza		0,779	0,730	0,711	0,666
Percentiles	**5**	-1,802	-1,495	-1,155	-1,164
	10	-1,511	-1,210	-0,920	-0,956
	25	-0,954	-0,696	-0,306	-0,450
	50	-0,396	-0,109	0,236	0,087
	75	0,254	0,447	0,812	0,682
	90	0,888	1,023	1,179	1,202
	95	1,127	1,351	1,699	1,519

Tabla VI.8 Descriptivos en la A1 de los casos perdidos y eliminados en el resto de aplicaciones

Los casos perdidos, debido a la mortalidad experimental, entre A1 y A2 son 658 y tienen una media inferior a la media global, un valor de -0,355. Entre A2 y A3 se producen las dos situaciones mencionadas. Por un lado, se eliminan 2.040 sujetos de la muestra con la reducción intencional, 238 de ellos eran estudiantes que no fueron medidos en A1, es decir, la primera prueba que realizaron fue la de A2. Estos estudiantes tienen una puntuación en A1 de 0,1 puntos por debajo de la media global. Por otro lado , debido a la mortalidad experimental, se pierden 269 sujetos entre estas dos aplicaciones, 17 de ellos tampoco comenzaron en A1, y su media es de 0,23. Finalmente, entre A3 y A4, se pierden otros 219 por la mencionada mortalidad experimental y su media en A1 fue de 0,115.

Si se comparan las medias en rendimiento de A1 de los sujetos perdidos y eliminados con la de los estudiantes que permanecen en la aplicación, como muestra la Tabla VI.9, se observan diferencias significativas en algunos casos.

Perdidos A1-A2		N	Media	DT	ET	Dif. M	ET Dif	T	Sig
A1	**No**	4448	0,112	0,853	0,013	0,466	0,037	12,701	0,00
	Si	658	-0,355	0,883	0,034				
Eliminados A2-A3									
A1	**No**	2646	0,259	0,819	0,016	0,365	0,026	14,213	0,00
	Si	1802	-0,105	0,855	0,020				
Perdidos A2-A3									
A1	**No**	4837	0,042	0,871	0,013	-0,189	0,054	-3,462	01
	Si	269	0,230	0,843	0,051				
Perdidos A3-A4									
A1	**No**	4887	0,049	0,873	0,012	-0,067	0,057	-1,177	0,240
	Si	219	0,115	0,816	0,055				

Tabla VI.9. Diferencia de medias en las puntuaciones de la 1ª Aplicación entre los casos eliminados o perdidos y la muestra inicial.

La prueba de Levene realizada previamente nos permite asumir la igualdad de varianzas entre los grupos en todos los casos. Si se pone atención en las medias de rendimiento de ambos grupos, los casos perdidos entre A1 y A2 obtienen puntuaciones que difieren significativamente de aquellos que permanecen en la muestra. La puntuación promedio en A1 aumenta más de 0,1 puntos (la media global es de 0,05 aproximadamente) respecto al promedio estimado con todos los sujetos. Esto se debe a que los casos perdidos entre las dos primeras aplicaciones tienen un rendimiento medio de 0,3 puntos por debajo de esa media global y al eliminarlos, el promedio aumenta.

Un fenómeno similar ocurre al analizar los resultados considerando a los sujetos eliminados entre A2 y A3 debido a la reducción muestral. Con estos sujetos eliminados de la muestra, la media del grupo asciende a 0,25. Este aumento está motivado por la puntuación más baja que obtiene este grupo de estudiantes en A1 (-0,105). En cambio, los casos perdidos en A2 tienen un rendimiento superior a la media pero, al extraerlos, aunque la diferencia entre las medias de ambos grupos es significativa, no produce apenas cambios en el rendimiento de los sujetos que permanecen. Una situación similar ocurre con los casos perdidos en A4 pero sin encontrar significatividad de la diferencia de medias

En la evaluación longitudinal no solo se pierden casos, se ha comprobado que en cada aplicación pueden aparecer nuevos sujetos que responden a las pruebas de rendimiento. También se ha llevado a cabo un análisis de los resultados que han obtenido en las aplicaciones en las que se incorporaron por primera vez.

Casos Nuevos		A2	A3	A4
N		434	91	56
Media		0,182	0,755	0,497
DT		0,836	0,678	0,660
Varianza		0,699	0,460	0,435
Percentiles	**5**	-1,080	-0,319	-0,584
	10	-0,890	-0,187	-0,411
	25	-0,497	0,227	-0,043
	50	0,156	0,719	0,543
	75	0,792	1,305	1,079
	90	1,350	1,638	1,349
	95	1,611	1,761	1,467

Tabla VI.10 Resultados de los casos nuevos de cada aplicación.

En A2 y A3 los nuevos casos que se incorporan tienen medias inferiores a la media global, con una diferencia más pronunciada en A3. Mientras que las medias de esos nuevos sujetos es de 0,182 y 0,755 respectivamente, las medias generales son de 0,482 y 1,299. En A4 estos sujetos obtienen casi 1,2 puntos menos de media. Los 56 nuevos casos que se incorporan en la última aplicación es de 0,497, el promedio general es de 1,659.

Otra información reseñable de estos nuevos casos es que de los 434 casos nuevos en A2, 255 se pierden en A3 y solo 17 vuelven a estar en A4, por tanto, 238 sujetos nuevos estuvieron dentro del grupo de eliminados con la reducción muestral, como ya se mencionó en el análisis de valores perdidos (Tabla VI.8).

Es una condición necesaria en el desarrollo de modelos de crecimiento poseer más de dos puntuaciones del rasgo evaluado[76]. Por este motivo, se incluyeron, en la muestra final, sujetos que al menos tuvieran tres puntuaciones. Esto solo es posible con los siguientes patrones de respuesta:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{bmatrix} \begin{bmatrix} 10111 \\ 11011 \\ 11101 \\ 11110 \end{bmatrix}$$

Como se observa en Tabla VI.11, El total de sujetos con al menos tres mediciones es de 2964. El número de estudiantes con solo dos puntuaciones de rendimiento también es elevado, un 35,5% del total. Los estudiantes con una única medida del logro son el grupo menos representativo en esta muestra.

[76]Más información en el apartado V.1.1

Mediciones	N	%	% acumulado
4	2158	37,95	37,95
3	806	14,17	52,12
2	2024	35,59	87,71
1	699	12,29	10,00
Total	5687	10,00	

Tabla VI.11. Nº de estudiantes en función del nº de mediciones recibidas.

El porcentaje de estudiantes a eliminar sería del 47,88% del tamaño muestral inicial. Si se analizan los resultados de rendimiento una vez llevada a cabo la reducción se obtienen los valores que se incluyen en la Tabla VI.12.

Muestra reducida		A1	A2	A3	A4
N	**Válidos**	2809	2801	2695	2745
	Perdidos	155	163	269	219
Media		0,244	0,650	1,331	1,702
DT		0,822	0,776	0,664	0,697
Varianza		0,675	0,602	0,441	0,486
Percentiles	**5**	-1,077	-0,676	0,211	0,488
	10	-0,822	-0,369	0,463	0,794
	25	-0,340	0,151	0,881	1,269
	50	0,226	0,649	1,347	1,739
	75	0,812	1,181	1,798	2,179
	90	1,360	1,622	2,172	2,565
	95	1,603	1,934	2,387	2,770

Tabla VI.12. Estadísticos descriptivos de la escala inicial con la muestra depurada.

Si comparamos estos resultados con los promedios obtenidos por la muestra completa (ver Tabla VI.7) se observa un aumento de las medias en las cuatro aplicaciones con mayor fuerza en las dos primeras aplicaciones con aumentos de 0,25 puntos aproximadamente. En A3 y A4 el aumento es de solo 0,05 puntos de rendimiento.

Por tanto, considerando estos resultados, para elaborar los distintos modelos de valor añadido lo más conveniente es eliminar a los estudiantes que formaron parte del proceso de reducción muestral en A3 también de las dos primeras aplicaciones para evitar esta interferencia en las medias globales de las aplicaciones.

VI.4 Presentación de los estudios empíricos

El análisis de las cuestiones metodológicas ligadas al desarrollo de los modelos de valor añadido es el origen de los estudios empíricos que se llevan a cabo en este trabajo. En primer lugar, las características de los datos de rendimiento de la evaluación de la Comunidad de Madrid, con un carácter longitudinal, requieren que se lleve a cabo un proceso de equiparación que ponga las cuatro puntuaciones de logro en una escala común para poder ser comparadas. Las evaluaciones que se realizan sobre el mismo constructo o dominio pero que aumentan progresivamente en complejidad y dificultad se adecúan a las propiedades de las escalas verticales. Las especificidad del diseño, con dos mediciones en cada curso académico, permite llevar a cabo el proceso de varias formas. Con la modificación de determinados parámetros durante la equiparación y estimación de las puntuaciones de logro es posible obtener resultados distintos. Una comparación de estos aspectos se lleva a cabo en este primer estudio empírico.

El segundo trabajo empírico está directamente vinculado a la elaboración del modelo de valor añadido. Utilizar un modelo de multinivel de crecimiento para el análisis y contar con cuatro mediciones de rendimiento, son factores que permiten llevar a cabo el proceso desde diferentes perspectivas. Por ejemplo, es posible utilizar únicamente dos mediciones para llevar a cabo un análisis utilizando la primera ocasión de medida como principal covariable y la última como variable dependiente. También es posible construir un modelo de crecimiento con las tres últimas mediciones como criterio y la primera toma de datos como covariable. La trayectoria de crecimiento puede modificarse empleando la distancia real en meses entre aplicaciones o hacer ajustes basados en la realidad empírica observada, como el paso del verano entre aplicaciones o la recogida de datos de la tercera aplicación cuando han trascurrido dos meses desde el inicio del curso. Otro aspecto es la elección del estatus inicial, modificarlo tiene efectos en la relación existen entre punto inicial y crecimiento (Rogosa, 1995). Se llevan a cabo comparaciones de los distintos modelos y su efecto en los residuos asociados a las escuelas, es decir, las estimaciones de VA dentro de esta perspectiva de análisis.

En los anexos se incluyen resultados que complementan los estudios empíricos que se llevan a cabo. El Anexo I presenta los resultados de los análisis de los ítems que formaron parte de las distintas pruebas aplicadas en la evaluación longitudinal, desde la teoría clásica de los test y desde la teoría respuesta al ítem, junto con un estudio de los supuestos de las puntuaciones de rendimientos estimadas (normalidad, homocedasticidad) y también de la escala vertical (propiedad de intervalo). El Anexo II presenta una metodología alternativa para el cálculo de las distancias horizontales empleadas para el análisis de la escala vertical en el estudio empírico que se presenta en el Capítulo VII. El último anexo, el tercero, se dedica a la sintaxis del software utilizado en los análisis de datos.

Capítulo VII: Comparación empírica de metodologías de equiparación para la construcción de una escala vertical de rendimiento en matemáticas.

Las estimaciones de VA pueden ser una buena fuente de información para la evaluación de escuelas. Para su estimación es necesario trabajar con datos del rendimiento individual de los estudiantes como materia prima. Esta información, recopilada mediante test estandarizados que producen puntuaciones medidas a nivel de intervalo, debe modelarse para obtener los resultados finales del VA.

Para poder estimar el VA de una escuela, docente o programa es necesario medir el cambio en el aprendizaje de los estudiantes y esto no sería posible si no se cumplen dos requisitos indispensables (Thum, 2003):

- La medida debe detectar cambio en un determinado grado o cantidad en el constructo medido. La noción de cambio tienen poco sentido si el constructo que se mide es diferente entre ocasiones de medida. Por tanto, es necesario que dicho constructo sea cualitativamente constante para que la escala utiliza identifique esos cambios.
- Las herramientas empleadas para recoger la información de rendimiento deben permitir establecer una escala común entre mediciones. Medir la cantidad de cambio no es posible si el instrumento de medida o la propia escala cambia de forma desconocida.

Por tanto, medir el aprendizaje como un proceso de cambio implica tomar diferentes mediciones del mismo constructo que va a ser evaluado en diferentes momentos temporales, en este caso matemáticas, y con distinto nivel de dificultad o complejidad entre aplicaciones. Pero no basta solo con eso, es necesario que los instrumentos de medida empleados sean capaces de medir la evolución de ese constructo a lo largo del tiempo e identificar qué cantidad de cambio se produce entre las distintas ocasiones de medida. Existen diferentes aproximaciones para medir el cambio en el aprendizaje a través de tests estandarizados y el VA es una de las metodologías de que utiliza medidas de cambio en el logro de los estudiantes como elemento de información.

Si los MVA utilizan medidas de ganancia, es necesario que cuenten con este tipo de escala, ya que al calcular esa diferencia se necesita, además de medir el mismo constructo, tener las mismas unidades de medida (Reckase, 2008). También es un requisito cuando se emplea un modelo longitudinal para estimar una pendiente de crecimiento vinculada al tiempo y el residuo asociado a cada escuela a través de un análisis de regresión multinivel (Bryk, Thum, Easton & Luppescu, 1998; Ponisciak & Bryk, 2005; Zvoch & Stevens, 2006; Briggs, Weeks & Wiley, 2008; Castro, Ruíz & López, 2009). Este último modelo es uno de los que se desarrollan en el segundo estudio empírico.

Aunque no todos los modelos de VA requieren de una escala vertical debido a sus características específicas, como los que únicamente utilizan dos puntuaciones del logro académico para desarrollar un modelo de regresión multinivel entre el pretest y el postest (Demie, 2003; Ray, McCormack & Evans, 2009), tampoco es un requisito en el modelo de Tennessee (TVASS) (Sanders & Horn, 1994; Sanders, Saxton & Horn, 1997) o en el modelo de percentiles de crecimiento de Betebenner (2009).

Este requisito metodológico abre un campo muy amplio de posibilidades relacionadas con la elaboración de la escala y se ha demostrado que el tipo de métrica importa (Yen, 1986; Chin, Kim & Nering, 2006; Jungnam, 2007; Briggs, Weeks & Wiley, 2008; Briggs & Weeks, 2009; Briggs & Betebenner, 2009). Por tanto, es necesario un estudio empírico que compare los posibles efectos que pueden tener sobre la escala, las decisiones metodológicas tomadas en el proceso

de elaboración como, por ejemplo, el diseño de los test, el tipo de equiparación de puntuaciones o la forma de estimar el constructo evaluado.

VII.1 Problema de Investigación

El propósito de este estudio es comprobar la manera en la que puede afectar a la habilidad estimada de un estudiante, las distintas decisiones psicométricas implicadas en el proceso de elaboración de la escala vertical del rendimiento académico. Utilizando la información proporcionada por las respuestas de los alumnos a los distintos test elaborados ad hoc para la evaluación del VA de las escuelas. Se utilizan los datos de la evaluación longitudinal realizada sobre una muestra de centros de la Comunidad de Madrid[77].

El diseño de recogida de información que se empleó en esta evaluación puede considerarse un diseño mixto porque utiliza ítems comunes para llevar a cabo una doble equiparación: horizontal y vertical. Se elaboraron dos formas para ser aplicadas en cada una de las cuatro mediciones. En cada aplicación las formas (A y B) comparten una proporción de ítem comunes que sirven para llevar a cabo la equiparación horizontal, también incorporan otro grupo de ítems comunes con las formas de la siguiente aplicación utilizando un diseño cruzado, es decir, los ítems que en una aplicación se sitúan en la forma A en la aplicación siguiente pasan a la forma B. Este último grupo de ítems se utiliza para llevar a cabo la equiparación horizontal. Además, las formas A y B se re partieron de forma aleatoria entre los estudiantes que participaron en el estudio por lo que, en teoría, se cuenta con grupos equivalentes además de los mencionados ítems de anclaje.

Este sistema de ítems comunes combinado con el diseño de grupos equivalentes permite probar diferentes metodologías de calibración de los parámetros de los ítems en el ámbito de la equiparación horizontal. Los ítems comunes también permiten probar distintas formas de calibración vertical.

Por tanto, considerando esa doble perspectiva horizontal y vertical del proceso, el **primer problema** es**:**

[77]El tipo de diseño utilizado para la construcción de los instrumentos de medida y las características de la muestra se especifican en el Capítulo VI.

¿Cuál es la metodología de equiparación horizontal adecuada a los datos de la evaluación?

Este problema ponen la atención en el proceso de equiparación horizontal. Se comprueba como varían las puntuaciones de rendimiento si se lleva a cabo una calibración por separado sin ningún tipo de transformación, es decir, sin realizar el proceso de escalamiento de las puntuaciones de ambas formas ya que se ha utilizado un diseño de grupos equivalentes para la recogida de información. O si, en cambio, se deben utilizar los procesos de calibración conjunta, fija y por separado. Teóricamente al utilizar un diseño alternado, denominado en espiral, para el reparto de los test, de manera que las formas se alternan en el momento de elaborar los paquetes para las aulas y se reparten alternativamente se contaría con grupos que son aleatoriamente equivalentes y por tanto comparables. Si esto se consigue, las diferencias de rendimiento entre los grupos puede atribuirse a las diferencias en dificultad entre las diferentes instrumentos de evaluación aplicados (Kolen & Brennan, 2004). Se comparan un total de 7 escalas en cada una de las ocasiones de medida:

- Calibración por separado per sin ningún tipo de transformación (CS).
- Calibración por separado utilizando las cuatro formas de estimación de las constantes A y B necesarias para la transformación: los métodos de transformación utilizados son: media/media (CSMM), media/sigma (CSMS), Haebara (CSH) y Stocking y Lord (CSSL).
- Calibración Conjunta (CC)
- Calibración Fija (CF)

El **segundo problema** es:

¿Cuál es la metodología de anclaje vertical adecuada a los datos de la evaluación?

Este problema se centra en el análisis de cómo afectan las decisiones tomadas en el proceso de anclaje vertical a las puntuaciones de la escala vertical elaborada. Se prueban tres maneras de llevar a cabo la calibración de los parámetros de los ítems (conjunta, fija y por separado), además de tres metodologías distintas de estimación de la habilidad del sujeto. El proceso de

calificación puede llevarse a cabo utilizando tres tipos de estimación principalmente: Máxima verosimilitud (ML), Bayes o esperada a posteriori (EAP) y Bayes Modal o máxima a posteriori (MAP) empleando todo el patrón de respuestas de los sujetos para llevar a cabo el proceso de estimación de la habilidad. Se han elaborado un total de 18 escalas verticales combinando estos dos factores. Por un lado seis tipos de calibración y por otro las tres formas mencionadas para estimar la habilidad.

A. Método de calibración vertical:
   - A.1 Calibración Conjunta
   - A.2 Calibración por separado (media/media (CSMM), media/sigma (CSMS), Haebara (CSH) y Stocking y Lord (CSSL))
   - A.3 Calibración fija

B. Método de estimación de la habilidad:
   - B.1 Máxima Verosimilitud (MV)
   - B.2 Empírica a Posteriori (EAP). Utilizando la distribución empírica estimada en la fase de calibración como distribución a priori (con el comando idist=3 en la sección SCORE de BILOGMG).
   - B.3 Máxima a Posteriori (MAP)

VII.2 Metodología

En este apartado se describen los diferentes procesos empleados para la construcción de la escala de rendimiento académico, en concordancia con los dos problemas planteados. Así como, los criterios que se utilizan para juzgar los distintos procedimientos

Conviene recordar que el primer problema se encuentra directamente vinculado con la metodología para llevar a cabo la equiparación horizontal de las dos formas distintas de test diseñadas para cada aplicación. El segundo problema planteado se vincula con las distintas metodologías de anclaje vertical, es necesario comparar los resultados producidos y averiguar si la forma de anclaje afecta a la estimación de las puntuaciones en el rasgo evaluado.

Para resolver ambos problemas se han implementado distintos modelos TRI de tres parámetros. Se ha utilizado este modelo psicométrico porque es uno de los más probados en la elaboración de escalas verticales de rendimiento (Kolen & Brennan, 2004; Chin, Kim & Nering, 2006; Jungnam, 2007; Tong & Kolen, 2007; Kang & Petersen, 2009).

Los distintos modelos TRI implementados y las estimaciones de los parámetros de los ítems y de las puntuaciones de logro se han llevado a cabo a través del programa BILOGMG 3.0 (Muraki, 1994). Conviene detallar que en la calibración por separado el cálculo de las constantes A y B que permiten llevar a cabo la transformación de la escala se ha llevado a cabo con el software ST 2.0 (Hanson y Zeng, 2004). Este programa permite calcular esas constante mediante los métodos media/media (Loyd & Hoover, 1980) y media/sigma (Marco, 1977) y los métodos de curva característica del ítem de Haebara (1980) y Stocking y Lord (1983).

VII.2.1 Problema 1. Comparación de procedimientos para la equiparación horizontal

El primer problema está formulado para averiguar si el diseño de los instrumentos para la recogida de información dentro de una misma aplicación permite contar con grupos equivalentes y si la implementación de algún tipo de equiparación produce variaciones de los resultados. Se han utilizado cuatro tipos de calibración distinta para construir 7 escalas de rendimiento en matemáticas en cada una de las ocasiones de medida:

- Calibración por separado pero sin ningún tipo de transformación de la habilidad entre formas (CS).
- Calibración por separado utilizando las cuatro formas de estimación de las constantes A y B necesarias para la transformación: los métodos de transformación utilizados son: Media/Media (CSMM), Media/sigma (CSMS), Haerbera (CSH) y Stocking y Lord (CSSL).
- Calibración Conjunta (CC)
- Calibración Fija (CF)

Para evitar la indeterminación de la escala que produce este tipo de modelos se asume una distribución normal de la habilidad con media cero y desviación típica igual a uno Además otras características coincidentes entre todos los tipos de equiparación son el número máximo ciclos del algoritmo EM que el programa utiliza, fue fijado a 50 (CYCLES=50) y el número de iteraciones Gauss-Newton que fue fijado a 25 (NEWTON=25). También el tipo de estimación es común, se empleó la que BILOG-MG utiliza por defecto, la estimación bayesiana esperada a posteriori (EAP)

De forma específica, en la calibración por separado (CS) los parámetros de los ítems de las dos formas en cada aplicación se estiman en ejecuciones distintas del software BILOGMG en una primera fase, es decir, dos estimaciones por aplicación. En los procesos de calibración por separado que requieren la transformación de la habilidad se han estimado las constantes A y B utilizando los puntos de cuadratura y pesos calculados en la primera fase del proceso. La estimación de esas constantes se ha llevado a cabo empleando el software S.T 2.0 (Handon y Zeng, 2004). Este programa emplea los parámetros de los ítems comunes entre formas y los datos de las habilidades estimadas con BILOGMG para calcular las constantes con los cuatro métodos de transformación mencionados. Los parámetros estimados se muestran en la tabla siguiente:

Aplicación 1	**SL**	**H**	**MM**	**MS**
B (Intercepto)	0,040	0,022	0,015	0,015
A (Pendiente)	0,950	0,961	1,034	1,032
Aplicación 2				
B	0,091	0,069	-0,0,001	0,048
A	0,986	0,995	1,070	0,970
Aplicación 3				
B	-0,055	-0,047	0,013	0,010
A	15	1,029	1,016	1,023
Aplicación 4				
B	-0,042	-0,033	-0,022	-0,049
A	11	0,995	0,995	1,048

Tabla VII.1. Intercepto y Pendiente para la calibración horizontal por separado en función de la metodología y la ocasión de medida

En la calibración conjunta (CC), los parámetros de las dos formas en cada aplicación se estiman con una única ejecución del software BILOGMG. Los estudiantes que responden a las dos formas se sitúan en una misma base de datos tratando los ítems que no corresponde responder a alguno de ellos (los específicos

de cada una de las formas) como perdidos por diseño e indicándolo en el programa con la opción NFname en el comando >INPUT.

Finalmente, para la calibración fija (CF), se estiman en primer lugar los parámetros de los ítems la forma A. Una vez calculados, se utilizan los de los reactivos comunes con la forma B como elementos fijos en la calibración de dicho instrumento utilizando dos opciones del programa: en el comando >GLOBAL la opción PRname para identificar el archivo con los valores de los parámetros de los ítems comunes en la forma y la opción FIX en el comando >TEST para identificar cuáles son los reactivos que deben ser fijados.

Para evaluar los resultados de este primer problema se han llevado a cabo dos grupos de análisis. En primer lugar, un grupo de análisis basados en la Teoría Clásica de los Test donde se calculan estadísticos comparados para las dos formas del test en cada aplicación y un análisis en profundidad de los índices de dificultad de los ítems comunes que comparten las dos formas de una misma ocasión de medida. En segundo lugar, los resultados producidos por los modelos TRI que incluyen:

A. Estudio de las medias y variabilidad. Se ha llevado a cabo un análisis de las medias y desviaciones típicas de las puntuaciones producidas por las dos formas del test en cada una de las cuatro aplicaciones producidas por los distintos métodos de calibración empleados. También se han calculado las diferencias entre las medias de las dos formas para comprobar que metodología produce menores diferencias medias.

B. Estudio de las distancias horizontales. Se analiza la distribución de puntuaciones estimadas por los distintos métodos de calibración. En primer lugar se calculan las funciones o curvas de distribución acumulada del rasgo en ambas formas para poder ser comparadas (de 0 a 100%). Y para estudiar las posibles diferencias en las distribuciones de los dos grupos se calculan las distancias horizontales en distintos puntos de la distribución, utilizando los percentiles como los autores Jungnam (2007) y Holland (2002). Además se añade como anexo otro procedimiento, generado específicamente para este trabajo, para calcular las distancias horizontales empleando la marca de clase de 100 intervalos que agrupan las puntuaciones de los sujetos entre distintos puntos de la distribución

determinados por las proporciones acumuladas ($0<\theta\leq1$, $1<\theta\leq2$,…, $99<\theta\leq100$). En los resultados del estudio se incluyen únicamente los resultados de las distancias horizontales empleando los percentiles de la distribución. Y en el Anexo II se incluyen los resultados utilizando las marcas de clase, además de explicar su proceso de cálculo y la comparación con los percentiles.

Se calculan las diferencias en 7 percentiles y marcas de clase específicos de la distribución (5, 10, 25, 50, 75, 90 y 95). En este caso y de forma opuesta a Holland (2002), no se puede asumir que las distribuciones de los grupos estén estocásticamente ordenadas, es decir, las curvas de distribución acumulada de ambos grupos deberían ser lo más semejantes posibles y, por este motivo, restar las puntuaciones del grupo A al B o al revés únicamente produce diferencias en el signo de la distancia y, por tanto su interpretación. Por ejemplo, la puntuación en el rasgo para los sujetos situados en el percentil 75 es de 1,8 para la forma A de un test y 2 para la B. La diferencia es -0,2, este valor es la distancia horizontal entre formas en el percentil 80 (Δ_{80}). Cuando las distancias son negativas los estudiantes que han respondido a la forma B obtienen una mejor puntuación. Es decir, como la escala sigue una distribución normal, por tanto, consta de valores negativos y positivos (entre -3 y +3 aproximadamente), y un valor mejor en valores inferiores a la media (valores negativos) es un valor más pequeño, más cercano a cero. En cambio, en valores por encima de la media, un mejor valor es una puntuación más alta.

Cuando las distancias son positivas son los de la Forma A los que obtienen un mejor resultado.

$$\Delta_p = \theta_A - \theta_B \qquad p = 1,2,3 \ldots 99 \qquad \textbf{Ec. VII.1}$$

Y la media, en valor absoluto, de esas distancias ($|\bar{\Delta}|$) desde el percentil 1 hasta el 99 refleja la separación existente en todo el rango de puntuaciones de la distribución.

$$|\overline{\Delta P}| = \frac{\sum_{p=1}^{p=99}|\Delta_p|}{99} \qquad \textbf{Ec. VII.2}$$

Las distancias se calculan en valor absoluto para la obtención de la media con la finalidad de comprobar las diferencias entre formas, sin importar en qué test se obtiene una puntuación mejor.

De esta forma, una distancia media menor indica menores diferencias entre las puntuaciones estimadas por las distintas metodologías. Se han construido una serie de gráficos para facilitar la interpretación de estas distancias.

VII.2.2 Problema 2. Comparación de procedimientos para el anclaje vertical

El segundo problema planteado es identificar qué tipo de variaciones se producen en la estimación de la escala vertical como consecuencia de emplear distintos tipos de calibración en el proceso de anclaje. Se elaboran un total de 18 escalas verticales con la combinación de, por un lado seis tipos de calibración y, por otro, tres formas distintas de estimar la habilidad:

A. Método de calibración vertical:
   - A.1 Calibración Conjunta
   - A.2 Calibración por separado (empleando los cuatro tipos de transformación CSMM, CSMS, CSSL, CSH)
   - A.3 Calibración fija

B. Método de estimación de la habilidad:
   - B.1 Máxima Verosimilitud (MV)
   - B.2 Empírica a Posteriori (EAP). Utilizando la distribución empírica estimada en la fase de calibración como distribución a priori (con el comando idist=3 en la sección SCORE de BILOGMG).
   - B.3 Máxima a Posteriori (MAP)

Para tratar con el segundo objetivo el proceso de calibración conjunta se ha ejecutado utilizando la opción NGroups en el comando >INPUT del software BILOGMG. Esta opción trata cada una de las mediciones como un grupo y estimando a la vez todos los parámetros construye una escala vertical en una sola ejecución del programa. Kolen y Brennan (2004) señalan que la calibración conjunta puede ser mejor en teoría para el proceso de anclaje vertical ya que

utiliza toda la información disponible de la estimación de los parámetros y se espera que produzcan resultados más estables. Se han construido cuatro grupos, uno por aplicación y se introduce la opción Reference=1 en el comando >CALIB para identificar la primera medición como inicio de la escala, el resto deben situarse en esa misma escala.

Para la calibración por separado se han estimado los parámetros de los ítems en cada aplicación en diferentes ejecuciones del programa BILOGMG. En realidad, es un diseño que combina la calibración conjunta para estimar las dos formas en cada una de las mediciones y la calibración por separado para conseguir la escala vertical. Una vez estimados esos parámetros se utilizan los ítems comunes para estimar las constantes A y B en cada par de aplicaciones consecutivas con el software S.T, es decir, 1ª y 2ª aplicación, 2ª y 3ª aplicación y 3ª y 4ª aplicación. Por tanto, para situar la habilidad de la 2ª aplicación en la misma escala que la 1ª, que se utiliza como base, únicamente es necesaria una transformación. Pero para situar la 4ª aplicación en la misma escala que la 1ª son necesarias tres transformaciones, primero utilizando A y B para situarla en escala de la 3ª, después empleando las constantes A y B de la 2ª y 3ª aplicación y la tercera transformación empleando las constantes obtenidas en el proceso realizado con la 1ª y 2ª aplicación. Los valores de esas constantes son los siguientes:

	A2 en A1	SL	H	MM	MS
	B (Intercepto)	0,450	0,430	0,467	0,475
	A (Pendiente)	0,923	0,897	0,883	0,965
	A3 en A2				
EAP	B	0,837	0,816	0,867	0,814
	A	0,758	0,736	0,892	0,693
	A4 en A3				
	B	0,444	0,452	0,346	0,339
	A	0,979	1,018	1,002	0,951
	A2 en A1	**SL**	**H**	**MM**	**MS**
	B (Intercepto)	0,451	0,433	0,467	0,475
	A (Pendiente)	0,921	0,888	0,883	0,965
	A3 en A2				
MAP	B	0,838	0,814	0,867	0,814
	A	0,761	0,723	0,892	0,693
	A4 en A3				
	B	0,449	0,492	0,346	0,339
	A	0,991	0,993	1,002	0,951
	A2 en A1	**SL**	**H**	**MM**	**MS**
	B (Intercepto)	0,449	0,431	0,467	0,475
	A (Pendiente)	0,921	0,892	0,883	0,965
	A3 en A2				
MVL	B	0,832	0,825	0,867	0,814
	A	0,754	0,713	0,892	0,693
	A4 en A3				
	B	0,448	0,482	0,346	0,339
	A	0,988	0,992	1,002	0,951

Tabla VII.2. Intercepto y Pendiente para la calibración vertical por separado en función de la metodología y las ocasiones de medida equiparadas.

Finalmente, para la calibración fija de la misma manera que en la anterior se utiliza la calibración conjunta para la equiparación horizontal de ambas formas. En primer lugar, se estiman los parámetros de los ítems de la 1ª aplicación y se seleccionan los de los ítems de anclaje con la 2ª aplicación. Estos permanecen fijos en la calibración de la 2ª ocasión de medida con las opciones del programa: en el comando >GLOBAL la opción PRname para identificar el archivo con los valores de los parámetros de los ítems comunes en la forma, la opción FIX en el comando >TEST para identificar cuáles son los reactivos que deben ser fijados y, además la opción NOadjust en el comando >CALIB para evitar transformaciones de la escala durante el proceso de ciclos EM.

Los seis métodos distintos de calibración se combinan con las tres metodologías de estimación de la habilidad mencionadas. Para modificar la forma

de estimar las puntuaciones de la escala se incluye en el comando >SCORE de BILOGMG la opción METHOD=1 para implementar MVL, METHOD=2 para EAP y METHOD=3 para MAP. En el caso concreto de la metodología de estimación empírica a posteriori (EAP) se ha decidido utilizar la distribución empírica calculada en la fase de calibración como distribución a priori a través de la opción IDIST=3 en el comando >CALIB. Por defecto BILOG utiliza la distribución normal como distribución a priori, pero los resultados alcanzados con este procedimiento, en el proceso de calibración conjunta, indican una falta de normalidad en la distribución estimada a posteriori como muestran los gráficos siguientes

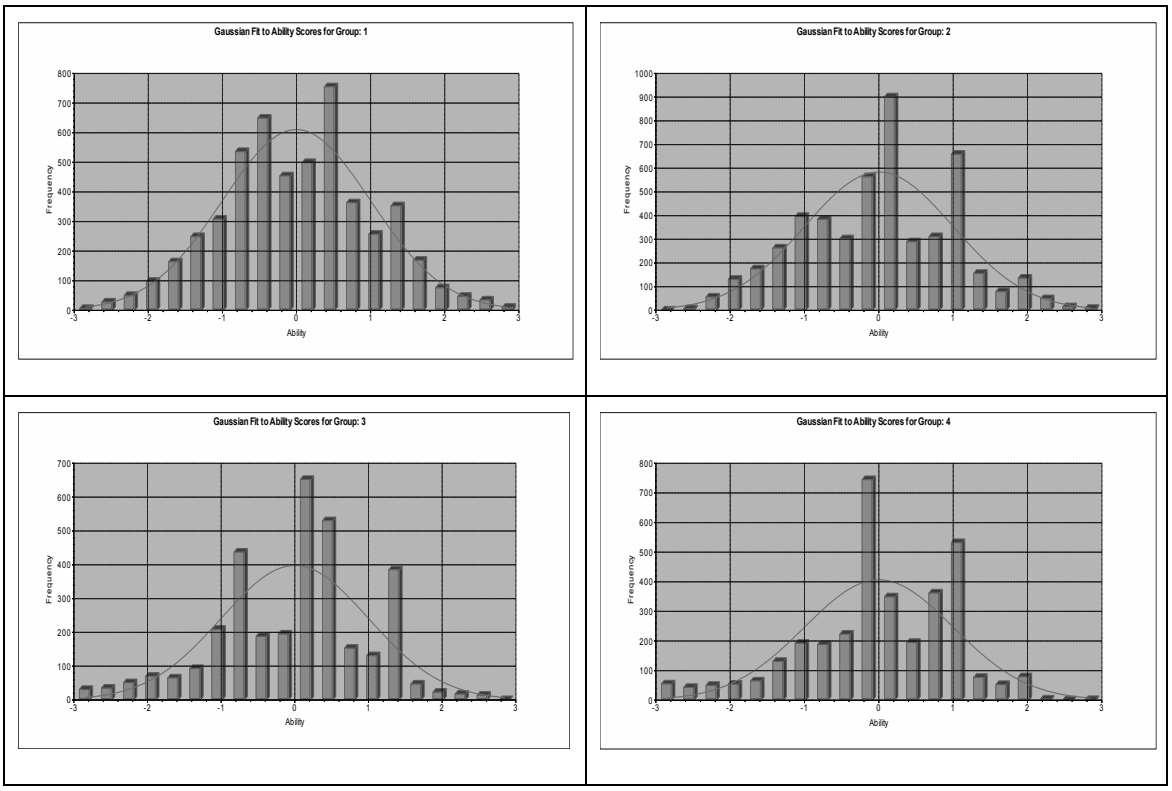


Gráfico VII.1. Distribución a posteriori de la habilidad con EAP y distribución normal a priori

En cambio, al utilizar la distribución empírica estimada en la fase de calibración como distribución a priori las características de la distribución posteriori cambian, ajustándose, en mayor medida, a la normalidad, como puede observarse en los gráficos que se presentan a continuación:

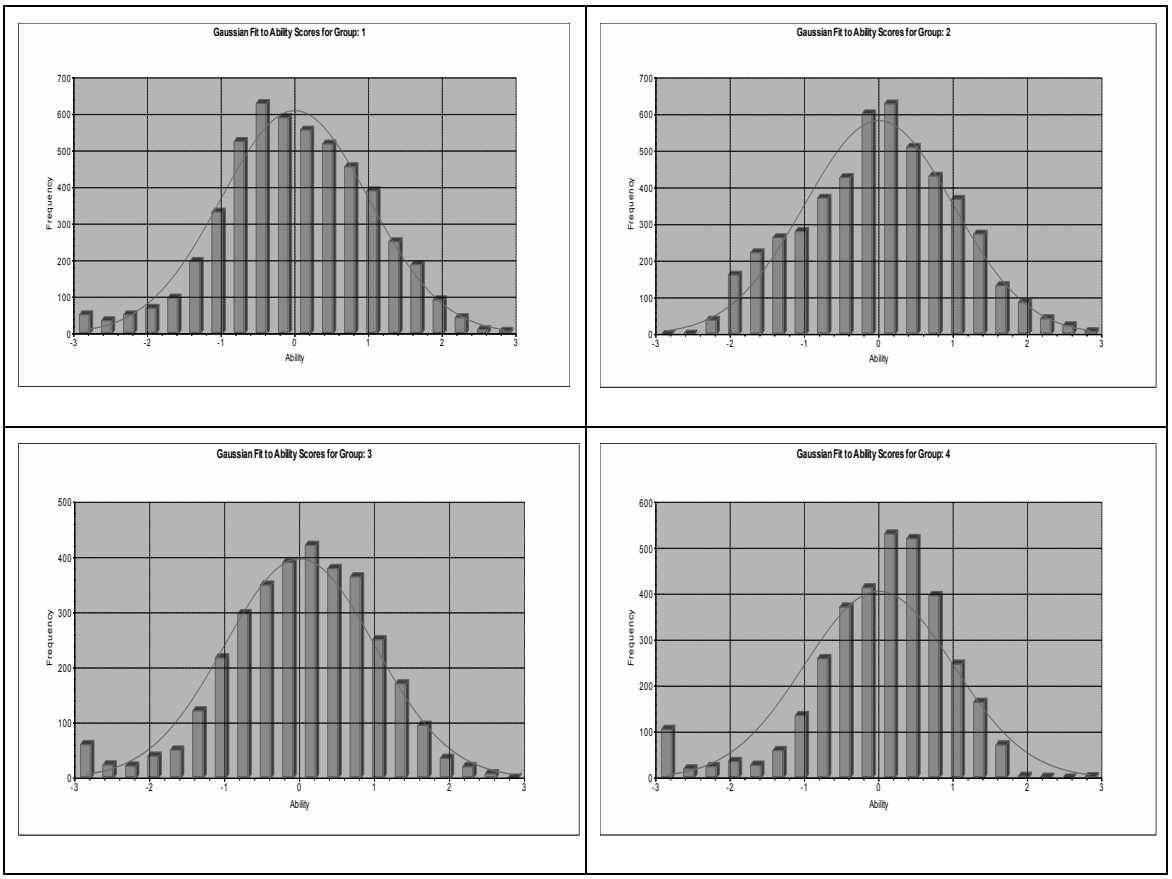


Gráfico VII.2. Distribución a posteriori de la habilidad con EAP utilizando la distribución empírica estimada en la fase de calibración como distribución a priori.

Para evaluar las diferentes escalas verticales elaboradas en el segundo objetivo del trabajo los criterios de comparación se centran en el estudio del crecimiento. En primer lugar, como fase de análisis previo, se estudian los índices de dificultad de los ítems de anclaje entre aplicaciones consecutivas según la TCT. Observando estos parámetros se comprueba si la proporción de respuestas correctas de dichos reactivos aumenta entre aplicaciones. Si ocurre así, es que a los estudiantes les resulta menos complicado responder correctamente a esos ítems en las aplicaciones superiores.

En segundo lugar, y entrando de lleno en las escalas verticales estimadas con TRI, se estudian varios aspectos:

A. Estudio del crecimiento y la variabilidad. Se estudia el cambio producido en la habilidad entre ocasiones de medida. Kolen y Brennan definen el crecimiento grado a grado o, en este caso, entre inicio y final de un mismo curso, como el cambio que se produce en el contenido enseñado entre un grado y el siguiente (2004). Analizando las diferencias entre las medias de aplicaciones

consecutivas es posible conocer como son los patrones de crecimiento, es decir, como evolucionan las diferencias aplicación a aplicación. Se estudia si la ganancia crece a medida que se avanza en el curso o, al contrario, se hace más pequeña.

La comparación de la variabilidad grado a grado. Se utilizan las desviaciones típicas en cada aplicación, es otro de los aspectos empleados en la investigación sobre escalas verticales ya que se han encontrado patrones de dispersión distintos entre métodos de escalamiento diferentes. Yen (1986) y Tong y Kolen (2007) comparan el método Thurstone con métodos TRI de escalamiento y encuentran que en el primero la dispersión aumenta con la edad, en cambio, utilizando TRI ocurre lo opuesto sobre todo con un diseño de ítems comunes.

B. Estudio del tamaño del efecto (Yen, 1986). Este parámetro permite comparar el crecimiento año a año o medición a medición, a lo largo de una escala ajustando por la variabilidad de la propia escala. Lo que realiza Yen es una estandarización de la diferencia de medias entre dos cursos o aplicaciones consecutivas, utilizando la raíz cuadrada de la media varianzas de cada aplicación. Puede observarse en la ecuación Ec. VII.3.

$$\text{Tamaño del efecto} = \frac{\overline{\theta}_{\text{superior}} - \overline{\theta}_{\text{inferior}}}{\sqrt{\frac{\sigma^2_{\text{superior}} + \sigma^2_{\text{inferior}}}{2}}} \qquad \textbf{Ec. VII.3}$$

Donde $\overline{\theta}_{\text{superior}}$ es la media de las puntuaciones de la escala en el curso superior, $\overline{\theta}_{\text{inferior}}$ es la media del curso inferior. $\sigma^2_{\text{superior}}$ y $\sigma^2_{\text{inferior}}$ representan las varianzas en cada una de las aplicaciones. Un tamaño del efecto mayor indica más crecimiento y una mayor separación entre los resultados de las aplicaciones. Este análisis se ha llevado a cabo en diferentes investigaciones sobre los efectos que producen las variaciones en la metodología de anclaje vertical (Kolen & Brennan, 2004; Chin, Kim & Nering, 2006; Jungnam, 2007; Briggs, Weeks & Wiley, 2008; Briggs & Weeks, 2009). La interpretación del tamaño del efecto también puede llevarse a cabo de la misma manera que una puntuación típica (z), identificando que proporción de estudiantes del grupo superior se encuentra por encima de la media del grupo inferior.

C. Estudio distancias horizontales entre los percentiles y las marcas de clase de los intervalos de puntuaciones que los propios percentiles determinan. De la misma forma que en el problema 1, en el apartado de resultados se incluyen solo las distancias horizontales empleando percentiles. Para comprobar el otro procedimiento debe consultarse el Anexo II.

Se compara cada par de aplicaciones consecutivas mientras que en el primer objetivo se comparan las dos formas dentro de una misma aplicación. Otra diferencia es que en este caso la distancia media no se calcula empleando el valor absoluto de cada distancia en los 99 percentiles ya que distancias negativas en este caso indican que no ha habido crecimiento y lo esperado es que las distancias entre aplicaciones consecutivas sean positivas.

$$\Delta_p = \theta_{superior} - \theta_{inferior} \qquad p = 1,2,3 \dots 99 \qquad \textbf{Ec. VII.4}$$

Por tanto una distancia horizontal es la diferencia en un determinado percentil entre la puntuación del curso superior y la del inferior. Y la distancia horizontal media se obtendría de la siguiente manera:

$$\overline{\Delta P} = \frac{\sum_{p=1}^{p=99} \Delta_p}{99} \qquad \textbf{Ec. VII.5}$$

VII.3 Resultados

La exposición de los resultados sigue el orden de los problemas formulados para dar respuesta a la cuestiones que plantean. Por consiguiente, en primer lugar, se comparan los procedimientos de equiparación horizontal y, en segundo lugar, los de anclaje vertical.

VII.3.1 Problema 1. Comparación de procedimientos para la equiparación horizontal

El tipo de diseño específico empleado en este estudio que combina una aplicación en espiral de dos formas elaboradas en cuatro aplicaciones para contar con grupos equivalentes, además de incluir ítems comunes en las dos formas de

cada una de las aplicaciones, permite comprobar diferentes metodologías para llevar a cabo la calibración en el proceso de equiparación horizontal. Se han elaborado cuatro modelos TRI de tres parámetros distintos que varían en su manera de llevar a cabo la calibración:

- Calibración por separado per sin ningún tipo de transformación (CS).
- Calibración por separado utilizando las cuatro formas de estimación de las constantes A y B necesarias para la transformación. Los métodos de transformación utilizados son: media/media (CSMM), media/sigma (CSMS), Haebara (CSH) y Stocking y Lord (CSSL).
- Calibración Conjunta (CC)
- Calibración Fija (CF)

Aunque se han construido cuatro modelos TRI, son un total de site resultados distintos a comparar ya que la transformación de los parámetros en la calibración por separado se lleva a cabo de cuatro maneras distintas, junto con los resultados sin ningún tipo de transformación de la habilidad. La calibración por separado sin llevar a cabo la transformación es posible realizarla debido al diseño de grupos equivalentes y es útil para conocer si realmente se producen esos resultados que el tipo de diseño reclama teóricamente, es decir, si las puntuaciones de los sujetos en ambas formas diseñadas son equivalentes. Antes de estudiar los resultados de los diferentes modelos TRI conviene observar, de forma preliminar, los resultados desde la aproximación de la Teoría Clásica de los Test:

VII.3.1.1 Análisis desde la TCT

A continuación se presentan los estadísticos descriptivos para todas las formas construidas en las cuatro aplicaciones que incluye este estudio.

	A1		A2		A3		A4	
	Forma A	**Forma B**	**Forma A**	**Forma B**	**Forma A**	**Forma B**	**Forma A**	**Forma B**
Nº de ítems	37	38	38	36	39	39	40	40
Nº de sujetos	2.574	2.532	2.443	2.438	1.665	1.662	1.69	1.703
Mínimo	4	5	3	4	4	5	2	0
Máximo	36	38	37	36	39	38	39	39
Nº medio de aciertos	21,57	21,52	20,43	19,42	22,16	22,09	22,35	21,92
Error típico de la media	0,12	0,11	0,13	0,13	0,15	0,15	0,14	0,15
Varianza	34,78	29,38	41,77	40,16	39,17	37,85	34,56	36,92
DT	5,90	5,42	6,46	6,34	6,26	6,15	5,88	6,08
Facilidad media	0,58	0,57	0,54	0,54	0,57	0,57	0,56	0,55
Rbp media	0,35	0,31	0,36	0,38	0,36	0,35	0,33	0,33
Fiabilidad (α de Cronbach)	0,80	0,76	0,82	0,83	0,82	0,81	0,78	0,79
Error típico de medida	2,61	2,66	2,73	2,64	2,66	2,65	2,73	2,77

Tabla VII.3. Análisis de las pruebas desde la Teoría Clásica de los Test

Conviene mencionar que la diferencia entre el número total de ítems entre las formas de las aplicaciones se debe a los procesos de depuración ya mencionados. La distinta longitud entre las formas en la 1º y 2ª aplicación puede evidenciar la necesidad de utilización de algún tipo de método de calibración horizontal, aunque el diseño inicialmente permita contar con grupos equivalentes sin llevar a cabo transformación alguna.

Los estadísticos calculados muestran similitud entre las formas. La facilidad media que muestra la proporción media de aciertos es prácticamente la misma entre formas, también entre aplicaciones es similar, y los errores típicos de medida también son parecidos. Los índices de fiabilidad de las pruebas se encuentran alrededor de un 0,8, la forma B de la 1ª aplicación es la que peor resultado presenta con 0,76 aunque sigue siendo un valor aceptable para este tipo de test de rendimiento. Los índices de discriminación (r_{bp} media) son de 0,35 aproximadamente. El de la forma B desciende a 0,31. En ambos casos se pueden considerar buenos índices de homogeneidad.

En la primera aplicación se observa que las puntuaciones de la forma A se encuentran un poco más dispersas que las de la forma B. También existe entre estas formas una diferencia en la fiabilidad estimada. La forma A es 0,4 puntos superior que la de la forma B en el valor de alpha, incluso cuando esta última tiene un ítem más.

Aunque los datos calculados desde la TCT pueden sugerir inicialmente resultados equivalentes entre las formas de cada aplicación, esas diferencias encontradas entre la discriminación y fiabilidad, sobre todo en la 1º medición, pueden plantear la necesidad de llevar a cabo una transformación de la habilidad para asegurar esa equivalencia de puntuaciones.

Si se analizan los parámetros de dificultad estimados desde la TCT en los ítems comunes que comparten las dos formas de una misma aplicación no se observan grandes diferencias, como muestra la tabla siguiente (Tabla VII.4):

	A1		**A2**		**A3**		**A4**	
ítems comunes	**Forma A**	**Forma B**	**Forma A**	**Forma B**	**Forma A**	**Forma B**	**Forma A**	**Forma B**
1	0,93	0,93	0,71	0,72	0,48	0,50	0,26	0,23
2	0,90	0,89	0,53	0,52	0,43	0,39	0,25	0,22
3	0,51	0,52	0,67	0,70	0,72	0,69	0,48	0,47
4	0,54	0,55	0,62	0,63	0,67	0,65	0,34	0,33
5	0,51	0,51	0,70	0,73	0,90	0,89	0,23	0,23
6	0,39	0,40	0,40	0,42	0,82	0,79	0,76	0,74
7	0,48	0,49	0,75	0,78	0,19	0,17	0,55	0,57
8	0,76	0,75	0,75	0,77	0,61	0,58	0,87	0,86
9	0,32	0,32	0,53	0,53	0,54	0,56	0,64	0,64
10	0,40	0,41	0,41	0,41	0,76	0,76	0,32	0,28
11	0,70	0,69	0,21	0,21	0,21	0,23	0,38	0,40
12	0,26	0,26	0,68	0,71	0,63	0,61	0,82	0,81
13	0,26	0,27	0,60	0,61	0,30	0,30	0,58	0,54
14	0,71	0,72	0,43	0,45	0,39	0,37	0,30	0,31
15	0,44	0,42	0,23	0,22	0,10	0,11	0,63	0,61
16	0,92	0,93	0,28	0,30	0,51	0,54	0,44	0,45
17	0,60	0,62	0,56	0,59	0,74	0,73	0,82	0,83
18	0,87	0,87	0,50	0,53	0,62	0,60	0,54	0,52
19	0,75	0,77			0,29	0,31	0,28	0,27
20							0,40	0,41
21							0,48	0,48
22							0,59	0,57
23							0,82	0,82
24							0,69	0,69
25							0,72	0,73
26							0,61	0,60
27							0,52	0,51
28							0,76	0,77
29							0,36	0,36
30							0,57	0,56
Media	0,59	0,60	0,53	0,55	0,52	0,51	0,53	0,53
Correlación Pearson	0,999		0,998		0,997		0,997	

Tabla VII.4. Índices de dificultad TCT (% Correctas) de los ítems comunes entre formas

Los resultados no reflejan diferencias superiores al 3% en la proporción de respuestas correctas entre los estudiantes de los dos grupos en estos ítems comunes. Y la media de esa proporción, como es de esperar, es similar entre formas. La mayor distancia se encuentra en la segunda aplicación donde hay un 2% más de estudiantes que responden, en términos medios, correctamente a esos ítems comunes. Las correlaciones entre parámetros de dificultad de los dos

instrumentos son superiores a 0,99. El comportamiento de los ítems comunes es bueno y representan más del 40% de las pruebas alcanzando una proporción mayor en la última aplicación. Estos factores indican que no debería haber problemas si se utilizan en los diferentes métodos de calibración horizontal.

En resumen, los resultados iniciales, llevados a cabo desde la TCT, pueden indicar que los grupos que han respondido a cada una de los instrumentos son equivalentes. Aunque puede ser adecuado emplear algún tipo de calibración para evitar problemas que puedan surgir por esas diferencias en longitud, fiabilidad y discriminación de alguna de las formas. Además unos buenos ítems de anclaje pueden aumentar la fiabilidad del proceso de transformación del rasgo.

VII.3.1.2 Análisis desde la TRI

Una vez analizados los resultados preliminares calculados desde la TCT es el momento de comenzar con los modelos TRI. El foco de este análisis son las estimaciones del rasgo, en este caso rendimiento en matemáticas, de cada uno de los estudiantes, obtenidas a través de las diferentes metodologías de calibración y su posición a lo largo de la distribución (percentil).

A. Análisis de las medias y la variabilidad

En primer lugar, se ha llevado a cabo un estudio de las puntuaciones medias estimadas con los diferentes tipos de calibración en los dos grupos y también la desviación típica para comprobar la dispersión de las puntuaciones. La información se presenta en tres tablas distintas. La primera incluye los resultados de los distintos tipos de transformación en la calibración por separado y, por supuesto, también se incorporan los resultados sin llevar a cabo ningún tipo de transformación. La segunda tabla muestra de ambos grupos empleando la metodología de calibración conjunta. Por último, los resultados de la calibración fija, están reflejados en la tercera tabla. Los resultados de la forma A en la calibración por separado y fija son los mismos ya que con ambos procesos dicha forma se calibra de forma independiente utilizando el mismo procedimiento.

Al analizar las cinco metodologías de calibración por separado, cuatro de ellas transforman la puntuación de la Forma B del test para situarla en la misma escala que la Forma A, se observa que los resultados varían entre aplicaciones

En la primera aplicación las medias de la Forma B son ligeramente superiores a las de la Forma A. El método de transformación CSSL es el que produce las mayores diferencias entre las medias y la calibración por separado sin ningún tipo de transformación es la metodología que menos diferencia produce en las medias de ambos grupos. La tendencia se repite en la segunda aplicación, aunque en este caso la diferencia entre las medias es la misma utilizando el método CSMM y sin llevar a cabo ninguna transformación.

En la tercera aplicación la transformación CSMM y CSMS producen puntuaciones medias ligeramente superiores en la Forma B, aunque en el último caso las diferencias son casi inexistentes. La calibración por separado sin transformación es la que produce menores diferencias entre las medias. Igual que en las dos aplicaciones anteriores, la metodología SL es la que mayores diferencias produce pero, en este caso, la puntuación media de la forma A es superior a la de la Forma B.

Finalmente, en la cuarta aplicación, todos los métodos de transformación empleados en la calibración por separado producen puntuaciones medias más bajas en la Forma B. Sin utilizar la transformación, las medias de ambos grupos coinciden.

Las puntuaciones medias indican que, en la calibración por separado, no emplear ningún tipo de transformación produce menores diferencias entre las medias de las dos formas en las cuatro aplicaciones. En cuanto a los métodos de transformación empleados, el CSMM es el que muestra resultados más parecidos en las puntuaciones medias, seguido de CSMS.

Respecto a las desviaciones típicas, no se refleja ninguna tendencia reseñable entre métodos de transformación. Los valores fluctúan entre metodologías y aplicaciones. Por ejemplo, en la 1ª aplicación la CSSL Y CSH producen puntuaciones más dispersas y en la 2ª CSMM y CS como puede comprobarse en la Tabla VII.5.

	A1			A2			A3			A4	
	Media	DT		Media	DT		Media	DT		Media	DT
Forma A			**Forma A**			**Forma A**			**Forma A**		
CS	0,002	0,915	CS	0,002	0,921	CS	0,001	0,928	CS	0,002	0,922
Forma B			**Forma B**			**Forma B**			**Forma B**		
CS	0,003	0,902	CS	0,004	0,928	CS	0,001	0,930	CS	0,002	0,923
CSSL	0,043	0,856	CSSL	0,095	0,914	CSSL	-0,054	0,934	CSSL	-0,040	0,924
CSH	0,025	0,867	CSH	0,073	0,923	CSH	-0,045	0,956	CSH	-0,031	0,919
CSMM	0,018	0,932	CSMM	0,004	0,993	CSMM	0,014	0,945	CSMM	-0,020	0,919
CSMS	0,019	0,931	CSMS	0,052	0,900	CSMS	0,012	0,950	CSMS	-0,047	0,968
A – B			**A – B**			**A – B**			**A – B**		
CS	-0,001		CS	-0,002		CS	0,000		CS	0,000	
CSSL	-0,041		CSSL	-0,093		CSSL	0,055		CSSL	0,042	
CSH	-0,023		CSH	-0,071		CSH	0,046		CSH	0,033	
CSMM	-0,016		CSMM	-0,002		CSMM	-0,013		CSMM	0,021	
CSMS	-0,016		CSMS	-0,050		CSMS	-0,011		CSMS	0,049	

Tabla VII.5. Medias, Desviaciones Típicas y diferencia de medias en la Calibración por Separado sin equiparación (CS) y utilizando 4 formas de transformación del rasgo: Stocking-Lord (CSSL), Haerbera (CSH), Media-Media (CSMM) y Media-Sigma (CSMS)

Si se emplea la CC, en las dos primeras aplicaciones las medias de la forma A son menores que las de la B. En cambio, en las dos últimas la tendencia cambia. Las mayores diferencias se producen en la segunda aplicación, aunque siguen siendo superadas por las producidas por los métodos SL y H en la calibración por separado.

	A1			A2			A3			A4	
	Media	DT		Media	DT		Media	DT		Media	DT
Forma A			Forma A			Forma A			Forma A		
CC	-0,004	0,921	CC	-0,028	0,919	CC	0,022	0,925	CC	0,013	0,924
Forma B			Forma B			Forma B			Forma B		
CC	0,009	0,897	CC	0,035	0,930	CC	-0,020	0,934	CC	-0,010	0,923
A – B			**A – B**			**A – B**			**A – B**		
CC	-0,013		CC	-0,063		CC	0,042		CC	0,023	

Tabla VII.6. . Medias, Desviaciones Típicas y diferencia de medias en la Calibración por Conjunta (CC)

Los resultados producidos por la calibración fija muestran también muy poca diferencia entre las medias de ambas formas. Las diferencias son menos que la calibración por separado con transformación.

	A1			A2			A3			A4	
	Media	DT		Media	DT		Media	DT		Media	DT
Forma A			**Forma A**			**Forma A**			**Forma A**		
CF	0,002	0,915	CF	0,002	0,921	CF	0,001	0,928	CF	0,002	0,922
Forma B			**Forma B**			**Forma B**			**Forma B**		
CF	0,006	0,900	CF	0,012	0,924	CF	-0,008	0,934	CF	0,002	0,923
A – B			**A – B**			**A – B**			**A – B**		
CF	-0,004		CF	-0,010		CF	0,008		CF	0,000	

Tabla VII.7. Medias, Desviaciones Típicas y diferencia de medias en la Calibración Fija (CF)

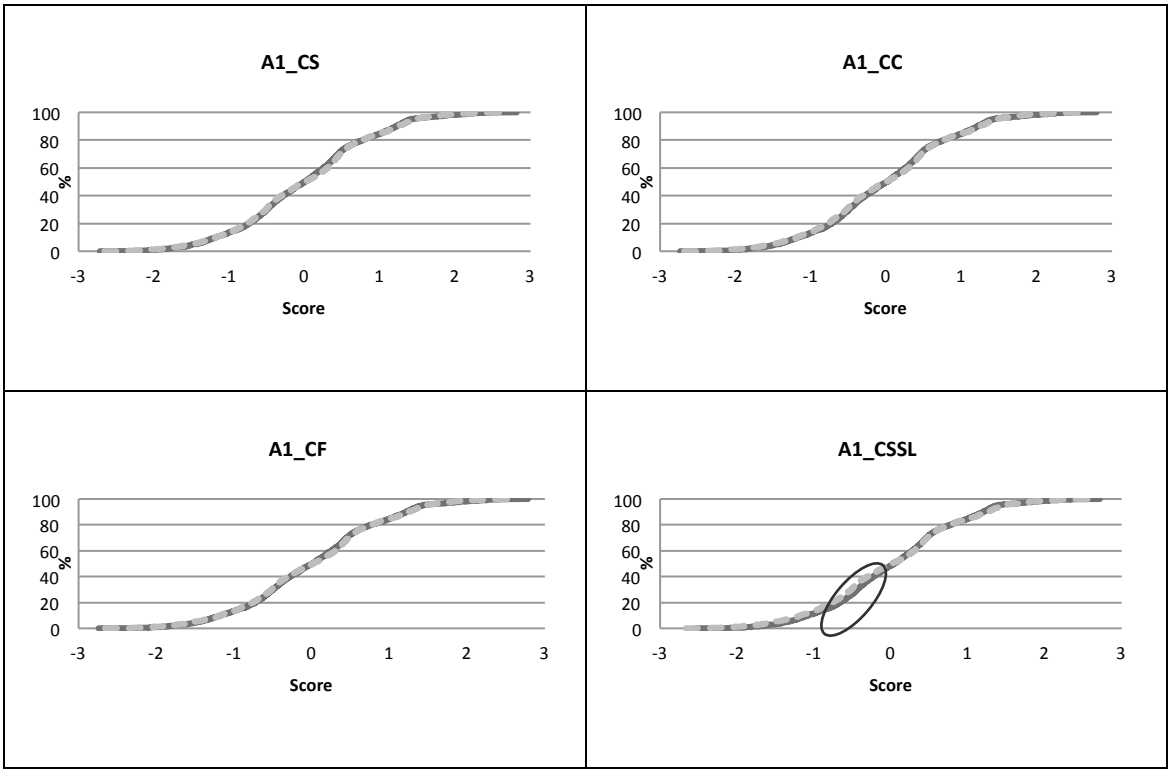
En resumen, el análisis de las puntuaciones medias y las diferencias entre las estimaciones de las dos formas de cada aplicación indican que la CS sin

transformación produce estimaciones medias del rasgo más cercanas entre sí que el resto de procesos de calibración horizontal. La siguiente metodología con una menor diferencia en las puntuaciones medias es la CF, seguidas de la CC y la CCMM.

B. Estudio de las distancias horizontales

Si se observan en las curvas construidas a partir de la distribución acumulada de las puntuaciones del rasgo, estimadas con cada uno de los métodos de calibración empleados, es posible comprobar cuál de ellos produce una mayor distancia entre los resultados de los dos grupos. Los siguientes gráficos muestran las curvas mencionadas para cada aplicación y método de calibración por separado. La línea discontinua hace referencia a la forma A y la continua a la B y una mayor separación entre ellas indica que las puntuaciones del rasgo de ambos grupos son distintas

En la primera aplicación (A1), de acuerdo con los gráficos (ver Gráfico VII.3), los métodos de transformación SL y H en la calibración por separado producen una mayor distancia entre los valores del rasgo situados entre los percentiles 20 y 40 aproximadamente. El resto de métodos de calibración producen curvas bastante parecidas.



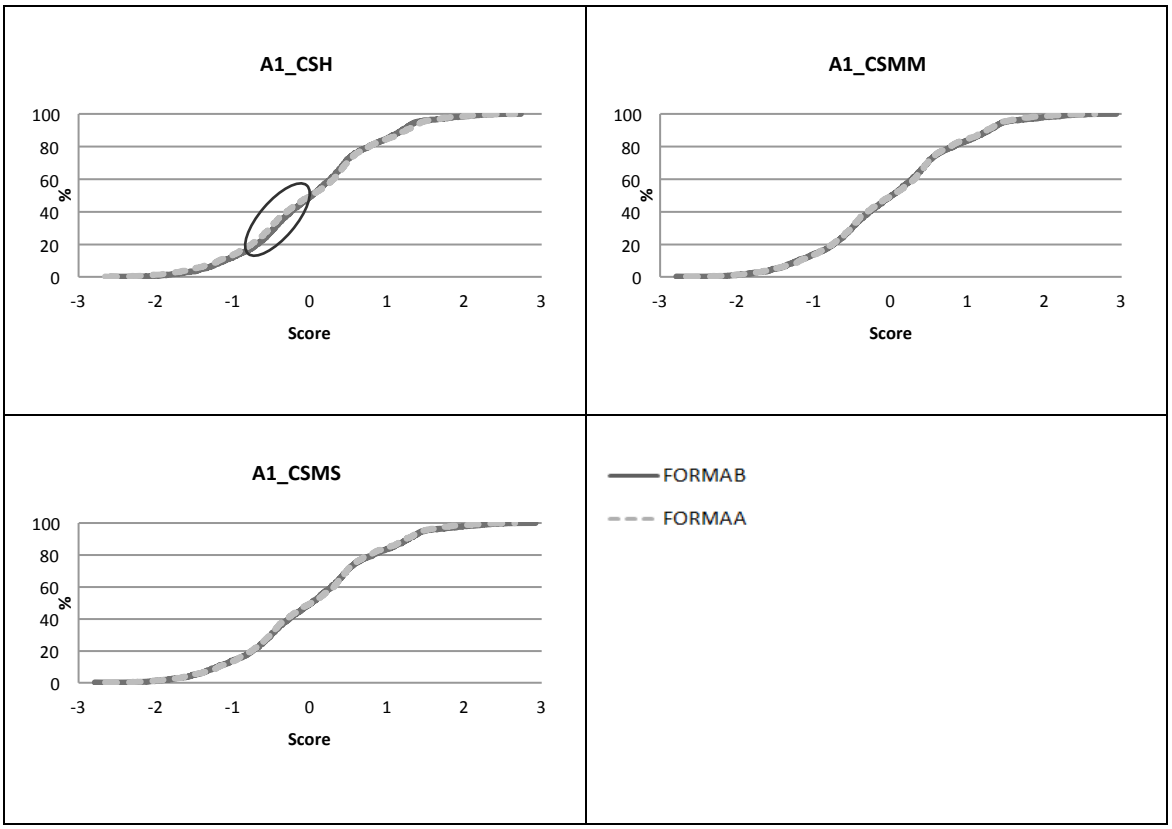
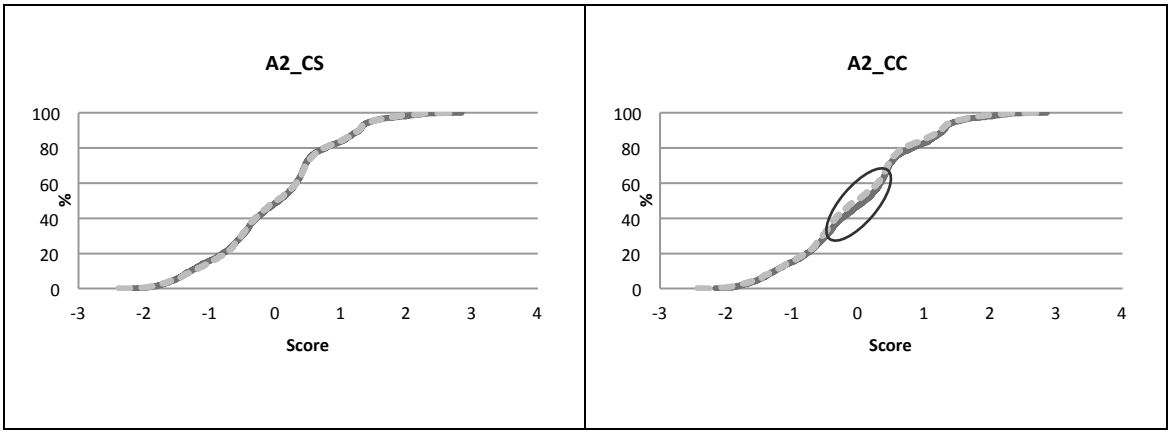


Gráfico VII.3. Curvas de Distribución Acumuladas para las dos formas del test, diferenciando los distintos tipos de calibración horizontal en la A1.

En la aplicación número dos (A2), la CC hace variar ligeramente valores del rasgo cercanos a la media, entre los percentiles 35 y 55. Los métodos SL y H producen estimaciones de la habilidad ligeramente distintas entre los percentiles 20 y 80 aproximadamente. También es conveniente mencionar que la transformación MM produce diferencias en los valores bajos del rasgo. En la CS y la CF las curvas de ambas formas son casi superpuestas



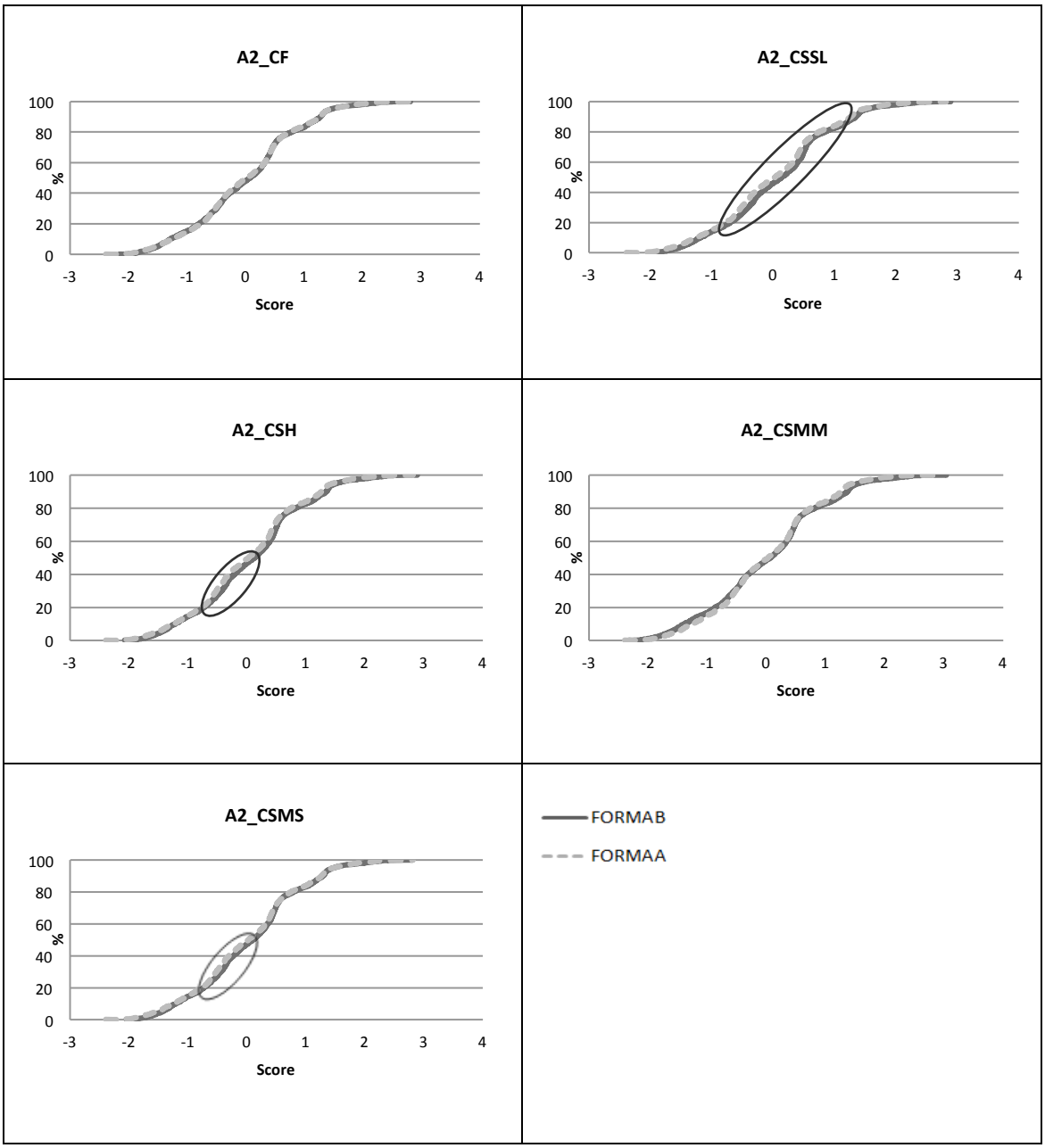


Gráfico VII.4. Curvas de Distribución Acumuladas para las dos formas del test, diferenciando los distintos tipos de calibración horizontal en la Aplicación 2.

En la tercera aplicación (A3), llevada a cabo en noviembre de 2006, las diferencias entre las posiciones que ocupan los sujetos en la distribución del rasgo estimado en ambas formas no son apreciables en el gráfico.

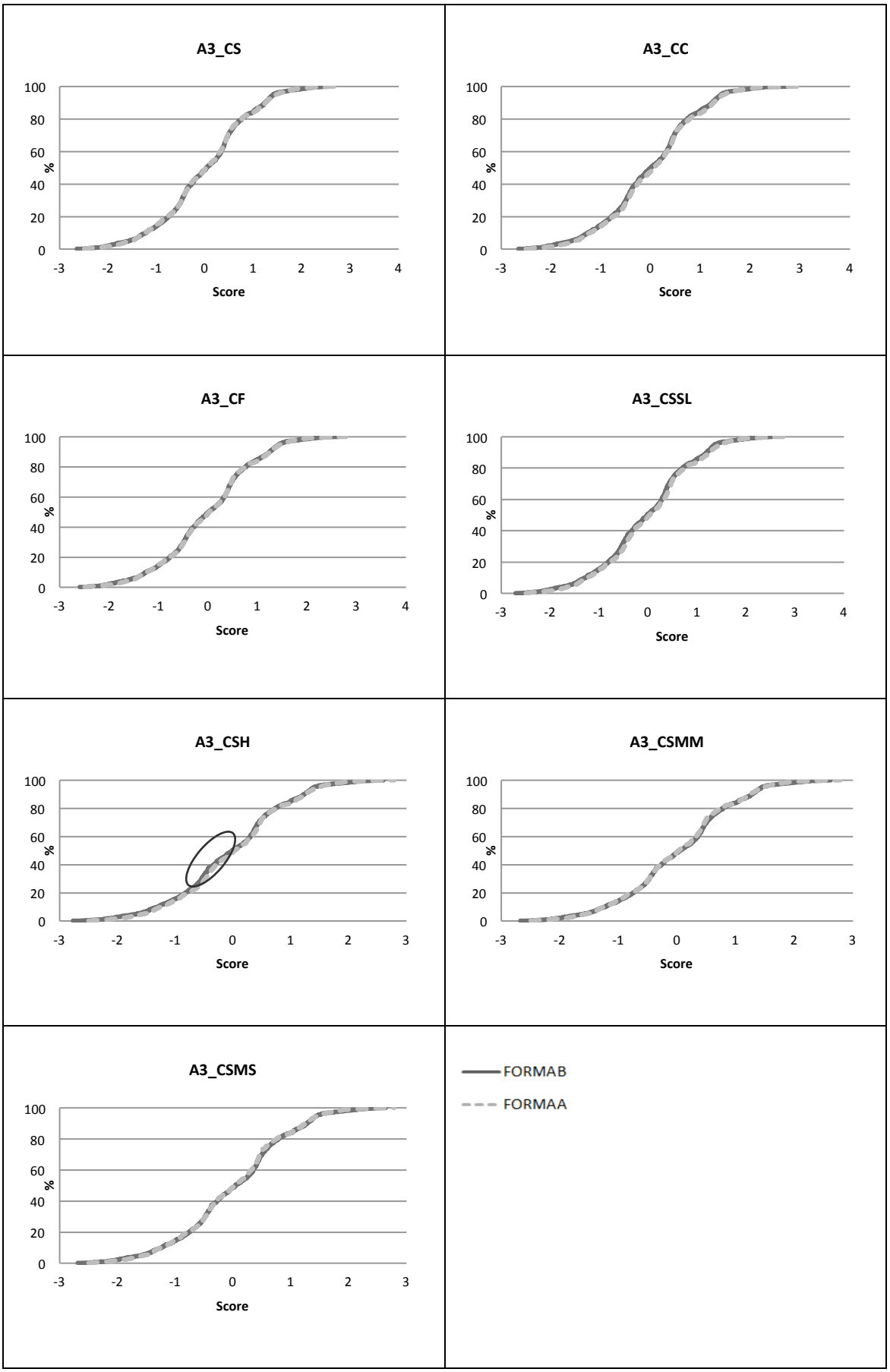
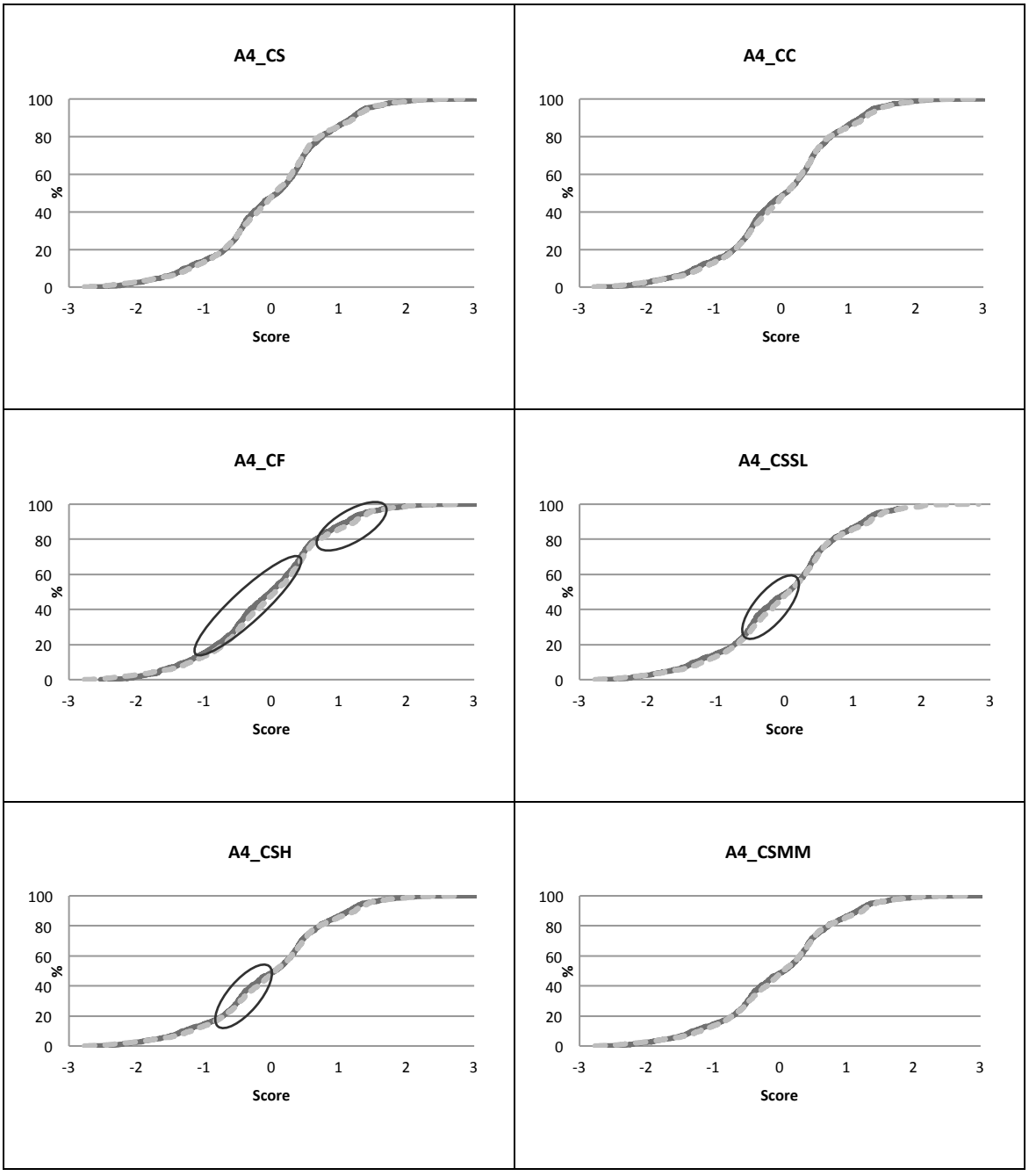


Gráfico VII.5. Curvas de Distribución Acumuladas para las dos formas del test, diferenciando los distintos tipos de calibración horizontal en la Aplicación 3.

Finalmente, en la aplicación número cuatro (A4), la CF produce diferencias casi a lo largo de todos los valores del rasgo. Los métodos de transformación CSSL, CSH y CSMS también estiman puntuaciones de rendimiento ligeramente distintas para las mismas posiciones percentílicas. En el caso de SL y H estas distancias se dan, principalmente, entre el percentil 25 y 45, en cambio para MS las diferencias se encuentran entre el 10 y el 50 aproximadamente.



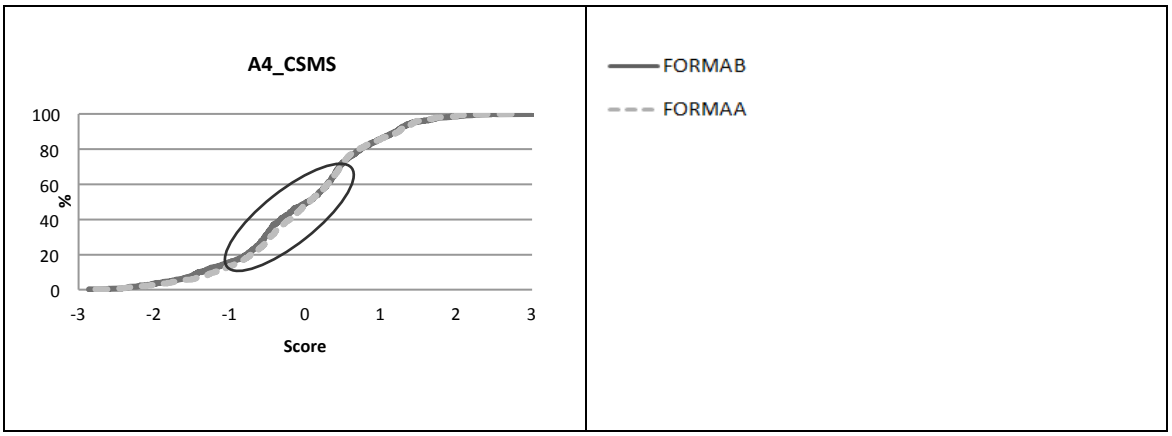


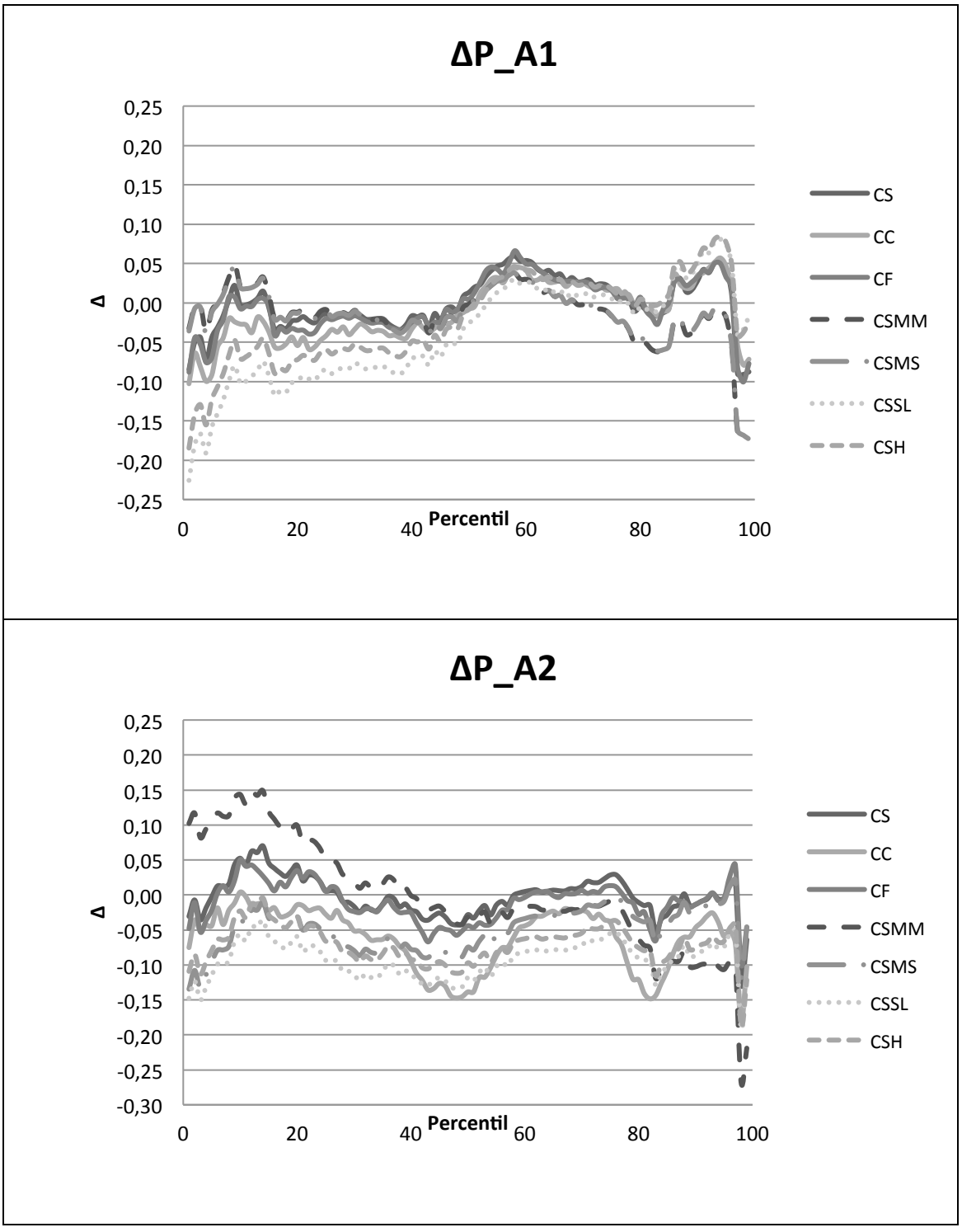
Gráfico VII.6. Curvas de Distribución Acumuladas para las dos formas del test, diferenciando los distintos tipos de calibración horizontal en la Aplicación 4.

Todos los gráficos muestran la casi total superposición de ambas curvas y es arriesgado concluir que existen diferencias, considerando además que los datos empleados pertenecen a una investigación empírica y no a una simulación y que puede haber pequeñas variaciones muestrales. Como consecuencia de esto se han construido cuatro gráficos resumen, uno por aplicación, para hacer visibles las distancias existentes entre las dos formas en cada método de calibración, aunque estas sean pequeñas. Se ha empleado para ello, siguiendo a Holland (2002), las distancia horizontales[78], que se han denominado Deltas (Δ) en este trabajo.

Si existen pequeñas distancias en el rasgo estimado en las dos formas a través de las distintas metodologías de calibración, aunque es difícil apreciarlo a simple vista. En los gráficos anteriores ya que las diferencias entre las puntuaciones del rasgo en los diferentes tramos de la distribución acumulada alcanzan valores máximos de 0,3 en los extremos de la distribución e inferiores a 0,05 en el centro de la distribución.

Por tanto, con la finalidad de comprobar al detalle la separación o distancia existente entre las puntuaciones estimadas para los estudiantes en las dos formas de las pruebas de matemáticas, se han elaborado cuatro gráficos resumen (ver Gráfico VII.7), uno para cada aplicación, que incluyen las distancias horizontales en cada uno de los 99 percentiles de la distribución, diferenciando cada uno de los métodos de equiparación.

[78]Se explican con más detalle en la sección de metodología de este capítulo (ver apartado VII.2.2) y de forma más extensa en el Anexo II.


ΔP_A1
0,25
0,20
0,15
0,10
0,05
0,00
-0,05
-0,10
-0,15
-0,20
-0,25
Δ
0
20
40
60
80
100
Percentil
CS
CC
CF
CSMM
CSMS
CSSL
CSH
ΔP_A2
0,25
0,20
0,15
0,10
0,05
0,00
-0,05
-0,10
-0,15
-0,20
-0,25
-0,30
Δ
0
20
40
60
80
100
Percentil
CS
CC
CF
CSMM
CSMS
CSSL
CSH

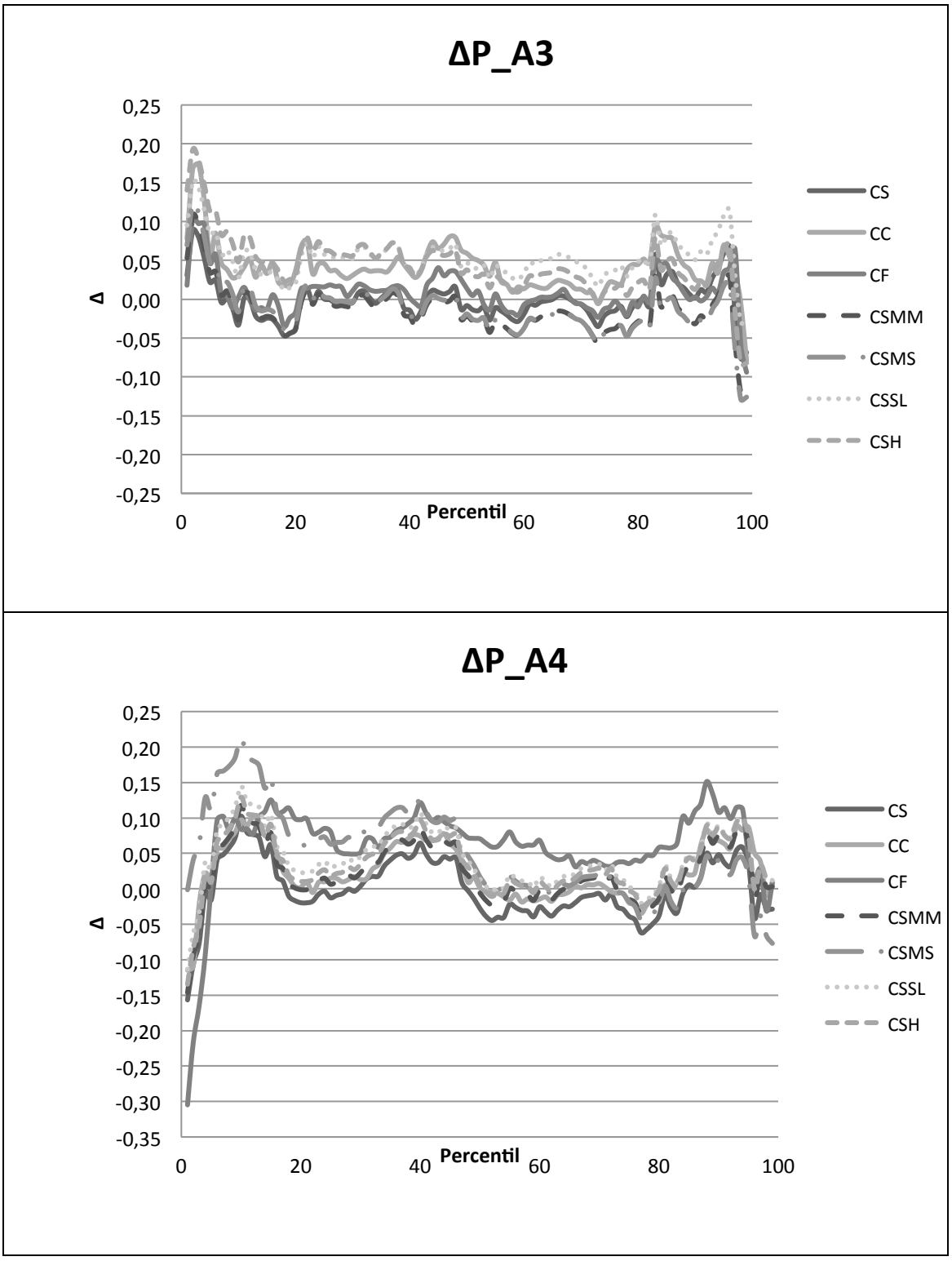


Gráfico VII.7. Distancias horizontales en las 99 Percentiles, en función de la metodología de calibración horizontal, en cada una de las aplicaciones por separado.

Cuando la distancia es negativa los estudiantes que han respondido a la Forma B obtienen una mejor puntuación en el rasgo estimado que los estudiantes del mismo percentil en la Forma A. Si esa distancia es positiva ocurre lo contrario. En cualquiera de los dos casos, es decir, una acumulación de distancias de un signo

determinado, señalarían que una de las formas obtendría puntuaciones distintas y con una tendencia específica. Lo óptimo es que se produzcan las menores distancias posibles y, en el caso de encontrarlas, que no exista tendencia en las mismas o, al menos, que no sean muy diferentes a las producidas por la metodología de equiparación horizontal que no realiza ningún tipo de transformación en el rasgo (CS).

La tendencia común que se observa es esa acumulación de las distancias más pronunciadas en los extremos de la distribución. En cambio, a medida que se llega al centro de la distribución los valores son próximos a cero. Los distintos métodos de calibración producen patrones similares en las distancias que se ve reflejado en los picos de las distintas líneas del gráfico.

En las estimaciones producidas por los cuatro métodos de transformación en la calibración por separado, trazados con líneas discontinuas en los gráficos se encuentran, en la mayor parte de los casos, las mayores distancias. Aunque las diferencias varían ligeramente entre aplicaciones.

Respecto a A1, la mayor parte de los métodos de calibración, excepto CSMM y CSMS, producen distancias negativas en los percentiles bajos (entre 10 y 45) de la distribución, es decir, puntuaciones más altas para la forma B. La CSSL y CSH son los métodos que mayores deltas producen en esta parte de la distribución. Entre el percentil 50 y 65 aproximadamente pasan a ser los estudiantes de la forma A los que obtienen mejores puntuaciones. Y entre los percentiles más altos (entre 80 y 95) la CSMM y CSMS producen distancias negativa, al contrario que el resto de metodologías.

En A2 entre los percentiles 1 y 20 aproximadamente las distancias obtenidas con la metodología de CSMM son positivas y en torno a 0,1, las más altas en comparación con el resto. Entre el percentil 50 y 70 todas las distancias se encuentran por debajo de 0,1, las metodologías CS y CF casi no producen distancias entre las formas, el resto son negativas, es decir, con valores superiores en el rasgo de los estudiantes que contestaron a la forma B. En esta aplicación el patrón de distancias entre CS y CF y muy parecido, el de la CC es similar al de las dos anteriores pero con valores negativos, por tanto, estimando ligeramente una mayor puntuación para los estudiantes de la forma B, similar a la CSSL y CSH.

En A3 las distancias solo superan el 0,1 en los percentiles más extremos de la distribución. El patrón es bastante similar entre las siete metodologías de calibración.

Finalmente, en A4, entre los percentiles 5 y 10 la calibración CSMS produce distancias positivas superiores a 0,1. Entre el percentil 45 y 80 la CF muestra un patrón distinto al resto, con distancias positivas mientras que el resto son negativas o casi inexistentes

La siguiente tabla (Tabla VII.8) analiza las distancias en los 7 percentiles concretos mencionados en el apartado de metodología (5, 10, 25, 50, 75, 90 y 95). Se han marcado en negrita las distancias iguales o superiores a ±0,1. Conviene recordar que valores negativos indican una mejor puntuación en el rasgo para los estudiantes que respondieron a la Forma B y valores positivos lo opuesto.

	Aplicación 1	**CS**	**CC**	**CF**	**CSMM**	**CSMS**	**CSSL**	**CSH**
Percentiles	5	-0,04	-0,09	-0,07	-0,01	-0,01	**-0,16**	**-0,12**
	10	0	-0,03	-0,01	0,02	0,02	**-0,10**	-0,07
	25	-0,01	-0,04	-0,02	-0,01	-0,01	-0,08	-0,06
	50	0,01	-0,01	0	0	0	-0,03	-0,01
	75	0,02	0,02	0,02	-0,02	-0,02	0,01	0,02
	90	0,03	0,03	0,03	-0,03	-0,03	0,05	0,05
	95	0,04	0,05	0,03	-0,02	-0,02	0,07	0,08
	Aplicación 2	**CS**	**CC**	**CF**	**CSMM**	**CSMS**	**CSSL**	**CSH**
Percentiles	5	0	-0,04	-0,02	**0,11**	**-0,10**	**-0,11**	-0,08
	10	0,05	0	0,05	**0,14**	-0,03	-0,06	-0,02
	25	0,01	-0,03	0,01	0,05	-0,06	-0,09	-0,07
	50	-0,03	**-0,14**	-0,04	-0,03	-0,07	**-0,12**	**-0,10**
	75	0,03	-0,04	0,01	-0,01	0	-0,06	-0,04
	90	-0,01	-0,05	-0,01	**-0,10**	-0,02	-0,09	-0,08
	95	0	-0,06	-0,01	**-0,11**	-0,01	-0,07	-0,07
	Aplicación 3	**CS**	**CC**	**CF**	**CSMM**	**CSMS**	**CSSL**	**CSH**
Percentiles	5	0,02	0,05	0,05	0,03	0,05	0,08	**0,11**
	10	-0,03	0,03	0	-0,03	-0,02	0,03	0,05
	25	0	0,05	0,02	0	0	0,06	0,06
	50	-0,01	0,06	0,01	-0,02	-0,02	0,05	0,04
	75	-0,02	0,02	-0,01	-0,04	-0,04	0,04	0,01
	90	0	0,03	0	-0,03	-0,04	0,05	0,01
	95	0,06	0,07	0,03	0,02	0,02	**0,11**	0,07

	Aplicación 4	CS	CC	CF	CSMM	CSMS	CSSL	CSH
Percentiles	5	-0,02	0,02	0,01	0	**0,11**	0,03	0,01
	10	**0,10**	**0,10**	0,08	**0,12**	**0,21**	**0,15**	**0,13**
	25	-0,01	0,02	0,08	0,01	0,06	0,03	0,02
	50	-0,03	0	0,07	-0,01	0,02	0,01	0,01
	75	-0,04	-0,02	0,04	-0,01	-0,02	0	0
	90	0,05	0,07	**0,11**	0,08	0,04	0,09	0,09
	95	0,04	0,09	0,05	0,06	0,02	0,08	0,08

Tabla VII.8. Distancias Horizontales en siete puntos de la distribución (Percentiles) en función de la metodología de calibración horizontal empleada, en cada una de las aplicaciones.

Como puede verse en la tabla VII.8, no existe tendencia definida en las distancias horizontales estimadas, es decir, no se acumulan distancias a favor de alguna de las dos formas en los distintos tramos de la distribución analizada. Si analizados cada aplicación por separado se puede mencionar que en la primera aplicación, tomando la CS como referencia, son la CC y CF las que producen resultados más parecidos. En la segunda aplicación, solo la CF sigue un patrón similar a la CS. En la tercera, aunque las distancias son similares entre las distintas metodologías, la CSH muestra unas mayores distancias con respecto a la CS en los percentiles por debajo de la mediana (percentil 50). En la cuarta aplicación, CC y CS son bastantes similares pero en el extremo superior de la distribución del rasgo (P90 y P95) es CSMS la que muestra unas distancias similares a CS.

Las medias, en términos absolutos, de esas distancias entre las formas de cada aplicación y para cada uno de los métodos de calibración están resumidas en la Tabla VII.9:

		CS	CC	CF	CSMM	CSMS	CSH	CSSL
$\lvert\overline{\Delta P}\rvert$	**A1**	0,028	0,034	0,029	0,026	0,026	0,059	0,050
	A2	0,022	0,062	0,023	0,061	0,051	0,093	0,071
	A3	0,019	0,043	0,019	0,026	0,027	0,055	0,049
	A4	0,034	0,035	0,034	0,035	0,060	0,048	0,042

Tabla VII.9. Distancias Horizontales Medias en función de la metodología de calibración horizontal y la aplicación.

Si se observan las distancias medias, el dato destacable es que todas son inferiores a 0,01. La mayor es la producida por la metodología CH en la segunda aplicación, con un valor por encima de 0,09. La metodología CS y CF produce distancias inferiores al 0,04 en todas las aplicaciones. Pero todas las distancias

horizontales medias entre las dos formas del instrumento de medida, producidas por los procedimientos de equiparación horizontal, son inferiores a 0,1.

VII.3.2 Problema 2. Comparación de procedimientos para el anclaje vertical

La elaboración de una escala vertical es esencial para analizar el VA de las escuelas empleando un diseño longitudinal de medida del rendimiento de los estudiantes y se utiliza en muchas de las evaluaciones que implementan este tipo de modelos (Bryk, Thum, Easton & Luppescu, 1998; Ponisciak & Bryk, 2005; Zvoch & Stevens, 2006; Briggs, Weeks & Wiley, 2008; Castro, Ruíz & López, 2009).

Las decisiones tomadas en diferentes aspectos metodológicos relacionados con la metodología de anclaje vertical pueden afectar a las estimaciones finales que componen la escala vertical y, por tanto, a las estimaciones de VA que producen distintos modelos de VA (Briggs, Weeks & Wiley, 2008; Briggs & Weeks, 2009). Con los datos procedentes de los ocho instrumentos aplicados en las cuatro aplicaciones se han elaborado varios modelos TRI de tres parámetros que difieren en el tipo de calibración de los parámetros y la forma de estimar la habilidad de los sujetos:

A. Método de calibración vertical:
   - A.1 Calibración Conjunta
   - A.2 Calibración por separado (empleando los cuatro tipos de transformación de los parámetros)
   - A.3 Calibración fija

B. Método de estimación de la habilidad:
   - B.1 Máxima Verosimilitud (MV)
   - B.2 Empírica a Posteriori (EAP). Utilizando la distribución empírica estimada en la fase de calibración como distribución a priori (con el comando idist=3 en la sección SCORE de BILOGMG).
   - B.3 Máxima a Posteriori (MAP)

VII.3.2.1 Análisis desde la TCT

De la misma forma que en el primer objetivo, antes de comenzar con la comparación de los resultados de los modelos TRI, conviene observar la dificultad de los ítems de anclaje calculado bajo los supuestos de la TCT (porcentaje de respuestas correctas) que reflejan la tendencia de crecimiento entre aplicaciones consecutivas. Estos resultados se muestran en la siguiente tabla (Tabla VII.10).

	Dificultad TCT							
	A1	**A2**		**A2**	**A3**		**A3**	**A4**
MO5B20_J6A21	0,795	0,824	**MJ6B27_N6A20**	0,587	0,700	**MN6B30_J7A1**	0,692	0,740
MO5B21_J6A22	0,437	0,501	**MJ6B28_N6A21**	0,566	0,826	**MN6B31_J7A2**	0,913	0,946
MO5B22_J6A23	0,566	0,653	**MJ6B29_N6A22**	0,577	0,776	**MN6B32_J7A3**	0,741	0,813
MO5B23_J6A24	0,458	0,507	**MJ6B30_N6A23**	0,464	0,691	**MN6B33_J7A4**	0,304	0,274
MO5B24_J6A25	0,702	0,763	**MJ6B31_N6A24**	0,323	0,329	**MN6B34_J7A5**	0,373	0,361
MO5B25_J6A26	0,661	0,754	**MJ6B32_N6A25**	0,519	0,780	**MN6B35_J7A6**	0,814	0,864
MO5B26_J6A27	0,211	0,262	**MJ6B33_N6A26**	0,411	0,495	**MN6B36_J7A7**	0,404	0,503
MO5B27_J6A28	0,431	0,543	**MJ6B34_N6A27**	0,466	0,711	**MN6B37_J7A8**	0,356	0,654
MO5B28_J6A29	0,505	0,639	**MJ6B35_N6A28**	0,486	0,554	**MN6B38_J7A9**	0,270	0,296
MO5B29_J6A30	0,145	0,169	**MJ6B36_N6A29**	0,484	0,720	**MN6B39_J7A10**	0,809	0,876
MO5A20_J6B19	0,863	0,900	**MJ6A29_N6B20**	0,659	0,809	**MN6AB19_J7B1**	0,294	0,439
MO5A21_J6B20	0,369	**0,303**	**MJ6A30_N6B21**	0,542	0,728	**MN6A30_J7B2**	0,771	0,764
MO5A22_J6B21	0,263	0,351	**MJ6A31_N6B22**	0,409	0,543	**MN6A30_J7B4**	0,715	0,719
MO5A23_J6B22	0,664	0,779	**MJ6A32_N6B23**	0,567	0,665	**MN6A30_J7B5**	0,511	0,574
MO5A24_J6B23	0,618	0,672	**MJ6A33_N6B24**	0,342	0,591	**MN6A30_J7B6**	0,631	0,728
MO5A26_J6B25	0,285	0,314	**MJ6A35_N6B26**	0,475	0,555	**MN6A30_J7B8**	0,215	0,316
MO5A27_J6B26	0,484	0,581	**MJ6A36_N6B27**	0,639	0,804	**MN6A30_J7B9**	0,422	0,537
			MJ6A37_N6B28	0,607	0,718	**MN6A30_J7B10**	0,801	0,864
			MJ6A38_N6B29	0,677	0,853			
Promedio	**0,497**	**0,560**		**0,516**	**0,676**		**0,558**	**0,626**

Tabla VII.10. Índices de dificultad TCT de los ítems de anclaje entre aplicaciones

Los resultados desde la TCT señalan una tendencia lógica con el aumento de la proporción de respuestas correctas entre los ítems de anclaje de aplicaciones consecutivas. Únicamente un ítem parece ser más difícil en la A2 que en la primera, en esta un 6,6% menos de estudiantes aciertan el ítem. En términos promedios el mayor salto se produce entre A2 y A3 con un 16% más de estudiantes que aciertan los ítems de anclaje, mientras que entre A1 y A2 y entre A3 y A4 aumenta un 6,3% y 6,8% respectivamente. Este es aproximadamente 2,5 veces más en ese cambio en la proporción de aciertos. Por tanto parece lógico que los estudiantes tengan un mayor incremento entre esas dos tomas de datos.

Este aumento de la facilidad de los ítems comunes en el paso de la segunda a la tercera aplicación, que corresponde con el periodo de verano y el primer mes de cursos, puede deberse a la conjunción de dos aspectos fundamentales. Por un lado, el periodo entre aplicaciones es más corto (5 meses) y, por otro, el repaso inicial que se produce durante las primeras semanas de curso. Estos son los factores podrían facilitar el recuerdo y la resolución de estos ítems.

Esta circunstancia, es una de las causas que justifica la realización de uno de los objetivos del segundo estudio empírico. Es decir, si el modelo de crecimiento empleado debe considerar la misma distancia entre aplicaciones o se debe ajustar a la distancia real entre aplicaciones. También es posible que sea necesario paliar ese efecto de recuerdo entre la segunda y tercer aplicación ponderando el crecimiento entre esas etapas mediante algún parámetro que tenga en cuenta ese cambio en la dificultad.

VII.3.2.2 Análisis desde la TRI

Los resultados de la estimación de los modelos de Teoría Respuesta al Ítem se presentan agrupados en tres apartados. En primer lugar los datos de crecimiento y variabilidad donde se analizan las medias y desviaciones típicas estimadas con los distintos modelos. En segundo lugar, los tamaño del efecto calculado que señalan el cambio de aplicación a aplicación ponderado por la variabilidad de la escala. Y, finalmente, las distancias horizontales calculadas empleando los percentiles.

A. Crecimiento y variabilidad

En primer lugar, conviene analizar las medias y desviaciones típicas producidas por las diferentes metodologías de calibración y de calificación de los sujetos (Tabla VII.11).

	A1			A2			A3			A4		
		Media	DT		Media	DT.		Media	DT		Media	DT.
EAP	CC	-0,003	0,910	CC	0,428	0,928	CC	1,216	0,725	CC	1,553	0,831
	CF	-0,001	0,917	CF	0,267	0,898	CF	0,611	0,846	CF	0,631	0,901
				CSSL	0,449	0,859	CSSL	1,221	0,656	CSSL	1,531	0,643
	CS	-0,001	0,917	CSH	0,429	0,835	CSH	1,161	0,619	CSH	1,457	0,630
				CSMM	0,466	0,822	CSMM	1,232	0,739	CSMM	1,503	0,740
				CSMS	0,474	0,899	CSMS	1,259	0,627	CSMS	1,484	0,596
		Media	DT		Media	DT.		Media	DT		Media	DT.
MAP	CC	0,051	0,871	CC	0,482	0,816	CC	1,256	0,673	CC	1,615	0,726
	CF	0,052	0,884	CF	0,289	0,867	CF	0,586	0,839	CF	0,589	0,886
				CSSL	0,500	0,816	CSSL	1,247	0,636	CSSL	1,559	0,630
	CS	0,052	0,884	CSH	0,480	0,787	CSH	1,179	0,582	CSH	1,492	0,579
				CSMM	0,513	0,783	CSMM	1,260	0,714	CSMM	1,529	0,717
				CSMS	0,526	0,855	CSMS	1,283	0,606	CSMS	1,506	0,577
		Media	DT		Media	DT.		Media	DT		Media	DT.
MVL	CC	-0,020	1,172	CC	0,382	1,154	CC	1,216	0,885	CC	1,557	0,971
	CF	-0,021	1,201	CF	0,222	1,213	CF	0,599	1,081	CF	0,618	1,185
				CSSL	0,403	1,487	CSSL	1,205	0,821	CSSL	1,517	0,821
	CS	-0,021	1,201	CSH	0,385	1,112	CSH	1,157	0,791	CSH	1,464	0,765
				CSMM	0,422	1,100	CSMM	1,220	0,932	CSMM	1,493	0,944
				CSMS	0,426	1,200	CSMS	1,249	0,790	CSMS	1,477	0,760

Tabla VII.11. Medias y desviaciones típicas del rasgo producidas por los diferentes métodos de calibración y calificación.

Por un lado, si se pone la atención en la metodología de anclaje vertical, la CF tiene las medias más bajas en las cuatro aplicaciones, independientemente del tipo de procedimiento empleado en la calificación. Además de no producirse casi cambio entre las dos últimas aplicaciones. El resto de métodos de anclaje muestran patrones similares a lo largo de las cuatro aplicaciones.

De los distintos tipos de calibración por separado, la transformación media/sigma (CSMS) tiene medias ligeramente superiores en la segunda y tercera aplicación. La CSSL las obtiene en la última aplicación.

Puede observarse también ese incremento mayor que se produce entre la segunda y tercera toma de datos, que ya se esperaba al observar la dificultad TCT de los ítems comunes.

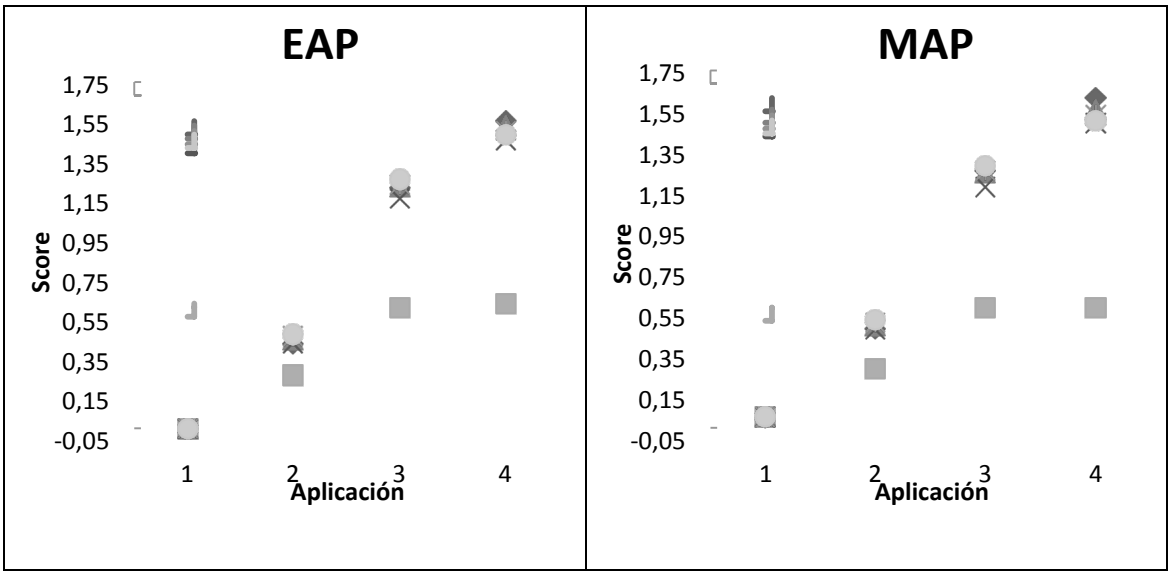
Por otro lado, si estudiamos los distintos métodos de calificación, el MAP estima medias algo mayores que el resto, independientemente del tipo de calibración, excepto en la CF donde el método de MVL y EAP produce mayores medias que en la MAP en la tercera y cuarta aplicación. No obstante los patrones son bastante similares.

Respecto a la dispersión, los métodos de calibración separada tienen, de forma general, desviaciones típicas más bajas en las distintas aplicaciones y tipos de calificación. La CF tiene menos dispersión cuando se emplea el tipo bayesianos (EAP y MAP) de estimación del rasgo de los sujetos.

Si se comparan las metodologías de calificación, la estimación MAP produce puntuaciones con menores desviaciones típicas, seguida a poca distancia por EAP y, finalmente el método MVL donde sí se produce un cambio destacable en la varianza de las puntuaciones.

Conviene mencionar que la reducción de la dispersión a medida que avanzan las aplicaciones se debe a las características específicas de la muestra y no a una influencia concreta del tipo de anclaje vertical o de estimación del ras. El motivo es una reducción de la muestra entre la segunda y tercera aplicación.

A continuación se incluye un gráfico por tipo de estimación del rasgo que incorpora las medias de las cuatro aplicaciones en función del tipo de anclaje utilizado.



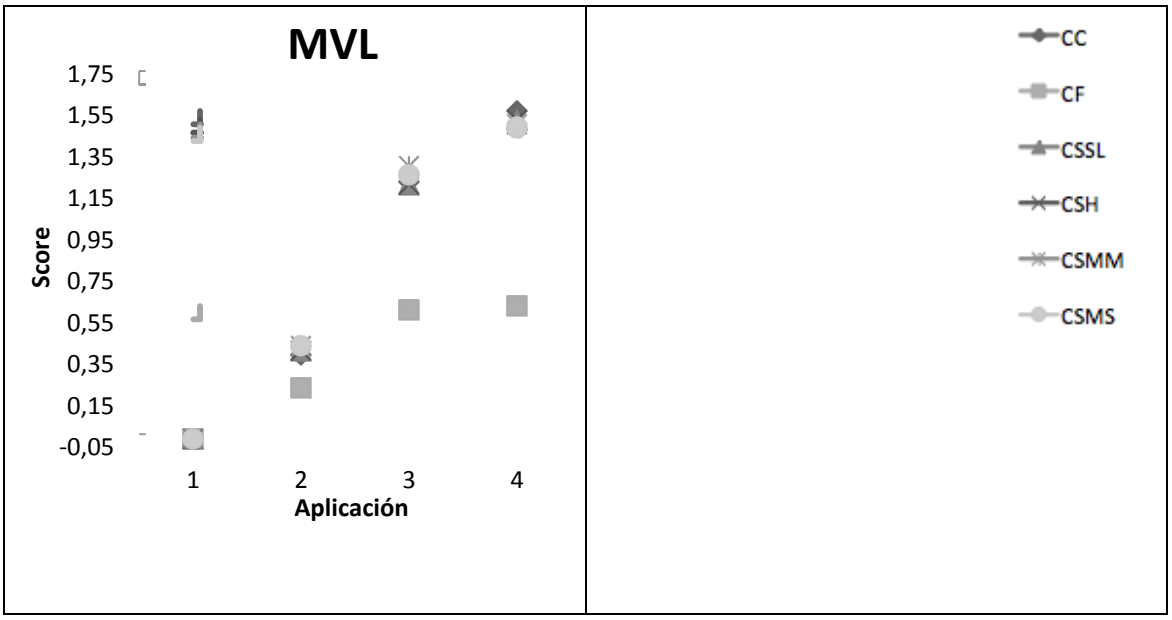


Gráfico VII.8. Puntuaciones medias en las cuatro aplicaciones estimadas, en función del método de calibración. Cada gráfico muestra un método de estimación del rasgo distinto.

En los tres gráficos, la CF muestra el crecimiento más suave entre aplicaciones, casi sin cambio en la última aplicación. Este tipo de anclaje muestra problemas, desmarcándose del resto de metodologías. Puede ser debido a que utilizando la CF el error de equiparación acumulado produzca una distorsión de los datos ya que ni el crecimiento es tan brusco entre A2 y A3, ni hay cambio en las dos últimas.

La CC obtiene la puntuación media mayor en la última aplicación. En cambio la CSH tiene la media más baja en la 3ª aplicación con las metodologías EAP y MAP, sin embargo con MLV es la más alta.

Por tanto, los resultados producidos por la CC parecen estar determinados, en menor medida, por el tipo de estimación del rasgo ya que muestra un patrón similar en las tres metodologías.

Para ayudar a cuantificar el cambio entre ocasiones de media, se incluye en la siguiente tabla (Tabla VII.12) las diferencias entre las medias de aplicaciones consecutivas:

	A1-A2		A2-A3		A3-A4	
EAP	CC	0,430	CC	0,789	CC	0,336
	CF	0,268	CF	0,344	CF	0,020
	CSSL	0,450	CSSL	0,772	CSSL	0,310
	CSH	0,430	CSH	0,732	CSH	0,297
	CSMM	0,467	CSMM	0,766	CSMM	0,271
	CSMS	0,475	CSMS	0,785	CSMS	0,225
	A1-A2		**A2-A3**		**A3-A4**	
MAP	CC	0,430	CC	0,774	CC	0,359
	CF	0,237	CF	0,297	CF	0,003
	CSSL	0,448	CSSL	0,748	CSSL	0,311
	CSH	0,428	CSH	0,699	CSH	0,313
	CSMM	0,461	CSMM	0,747	CSMM	0,269
	CSMS	0,474	CSMS	0,758	CSMS	0,222
	A1-A2		**A2-A3**		**A3-A4**	
MVL	CC	0,402	CC	0,834	CC	0,341
	CF	0,243	CF	0,377	CF	0,019
	CSSL	0,424	CSSL	0,802	CSSL	0,312
	CSH	0,406	CSH	0,772	CSH	0,307
	CSMM	0,443	CSMM	0,798	CSMM	0,273
	CSMS	0,447	CSMS	0,823	CSMS	0,228

Tabla VII.12. Diferencia de medias entre aplicaciones consecutivas, en función del método de calibración y de calificación.

Entre la primera y segunda aplicación la metodología CSMS produce un incremento ligeramente superior al resto, los valores oscilan alrededor de 0,5, medio punto en el rasgo. A excepción de la CF que destaca del mostrando la diferencia más baja, prácticamente la mida (0,25). El patrón es similar entre procedimientos de estimación del rasgo.

El segundo cambio, entre la segunda y tercera toma de datos, es el más alto en todas las metodologías debido a los aspectos ya mencionados, aproximadamente 0,85 en MVL y 0,75 en los procedimientos bayesianos. Aun así, la CF, con la metodología de estimación MAP, muestra una diferencia similar a la que se produce entre las dos primeras aplicaciones. El procedimiento de estimación MVL, produce el mayor cambio en todas las metodologías, llegando al doble del que se producía entre las primeras tomas de datos.

En la tercera aplicación, las diferencias toman valores más bajos que entre las dos primeras, oscilan entre 0,2 y 0,3. Casi inexistente en la CF y con el valor más alto para la CC, alcanzando el 0,35.

B. Tamaños del efecto

El tamaño del efecto cuantifica las diferencias entre aplicaciones una vez estandarizas. Los resultados varían ligeramente respecto a las diferencias de medias como refleja la Tabla VII.13.

	A1-A2		A2-A3		A3-A4	
EAP	CC	0,662	CC	1,340	CC	0,610
	CF	0,417	CF	0,558	CF	0,033
	CSSL	0,716	CSSL	1,429	CSSL	0,674
	CSH	0,693	CSH	1,408	CSH	0,672
	CSMM	0,758	CSMM	1,385	CSMM	0,518
	CSMS	0,740	CSMS	1,433	CSMS	0,520
	A1-A2		**A2-A3**		**A3-A4**	
MAP	CC	0,721	CC	1,463	CC	0,725
	CF	0,383	CF	0,493	CF	0,005
	CSSL	0,744	CSSL	1,445	CSSL	0,695
	CSH	0,722	CSH	1,428	CSH	0,762
	CSMM	0,781	CSMM	1,409	CSMM	0,531
	CSMS	0,770	CSMS	1,445	CSMS	0,531
	A1-A2		**A2-A3**		**A3-A4**	
MVL	CC	0,488	CC	1,147	CC	0,520
	CF	0,285	CF	0,465	CF	0,023
	CSSL	0,444	CSSL	0,944	CSSL	0,537
	CSH	0,497	CSH	1,266	CSH	0,391
	CSMM	0,544	CSMM	1,204	CSMM	0,306
	CSMS	0,527	CSMS	1,146	CSMS	0,416

Tabla VII.13. Tamaños del efecto en función del método de calibración y calificación.

El patrón es parecido al mostrado por las medias, un crecimiento similar entre la primera y segunda aplicación y la tercera y cuarta toma de datos. Los valores oscilan en torno al 0,5 en el procedimiento de MVL y 0,75 en los bayesianos. Y un mayor cambio entre la 2º y 3ª aplicación, alcanzando el 1,4 en MAP y EAP y 1,2 en MLV

El procedimiento MAP obtiene los mayores tamaños del efecto en las tres aplicaciones. Además, con esta metodología, el cambio entre las dos últimas aplicaciones es ligeramente superior a la de las dos primeras con las metodologías CC y CSH.

No se encuentran patrones decrecientes en los tamaños del efecto a medida que aumenta el curso como los encontrados por Jungnam (2007). Parece que si existe cierta disminución si se compara el cambio entre la 1ª y 2ª con la 3ª y 4º pero se rompe la tendencia con la CC y CSH.

Para comprobar si el tipo de calibración o de calificación puede afectar de forma distinta al crecimiento en tramos diferentes de la distribución del rasgo se incluyen las distancias horizontales entre aplicaciones consecutivas.

C. Distancias horizontales

		A1-A2	CC	CF	CSMM	CSMS	CSSL	CSH
EAP	Percentiles	5	0,444	0,257	0,614	0,495	0,535	0,555
		10	0,302	0,219	0,532	0,437	0,465	0,478
		25	0,430	0,261	0,506	0,459	0,462	0,459
		50	0,475	0,311	0,494	0,503	0,477	0,456
		75	0,388	0,269	0,398	0,459	0,406	0,369
		90	0,364	0,247	0,349	0,455	0,379	0,328
		95	0,364	0,221	0,309	0,441	0,352	0,292
		A2-A3	**CC**	**CF**	**CSMM**	**CSMS**	**CSSL**	**CSH**
	Percentiles	5	1,054	0,329	0,851	1,192	1,063	1,047
		10	1,146	0,468	0,925	1,193	1,089	1,064
		25	0,944	0,451	0,859	1,008	0,948	0,915
		50	0,804	0,359	0,763	0,779	0,767	0,726
		75	0,687	0,285	0,713	0,602	0,637	0,587
		90	0,561	0,222	0,634	0,414	0,489	0,433
		95	0,468	0,161	0,597	0,314	0,413	0,352
		A3-A4	**CC**	**CF**	**CSMM**	**CSMS**	**CSSL**	**CSH**
	Percentiles	5	-0,044	-0,162	0,170	0,200	0,248	0,194
		10	0,324	-0,039	0,285	0,278	0,341	0,296
		25	0,399	0,013	0,312	0,280	0,355	0,325
		50	0,412	0,041	0,296	0,245	0,331	0,318
		75	0,360	0,063	0,255	0,190	0,286	0,290
		90	0,366	0,084	0,257	0,174	0,279	0,297
		95	0,372	0,128	0,267	0,171	0,283	0,309
		A1-A2	**CC**	**CF**	**CSMM**	**CSMS**	**CSSL**	**CSH**
MAP	Percentiles	5	0,518	0,243	0,633	0,527	0,565	0,592
		10	0,457	0,212	0,559	0,475	0,501	0,520
		25	0,445	0,226	0,504	0,464	0,466	0,467
		50	0,446	0,264	0,480	0,492	0,466	0,446
		75	0,396	0,238	0,399	0,462	0,408	0,368
		90	0,355	0,214	0,330	0,438	0,360	0,303
		95	0,345	0,183	0,283	0,416	0,325	0,257
		A2-A3	**CC**	**CF**	**CSMM**	**CSMS**	**CSSL**	**CSH**
	Percentiles	5	0,972	0,253	0,817	1,129	1,005	1,000
		10	0,983	0,360	0,863	1,109	1,009	0,992
		25	0,904	0,374	0,820	0,959	0,900	0,868
		50	0,789	0,328	0,750	0,760	0,750	0,702
		75	0,677	0,273	0,702	0,587	0,625	0,559
		90	0,568	0,204	0,631	0,409	0,488	0,409
		95	0,492	0,166	0,600	0,315	0,418	0,331

		A3-A4	CC	CF	CSMM	CSMS	CSSL	CSH
	Percentiles	5	0,156	-0,125	0,193	0,214	0,257	0,260
		10	0,292	-0,056	0,270	0,264	0,322	0,320
		25	0,363	-0,021	0,292	0,263	0,337	0,335
		50	0,394	0,009	0,288	0,238	0,328	0,328
		75	0,386	0,029	0,258	0,192	0,296	0,300
		90	0,394	0,080	0,273	0,186	0,305	0,309
		95	0,395	0,116	0,282	0,183	0,311	0,315
		A1-A2	**CC**	**CF**	**CSMM**	**CSMS**	**CSSL**	**CSH**
	Percentiles	5	0,308	0,182	0,541	0,279	0,437	0,483
		10	0,432	0,270	0,592	0,423	0,517	0,541
		25	0,468	0,299	0,537	0,464	0,492	0,494
		50	0,448	0,299	0,480	0,490	0,464	0,443
		75	0,369	0,227	0,360	0,448	0,371	0,331
		90	0,314	0,187	0,277	0,429	0,311	0,253
		95	0,331	0,169	0,256	0,455	0,305	0,236
		A2-A3	**CC**	**CF**	**CSMM**	**CSMS**	**CSSL**	**CSH**
MVL	Percentiles	5	1,435	0,680	1,223	1,665	1,495	1,514
		10	1,159	0,564	0,997	1,308	1,186	1,194
		25	0,945	0,448	0,846	1,007	0,939	0,929
		50	0,788	0,359	0,749	0,757	0,743	0,712
		75	0,663	0,288	0,705	0,571	0,606	0,553
		90	0,543	0,224	0,638	0,380	0,458	0,388
		95	0,457	0,188	0,592	0,259	0,363	0,283
		A3-A4	**CC**	**CF**	**CSMM**	**CSMS**	**CSSL**	**CSH**
	Percentiles	5	0,087	-0,175	0,177	0,217	0,248	0,243
		10	0,298	-0,062	0,284	0,283	0,335	0,325
		25	0,369	-0,024	0,298	0,271	0,340	0,332
		50	0,399	0,036	0,292	0,241	0,328	0,322
		75	0,384	0,099	0,271	0,199	0,303	0,300
		90	0,381	0,140	0,278	0,184	0,303	0,302
		95	0,358	0,159	0,294	0,184	0,313	0,312

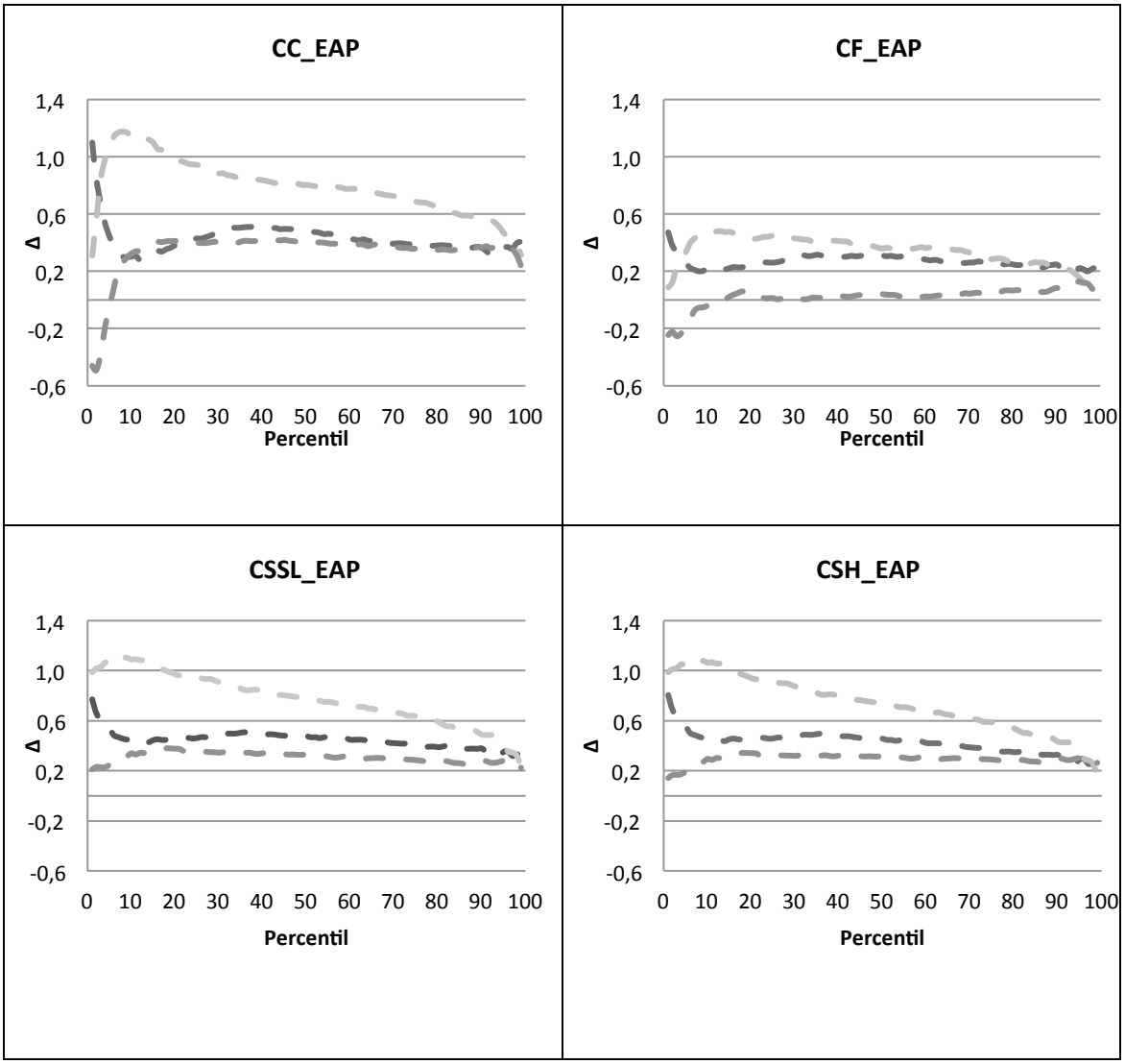
Tabla VII.14. Distancias Horizontales en 7 puntos específicos (percentiles) de la distribución, en función de la metodología de calibración vertical y el método de calificación.

Existe cierta tendencia de disminución de la distancia a medida que se asciende a lo largo de la distribución. La tendencia se rompe en la distancia entre las dos últimas aplicaciones con la CC. La CF, con los procedimientos de estimación EAP y MVL, tienen distancias negativas en el extremo inferior de la distribución (percentiles 5 y 10), también ocurre, aunque con una distancia inferior a 0,05, en el percentil 5 de la CF.

Los distintos tipos de calibración por separado (CSMM, CSMS, CSSL y CSH) muestran un cambio mucho mayor, entre la 1ª y 2ª aplicación, con la estimación MVL, siendo los sujetos que más crecen.

La CC parece se tener el patrón más estable a lo largo de toda la distribución del rasgo, suavizando la tendencia a la diminución de la distancia en la parte superior del rasgo e incluso cambiando la tendencia, principalmente con el método de estimación MAP.

A continuación se incluyen los gráficos con las distancias horizontales calculadas en los 99 percentiles de la distribución (Gráfico VII.9, Gráfico VII.10, Gráfico VII.11).



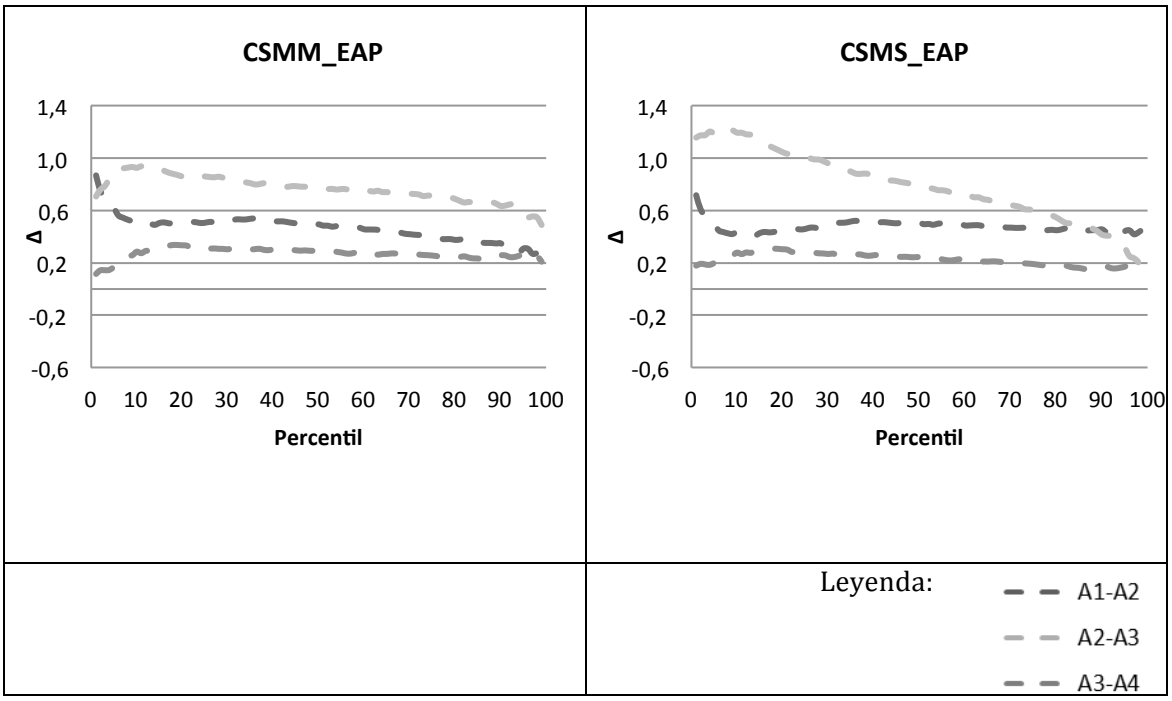
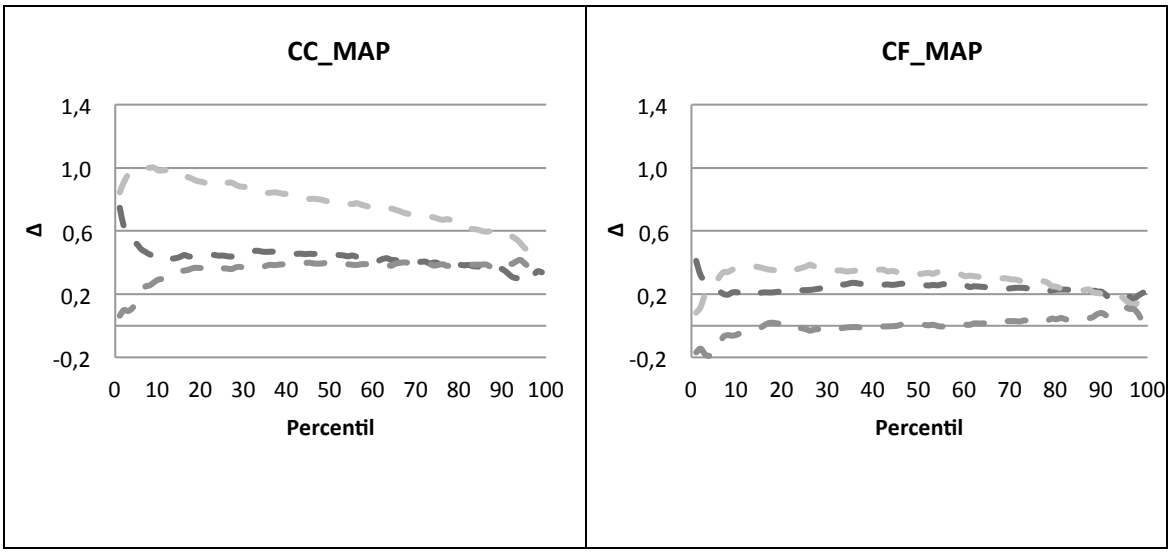


Gráfico VII.9. Distancias horizontales en los 99 percentiles, en función de la metodología de calibración horizontal, en cada una de las aplicaciones por separado. Método de Estimación EAP



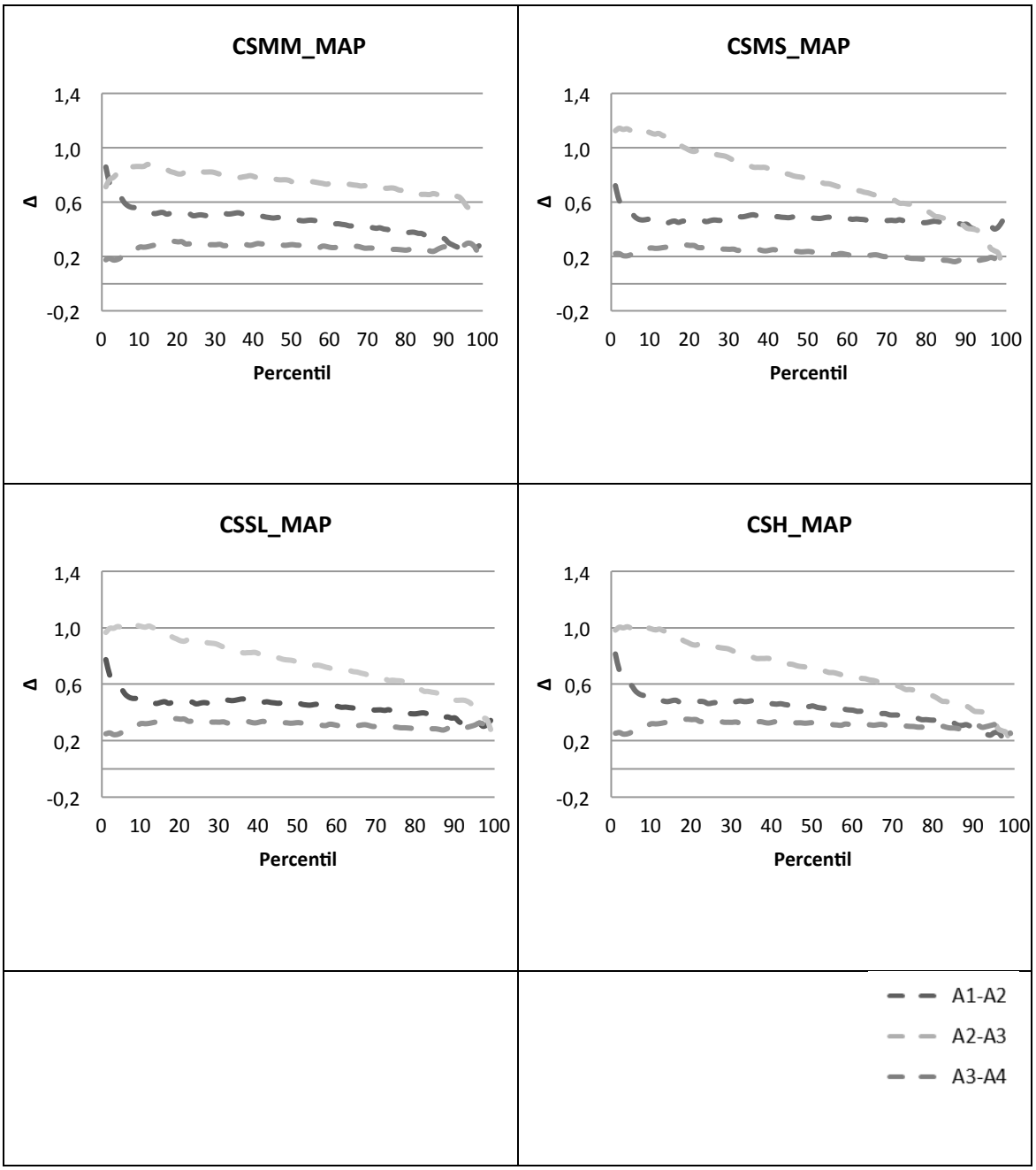
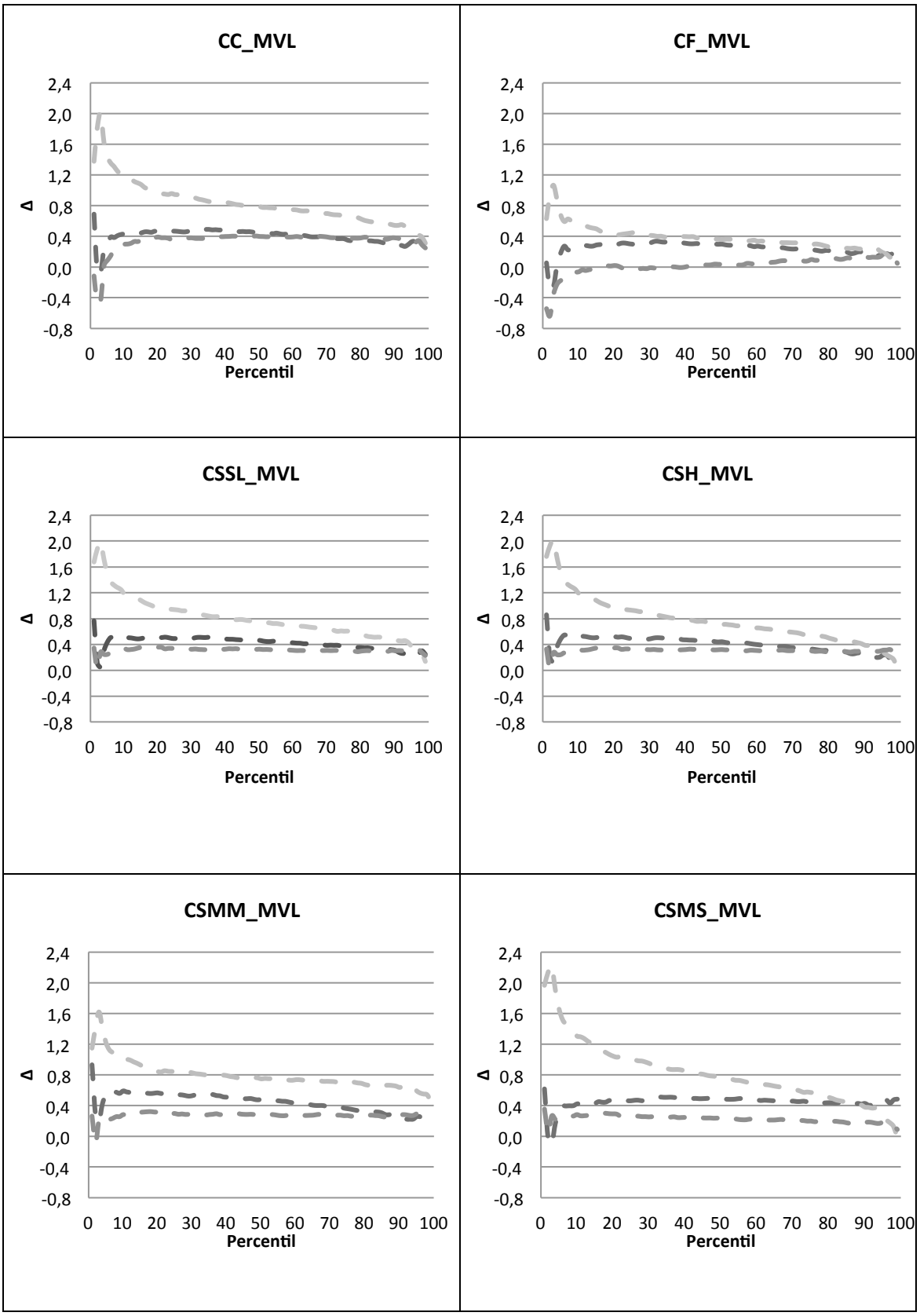


Gráfico VII.10. Distancias horizontales en los 99 percentiles, en función de la metodología de calibración horizontal, en cada una de las aplicaciones por separado. Método de Estimación MAP


CC_MVL
CF_MVL
CSSL_MVL
CSH_MVL
CSMM_MVL
CSMS_MVL
Δ
Percentil
2,4
2,0
1,6
1,2
0,8
0,4
0,0
-0,4
-0,8
0 10 20 30 40 50 60 70 80 90 100

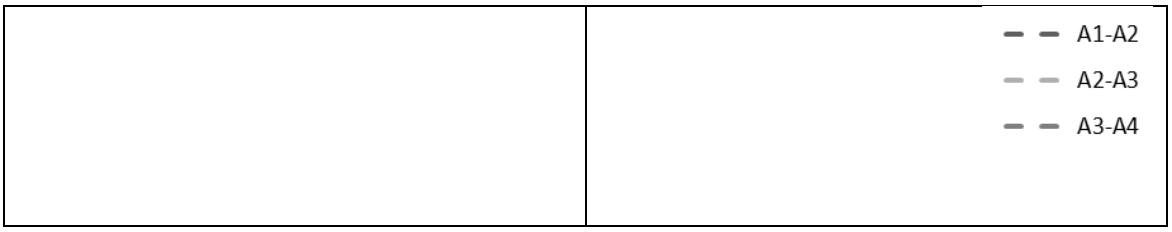


Gráfico VII.11. Distancias horizontales en los 99 percentiles, en función de la metodología de calibración horizontal, en cada una de las aplicaciones por separado. Método de Estimación MVL

Finalmente, para resumir la información de las distancias horizontales, la siguiente tabla incorpora las medias calculadas en cada una de las metodologías de anclaje y de calificación.

		CC	CF	CSMM	CSMS	CSSL	CSH
EAP	A1-A2	0,426	0,266	0,465	0,473	0,447	0,427
	A2-A3	0,795	0,347	0,768	0,788	0,775	0,734
	A3-A4	0,341	0,021	0,272	0,226	0,310	0,297
MAP	A1-A2	0,428	0,236	0,460	0,472	0,446	0,426
	A2-A3	0,776	0,30,00	0,749	0,760	0,749	0,701
	A3-A4	0,360	0,0,003	0,270	0,223	0,312	0,313
MVL	A1-A2	0,401	0,244	0,443	0,446	0,423	0,406
	A2-A3	0,835	0,383	0,799	0,823	0,803	0,772
	A3-A4	0,343	0,019	0,274	0,228	0,313	0,308

Tabla VII.15. Distancias Horizontales medias en función de la metodología de anclaje vertical y de estimación del rasgo.

La metodología MAP produce las mayores distancias entre las dos primeras aplicaciones y las dos últimas, siendo la CC la que obtiene mayores distancias. La CF, como siempre, sique un patrón distinto al resto. En cambio, la estimación EAP produce el mayor cambio entre A2 y A3, también es la CC la que obtiene mayor valor.

En definitiva, la metodología CC muestra los resultados más estables en comparación con el resto de metodologías de equiparación vertical. El procedimiento de CF parece tener problemas a medida que se avanza en las aplicaciones, sin llegar a producir crecimiento entre A3 y A4. Los cuatro tipos de calibración por separado muestran una tendencia a la disminución del crecimiento a media que aumenta el rasgo, esta tendencia se suaviza en la CC.

Capítulo VIII: Comparación empírica de modelos de Valor Añadido: tiempo, ocasiones de medida y relación entre estatus inicial y crecimiento.

En el Capítulo V de esta tesis se analizan las principales aproximaciones para llevar a cabo los análisis del Valor Añadido (VA en adelante). Estos análisis de denominan Modelos de Valor Añadido[79] (MVA en adelante). Elegir una determinada perspectiva u otra y la toma de decisiones en diferentes cuestiones metodológicas que implican los distintos análisis puede afectar a los resultados (McCaffrey, Lockwood, Koretz & Hamilton, 2003; Doran, 2003; Lockwood, Louis & McCaffrey, 2003; McCaffrey, Koretz, Louis & Hamilton, 2004; Lockwood et al., 2007; Armein-Beardsley, 2008; Braun, Chudowsky & Koenig, 2010).

Otro gran volumen de estudios se encarga de analizar los diferentes MVA y que influencia tienen en las estimaciones finales de los efectos de las escuelas o los docentes (Ballou, Sanders & Wright, 2004; Tekwe, et al., 2004; McCaffrey, Koretz, Louis & Hamilton, 2004; Choi, Goldschmidt & Yamashiro, 2006; Sanders, 2006; Lockwood et al., 2007). No es posible extraer la conclusión de que exista un MVA único adecuado para todas las situaciones de evaluación del rendimiento de las escuelas, sino aquel que es más apropiado en un contexto determinado. Para evitar sesgos y obtener una buena medida del VA en educación, debe cumplir ciertos requisitos que pueden variar entre modelos y que dependerán de las decisiones tomadas durante la planificación y desarrollo de este tipo de estimaciones.

[79]La definición de los modelos de valor añadido se incluye en el apartado I.2.3.1.2

Entre los factores que determinan la utilización de un determinado modelo de VA se encuentran, por un lado, los objetivos que persigue la evaluación y los usos que tendrán los resultados obtenidos. Es decir, las estimaciones de VA de la escuela servirán para penalizar o incentivar su trayectoria (rendición de cuentas), se utilizarán para identificar posibles situaciones de riesgo e intentar implementar programas que mejoren dicha situación o simplemente tienen un carácter informativo o con una finalidad investigadora. Por otro lado, un segundo grupo de factores, se encuentra vinculado a la metodología utilizada para la elaboración del MVA. Los requisitos metodológicos de cada modelo pueden ser un elemento de información útil en el momento de decantarse por un uno de ellos. Las cuestiones metodológicas principales se resumen a continuación:

- Seleccionar el tipo de análisis adecuado para el análisis del cambio en aprendizaje con el objetivo de estimar esos efectos asociados a las escuelas. Las dos aproximaciones básicas para el análisis del VA son la utilización de una medida de ganancia como variable de resultados (ganancia bruta, residual o estimada) o una medida de crecimiento. Dentro de esta última perspectiva, el estudio del crecimiento puede abordarse desde el análisis multinivel, que estima un estatus inicial y una pendiente de crecimiento entre las distintas aplicaciones, o desde los modelos lineales mixtos como en el modelo EVAAS o el de persistencia[80].
- La elaboración final del MVA depende de las decisiones que se tomen en distintos aspectos como: el anidamiento de los datos o la posibilidad de que los elementos de análisis (estudiantes) cambien entre unidades de agrupación; considerar los efectos de las escuelas fijos o aleatorios y también si persisten o disminuyen con el tiempo; o el carácter univariante o multivariante de la variable dependiente.
- Elaborar una escala adecuada para poder medir el crecimiento a lo largo del tiempo. Las escalas verticales son una de las opciones para conseguir este requisito y la manera en que se construyen puede afectar a los valores estimados en el rasgo del estudiante y la

[80]Estos modelos se tratan con detalle en el capítulo V (Apartado V.2.3.2)

trayectoria de crecimiento. Un análisis de estos aspectos se ha llevado a cabo en el capítulo anterior.

- Seleccionar una medida adecuada de tiempo en el caso de utilizar modelos de crecimiento. El número de aplicaciones, la selección de un nivel de partida determinado y la distancia temporal entre las distintas tomas de datos son factores decisivos en el proceso. Por ejemplo, contar con medidas al inicio y final de curso en dos grados consecutivos conlleva analizar el periodo de verano. Esta característica puede determinar la elaboración de la variable tiempo.
- Utilizar modelos contextualizados. La introducción o no de covariables de contexto es otra decisión importante en el desarrollo de un determinado modelo de VA. El poder explicativo de estas variables puede verse reducido en los estudios longitudinales porque las diferentes puntuaciones de rendimiento del estudiante pueden ejercer ese control (Sanders, 2006; Lissitz, Doran, Schafer & Willhoft, 2006; Sanders & Wright, 2008).

Debido a la cantidad de factores que pueden variar cuando se opta por un tipo de análisis del VA u otro parece oportuno comprobar:

- La cantidad de varianza y el sesgo de las estimaciones producidas por los modelos. Los modelos estimados pueden variar en los errores estándar asociados a las puntuaciones de VA de los centros educativos. Si se pretende identificar escuelas significativamente diferentes de la media conviene que estos errores sean pequeños. Errores demasiado grandes pueden contribuir al sesgo de los resultados. Se recomienda utilizar los intervalos de confianza de las estimaciones para asegura las diferencias entre escuelas en los resultados obtenidos.
- Los datos perdidos. Es conveniente probar si afecta a los resultados eliminar los casos perdidos o asumir que tienen una distribución similar a la de la puntuación de rendimiento. Los resultados deben acompañarse de cuestiones destacables en este aspecto, por ejemplo, altos índices de ausencia el día de la prueba en determinadas zonas o

sujetos extraídos de los análisis (Dermitas, 2004; Zaidman-Zait & Zumbo, 2005).

- El tamaño mínimo de los conglomerados. Centros con menos de 20 individuos pueden tener asociados errores de estimación alto y, en consecuencia, tenderán a no diferenciarse significativamente de la media si se utilizan estimadores BLUP para los efectos aleatorios de las escuelas (Lockwood, Louis & McCaffrey, 2003).
- Estabilidad de los resultados. Puede ser conveniente llevar a cabo análisis anuales que puedan identificar cambios en las estimaciones de VA en una escuela determinada a lo largo de la evaluación. Se necesita encontrar cierta estabilidad de los resultados para poder asegura los efectos que las escuelas tienen en el crecimiento académico de sus estudiantes. En el otro extremo, se encuentra la necesidad de detectar prematuramente escuelas con problemas o en situaciones de riesgo para poder actuar de forma inmediata. Por lo tanto, debe buscarse cierto equilibrio entre estas dos situaciones. Si se pretende llevar a cabo actuaciones sobre los centros los resultados deben observarse al menos a lo largo de tres años (OCDE, 2008) pero sin dejar de lado los resultados anuales. Actuando de esta forma se aumenta la fiabilidad de los resultados.
- Finalmente, conviene que los resultados de VA se acompañen de ciertos datos asociados a sus estimaciones como, por ejemplo, puntuaciones poco usuales o casos extremos, también si se encuentran situaciones inesperadas de las características de las distribuciones. Todo esto con el objetivo de aumentar la credibilidad de los análisis estadísticos y que los resultados sean totalmente fiables y reflejen la situación escolar real.

Todos las cuestiones mencionadas son motivos suficientes para llevar a cabo un estudio piloto que permita probar diferentes modelos de VA. La elección de un modelo u otro puede afectar a los resultados finales de VA y existen ventajas e inconvenientes asociados a cada uno de ellos. Por ejemplo, los modelos más complejos pueden tener mejores propiedades estadísticas pero pierden

transparencia dificultando el entendimiento de los resultados a las audiencias sin conocimientos estadísticos que utilizan los resultados. Lockwood et al. (2007) prueban diversos modelos de VA y encuentran diferencias sustanciales, sobre todo entre los modelos de ganancia y el modelo de persistencia que utiliza modelos lineales mixtos para la estimación.

VIII.1 Problema de investigación

La finalidad principal de este trabajo es la toma decisiones en distintos aspectos metodológicos vinculados a la elaboración de los modelos estadísticos utilizados para estimar el VA de los centros educativos. Por tanto, se trata de dar respuesta a la cuestión:

¿Qué MVA es el más adecuado a los datos de la evaluación?

La estructura de los datos empleados en esta tesis, con cuatro mediciones del rendimiento en matemáticas a lo largo de dos cursos académicos, permite la formulación de modelos distintos. La toma de decisiones en los siguientes aspectos son problemas específicos que se pretenden resolver en este estudio:

- Determinar qué medida de tiempo es la más adecuada para representar el modelo de crecimiento de los datos.
- Analizar los efectos de la relación ente estatus inicial y cambio en aprendizaje, principalmente el efecto de regresión. Cambiar el punto de partida o introducir el rendimiento previo como covariable son factores que pueden modificar esa correlación.
- Comparar las estimaciones asociadas a las escuelas producidas por el MVA que utiliza el análisis multinivel longitudinal con la de los modelos basados en puntuaciones de ganancia.

En primer lugar, en los modelos de crecimiento, la trayectoria de cambio es una función del tiempo. Utilizar el número de aplicaciones (0, 1, 2, 3…n) es lo más común pero, si se procede de esta forma, se asume una distancia similar entre las aplicaciones. Sin embargo, las características del diseño de recogida de datos de este trabajo no se adecúa a esa situación. Las cuatro mediciones llevadas a cabo no

se realizaron en el mismo momento temporal, se planificaron con el objetivo de contar con puntuaciones de rendimiento al inicio y final de los dos cursos evaluados. Por lo que utilizar una medida de tiempo, por ejemplo, los meses transcurridos entre aplicaciones, parece más acorde con la situación de evaluación.

Además, los análisis de los ítems comunes entre aplicaciones desde la teoría clásica y los resultados del estudio de anclaje vertical revelan un cambio en rendimiento entre la segunda y tercera aplicación fuera de lo normal. Por un lado, los ítems comunes resultan más fáciles entre estas dos aplicaciones que entre el resto. La facilidad[81] aumenta un 6,3% entre la 1ª y 2ª aplicación, un 6,8 entre la 3ª y 4ª y un 16% entre la 2ª y la 3ª. El aumento es de más del doble, son exactamente 2,5 veces más. Algunas características educativas de la etapa evaluada pueden influir en este incremento, por ejemplo, el momento en el que se llevó a cabo la evaluación en esa tercera aplicación. Se recogió la información en noviembre, con el curso iniciado dos meses antes, momento en el que se repasan los contenidos del curso anterior. También es posible que algunos estudiantes realicen algún tipo de repaso durante el periodo de verano, por ejemplo, en academias o clases particulares.

Por tanto, analizar los posibles efectos provocados por la utilización de diferentes medidas de tiempo, así como una posible corrección entre la segunda y tercera aplicación, son aspectos que deben abordarse en este estudio.

En segundo lugar, la relación entre el estatus inicial y el crecimiento es otro de los aspectos que necesita ser estudiado en profundidad en los análisis de ganancia y crecimiento y, por consiguiente, también en los análisis del VA. Esta relación ha sido analizada con el objetivo de identificar y paliar el efecto de regresión hacia la media (ERM en adelante) (Nesselroade, Stigler & Baltes, 1980; Rogosa, 1995; Marsh & Hau, 2002; Rocconi & Ethington, 2006; Castro, Ruíz & López, 2009) o para el estudio del efecto diferencial que el estatus inicial tiene en la tasa de crecimiento (Seltzer, Choi & Thum, 2002; Choi, Seltzer, Herman & Yamashiro, 2007).

[81]Facilidad como proporción de estudiantes que aciertan los ítems comunes

Las características de la etapa educativa[82] evaluada puede provocar que las escuelas tengan poblaciones de estudiantes distintas, es decir, cuando los estudiantes cambian de la etapa primaria a secundaria obligatoria se dan dos posibilidades distintas: permanecer en la misma escuela o cambiar de centro educativo debido a que la escuela de educación primaria no oferta la siguiente etapa. Por tanto, puede haber centros con más proporción de estudiantes nuevos y los resultados que obtengan es fruto de lo que hicieron sus anteriores escuelas. Otro factor puede ser la falta de experiencia en este tipo de evaluaciones por parte de los estudiantes muestreados. Estos son motivos suficientes para llevar a cabo un análisis en profundidad de la relación y averiguar su efecto sobre las estimaciones de VA.

Rogosa (1995) muestra como la variación en la elección del punto inicial puede cambiar el sentido y la intensidad de esta relación. La utilización de un determinado momento como referente inicial puede tener efectos en la evaluación y la interpretación de los modelos. No obstante, la influencia del punto de partida en la estimación del crecimiento es inversamente proporcional al número de ocasiones de medidas con las que se cuente (Stevens & Zvoch, 2006). En los modelos de crecimiento para el análisis de los efectos escolares se ha intentado paliar ese ERM utilizando el rendimiento previo como covariable de dos formas distintas:

A. Utilizar la primera medición como covariable y eliminando un punto de la función de tiempo (Marsh & Hau, 2002), es decir, la variable dependiente tiene una ocasión de medida menos.
B. Introducir el coeficiente centrado en torno a la media global como predictor entre niveles (tiempo-estudiante) (Castro, Ruíz & López, 2009) pero sin eliminar esa ocasión de medida de la variable dependiente.

Ambas metodologías realizan ajustes de los datos por lo que pueden modificar los resultados finales, es decir, el residuo de crecimiento estimado para las escuelas pero ¿qué tipo de modificaciones introduce sobre la estimación de VA

[82]La primera toma de datos de rendimiento de los estudiantes fue al comienzo del primer curso de Educación Secundaria Obligatoria (más información en el Apartado VI.1).

de las escuelas?. Si la finalidad es paliar los efectos de la relación entre el estatus inicial y el crecimiento, llevar a cabo un ajuste de los residuos las escuelas asociados a los coeficientes de estatus inicial y crecimiento, a través del análisis de regresión simple, puede ser otra estrategia óptima.

Y, en tercer lugar, el debate sobre la definición del cambio en aprendizaje como ganancia o crecimiento se ve reflejado en modelos estadísticos distintos que se adecúan a los diferentes puntos de vista. Los modelos de ganancia tienen propiedades estadísticas problemáticas porque los ajustes hechos para la variación entre escuelas con los estudiantes es débil (OCDE, 2008). La misma opinión tiene Willet destacando que los modelos con dos únicas tomas de datos pierden fuerza si se introducen predictores del cambio (1989a; 1994). Los diseños longitudinales, al permitir evaluar la trayectoria del crecimiento de los alumnos durante un periodo de tiempo, son considerados por algunos autores como los más adecuados para evaluar el progreso de los alumnos y la eficacia de las escuelas (Singer & Willett, 2003; Stevens & Zvoch, 2006; Thum, 2009).

Los análisis del cambio a través de modelos longitudinales de análisis parecen más adecuados para los objetivos que persigue una evaluación basada en las estimaciones de VA. El crecimiento en aprendizaje se incorpora en los MVA de dos formas distintas: mediante el análisis multinivel longitudinal (Bryk & Raudenbush, 2002) o los modelos lineales mixtos (Sanders & Horn, 1994). Elegir una perspectiva u otra conlleva asumir ciertos supuestos en los datos y debe ser, por tanto, el modelo el que se adecúe a las características de los datos y las necesidades de la evaluación y no al revés. Sin embargo, no siempre es factible contar con datos longitudinales del rendimiento académico o cualquier otro constructo.

En ocasiones, el MVA debe adaptarse a los datos disponibles, es decir, no se planifica con anterioridad el diseño de recogida de información. Por ejemplo, la tendencia en Europa no es la evaluación longitudinal como ocurre en EE.UU, sino que se recoge información en dos puntos concretos de la educación obligatoria. En España, todas las Comunidades Autónomas deben evaluar a sus alumnos en 4º de Educación Primaria y 2º de Educación Secundaria a través de las evaluaciones de diagnóstico pero su diseño no está destinado a medir el VA y no se ha pensado la

forma de seguir a los estudiantes en esos dos cursos evaluados. En cambio, en Inglaterra, se recogen datos al final de las cuatro etapas clave de la educación obligatoria (7, 11, 14 y 16 años) y se han llevado a cabo varios estudios piloto que prueban modelos basados en la ganancia. Se decantan por un modelo de ganancia residual multinivel con las puntuaciones de los estudiantes al final de las etapas clave 2 y 4, también incluyen covariables de contexto en los análisis del VA (Ray A. , 2006). En Polonia también se han probado MVA con dos únicas puntuaciones de test al final de la educación primaria y del primer ciclo de secundaria obligatoria (Jakubowski, 2008). De forma similar a Inglaterra se estiman modelos de ganancia residual de efectos fijos y aleatorios sin encontrar grandes diferencias. Sin embargo, si se incluye el nivel socioeconómico de los estudiantes como predictor, los resultados no son tan parecidos.

Es conveniente, por tanto, probar distintas perspectivas de análisis del VA, tanto modelos de ganancia como de crecimiento, y estudiar los posibles efectos en las estimaciones de VA.

VIII.2 Metodología

Los datos utilizados en este trabajo permiten el estudio de modelos distintos de análisis del VA. Es posible variar el número de ocasiones de medida, el punto de partida y la trayectoria de crecimiento. Estos factores pueden afectar a las estimaciones finales del VA de las escuelas y, por tanto, deben tratarse con cautela.

Estos datos, extraídos de la mencionada evaluación de la Comunidad de Madrid[83], utiliza dos mediciones del logro, al inicio y final de cada curso, durante dos años académicos en diferentes cohortes de estudiantes (Gaviria, Biencinto & Navarro, 2009; Castro, Ruíz & López, 2009; Lizasoain & Joaristi, 2009; Blanco, González & Ordóñez, 2009). La primera toma de datos se lleva a cabo sobre estudiantes que comienzan el primer curso de Educación Secundaria Obligatoria y finaliza cuando terminan el primer ciclo de esa etapa un curso después. Las características de la etapa, principalmente en la primera toma de datos (1º de

[83]Más información en el capítulo VI

ESO), pueden producir dos situaciones distintas: estudiantes que comienzan la secundaria obligatoria en un centro nuevo porque su escuela de primaria no ofertaba esa siguiente etapa obligatoria, en cambio, otros pueden ser antiguos alumnos de los centro. Por tanto, esa medición debe tomarse con cautela porque lo que los alumnos saben puede ser resultado de lo que hicieron en sus anteriores escuelas.

No existe un acuerdo generalizado sobre cuál es la técnica más adecuada para el cálculo del VA, sin embargo se recomienda el análisis jerárquico de los datos para llevar a cabo la estimación estadística de un modelo que utiliza datos longitudinales de rendimiento (Goldstein, 1986).

Los modelos multinivel o modelos jerárquicos lineales son una extensión de los modelos lineales mixtos como se estudió en el capítulo 3. Ambos se utilizan para llevar acabo análisis del VA. Este tipo de análisis de regresión permite separar la varianza de los distintos niveles estudiados. Y así averiguar que parte de la variación se debe a las escuelas.

Los modelos también estiman residuos asociados a las escuelas que se emplean como elemento principal para el cálculo del VA. Los MVA producirán estimaciones distintas de esos residuo porque, aunque la variable dependiente sea la misma, varían en la consideración del cambio, el número de mediciones, la función temporal, etc.

Todos los modelos que se han desarrollado en este trabajo son análisis de regresión multinivel o, de forma más general, modelos lineales mixtos que incluyen coeficientes fijos y aleatorios. Es posible realizar una categorización en función de los niveles construidos:

- Modelos de 3 niveles. Estos análisis consideran que el cambio en aprendizaje se refleja en una trayectoria de crecimiento en función del tiempo y un estatus inicial de partida que es posible estimar a partir de las diferentes puntuaciones de rendimiento de un estudiante. Estas puntuaciones son el primer nivel de agregación, es decir, el tiempo. Las trayectorias pueden variar entre estudiantes (nivel 2) y entre escuelas (nivel 3). Los residuos asociados a las escuelas son las estimaciones de VA.

- Modelos de 2 niveles. De forma distinta a los anteriores, estos no estiman una pendiente de crecimiento asociada al tiempo sino que calculan los cambios anuales en aprendizaje. En esta categoría se puede llevar a cabo una diferenciación entre modelos univariantes que utilizan la puntuación de ganancia o el postest como variable dependiente en los análisis (ganancia bruta y ganancia residual), y los modelos multivariantes que utilizan todo el vector de puntuaciones de rendimiento como variable criterio en el modelo pero estiman coeficientes en cada ocasión de medida (modelo EVAAS y ganancia estimada).

Todos los modelos se han desarrollado con el software MLWin 2.21[84], aunque también es posible llevar a cabo el proceso de estimación con IBM SPSS 19 (IBM, 2010). No obstante, este último software no estima todavía residuos de tercer nivel. Por este motivo se utilizó MLWin pero llevar a cabo la estimación de los coeficientes fijos y aleatorios de los modelos y también los residuos asociados a las escuelas. Además se incluye en el Anexo III[85] la sintaxis para llevar a cabo la estimación con IBM SPSS.

No se prueban modelos de clasificación cruzada porque la estructura de los datos no estaba diseñada para diferenciar entre docentes o aulas, únicamente escuelas. Por tanto, los cambios de estudiantes entre aulas no pueden ser tenidos en cuenta. Por el mismo motivo, al centrar la atención en los centros educativos, tampoco se han probado modelos que permitan una disminución de los efectos entre aplicaciones.

El modelo utilizado como base de los análisis es un modelo longitudinal de crecimiento multinivel o modelo de curva de crecimiento[86]. En resumen, las características básicas de este MVA son las siguientes:

[84]El programa MLWin (Rasbash et al., 2009) es propiedad del Centro de Análisis Multinivel de la Universidad de Bristol (http://www.bristol.ac.uk/cmm/)

[85]Este anexo incluye la sintaxis para estimar modelos multinivel utilizando modelos lineales mixtos con IBM SPSS.

[86]Descrito en el Apartado V.2.3.1.

- Es un modelo de crecimiento, por tanto, estima un estatus inicial y una pendiente de crecimiento para analizar la trayectoria de cambio en aprendizaje a lo largo de las distintas mediciones.
- Considera el anidamiento completo de los datos en tres niveles. Las diferentes mediciones a lo largo del tiempo (nivel 1), los efectos diferenciales entre estudiantes (nivel 2) y las variaciones entre escuelas (nivel 3).
- El modelo de base no incluye ningún predictor, solo el estatus inicial y la pendiente de crecimiento. No obstante, algunas variaciones del modelo incluirán el rendimiento previo de maneras distintas pero no se incluye ninguna variable del contexto socioeconómico del estudiante o de la escuela para ajustar los resultados. Se ha demostrado que introducir el rendimiento previo captura los efectos que el nivel socioeconómico trata de medir en este tipo de análisis (Hibpshman, 2004). Además, existen estudios que prueban modelos similares con y sin predictores que no encuentran cambios sustanciales en las estimaciones finales (Ballou, Sanders & Wright, 2004).
- La estimación del VA es el residuo de regresión asociado a la pendiente de crecimiento de las escuelas y se define como la diferencia en crecimiento de un determinado centro educativo respecto a la media global de crecimiento de toda la muestra. Si se incluye el rendimiento previo como predictor la interpretación cambia ligeramente.

Las características específicas de los datos[87] son adecuadas para este tipo de análisis. Las puntuaciones de rendimiento de los estudiantes se recogen al inicio y final de curso durante dos grados consecutivos por lo que la distancia entre aplicaciones no es la misma. Además, los rasgos de la etapa evaluada (1º y 2º de ESO) puede que determinen un cambio en el punto de partida.

[87]Ver apartado IV.1 para más detalle.

	2005			2006											2007					
1º ESO	O	N	D	E	F	Mr	Ab	My	Jn	Jl	A									
2º ESO												S	O	N	D	E	F	Mr	Ab	My	Jn

Tabla VIII.1. Recogida de información de rendimiento en matemáticas.

En la Tabla VIII.1 se puede per observar el momento de administración de los diferentes instrumentos de medida. La separación entre las mediciones de inicio y final de curso en 1º de ESO es de 8 meses, mientras que en 2º de ESO es de siete. El periodo de verano es de cinco meses porque la tercera aplicación tuvo que retrasarse lo que condiciona la tercera medición que se lleva a cabo con el curso ya empezado. Debido a esta característica del diseño de recogida de información se utiliza una medida de tiempo con la distancia mensual entre aplicaciones como referente.

$$T(0, 8, 13, 20)$$

Las cuatro puntuaciones de rendimiento se encuentran escaladas verticalmente con una metodología TRI de calibración conjunta con estimaciones bayesianas (MAP) del rasgo de los sujetos, es decir, el rendimiento en matemáticas. La selección de esta escala es fruto del estudio de comparación llevado a cabo en el capítulo V de esta tesis. Esta escala vertical es adecuada para la estimación de los modelos de crecimiento. El vector con las cuatro puntuaciones se utiliza como variable dependiente en el análisis multinivel y componen el primer nivel de análisis, el crecimiento en función del tiempo.

$$Y_{tij} = \beta_{0ij} + \beta_{1ij}(T) + e_{tij} \qquad \textbf{Ec. VIII.1}$$

Utilizando notación matricial para este primer nivel, la estructura de los distintos coeficientes queda más clara:

$$\begin{bmatrix} Y_{1ij} \\ Y_{2ij} \\ Y_{3ij} \\ Y_{4ij} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 8 \\ 1 & 13 \\ 1 & 20 \end{bmatrix} \begin{bmatrix} \beta_{0ij} \\ \beta_{1ij} \end{bmatrix} + e_{tij} \qquad \textbf{Ec. VIII.2}$$

β_{1ij}Es la tasa de crecimiento del estudiante *i* de la escuela *j* y β_{0ij} es el estatus inicial, es decir, cuando β_{1ij} es igual a cero que, en este caso, se produce en

la primera ocasión de media. Esta función del tiempo puede modificarse para ponderar el efecto del crecimiento entre aplicaciones o cambiar el punto de partida. Si formulamos los tres niveles en una misma ecuación (Ec. VIII.3):

$$Y_{tij} = \beta_{00} + \beta_{10}(T) + u_{0j} + u_{1j}(T) + r_{0ij} + r_{1ij}(T) + e_{tij} \qquad \textbf{Ec. VIII.3}$$

β_{00} y β_{10} son las medias globales en estatus inicial y la tasa de crecimiento, respectivamente. El resto de parámetros son los residuos aleatorios de cada nivel (*r* para los estudiantes y *u* para las escuelas). u_{1j} es el residuo asociado a la trayectoria de crecimiento los centros educativos y puede considerarse el VA de una escuela *j* y, como puede verse la ecuación, el término se encuentra vinculado al tiempo.

Esto quiere decir que el VA de una escuela en la segunda ocasión de medida, al final del primer curso de secundaria obligatoria es igual $au_{1ij} * (13)$ y la puntuación estimada en esa misma ocasión de medida para una escuela es resultado de sumar al estatus inicial, la pendiente de crecimiento y sus residuos correspondientes:

$$Y_{3ij} = \beta_{00} + \beta_{10}(13) + u_{0j} + u_{1j}(13) \qquad \textbf{Ec. VIII.4}$$

En este estudio hay dos grupos de coeficientes a estimar: Por un lado, el estatus inicial de rendimiento en matemáticas de las escuelas (β_{00}) y su crecimiento en función del tiempo (β_{10}) que componen la parte fija de la ecuación de regresión. Y, por otro, los términos aleatorios, es decir, las varianzas y covarianzas de esos dos coeficientes de regresión entre estudiantes y escuelas ($r_{0ij}, r_{1ij}, u_{0j}, u_{1j}$) y, por último, el residuo aleatorio de primer nivel (e_{tij}).

En cada uno de los tres problemas planteados se formulan modelos distintos. Con el objetivo de identificar los distintos modelos que se han elaborado se utiliza la denominación abreviada M1_1. La M es la inicial de modelo, el primer digito indica el orden y el segundo dígito el problema en el que se desarrollan. En este ejemplo, es el modelo 1 que se ha implementado en el primer problema planteado.

VIII.2.1 Problema 1. Selección de una medida adecuada de tiempo

El primer problema o, más bien, cuestión metodológica que debe resolverse para desarrollar un modelo multinivel longitudinal es:

¿Qué medida de tiempo es la más adecuada en el modelo de crecimiento?

Para elaborar las distintas variables de tiempo que van a determinar la trayectoria decrecimiento en rendimiento de los estudiantes se han tenido en cuenta dos factores determinantes:

- El intervalo temporal entre las distintas aplicaciones de medida. El modelo multinivel de crecimiento base utiliza la distancia, en meses, entre las mediciones. También se ha formulado otro que considera la misma distancia entre aplicaciones.
- El aumento de la facilidad en los ítems comunes entre A2 y A3. Se ha detectado un cambio en la proporción de respuestas correctas con respecto a los que se producen entre el resto de aplicaciones consecutivas. El aumento es 2,5 veces superior, concretamente el aumento en porcentaje es de un 9,5% más de estudiantes que responden correctamente a esos ítems. Se ha utilizado un parámetro de ponderación que modifica el cambio entre A2 y A3 de dos maneras distintas: Multiplicando la distancia entre aplicaciones por ese elemento de ponderación o dividiéndola.

En total, se han construido seis modelos distintos con distintos valores para la función de tiempo. Las distintas aproximaciones parten de dos modelos generales y se realizan cambios añadiendo un factor de ponderación en el crecimiento entre A2 y A3. Un primer modelo (M1_1) que utiliza la distancia, en meses, transcurrida entre las distintas aplicaciones, y otro (M3_1) que considera la misma distancia entre las distintas mediciones.

El criterio de ponderación se basa en el aumento detectado en la proporción de respuestas correctas en los ítems comunes, en comparación con el resto. Ese aumento es más del doble, exactamente son 2,5 veces más. Por tanto, es valor temporal de cambio entre A2 y A3 se pondera de dos formas distintas: dividiendo y multiplicando ese valor por el criterio de ponderación, que se ha denominado

lambda (λ). Por tanto, si el punto inicial de crecimiento se establece siempre con valor cero y k es el criterio temporal utilizado (meses o aplicaciones), las distintas funciones de tiempo ponderadas se calculan como se muestra en Ec. VIII.5:

$$t1 = 0; t2 = t1 + k; t3 = t2 + (k * \lambda); t4 = t3 + k$$

$$t1 = 0; t2 = t1 + k; t3 = t2 + ({}^{k}/_{\lambda}); t4 = t3 + k \qquad \textbf{Ec. VIII.5}$$

$$\lambda = 2{,}5$$

Las seis funciones de tiempo resultantes que van a incorporarse en los distintos modelos son:

M1_1:	T_1(0 8 13 20)
M2_1:	T_2(0 8 $\left(8 + {}^{5}/_{2,5} = 10\right)$ 17)
M3_1:	T_3(0 8 $(8 + 5 * 2{,}5 = 20{,}5)$ 27,5)
M4_1:	T_4(0 1 2 3)
M5_1:	T_5(0 1 $\left(1 + {}^{1}/_{2,5} = 1{,}4\right)$ 2,4)
M6_1:	T_6(0 1 $(1 + 1 * 2{,}5 = 3{,}5)$ 4,5)

Tabla VIII.2. Valores de la función temporal en los distintos modelos de crecimiento.

VIII.2.2 Problema 2. Relación entre estatus inicial y crecimiento y efecto de regresión hacia la media.

El segundo problema planteado analiza un aspecto metodológico de los modelos de cambio que puede afectar a las estimaciones del VA de las escuelas: la correlación entre estatus inicial y crecimiento. Se plantea de la siguiente forma:

¿Cómo afecta la relación entre el punto inicial y el crecimiento a la estimación del VA de las escuelas?

El tipo de relación entre estatus inicial y cambio en aprendizaje es un indicador de posibles artefactos del diseño que pueden producirse darse en este tipo de análisis, principalmente el mencionado Efecto de Regresión hacia la Media[88] (ERM en adelante). Además, existen factores que pueden alterar esa relación como elegir un punto de partida determinado o introducir el rendimiento previo como covariable en los análisis. Para estudiar sus efectos sobre la

[88]Mas información en el apartado IV.3.2

estimación del residuo de las escuelas, es decir, el VA, se han elaborado modelos diferentes que tratan de ajustar esa relación y que varían en dos grandes aspectos:

- La selección de la aplicación que se considera el punto de partida o el estatus inicial de rendimiento.
- La inclusión del rendimiento previo como predictor en el modelo como las dos aproximaciones mencionadas que tratan de paliar las consecuencias del ERM en las estimaciones (Marsh & Hau, 2002; Castro, Ruíz & López, 2009).

La combinación de estos factores se ha utilizado para construir cinco modelos distintos. El primer modelo (M1_2) es el modelo utilizado como referente, el mismo que en el primer problema (M1_1), es el que utiliza la distancia, en meses, transcurrida entre las distintas aplicaciones como función de tiempo (T (0, 8, 13, 20)).Los modelos M2_2 y M3_2 son variaciones del primero para seleccionar la A2 y A3, respectivamente, como puntos de partida. Los modelos cuatro y cinco (M4_2 y M5_2) incluyen el rendimiento previo como predictor de dos maneras distintas:

- M4_2: Introduce el rendimiento en A1 centrado en torno a la media global de A1como predictor de la tasa de crecimiento, es una variable entre niveles (tiempo-estudiante):

$$Y'_{1i} = (Y_{1i} - \overline{Y}_1)(T) \qquad \textbf{Ec. VIII.6}$$

Donde Y_{1i} es la puntuación inicial del estudiante i, $\overline{Y}_1$ es la media global de todos los sujetos en la primera ocasión de medida y está multiplicando a T, que es la función temporal, la misma que en el M1_2.La ecuación final multinivel es la siguiente:

$$Y_{tij} = \beta_{0,00} + \beta_{10}(T) + \beta_{20}(Y'_{1i}) + r_{0ij} + r_{1ij}(T) + u_{0j} + u_{1j}(T) + e_{t} \qquad \textbf{Ec. VIII.7}$$

- M5_2: utiliza la primera medición, también centrada respecto a la media global, como covariable en el modelo pero, esa medida, no forma parte de la función de crecimiento, es decir, no es un punto de la función de tiempo. Por tanto, esa función solo tiene tres puntos en este modelo. La ecuación multinivel sería igual a la anterior pero con solo tres mediciones en el primer nivel, por tanto, la función de tiempo

cambia y también la covariable de rendimiento previo que ya no se encuentra vinculada al tiempo.

Las características de estos cinco modelos se incluyen en la tabla siguiente (Tabla VIII.3):

M1_2:	T_1(0 8 13 20)
M2_2:	T_2(−8 0 5 12)
M3_2:	T_3(−13 5 0 7)
M4_2:	T_4(0 8 13 20); Y'_{1i}
M5_2:	T_5(0 5 12); Y_{1i}

Tabla VIII.3. Características de los cinco modelos elaborados en el problema 2.

El estatus inicial puede determinar el crecimiento de los estudiantes y por tanto las estimaciones del VA de las escuelas. El ERM es un artefacto de diseño que se produce cuando la relación es negativa pero en los modelos multinivel longitudinales es muy difícil distinguir ese efecto de otros factores como los errores de medida de las variables utilizadas que pueden ser mayores en la A1, el encogimiento producido por las estimaciones bayesianas de los efectos aleatorios o el propio diseño de la escala vertical con los efectos, ya mencionados, de suelo y techo. Por tanto, si determinar que parte del cambio de un estudiante se debe a al ERM es muy complicado con datos empíricos, llevar a cabo ajustes con toda la muestra resulta arriesgado.

No obstante, el diseño longitudinal de medida utilizado para recoger los datos de rendimiento que emplea este trabajo si puede verse afectado por ese estatus inicial debido a las características de la etapa educativa evaluada. Con la finalidad de paliar ese efecto del punto inicial sin cambiar el orden del punto de partida o incluir covariables para ajustar el crecimiento de los estudiantes, se lleva a cabo una propuesta que utiliza directamente los residuos multinivel del estatus inicialu_{0j} y el crecimientou_{1j} asociados a las escuelas. A través de una regresión simple de la ecuación Ec. VIII.8 se ajusta el VA en función de los distintos puntos de partida de las escuelas:

$$E(u_{1j}|u_{0j}) \quad \text{ó} \quad u_{1j} = \alpha_{00} + \alpha_{10}(u_{0j}) + v_j \qquad \textbf{Ec. VIII.8}$$

α_{00} es el VA medio para todas las escuelas, por tanto, igual a cero porque la media de los residuos de crecimiento de las escuelas estimados con la metodología

multinivel son diferencias respecto a la media en crecimiento; α_{10}es el efecto del estatus inicial de la escuela sobre el VA. Un valor negativo indica un VA más alto para las escuelas que parten por debajo de la media, si fuera positivo serían las escuelas que tienen niveles iniciales más altos las que obtienen un mayor VA. De este ajuste se obtiene un residuo $v_j$que se considera la nueva estimación del VA de los centros educativos, sin el efecto de los puntos de partida. Esta nueva estimación está libre de los posibles efectos diferenciales que los distintos estatus iniciales de rendimiento pueden provocar. Los resultados de estos análisis se han denominado M6_2.

VIII.2.3 Problema 3: Comparación de modelos de ganancia y crecimiento

La manera de medir el cambio en aprendizaje a través de estimaciones de la ganancia o del crecimiento también conlleva formas distintas de analizar el VA de las escuelas. Es necesario llevar a cabo una comparación de las estimaciones producidas por MVA que varían en su aproximación al estudio del cambio. Por tanto, el último problema se plantea así:

¿Cómo afecta la forma de medir el cambio en aprendizaje a las estimaciones de VA de las escuelas?

La ventaja de contar con una estructura de datos de rendimiento que cuenta con cuatro puntuaciones de rendimiento en solo dos cursos académicos es que, además de modelos de crecimiento, también es posible calcular la ganancia entre ambos cursos utilizando las puntuaciones que se recogieron al final de cada uno de ellos. Los modelos elaborados para tratar este problema se pueden agrupar en dos grandes categorías:

- MVA basados en el crecimiento: son los que utilizan más de dos puntuaciones de rendimiento como variable dependiente en los análisis. Las dos principales aproximaciones, ya descritas[89], para tratar con esta estructura de datos son los modelos multinivel longitudinales (M1_3) y los modelos lineales mixtos (M2_3), aunque

[89]Más información sobre los MVA en el Capítulo V

una puede considerarse una extensión de la otra. Ambas metodologías utilizan toda la matriz de puntuaciones de los estudiantes como variable dependiente. Por tanto, ambos son modelos multivariantes. Sin embargo, mientras que el M1_3 estima un coeficiente para el estatus inicial y otro para el crecimiento, el M2_3 estima coeficientes para cada una de las mediciones de rendimiento y las utiliza para calcular ganancias entre cursos. No es posible estimar una pendiente de crecimiento entre mediciones pero con los modelos mixtos se pueden calcular ganancias entre-cursos, es decir, entre la puntuación final del primer curso (A2) y la puntuación final del segundo curso (A4). Y también es posible estimar las ganancias intra-cursos, es decir, entre la primera y segunda aplicación y entre la tercera y la cuarta aplicación.

- MVA basados en la ganancia: solo emplean dos mediciones para estimar el cambio en aprendizaje y hay tres maneras de realizarlo: ganancia bruta, residual y estimada. En esta sección solo la ganancia estimada emplea un modelo multivariante con las dos puntuaciones de rendimiento, es un modelo bivariado (M3_3). El resto son modelos univariantes (M4_3 y M5_3).

El primer modelo estimado (M1_3) tiene tres niveles de agregación (tiempo, estudiante y escuela), el resto solo estiman variación aleatoria en dos niveles (estudiante y escuela). No obstante, para desarrollar el modelo de ganancia estimada (M3_3) con el software de análisis multinivel MLWin, es necesario un tercer nivel de agregación que defina la estructura multivariada de la variable dependiente pero sin varianza residual. Por tanto, solo hay varianza entre estudiantes y escuelas no entre las puntuaciones anidadas en los estudiantes.

Debido a esta variación de niveles entre modelos y para facilitar la interpretación los residuos aleatorios se denominan siempre de la misma manera. Es decir, cuando el modelo incluya varianza aleatoria en el nivel 1 (tiempo), su residuo es e_{tij}. Si el residuo está asociado al estudiante, independientemente de ser el primer o segundo nivel, es r_{ij}. Y si es de la escuela u_j.

El modelo de base (M1_3), es el mismo que en los anteriores problemas. Un modelo multinivel longitudinal con tres niveles de agregación y con los valores de la variable tiempo T(0, 8, 13, 20). Los coeficientes, fijos y aleatorios, a estimar para las escuelas aparecen en la siguiente fórmula (Ec. VIII.9):

$$\begin{bmatrix} Y_{1ij} \\ Y_{2ij} \\ Y_{3ij} \\ Y_{4ij} \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 8 \\ 1 & 13 \\ 1 & 20 \end{bmatrix} + \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \qquad \textbf{Ec. VIII.9}$$

El residuo de tercer nivel asociado a la pendiente de crecimiento (u_{1j}) es el VA de la escuela. También considera varianza aleatoria de estos ambos coeficientes entre estudiantes (r_{0ij}, r_{1ij}) e incluye el término de error residual (e_{tij}) que es un escalar. Se ha añadido a la comparación de resultados de este problema los residuos de dos de los modelos que obtuvieron los mejores resultados en el problema anterior. Son el que utiliza la segunda ocasión de medida como punto de partida ($\beta_1 = -8{,}0{,}5{,}12$), que en este problema se denomina M1.1_3, y el que lleva a cabo el ajuste mediante una análisis de regresión simple entre los residuos de estatus inicial y crecimiento para estimar el nuevo residuo de las escuelas libre del efecto del estatus inicial, este modelo se ha denominado M1.2_3.

El modelo dos (M2_3) es un modelo lineal mixto que trata de aproximarse a los análisis del VA realizados por el modelo EVAAS pero ajustado a los datos de este estudio y con algunas diferencias:

- La primera de ellas es que el EVAAS pone la atención en los docentes y no en las escuelas. Los datos utilizados en este trabajo únicamente identifican escuelas.
- Una segunda distinción es la inclusión de efectos cruzados por pate del EVAAS, es decir, lo usual es que los estudiantes cambien de docente al cambiar de curso académico o incluso cambiar de profesor el mismo curso. Utilizando efectos cruzados es posible seguir a los estudiantes aunque cambien de docente, incluso de escuela. En este trabajo los datos están completamente anidados, estudiantes dentro de escuelas.

- Una tercera distinción, relacionada con la anterior, es la persistencia de los efectos cuando provienen de profesores distintos. Esos efectos permanecen constantes al cambiar de profesor o de escuela. Los datos de este trabajo evalúan un mismo centro en todas las mediciones por lo que no se distinguen efectos previos producidos por otras escuelas.
- Una cuarta diferencia es que el modelo EVAAS necesita que la distancia entre mediciones sea la misma ya que la base de comparación es la ganancia entre cursos evaluados. Sin embargo, la única forma de realizarlo es con las puntuaciones de A2 y A4 que se recogieron al final de los dos cursos evaluados. Pero si se procede de esta manera el modelo acaba convirtiéndose prácticamente en una ganancia estimada.

A pesar de las diferencias y con finalidad experimental se adapta la metodología EVAAS a los datos disponibles. Se incluyen las cuatro puntuaciones de rendimiento como variable dependiente de la misma forma que en el M1_3 pero en lugar de estimar un estatus inicial y una pendiente de crecimiento construye una ecuación de regresión con las puntuaciones de cada una de las aplicaciones para cada estudiante en el primer nivel. Por tanto:

$$\begin{bmatrix} Y_{1ij} \\ Y_{2ij} \\ Y_{3ij} \\ Y_{4ij} \end{bmatrix} = \begin{bmatrix} \beta_{0j} \\ \beta_{1j} \\ \beta_{2j} \\ \beta_{3j} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} + \begin{bmatrix} r_{ij} \end{bmatrix} \qquad \textbf{Ec. VIII.10}$$

Cada uno de los coeficientes β son las medias de los estudiantes de la escuela j en cada una de las aplicaciones y r_{ij} es el término residual. Este residuo de primer nivel no es un escalar, se incluye una matriz de varianzas-covarianzas no estructurada que permita la correlación entre las distintas puntuaciones de un estudiante.

El nivel 2 (escuelas) incluye varianza aleatoria entre escuelas de los coeficientes β.

$$\begin{bmatrix}\beta_{0j}\\ \beta_{1j}\\ \beta_{2j}\\ \beta_{3j}\end{bmatrix} = \begin{bmatrix}\beta_0\\ \beta_1\\ \beta_2\\ \beta_3\end{bmatrix}\begin{bmatrix}1&0&0&0\\ 0&1&0&0\\ 0&0&1&0\\ 0&0&0&1\end{bmatrix} + \begin{bmatrix}u_{0j}\\ u_{1j}\\ u_{2j}\\ u_{3j}\end{bmatrix} \qquad \textbf{Ec. VIII.11}$$

En este caso los coeficientes β son las medias globales en cada una de las mediciones llevadas a cabo (A1, A2, A3 y A4 respectivamente); y u es la variación residual de las escuelas asociada a esas medias. No hay pendiente de crecimiento pero el tipo de diseño utilizado para tomar las distintas mediciones, al inicio y al final de cada curso, permite calcular las ganancias en cada uno de los cursos por separado o entre el final de cada uno de los cursos.

Para calcular esas ganancias el modelo EVAAS asume que el efecto previo del profesor permanece con el estudiante cuando progresa[90]. Por tanto, el efecto de la escuela en la ganancia del estudiante es lo que queda una vez eliminado el efecto de la ganancia del estudiante y la ganancia en la media global. Por ejemplo, si se pretende observar la ganancia en el primer curso evaluado de una escuela *j*:

$$Y_{2j} - Y_{1j} = (\beta_1 - \beta_0) + Pu_{1j} \qquad \textbf{Ec. VIII.12}$$

Si únicamente quiere analizarse el efecto de la escuela en el primer curso evaluado, es decir, el VA, entonces debe observarse u_{1j}. P es una medida de proporción de escolarización en una escuela determinada, para el caso en el que el modelo diseñado permita la identificación de un estudiante cuando cambia de docente. En este caso permanece constante e igual a uno porque no se contemplan esos cambios.

La ganancia de una escuela entre los dos cursos evaluados se calcula considerando las puntuaciones al final de cada curso (A2 y A4), es igual:

$$Y_{4j} - Y_{2j} = (\beta_3 - \beta_1) + Pu_{3j} \qquad \textbf{Ec. VIII.13}$$

El residuo u_{3j} es considerado en el modelo EVAAS el VA entre dos cursos distintos, en concordancia con la persistencia de los efectos de docentes previos. No obstante, la simplicidad del modelo planteado en este trabajo, en comparación

[90]Más información en el apartado V.2.3.2.1.

con el EVAAS, y la estructura de los datos con puntuaciones al inicio y final de un mismo curso son factores que determinan un cambio en la consideración del VA calculado mediante sus residuos, sin tener en cuenta la persistencia de esos efectos:

$$Y_{4j} - Y_{2j} = (\beta_3 + Pu_{3j}) - (\beta_1 + Pu_{1j}) \qquad \textbf{Ec. VIII.14}$$

Por tanto, también se incluye la diferencia entre los residuos entre A3 y A4 ($Pu_{3j} - Pu_{1j}$) estimados con este modelo en los resultados de comparación (M2.1_3).

De este modo, también es posible calcular el VA de cada curso por separado. Con $Pu_{1j} - Pu_{0j}$ (M2.2_3) para el primer curso y $Pu_{3j} - Pu_{2j}$ (M2.3_3) para el segundo. Así se obvia el periodo de verano que puede analizarse de forma aislada ($Pu_{2j} - Pu_{1j}$ (M2.4_3)) y desde el inicio al final de la evaluación ($Pu_{3j} - Pu_{0j}$ (M2.5_3)). Estos modelos se han incorporado en la comparación de resultados.

Los análisis de VA desarrollados por el EVAAS son mucho más complejos. Por ejemplo, incluye las puntuaciones en más de una materia, cinco medidas de logro académico, efectos aleatorios cruzados, uno modelo para las escuelas, otro para los docentes y otro para el sistema educativo en general, etc. El M2_3 es solo una simple aproximación y, por tanto, sus resultados deben considerarse en el mismo sentido, es decir, una aproximación a los que produciría el modelo EVVAS. Uno de los factores que sí asemejan a ambos modelos es que incluye una matriz de varianzas-covarianzas sin estructura entre los residuos de primer nivel, es decir, entre las distintas puntuaciones de rendimiento en matemáticas de los estudiantes.

$$r = \begin{bmatrix} r_1 & & & \\ r_{21} & r_2 & & \\ r_{31} & r_{32} & r_3 & \\ r_{41} & r_{42} & r_{43} & r_4 \end{bmatrix} \qquad \textbf{Ec. VIII.15}$$

Esta estructura puede ser un mejor reflejo de la falta de homogeneidad encontrada entre las distintas mediciones del rasgo evaluado. El M1_3 considera una varianza común entre las aplicaciones que forman parte de crecimiento y puede ser un aspecto que aumente el sesgo.

El tercer modelo (M3_3) comienza con los análisis de la ganancia, concretamente la ganancia estimada. La estructura es similar al anterior pero con dos puntuaciones de logro únicamente, A2 y A4. Las puntuaciones al final de cada curso evaluado:

$$\begin{bmatrix} Y_{2ij} \\ Y_{4ij} \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \begin{bmatrix} 10 \\ 01 \end{bmatrix} + \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \qquad \textbf{Ec. VIII.16}$$

De la misma forma que el modelo anterior, los coeficientes β son las medias globales en el pretest y postest respectivamente; y u_{tj} los residuos aleatorios asociados a cada puntuación. La ganancia estimada es por tanto:

$$G_j = (\beta_1 + u_{1j}) - (\beta_0 + u_{0j}) \qquad \textbf{Ec. VIII.17}$$

Se utiliza la diferencia entre los residuos $(u_{1j} - u_{0j})$ como parámetro de comparación, una hipotética puntuación de VA[91] de la escuela. Para la estimación se utilizaron únicamente a los sujetos que contaban con las dos medidas de rendimiento.

El modelo número cuatro (M4_3) continua con el concepto de residuo como elemento para valorar el VA de un centro. Es la ganancia residual. Se lleva a introduce la puntuación en A2, centrada en torno a la gran media, como principal covariable en la regresión sobre la puntuación en A4:

$$Y_{4ij} = \beta_0 + \beta_2 Y_{2ij} + u_{0j} \qquad \textbf{Ec. VIII.18}$$

β_0 es la media global ajustada, una vez que se incluye el rendimiento en A2 como covariable. Por tanto, es igual a la puntuación media en el postest para aquellas escuelas con una puntuación media en el pretest; $\beta_1$es el efecto diferencial del rendimiento previo; y u_{0j} es el residuo ajustado, sin el efecto del rendimiento previo y en este modelo se considera la puntuación de VA de la escuela. Se incluye

[91]Los modelos basados en la ganancia están afectados en mayor medida por los factores contextuales que aquellos basados en el crecimiento (Ballou, Sanders & Wright, 2004; Tekwe et al., 2004). Por tanto, los MVA que utilizan únicamente dos mediciones del logro, en cualquiera de las tres versiones de la concepción de ganancia, deberían incluir factores del contexto socioeconómico de los estudiantes. En este trabajo los análisis están libres de covariables, excepto el caso del rendimiento previo del estudiante que se incluye en el modelo de ganancia residual (M4_3)

también una variación de este modelo con el rendimiento en A1 como segunda covariable (M4.1_3):

$$Y_{4ij} = \beta_0 + \beta_1 Y_{2ij} + \beta_2 Y_{1ij} + u_{0j} \qquad \textbf{Ec. VIII.19}$$

El quinto modelo (M5_3) analiza la ganancia bruta entre las dos puntuaciones de rendimiento recogidas al final de los cursos evaluados (A2 y A4). Para realizarlo se calcula la diferencia entre ambas puntuaciones de cada estudiante:

$$G_{ij} = Y_{4ij} - Y_{2ij} \qquad \textbf{Ec. VIII.20}$$

Una vez calculada se calculan medias agregadas de esas ganancias en cada centro educativo:

$$G_j = \frac{\sum_{i=1}^{i=n} G_{ij}}{n_j} \qquad \textbf{Ec. VIII.21}$$

G_j es la ganancia media de una escuela j, es decir, el sumatorio de las ganancias de cada uno de sus estudiantes divido por el número de alumnos (n_j). Una vez obtenidas se calculan las puntuaciones diferenciales de esta ganancia para cada escuela, así la media es la misma que los residuos de regresión estimados con los otros modelos. Esta puntuación diferencial es el parámetro de comparación. De forma alternativa se construye un modelo multinivel que utiliza las puntuaciones de ganancias de los estudiantes como variable dependiente (M5.1_3) para estimar el residuo asociado a las escuelas y transformar esa ganancia bruta en desviaciones respecto a la media estimada. Y, un segundo modelo alternativo (M5.2_3), que incluye también el rendimiento en A1 (centrado en torno a la media global) como covariable en el análisis multinivel. De esta forma, el M5.1_3:

$$G_{ij} = \beta_{0j} + r_{ij}$$
$$\beta_{0j} = \beta_0 + u_{0j} \qquad \textbf{Ec. VIII.22}$$

β_0 es la ganancia media global entre la A2 y A4. Y r y u los residuos aleatorios de esa media en ambos niveles (estudiante y escuelas). El modelo M5.2_3 incorpora la puntuación en A1 como covariable en los análisis:

$$G_{ij} = \beta_{0j} + \beta_1 Y_{1ij} + r_{ij}$$
$$\beta_{0j} = \beta_0 + u_{0j} \qquad \textbf{Ec. VIII.23}$$

En este caso β_0 es la ganancia media global ajustada, eliminando el efecto del rendimiento previo Y_{1ij}.

Finalmente, se ha elaborado otro modelo (M6_3) que utiliza únicamente las puntuaciones diferenciales medias de los estudiantes de una misma escuela en la última aplicación (A4) como medida de su eficacia. De esta forma se obtienen referencias de los resultados con un modelo de estatus.

En resumen, los modelos que van a compararse para resolver el problema planteado son un total de dieciséis. Seis son modelos generales que reflejan concepciones distintas de la evaluación basadas en la ganancia, el crecimiento o el estatus de rendimiento de las escuelas. Los otros diez restantes son variaciones de los generales, dos son variaciones del modelo multinivel de crecimiento que se estimaron en el problema dos, cinco son una variación del EVAAS que no considera la persistencia de los efectos de las escuelas y estima ganancias para las dos cursos por separado, otro incorpora una covariable más de rendimiento previo en el modelo de ganancia residual y, finalmente, dos de ellos son una variación de la ganancia bruta estimada con modelos multinivel. La Tabla VIII.4 resume estos catorce modelos y sus estimaciones o puntuaciones asociadas a las escuelas, a las que se ha denominado P:

M1_3:	Modelo Multinivel de Curva de Crecimiento $P_{M1\_3} = u_{1j}$; Considerando A2 como estatus inicial $P_{M1.1\_3} = u_{1j}$; Con ajuste de los residuos a posteriori $P_{M1.2\_3} = v_{1j}$
M2_3:	Modelo Lineal Mixto $P_{M2\_3} = u_{3j}$; $P_{M2.1\_3} = u_{3j} - u_{1j}$; Curso 1 $P_{M2.2\_3} = u_{1j} - u_{0j}$; Curso 2 $P_{M2.3\_3} = u_{3j} - u_{2j}$;Verano $P_{M2.4\_3} = u_{2j} - u_{1j}$; Total $P_{M2.5\_3} = u_{3j} - u_{0j}$
M3_3:	Ganancia Estimada=$P_{M3\_3} = u_{1j} - u_{0j}$
M4_3:	Ganancia Residual $P_{M4\_3} = u_{0j}$; con dos predictores de rendimiento previo $P_{M4.1\_3} = u_{0j}$
M5_3:	Ganancia Bruta=$P_{M5\_3} = G_j - \overline{G}$; Modelo multinivel con la ganancia bruta como variable dependiente=$P_{M5.1\_3} = u_{0j}$; incluyendo la puntuación en A1 como covariable=$P_{M5.2\_3} = u_{0j}$
M6_3:	Modelo de Estatus $P_{M6\_3} = \frac{\sum_{i=1}^{i=n}(Y_{4ij} - \overline{Y}_4)}{n_j}$

Tabla VIII.4. Modelos elaborados en el problema 3.

No todas estas aproximaciones pueden considerarse estimaciones del VA de una escuela. Únicamente los modelos basados en el crecimiento pueden reclamar esa propiedad. Los modelos de ganancia necesitan incluir covariables del contexto socioeconómico para ajustar los resultados y el modelo de estatus no refleja ese cambio en aprendizaje necesario en los MVA. Sin embargo, todos los modelos desarrollados se utilizan en sistemas de evaluación que analizan el rendimiento de las escuelas a través del logro de sus estudiantes.

VIII.2.4 Metodología para la comparación de los resultados

Los resultados se presentan de forma separada para cada uno de los problemas planteados. El orden es el siguiente:

1. Resultados Problema 1: Selección de una medida adecuada de tiempo.
2. Resultados Problema 2: Relación entre estatus inicial y crecimiento y efecto de regresión hacia la media
3. Resultados Problema 3: Comparación de modelos de ganancia y crecimiento.

En cada apartado se comparan los diferentes modelos desarrollados. Para valorar el ajuste global de los modelos se utiliza el estadístico de verosimilitud calculado con la finalidad de comprobar en qué medida el modelo es capaz de captar la variabilidad de los datos. Cuando se utilizan métodos de estimación de máxima verosimilitud, como es el caso, este índice de ajuste es menos dos veces el

logaritmo del valor que maximiza la función de verosimilitud en esa estimación. Este índice es conocido por su nombre inglés *"Deviance"* y los valores más bajos representan un mejor ajuste. No obstante, los índices de ajuste únicamente pueden compararse cuando los modelos son anidados, es decir, si uno puede obtenerse igualando a cero algún parámetro del otro.

En el primer problema planteado, todos los modelos estimados tienen los mismos parámetros pero varían en los valores de la variable tiempo que afecta a la pendiente de crecimiento.

En el segundo problema, los tres modelos iniciales (M1_2, M_2_2 y M3_2) también estiman los mismos parámetros pero los valores de la función de tiempo cambian para establecer diferentes puntos de partida. El M4_2 añade un único predictor y, por tanto, se encuentra anidado con los anteriores. No ocurre lo mismo con el último modelo de este problema (M5_2), al extraer la puntuación de logro en A1 para utilizarla como predictor cambia la variable dependiente que solo cuenta con tres ocasiones de medida.

Finalmente, en el tercer problema, están anidados entre sí, por un lado, los modelos de ganancia residual (M4_3 y M4.1_3) y, por otro, los dos modelos multinivel llevados a cabo con la puntuación de ganancia bruta (M5.1_3 y M5.2_3).

Además de este índice de ajuste global de los modelos se presentan más resultados que se agrupan en dos grandes bloques: Por un lado, el análisis de los coeficientes fijos y aleatorios y la relación entre el punto de partida y el cambio o crecimiento en aprendizaje. Y, por otro, un estudio en profundidad de los residuos de los distintos de las escuelas que, en algunos casos, se utilizan para el cálculo del VA. En esta última sección de resultados se incluye la relación de esos residuos entre modelos y una representación gráfica de los mismos.

Los aspectos analizados que incluyen estos dos grandes apartados de resultados, en cada problema, se detallan a continuación

VIII.2.4.1. Coeficientes estimados

En esta sección se analiza la significatividad de los distintos parámetros estimados en los modelos para las escuelas. Principalmente son dos: los fijos (β) y

los aleatorios (u), aunque también se estudia la varianza en los otros dos niveles (estudiantes y tiempo). Para comprobar esta significatividad es suficiente con calcular el cociente entre el valor estimado del parámetro por su error típico, conocido como test de Wald. Si este resultado es mayor que dos, el coeficiente será significativo (p<0,05) (Gaviria & Castro, 2005).

También se analizan otros aspectos como la correlación intraclase o autocorrelación (ρ) que calcula el grado de homogeneidad de los contextos. Asistir a una determinada escuela produce un efecto en sus estudiantes que les hace parecerse entre sí y diferenciarse de estudiantes de otras escuelas, en términos de rendimiento. Los modelos de regresión clásicos no tienen en cuenta este aspecto, esa parte de la varianza que se debida a las unidades estudiadas. En este caso, que parte de varianza del rendimiento en matemáticas se debe al alumno y cuánta a las escuelas.

$$\rho = \frac{u_0 + u_1 - 2\sigma_{u_0 u_1}}{(u_0 + u_1 - 2\sigma_{u_0 u_1}) + (r_0 + r_1 - 2\sigma_{r_0 r_1}) + e} \qquad \textbf{Ec. VIII.24}$$

Esta correlación intraclase únicamente se calcula en los análisis que consideran varianza en diferentes niveles de agregación y varía en función de los coeficientes aleatorios que incorpora cada uno de ellos.

Otro aspecto que se analiza es la correlación entre estatus inicial y crecimiento en el nivel de los estudiantes y las escuelas. Se calcula de forma simple utilizando las varianzas y covarianzas estimadas en los distintos modelos:

$$\text{Corr}_{u0u1} = \frac{\sigma_{u_0 u_1}}{\sigma_{u_0} * \sigma_{u_1}} \qquad \textbf{Ec. VIII.25}$$

La presentación de resultados del problema 3 cambia ligeramente porque los modelos no estiman los mismos coeficientes. Tampoco todos emplean el análisis multinivel para conseguir sus objetivos o, si lo hacen, no incluyen el tiempo como primer nivel de agregación. Por este motivo, no es posible calcular la correlación intraclase en todos los modelos de este problema y el cálculo de la correlación entre estatus y cambio también varía. Por ejemplo, la correlación

intraclase en el modelo de ganancia estimada (M3_3) se calcula de la siguiente forma:

$$\rho = \frac{u_{0j} + u_{1j} - 2\sigma_{u_0 u_1}}{(u_0 + u_1 - 2\sigma_{u_0 u_1}) + (r_0 + r_1 - 2\sigma_{r_0 r_1})} \qquad \textbf{Ec. VIII.26}$$

La manera de calcular la correlación entre estatus y cambio también es distinta en función de las distintas consideraciones de ese cambio que emplean los distintos modelos. En el modelo lineal mixto (M2_3) hay un mayor número de parámetros aleatorios, tanto en el nivel de los estudiantes como el delas escuelas. Por tanto, la ecuación incorporar esos valores.

En los modelos de ganancia residual (M4_3 y M4.1_3) es el residuo ajustado u_{0j} el que se relaciona con el estatus. Por último, en los modelos de ganancia bruta (M5_3, M5.1_3 y M5.2_3) la correlación se calcula de dos formas. En los dos modelos multinivel (M5.1_3 y M5.2_3), de la misma forma que en la ganancia residual, es el residuo ajustado u_{0j} el parámetro correlacionado con Y_{2j}. En el otro modelo es directamente la ganancia.

La mayor parte de modelos utilizan A2 como punto de partida. Los dos únicos modelos que utilizan A1 como estatus inicial son el modelo base M1_3 y su variación M1.2_3 que lleva a cabo el ajuste de los residuos a posteriori y que, por tanto, trata de eliminar el efecto de ese estatus inicial sobre el crecimiento. Mención a parte merecen las variantes del M2_3, donde el punto de partida cambia si se analiza un determinado curso o el periodo de evaluación completo.

En pequeña diferencia entre problemas en esta sección de resultados es que en los dos primeros, debido a que todos son modelos que estiman una pendiente de crecimiento, se incluye una tabla con las medias estimadas en cada una de las aplicaciones en función de los coeficientes de los modelos, es decir, teniendo en cuenta esa variación en la función de tiempo que se producen entre los modelos. Sin embargo, no se incluyen en el último problema porque la mayor parte son modelos de ganancia con dos únicas puntuaciones de logro.

VIII.2.4.2. Análisis de los residuos de las escuelas

Una vez analizados los coeficientes fijos y aleatorios de los distintos modelos, la atención se centra en los residuos estimados para las escuelas, que, en alguno de los casos, servirán como puntuaciones de VA. Para observar como fluctúan estas estimaciones del VA de las escuelas en función de las diferentes condiciones metodológicas se correlacionan (correlación de Pearson) los resultados producidos por los modelos de cada problema. Además se han elaborado rankings ordenados de las escuelas en función del residuo de crecimiento en los dos primeros problemas o de la puntuación utilizada para evaluar el rendimiento de las escuelas en los distintos modelos problema 3, no en todos ellos puede considerarse VA. También se llevan a cabo correlaciones no paramétricas (Spearman).

Los errores típicos de estimación de los residuos son otro de los elementos de análisis. Un modelo con errores más bajos puede ser tener una mayor capacidad para diferenciar mayor cantidad de escuelas distintas de la media.

Se lleva a cabo un estudio de esos residuos con mayor profundidad mediante dos tipos de gráficos. El primero relaciona el VA, la ganancia o la puntuación de rendimiento de las escuelas estimadas con los distintos modelos con los errores de estimación y el ranking. En cada problema se incluye un gráfico de cada modelo estimado donde aparecen las diferentes escuelas de la muestra ordenadas en función de la puntuación utilizada para valorar el logro de las escuelas, ya sea VA o no, y acompañadas de su intervalo de confianza al 95%. Para el estimar es intervalo se utiliza la error típico del residuo estimado, en el problema tres al contar con modelos que utilizan diferencias entre residuos ese error típico es también el de la diferencia:

$$\mathrm{ETdif} = \sqrt{\sigma_{\sigma_1}^2 + \sigma_{\sigma_2}^2 - 2r_{12}\sigma_{\sigma_1}\sigma_{\sigma_2}} \qquad \textbf{Ec. VIII.27}$$

El segundo tipo de gráfico es de dispersión. En ellos se relaciona el cambio en rendimiento con el estatus inicial de cada una de las escuelas. En el eje de abscisas se sitúa los puntos de partida de los centros educativos y en el eje de ordenadas el cambio o crecimiento. En los dos primeros problemas todos los

modelos estiman un estatus inicial y una pendiente de crecimiento para las escuelas a través de los residuos, estas son las variables que forman el gráfico de dispersión. Sin embargo, en el tercer problema el cambio en aprendizaje no se mide únicamente a través de la pendiente de crecimiento, cada modelo asume uno distinto. Por tanto, el gráfico de dispersión de cada modelo incluye como variable en el eje de ordenadas su propia definición de ganancia. También debe considerarse que los modelos pueden variar en sus puntos de partida pero para poder llevar a cabo una comparación, en este tercer problema, se utilizan las puntuaciones brutas iniciales en A1 o A2 o A3 dependiendo del modelo, como variable en el eje de abscisas.

Con el objetivo de detectar posibles cambios entre las posibles clasificaciones de los centros educativos, la información de los gráficos anteriores se resume en tablas de contingencia que comparan los resultados del modelo base con el resto. También existen dos tipos diferentes de tablas de contingencia dependiendo del gráfico al que hacen referencia.

El ranking de residuos de crecimiento representados junto su intervalo de confianza permite clasificar a las escuelas en tres grupos distintos: los centros educativos que no se diferencian significativamente de la media global, los que se encuentran significativamente por encima de esa media y, finalmente los que se encuentran por debajo. Se construyen tablas 3x3 que comparan los centros que se sitúan en esas categorías en el modelo de base y otro distinto. De esta manera es posible identificar cualquier cambio significativo en el ranking.

El gráfico de dispersión permite clasificar a las escuelas por cuadrantes. El primer cuadrante (superior-izquierda) agrupa a aquellos centros educativos que tienen un estatus inicial por debajo de la media pero un crecimiento por encima; un segundo cuadrante (inferior-izquierda) que sitúa a los que también empiezan con un nivel de logro inferior pero tampoco superan en nivel de crecimiento a esa media global; El tercero (superior-derecha) incluye escuelas que tienen un nivel inicial de rendimiento superior a la media global y crecen también a un mayor ritmo; y, por último, el cuarto cuadrante (inferior-derecha) está formado por los centros educativos que comienzan por encima de la media per su nivel de crecimiento no alcanza a la media global.

Junto con las tablas de contingencia se incluye el valor del estadístico chi-cuadrado (χ^2) estimado para probar la relación entre esas categorías entre los modelos probados. Además del valor de la probabilidad asociada a ese valor de χ^2.

VIII.3 Resultados

VIII.3.1 Problema 1. Selección de una medida adecuada de tiempo

El diseño específico de recogida de información con tomas de datos al inicio y final de curso tiene como consecuencia contar con mediciones del logro que no están igualmente distanciadas en el tiempo. Además, entre las distintas mediciones se encuentra un periodo de verano y un cambio en la facilidad de los ítems comunes entre A2 y A3, factores que pueden determinar que medida del tiempo se emplea en el modelo. Conviene recordar los distintos modelos estimados en este primer problema que varían en los valores de esa función de tiempo, son los siguientes:

M1_1:	T_1(0 8 13 20)
M2_1:	T_2(0 8 $\left(8+{}^{5}/_{2,5}=10\right)$ 17)
M3_1:	T_3(0 8 $(8+5*2,5=20,5)$ 27,5)
M4_1:	T_4(0 1 2 3)
M5_1:	T_5(0 1 $\left(1+{}^{1}/_{2,5}=1,4\right)$ 2,4)
M6_1:	T_6(0 1 $(1+1*2,5=3,5)$ 4,5)

Tabla VIII.5. Resumen de la variable Tiempo en los modelos del Problema 1.

Los dos factores que hacen variar la función de tiempo entre modelos son la distancia entre ocasiones de medida y el factor de corrección del aumento en la facilidad de los ítems comunes entre A2 y A3.

VIII.3.1.1. Coeficientes estimados

La elección de unos valores u otros en la función de crecimiento afecta a la estimación de los coeficientes fijos y aleatorios de los distintos modelos como se puede observar en la Tabla VIII.6. En esta tabla los coeficientes de cada modelo aparecen separados por columnas. Cada coeficiente se acompaña de su error típico (ET). La tabla separa los efectos fijos de los aleatorios y estos últimos están separados por niveles: el nivel tres hace referencia a las escuelas, el dos a los

estudiantes y el uno es el residual del modelo de regresión. Finalmente, en los dos últimos apartados de la tabla se muestran, junto con los resultados del ajuste global del modelo (*deviance*), la correlación intraclase o, más bien, la cantidad de varianza que se debe a los estudiantes (p_r) y cuánta a las escuelas (p_u). También se incluye la información de la relación entre el estatus inicial y el crecimiento en los niveles de estudiantes (p_r0*r1) y de escuelas (p_u0*_u1). Y la cantidad de varianza reducida en estos mismos niveles respecto al modelo nulo.

	M1_1		M2_1		M3_1		M4_1		M5_1		M6_1	
EFECTOS FIJOS												
	Coef.	**ET**	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**
β0	247,170	2,260	247,407	2,237	249,261	2,263	248,603	2,258	248,294	2,239	251,685	2,253
β1	4,192	0,060	4,869	0,073	2,922	0,040	27,711	0,387	34,845	0,510	17,133	0,230
EFECTOS ALEATORIOS												
NIVEL 3	**u0**	**u1**	**u0**	**u1**	**u0**	**u1**	**u0**	**u1**	**u0**	**u1**	**u0**	**u1**
u0	291,774		290,435		294,261		292,301		286,209		292,445	
ETu0	58,088		56,962		58,29		58,023		57,031		57,782	
u1	-2,676	0,163	-3,134	0,236	-1,854	0,07	-17,758	6,646	-21,060	11,684	-10,549	2,288
ETu1	1,189	0,041	1,407	0,06	0,791	0,018	7,646	1,696	9,875	2,935	4,521	0,596
NIVEL 2	**r0**	**r1**	**r0**	**r1**	**r0**	**r1**	**r0**	**r1**	**r0**	**r1**	**r0**	**r1**
r0	1154,862		837,937		1187,039		1174,263		1080,520		1158,456	
ETr0	41,048		26,355		39,393		39,925		37,028		38,292	
r1	-16,443	0,434	0*	0*	-13,393	0,356	-1209	27,283	-98,320	4,465*	-74,124	10,795
ETr0	1,526	0,088	-	-	0,998	0,039	9,665	3,637	8,268	6,500	5,700	1,352
NIVEL 1												
e	479,956		592,303		407,073		433,927		556,313		418,182	
ET	9,359		9,351		7,971		8,477		8,781		8,193	
Deviance	106188,561		107564,716		105277,603		105633,175		106914,723		105376,709	
p_u	0,151		0,172		0,155		0,151		0,156		0,154	
p_r	0,605		0,485		0,633		0,652		0,588		0,642	
p_u0*u1	-0,388		-0,379		-0,409		-0,403		-0,364		-0,408	
p_r0*r1	-0,734		-		-0,652		-0,670		-		-0,663	
				%		**%**		**%**		**%**		**%**
Reducción N3			-0,350	-0,118	0,750	0,252	37,174	12,504	42,724	14,371	18,542	6,237
Reducción N2			-350,245	-29,477	25,999	2,188	253,382	21,325	88,978	7,489	129,317	10,884

*Coeficiente no significativo

Tabla VIII.6. Coeficientes, errores típicos, ajuste, correlaciones intraclase y correlación entre estatus inicial y crecimiento en el Problema 1.

En primer lugar, el análisis de los coeficientes muestra valores similares para el estatus inicial medio (β_0). Aunque los modelos que emplean el número de ocasiones de medida en la función de tiempo (M4_1, M5_1 y M6_1) son ligeramente superiores. Si se lleva a cabo una comparación entre pares de modelos, aquellos que multiplican la distancia entre A2 y A3 por el factor de corrección (M3_1 y M6_1) tienen el punto de partida más alto respecto a los que dividen por ese factor y los que no llevan a cabo ninguna modificación. Como era de esperar las pendientes de crecimiento (β_1) medias son mayores a medida que se reduce la distancia entre ocasiones de medida. Los coeficientes son mayores en los modelos que utilizan el número de ocasiones de medida en la variable tiempo.

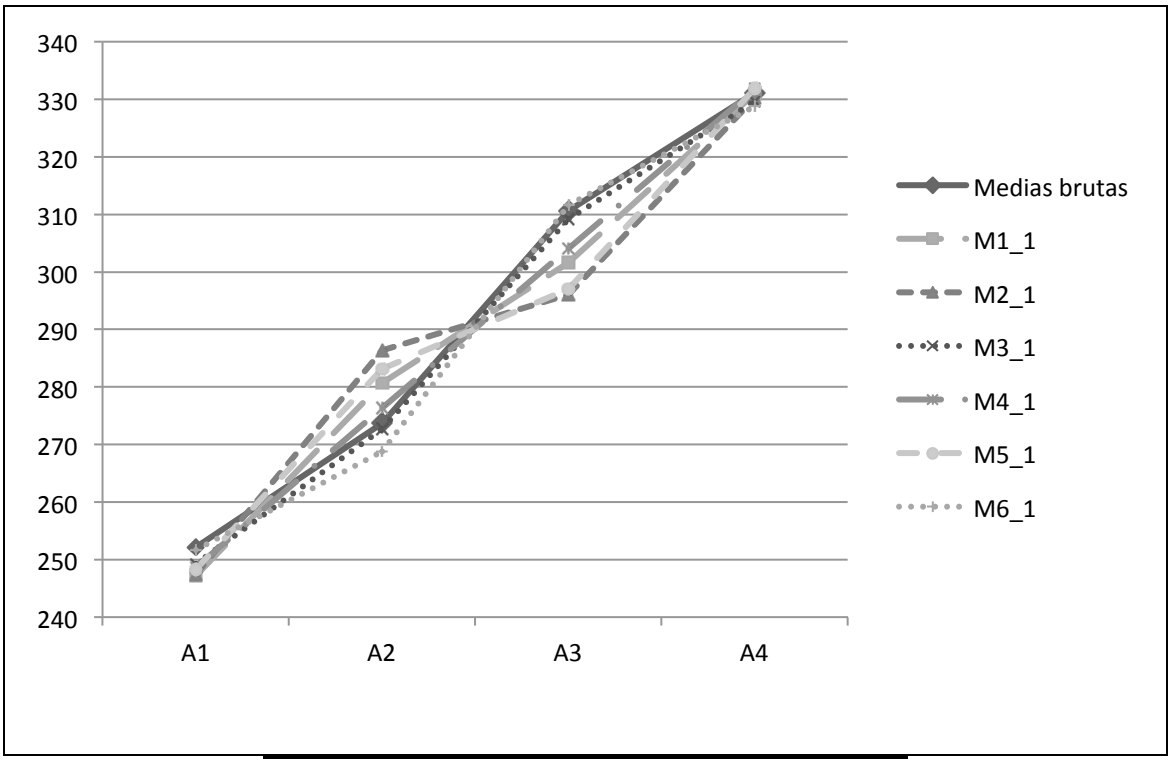
En segundo lugar, el análisis de los efectos aleatorios muestra algunos aspectos reseñables. El término residual de primer nivel (e) es menor cuando los modelos multiplican por el factor de corrección (M3_1 y M6_1), son también estos los que obtienen ligeramente mejor índice de ajuste con unos valores de *deviance* más bajos. En el nivel 2 (estudiantes) los modelos que dividen por el factor de corrección (M2_1 y M5_1) reducen la varianza en los puntos de partida (r_0), aunque en menor medida en el que utiliza el factor de corrección sobre el número de ocasiones de medida (M5_1). Estos modelos también eliminan la varianza significativa entre las pendientes de crecimiento (r_1) de los estudiantes. En el caso del modelo que utiliza la distancia en meses (M2_1) la covarianza entre el estatus inicial y el crecimiento deja de ser significativa. En cambio, la varianza de las pendientes de crecimiento (u_1) en el tercer nivel (escuelas) no deja de ser significativa al incluir el factor de corrección lo que indica que las escuelas crecen a ritmos distintos incluso con la corrección. La tendencia en este nivel de aleatorización es una mayor varianza en el crecimiento a medida que disminuye la distancia entre ocasiones de medida, es decir, los modelos que utilizan el número de ocasiones de medida obtienen mayores valores de varianza entre las pendientes de crecimiento de las escuelas.

En tercer y último lugar, las correlaciones analizadas también muestran algunas variaciones. De nuevo, el modelo M2_1 destaca por ser el que determina una mayor proporción de varianza debida a las escuelas, con un valor de la

autocorrelación (p_u) de 0,172, es decir, un 17,2% de la varianza total se debe a las escuelas. Un 48,5% se debe a la varianza entre estudiantes, el resto es el término residual de primer nivel (e), un 34,3%. También es el único modelo que consigue reducir la varianza explicada en el nivel 2 y 3 respecto al modelo de base (M1_1). En el resto de modelos la varianza residual de primer nivel oscila alrededor del 20%-25%

En resumen, los modelos que tratan de asemejarse a los resultados brutos aumentando el valor de la función de tiempo entre A2 y A3 multiplicándolo por el cambio en el nivel de facilidad de los ítems comunes (M3_1 y M6_1) parecen presentar unos mejores resultados de ajuste con menor *deviance* y menor valor del término residual de primer nivel. Pero esto no quiere decir que sean más adecuados para la evaluación. El modelo que tratan de paliar ese efecto del cambio en la facilidad incluyendo una reducción del término de la función de tiempo entre A2 y A3 y utiliza la distancia en meses entre aplicaciones (M2_1) reduce la varianza aleatoria entre estudiantes respecto al modelo base y uno de los objetivos del valor añadido es eliminar esos posibles efectos del contexto de los estudiantes. Sin embargo, estos modelos obtienen los valores más altos en el índice de ajuste.

Para poder extraer una conclusión es necesario observar el resto de resultados. A continuación se representa de forma numérica (Tabla VIII.7) y gráfica (Gráfico VIII.1) las medias estimadas en los distintos modelos en comparación con las medias brutas en cada una de las cuatro ocasiones de medida.



	A1	A2	A3	A4
Medias brutas	252,155	273,709	310,686	331,112
M1_1	247,170	280,706	301,666	331,010
M2_1	247,407	286,359	296,097	330,180
M3_1	249,261	272,637	309,162	329,616
M4_1	248,603	276,314	304,025	331,736
M5_1	248,298	283,141	297,078	331,921
M6_1	251,685	268,818	311,650	328,783

Gráfico VIII.1. y Tabla VIII.7. Medias brutas y medias estimadas con los seis modelos

Los resultados de comparación de las medias estimadas en las distintas ocasiones de medida indican tres patrones distintos de crecimiento relacionados con la utilización del factor de corrección. Dividir la distancia entre A2 y A3 por el factor de corrección (M2_1 y M5_1) suaviza el cambio entre esas dos aplicaciones que puede ser debido al cambio detectado en la facilidad de los ítems comunes u otros aspectos no controlables. En cambio, multiplicar por ese factor de corrección hace que las medias estimadas en estos modelos (M3_1 y M6_1) se asemejen a las medias brutas. Finalmente, los modelos que no introducen el factor de corrección también se parecen aunque utilizar meses en la función de tiempo suaviza el cambio entre A2 y A3, característica acorde con la realidad educativa de la evaluación, es decir, el análisis del cambio en el periodo de verano entre dos cursos académicos.

VIII.3.1.2. Análisis de los residuos de las escuelas

Los residuos asociados a la pendiente de crecimiento estimados con estos modelos en el tercer nivel de agregación, se considera el valor añadido de las escuelas. Uno de los objetivos de las evaluaciones basadas en estas estimaciones es detectar escuelas que diferencian significativamente de la media. En ese caso, los modelos con menores valores de los errores típicos asociados a esos residuos de las escuelas pueden ser más aconsejables.

Se han calculado correlaciones paramétricas (Pearson) entre los residuos de tercer nivel estimados con los distintos modelos de este primer problema (Tabla VIII.8). También se han calculado correlaciones entre los rankings de las escuelas, en este caso han sido no paramétricas, concretamente rho de Spearman (Tabla VIII.9). Estas correlaciones ayudan a conocer cuáles son las estimaciones de VA que más se parecen.

		u1_M1_1	u1_M2_1	u1_M3_1	u1_M4_1	u1_M5_1	u1_M6_1
u1_M1_1	**Pearson**	1	,994	,986	,996	,997	,966
	Sig		,000	,000	,000	,000	,000
u1_M2_1	**Pearson**		1	,967	,981	,996	,943
	Sig			,000	,000	,000	,000
u1_M3_1	**Pearson**			1	,994	,974	,995
	Sig				,000	,000	,000
u1_M4_1	**Pearson**				1	,990	,981
	Sig					,000	,000
u1_M5_1	**Pearson**					1	,952
	Sig						,000
u1_M6_1	**Pearson**						1

Tabla VIII.8. Correlaciones de Pearson entre las estimaciones de VA en el Problema 1.

Los valores de las correlaciones entre los residuos son altos y positivos, lo que señala que los residuos de todos los modelos van en la misma dirección. Los valores de Pearson más elevados se dan entre los modelos que no incluyen ningún tipo de corrección con el resto y son algo superiores con los modelos que dividen por el factor de corrección. El valor más bajo (0,943) se da entre el modelo que utiliza meses y divide por el factor de corrección (M2_1) y el que, en cambio, utiliza el número de ocasiones y multiplica por ese mismo factor (M6_1).

		Ru1_M1_1	Ru1_M2_1	Ru1_M3_1	Ru1_M4_1	Ru1_M5_1	Ru1_M6_1
Ru1_M1_1	**Spearman**	1	,990	,982	,993	,995	,961
	Sig	.	,000	,000	,000	,000	,000
Ru1_M2_1	**Spearman**		1	,965	,976	,991	,941
	Sig			,000	,000	,000	,000
Ru1_M3_1	**Spearman**			1	,993	,970	,992
	Sig				,000	,000	,000
Ru1_M4_1	**Spearman**				1	,986	,977
	Sig					,000	,000
Ru1_M5_1	**Spearman**					1	,946
	Sig						,000
Ru1_M6_1	**Spearman**						1

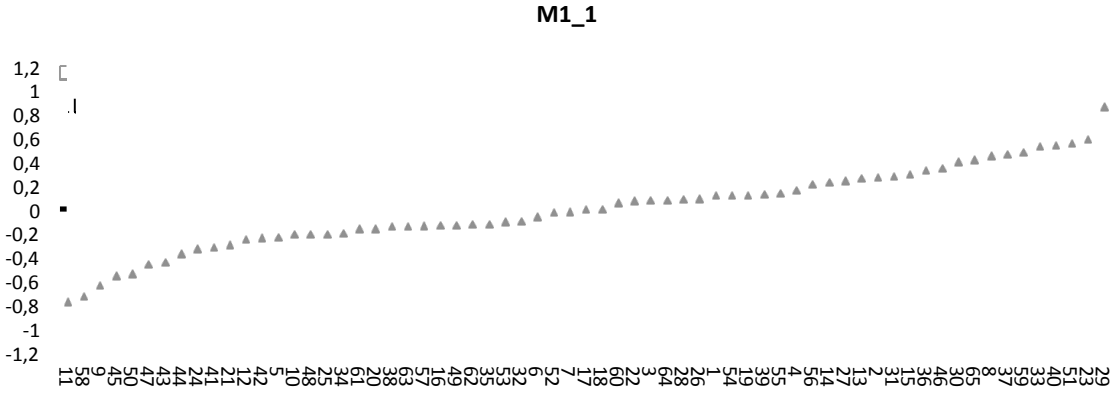
Tabla VIII.9. Correlaciones Rho de Spearman entre los rankings de escuelas del Problema 1.

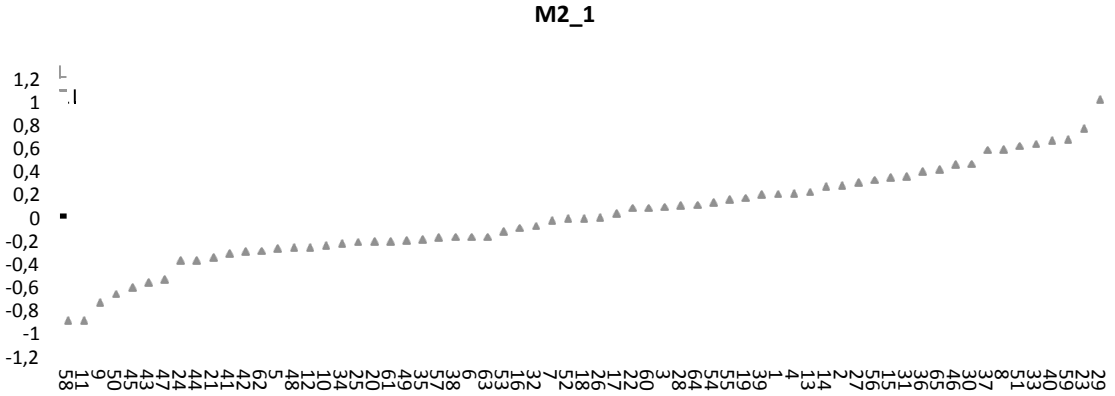
Una tendencia similar a la encontrada entre las correlaciones de los residuos de las escuelas se puede ver entre las distintas clasificaciones de esos centros educativos.

Los valores de las correlaciones señalan que los residuos estimados a través de los distintos modelos son similares. No obstante, puede haber ligeros cambios en las posiciones que ocupan las escuelas. Para comprobar este fenómeno se han construido dos tipos de gráficos. El primero representa el residuo de la pendiente de crecimiento asociado a cada escuela junto con su intervalo de confianza al 95% y el segundo representa a los centros educativos en cuatro cuadrantes distintos en función de su crecimiento y estatus inicial.

La información proporcionada por los gráficos se resume en tablas de contingencia para comparar los resultados obtenidos mediante el modelo base (M1_1) con el resto. Para los gráficos de ranking los residuos se agrupan en tres categorías que informan sobre si las escuelas son iguales o significativamente distintas de la media global (igual a la media, superior a la media y inferior a la media). Para el caso de los gráficos de dispersión se crean las categorías en función de los cuatro cuadrantes delimitados y ya mencionados. La información de estas tablas ya se ha especificado con más detalle en el apartado VIII.2.4 de este capítulo sobre la metodología de comparación de resultados.

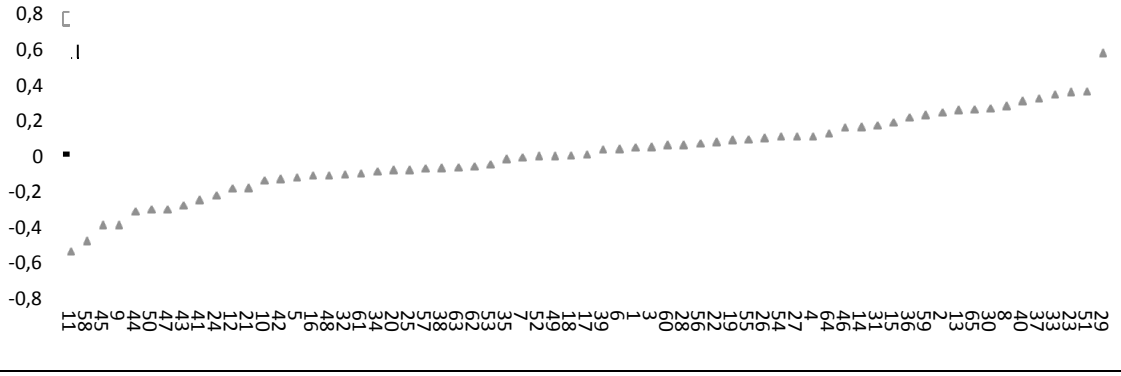
A continuación, la Figura VIII.1 incluye los resultados y cambios entre las posiciones que ocupan las escuelas en la ordenación determinada por el residuo de crecimiento de los seis modelos.





			R_cat_M2_1		
			Igual a la media	**Superior a la media**	**Inferior a la media**
R_cat_M1_1	**Igual a la media**	Recuento	11	0	0
		%	100%	0%	0%
	Superior a la media	Recuento	0	26	0
		%	0%	100%	0%
	Inferior a la media	Recuento	0	0	28
		%	0%	0%	100%

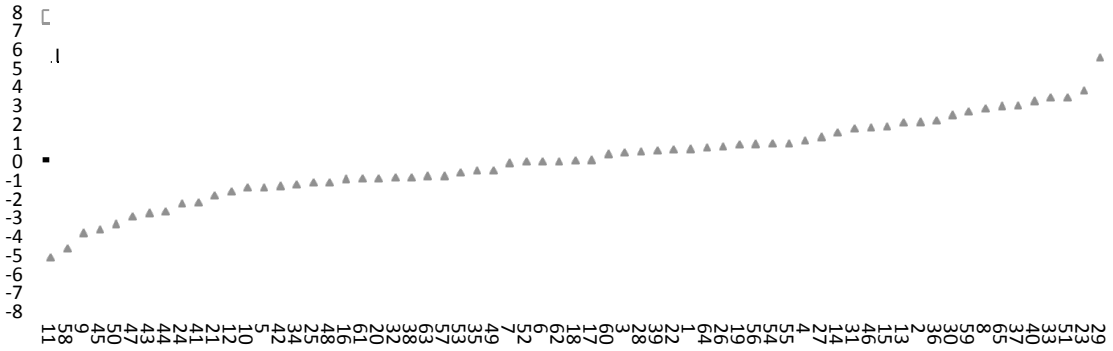
Chi-cuadrado de Pearson= 130,00 P=0,000



			R_cat_M3_1		
			Igual a la media	**Superior a la media**	**Inferior a la media**
R_cat_M1_1	**Igual a la media**	Recuento	7	3	1
		%	63,6%	27,3%	9,1%
	Superior a la media	Recuento	1	25	0
		%	3,80%	96,2%	0%
	Inferior a la media	Recuento	2	1	25
		%	7,10%	3,60%	89,3%

Chi-cuadrado de Pearson= 76,957 P=0,000

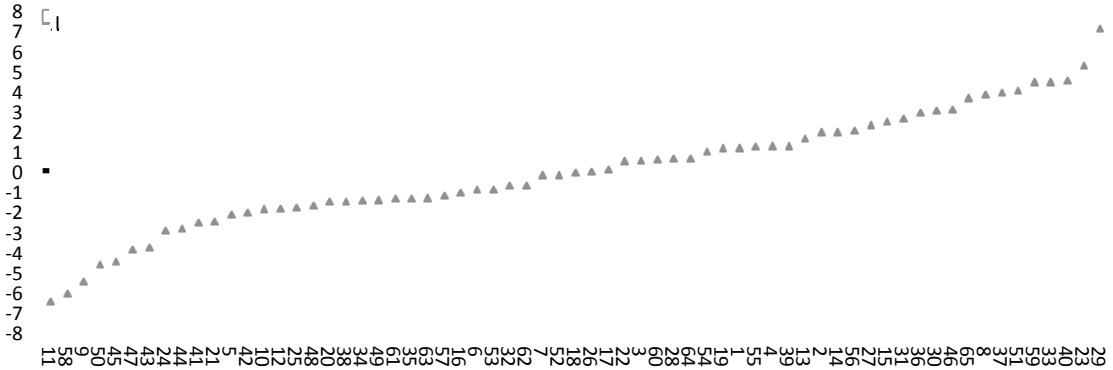
M4_1



			R_cat_M4_1		
			Igual a la media	**Superior a la media**	**Inferior a la media**
R_cat_M1_1	**Igual a la media**	Recuento	7	3	1
		%	63,64%	27,27%	9,09%
	Superior a la media	Recuento	0	26	0
		%	0%	100%	0%
	Inferior a la media	Recuento	2	0	26
		%	7,14%	0%	92,86%

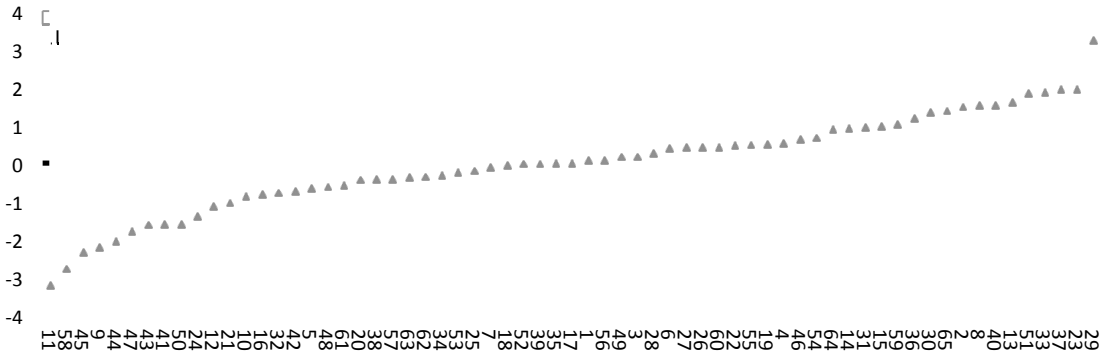
Chi-cuadrado de Pearson= 86,654 P=0,000

M5_1



			R_cat_M5_1		
			Igual a la media	**Superior a la media**	**Inferior a la media**
R_cat_M1_1	**Igual a la media**	Recuento	11	0	0
		%	100%	0%	0%
	Superior a la media	Recuento	0	26	0
		%	0%	100%	0%
	Inferior a la media	Recuento	1	0	27
		%	3,57%	0%	96,43%

Chi-cuadrado de Pearson= 122,455 P=0,000



			R_cat_M6_1		
			Igual a la media	**Superior a la media**	**Inferior a la media**
R_cat_M1_1	**Igual a la media**	Recuento	6	4	1
		%	54,55%	36,36%	9,09%
	Superior a la media	Recuento	4	22	0
		%	15,38%	84,62%	0%
	Inferior a la media	Recuento	4	1	23
		%	14,29%	3,57%	82,14%

Chi-cuadrado de Pearson= 55,522 P=0,000

Figura VIII.1. Grafico del Valor Añadido de las escuelas e Intervalo de Confianza al 95% y tablas de contingencia que reflejan los cambios en las posiciones respecto al modelo base en problema 1.

Son muy pocos los cambios en la posición que ocupan las escuelas en los rankings elaborados con las estimaciones de los seis modelos. No hay ningún cambio entre la clasificación del modelo base (M1_1) y el que divide la función de tiempo por el factor de corrección (M2_1), ambos detectan 11 escuelas igual a la media y, por tanto, 64 que no lo son. No obstante, hay algún cambio significativo con respecto a los modelos que utilizan el número de ocasiones de medida en la función de tiempo, con mayor impacto cuando la función de tiempo multiplica por el factor de corrección, independientemente si se utiliza meses u ocasiones de medida. Con respecto al M3_1, cuatro escuelas con un residuo igual a la media cambian, tres pasan a tenerlo superior a la media y uno de ellos inferior. También

tres centros educativos con resultados del residuo de crecimiento inferior a la centro pasan a tenerlo igual o superior. Este modelo estima tres centros más que el anterior por encima de la media y un total de 55 centros distintos de la media, uno más que los dos anteriores. El M4_1 es el modelo que identifica mayor número de centros con residuos de crecimiento significativamente distintos de la media, un total de 56. El resto de modelos que utilizan el número de ocasiones de medida (M5_1 y M6_1) identifican un menor número de escuelas distintas de la media, 53 y 51 respectivamente.

Los gráficos de dispersión (Gráfico VIII.2) muestran también pocos cambios de posición de las escuelas entre cuadrantes si se comparan con el modelo base de crecimiento.

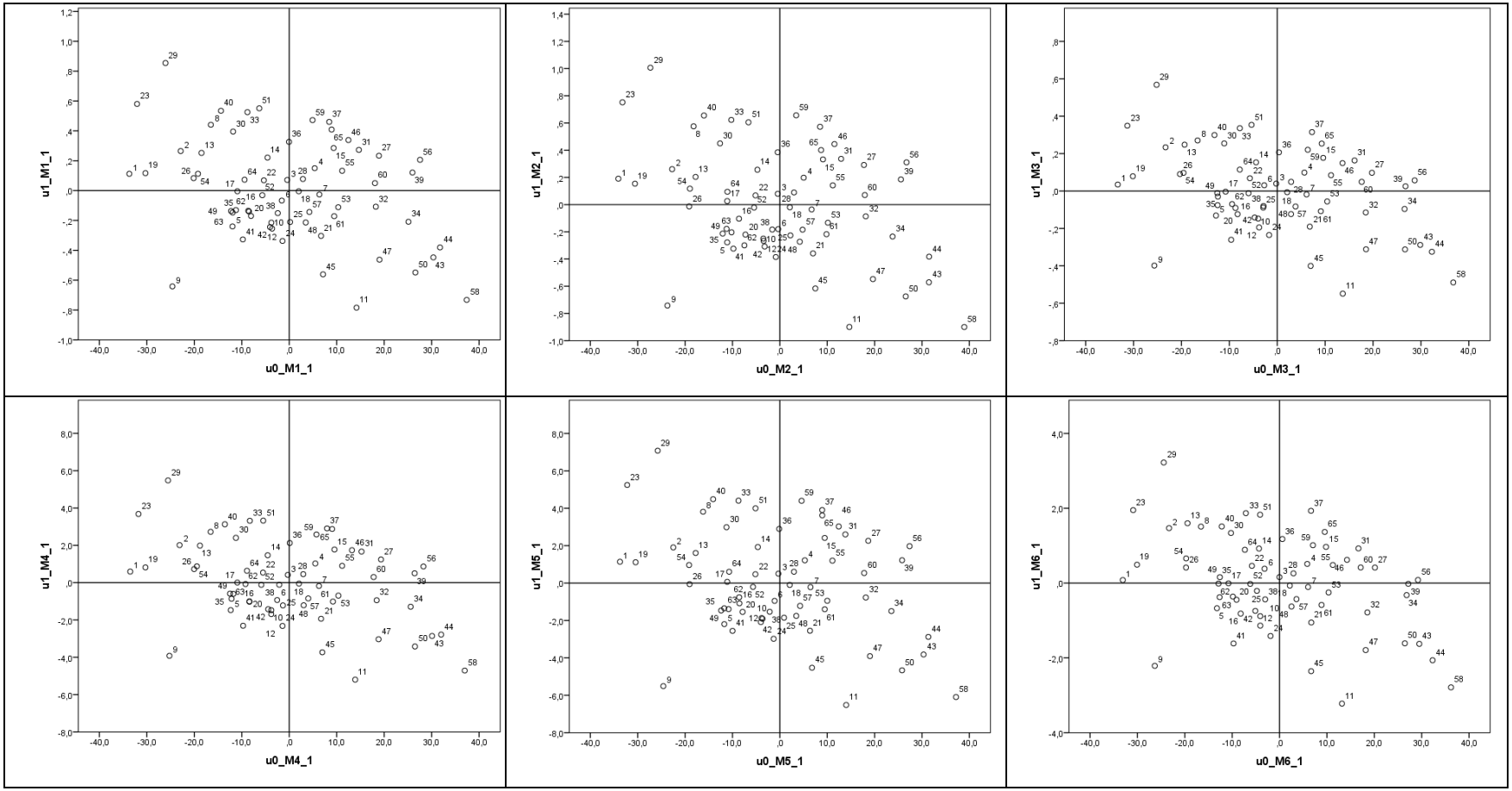


Gráfico VIII.2. Gráficos de dispersión de los residuos de las escuelas (u0*u1) en los modelos del problema 1.

Únicamente uno o dos centros cambian de cuadrante entre modelos como muestra la Tabla VIII.10. Por ejemplo, en la comparación del M1_1 y M2_1 una escuela (la número 26) con bajo estatus inicial y alto crecimiento pasa a tener bajo crecimiento y otra escuela cambia en el sentido opuesto (escuela 17).

			Disp_cat_M2_1			
			Bajo Estatus y Alto Crecimiento	**Bajo Estatus y Bajo Crecimiento**	**Alto Estatus y Alto Crecimiento**	**Alto Estatus y Bajo Crecimiento**
Disp_cat_M1_1	**Bajo Estatus y Alto Crecimiento**	Recuento	17	1	0	0
		%	94,4%	5,6%	,0%	,0%
	Bajo Estatus y Bajo Crecimiento	Recuento	1	16	0	0
		%	5,9%	94,1%	,0%	,0%
	Alto Estatus y Alto Crecimiento	Recuento	0	0	13	0
		%	,0%	,0%	100%	,0%
	Alto Estatus y Bajo Crecimiento	Recuento	0	0	0	17
		%	,0%	,0%	,0%	100%

Chi-cuadrado de Pearson= 180,981 P=0,000

			Disp_cat_M3_1			
			Bajo Estatus y Alto Crecimiento	**Bajo Estatus y Bajo Crecimiento**	**Alto Estatus y Alto Crecimiento**	**Alto Estatus y Bajo Crecimiento**
Disp_cat_M1_1	**Bajo Estatus y Alto Crecimiento**	Recuento	17	0	1	0
		%	94,4%	,0%	5,6%	,0%
	Bajo Estatus y Bajo Crecimiento	Recuento	1	16	0	0
		%	5,9%	94,1%	,0%	,0%
	Alto Estatus y Alto Crecimiento	Recuento	0	0	13	0
		%	,0%	,0%	100%	,0%
	Alto Estatus y Bajo Crecimiento	Recuento	0	1	0	16
		%	,0%	5,9%	,0%	94,1%

Chi-cuadrado de Pearson= 172,785 P=0,000

			Disp_cat_M4_1			
			Bajo Estatus y Alto Crecimiento	**Bajo Estatus y Bajo Crecimiento**	**Alto Estatus y Alto Crecimiento**	**Alto Estatus y Bajo Crecimiento**
Disp_cat_M1_1	**Bajo Estatus y Alto Crecimiento**	Recuento	17	0	1	0
		%	94,4%	,0%	5,6%	,0%
	Bajo Estatus y Bajo Crecimiento	Recuento	1	16	0	0
		%	5,9%	94,1%	,0%	,0%
	Alto Estatus y Alto Crecimiento	Recuento	0	0	13	0
		%	,0%	,0%	100%	,0%
	Alto Estatus y Bajo Crecimiento	Recuento	0	1	0	16
		%	,0%	5,9%	,0%	94,1%

Chi-cuadrado de Pearson= 172,785 P=0,000

			Disp_cat_M5_1			
			Bajo Estatus y Alto Crecimiento	**Bajo Estatus y Bajo Crecimiento**	**Alto Estatus y Alto Crecimiento**	**Alto Estatus y Bajo Crecimiento**
Disp_cat_M1_1	**Bajo Estatus y Alto Crecimiento**	Recuento	17	1	0	0
		%	94,4%	5,6%	,0%	,0%
	Bajo Estatus y Bajo Crecimiento	Recuento	1	16	0	0
		%	5,9%	94,1%	,0%	,0%
	Alto Estatus y Alto Crecimiento	Recuento	0	0	13	0
		%	,0%	,0%	100%	,0%
	Alto Estatus y Bajo Crecimiento	Recuento	0	0	0	17
		%	,0%	,0%	,0%	100%

Chi-cuadrado de Pearson= 180,981 P=0,000

			Disp_cat_M6_1			
			Bajo Estatus y Alto Crecimiento	**Bajo Estatus y Bajo Crecimiento**	**Alto Estatus y Alto Crecimiento**	**Alto Estatus y Bajo Crecimiento**
Disp_cat_M1_1	**Bajo Estatus y Alto Crecimiento**	Recuento	17	0	1	0
		%	94,4%	,0%	5,6%	,0%
	Bajo Estatus y Bajo Crecimiento	Recuento	2	15	0	0
		%	11,8%	88,2%	,0%	,0%
	Alto Estatus y Alto Crecimiento	Recuento	0	0	12	1
		%	,0%	,0%	92,3%	7,7%
	Alto Estatus y Bajo Crecimiento	Recuento	0	1	0	16
		%	,0%	5,9%	,0%	94,1%

Chi-cuadrado de Pearson= 158,274 P=0,000

Tabla VIII.10. Tablas de contingencia y χ^2 para la relación. Cambios en los cuadrantes del gráfico de dispersión respecto al modelo base en el problema 1.

En conclusión, introducir la corrección entre los meses transcurridos entre la aplicación A2 y A3 en el modelo que utiliza la función de tiempo (M5_1) no produce cambios sustanciales.

Recordemos que existen factores que pueden alterar ese cambio además del mencionado cambio en la facilidad de los ítems de anclaje que formaron parte de las pruebas en esas aplicaciones. Por ejemplo, la tercera medición se llevó a cabo en el mes de Noviembre, cuando el curso escolar comienza en Septiembre. Durante esos dos meses transcurridos los estudiantes ya han estado bajo la influencia de la acción de las escuelas y, normalmente, al inicio del curso se repasan los principales contenidos del curso anterior.

No obstante, multiplicar por ese mismo factor para situar el cambio entre esas aplicaciones en la misma tasa que los datos brutos (M3_1) parece producir más cambios respecto al modelo base, sobre todo, si el modelo utiliza el nº de aplicaciones en la función de tiempo (M6_1). En este caso se obtiene el valor de χ^2 más bajo (158,274)

Utilizar el factor de corrección con el nº de aplicaciones (M5_1) provoca que las clasificaciones de las escuelas sean más similares al modelo base que el M4_1 que no introduce ninguna corrección sobre esas aplicaciones. El valor de χ^2 es más alto en el M5_1, similares a los que se obtienen en la relación entre M1_1 y M2_1. Por tanto, dividir por el factor de corrección produce que los resultados del modelo que utiliza el número de ocasiones de medida se asemejen, en mayor medida, al modelo base (M1_1) que utiliza los meses como unidad de tiempo.

VIII.3.2 Problema 2: Relación entre estatus inicial y crecimiento y efecto de regresión hacia la media

Este problema estudia la relación existente entre el estatus inicial y la pendiente de crecimiento y sus posibles efectos en las estimaciones de VA. Esa relación puede ser un síntoma provocado por algún artefacto del diseño como el ERM, que suele producirse en los estudios de cambio en aprendizaje. Sin embargo, diferenciar el sesgo producido por ese artefacto concreto del diseño de otro producido, por ejemplo, por un mayor error de estimación la puntuación de rendimiento de alguna de las aplicaciones, las características de la propia escala

vertical o la utilización de estimadores BLUP es una tarea imposible. Y estos factores también pueden alterar esa relación entre estatus inicial y crecimiento

Otros factores característicos de la etapa educativa evaluada como el ya mencionado inicio de ciclo en secundaria o la falta de experiencia en este tipo de evaluaciones también pueden sesgar la primera toma de datos de rendimiento. Por tanto, cambiar el punto de partida y comprobar el efecto que produce tanto en la relación como las estimaciones de VA es una prueba necesaria. Los modelos que se prueban son los siguientes:

M1_2:	T_1(0 8 13 20)
M2_2:	T_2(−8 0 5 12)
M3_2:	T_3(−13 5 0 7)
M4_2:	T_4(0 8 13 20); Y'_{1i}
M5_2:	T_5(0 5 12); Y_{1i}

Tabla VIII.11. Resumen de los modelos estimados en el problema 2.

Antes de comenzar con la comparación de modelos se han estudiado las cuatro puntuaciones de rendimiento para observar patrones de correlación y otros factores que determinen la existencia de algún tipo de problema. Nesselroade, Stigler y Baltes (1980) destacan que los diseños con dos únicas mediciones pueden maximizar los problemas asociados al ERM. En los modelos con más de dos ocasiones de medida esos efectos dependerán de los patrones de correlación que se produzcan entre mediciones. Si el patrón de correlaciones es constante no se debería esperar ese efecto más allá de la segunda ocasión de medida. En cambio, un patrón decreciente en las correlaciones puede ser indicador de ese efecto a lo largo de todas las mediciones.

En la Tabla VIII.12 se pueden observar los patrones de correlación entre las puntuaciones de rendimiento de las cuatro aplicaciones.

		Rasgo_A1	Rasgo_A2	Rasgo_A3	Rasgo_A4
Rasgo_A1	Pearson	1	,728	,698	,651
	Sig. (bilateral)		,000	,000	,000
Rasgo_A2	Pearson		1	,739	,714
	Sig. (bilateral)			,000	,000
Rasgo_A3	Pearson			1	,734
	Sig. (bilateral)				,000
Rasgo_A4	Pearson				1

Tabla VIII.12. Correlaciones entre las cuatro medidas de rendimiento.

Existe una correlación imperfecta entre A1 y A2 pero su valor es similar al que se producen entre A2 y A3 y entre A3 y A4, aproximadamente 0,7. El valor de la correlación disminuye ligeramente a medida que aumenta la distancia con la siguiente ocasión de medida, aunque no es un cambio demasiado destacable. Esta estabilización del coeficiente a lo largo de las aplicaciones muestra que el posible ERM tendería a desaparecer. Por tanto, el modelo longitudinal no se vería muy afectado.

Aun así, esa correlación imperfecta entre las dos primeras aplicaciones podría indicar un posible ERM. Pero puede ser también que esa correlación se deba a una medición imperfecta en alguna de las aplicaciones. El análisis de los errores típicos de estimación pueden ser de ayuda para detectar alguna alteración.

Los errores típicos de estimación medios que se muestran en la Tabla VIII.13 indican un mayor sesgo en las estimaciones del rasgo de la primera ocasión de medida (A1). Con valores medios superiores.

	Media	**Varianza**
ET_Rasgo_A1	21,669	6,406
ET_Rasgo_A2	18,719	7,765
ET_Rasgo_A3	15,910	3,719
ET_Rasgo_A4	16,327	5,067

Tabla VIII.13. Media y Varianza de los errores típicos de estimación.

Por tanto, existen indicios que señalan una mayor incertidumbre en la estimación de la puntuación de rendimiento en la primera toma de datos. Características de la etapa educativa evaluada y factores estadísticos ponen de manifiesto los posibles problemas de tomar como referente ese punto de partida.

VIII.3.2.1. Coeficientes estimados

Rogosa (1995) demuestra que el efecto de regresión hacia la media (ERM) solo se produce cuando existe una correlación negativa entre estatus inicial y cambio pero variar la ocasión de medida que se utiliza como estatus inicial también puede cambiar la dirección de esa correlación.

Recordemos que los tres primeros modelos (M1_2, M2_3 y M3_2) cambian la posición del punto de partida. El primer modelo utiliza la primera aplicación (A1) como punto de partida, el segundo A2 y el tercero A3. M4_2 y M5_2 incluyen

el rendimiento previo como covariable de las dos formas descritas. En la Tabla VIII.14 se muestran lo coeficientes fijos y aleatorios estimados con los cinco modelos.

	M1_2		M2_2		M3_2		M4_2		M5_2	
EFECTOS FIJOS										
	Coef.	**ET**	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**
β0	247,170	2,260	280,753	2,083	301,679	2,055	247,216	2,231	278,264	1,354
β1	4,192	0,060	4,189	0,060	4,166	0,054	4,166	0,054	4,761	0,095
β2							-03	0,00	0,568	0,010
EFECTOS ALEATORIOS										
NIVEL 3	**u0**	**u1**	**u0**	**u1**	**u0**	**u1**	**u0**	**u1**	**u0**	**u1**
u0	291,774		254,159		249,339		283,273		10,00,038	
ETu0	58,088		49,269		49,269		56,282		20,812	
u1	-2,676	0,163	-1,371*	0,159	-0,554*	0,163	0,229*	0,13	-3,651	0,395
ETu1	1,189	0,041	1,045	0,04	11	0,041	0,971	0,033	1,178	0,1
NIVEL 2	**r0**	**r1**	**r0**	**r1**	**r0**	**r1**	**r0**	**r1**	**r0**	**r1**
r0	1154,862		919,988		80,00,725		1076,559		285,953	
ETr0	41,048		28,030		25,136		32,393		12,599	
r1	-16,443	0,434	-13,009	0,435	-10,813	0,434	0,00	0,00	0,00	0,00
ETr0	1,526	0,088	1,088	0,088	0,983	0,088	0,00	0,00	0,00	0,00
NIVEL 1										
e	479,956		479,944		479,951		469,094		450,013	
ET	9,359		9,358		9,359		7,565		9,785	
Deviance	106188,561		106190,461		106188,88		101724,909		72612,2	
p_u	0,151		0,151		0,161		0,155		0,128	
p_r	0,605		0,563		0,530		0,589		0,339	
p_u0*u1	-0,388		0,00		0,00		0,00		-0,581	
p_r0*r1	-0,734		-0,650		-0,580		-		-	
				%		**%**		**%**		**%**
Reducción N3			-42,971	-14,454	-47,787	-16,074	-13,886	-4,671	-189,554	-63,761
Reducción N2			-241,741	-20,345	-365,397	-30,753	-111,623	-9,394	-902,229	-75,934

*Coeficiente no significativo

Tabla VIII.14 Coeficientes, errores típicos, ajuste, correlación intraclase y correlación entre estatus inicial y crecimiento en el Problema 2.

El estatus inicial medio (β_0) lógicamente cambia en función de la aplicación utilizada como punto de partida. Así, en el M3_2 alcanza el mayor valor con 301 puntos aproximadamente. La puntuación inicial media es de 247 puntos cuando se utiliza A1 como punto de partida. Respecto a los niveles de crecimiento medio (β_1), los coeficientes estimados son similares entre modelos. Aquellos modelos que utilizan las cuatro puntuaciones temporales en su variable dependiente (M1_2 – M4_2) obtienen valores medios de crecimiento de 4,1 puntos en la escala. El último modelo (M5_2) tiene un coeficiente β_1 ligeramente superior (4,7 puntos).

El análisis de los efectos aleatorios revela aspectos interesantes. En efecto, como Rogosa (1995) demostró, cambiar el punto de partida modifica la relación existente entre estatus inicial y crecimiento de los modelos longitudinales. En este, caso, la covarianza negativa entre esos coeficientes asociados a las escuelas (u0,u1) no es significativa en M2_2 y M3_3. Cuando el punto de partida es la puntuación de rendimiento en A1 ese mismo coeficiente resulta significativo con un coeficiente de

-2,676 y error típico 1,189. El M4_3 también elimina esa covarianza negativa pero también reduce la varianza entre las pendientes de crecimiento (u1) de las escuelas respecto al modelo de base, un cambio de 0,16 a 0,13, un 21% aproximadamente de la varianza del crecimiento entre escuelas. La varianza del estatus inicial (u0) se reduce en A2 y A3 como muestran los coeficientes de los M2_2 y M3_2 respectivamente. Aspecto que ya se comprobó en el análisis de la homogeneidad de puntuaciones del capítulo IV y, como también describió Ballou (2009), es un aspecto característico de las escalas verticales construida bajo los supuestos de la Teoría Respuesta al Ítem.

Los efectos aleatorios de segundo nivel (estudiantes) sufren una reducción drástica en la varianza de la pendiente (r1) de crecimiento y la covarianza entre estatus y crecimiento (r0,r1) con los modelos que incluyen el rendimiento en A1 como predictor en el modelo. Respecto a la varianza entre estudiantes del estatus inicial (r1) la tendencia es similar a la de las escuelas (u1).

Los índices de ajuste global del modelo pueden compararse en los cuatro primeros (M1_2-M4_2) porque utilizan la misma variable dependiente y el mismo número de ocasiones de medida en la función de tiempo para la pendiente de crecimiento. En cambio el M5_2 únicamente consta de tres medidas del rendimiento en su variable dependiente y, por tanto, la varianza del rasgo no es la misma que en el resto de modelos. El M4_2 obtiene el mejor índice de ajuste, menor valor de *deviance*. Además es más simple porque algunos efectos aleatorios se fijan a cero, por tanto, el modelo tiene menos parámetros que el M1_2. Sin embargo, esos resultados no son concluyentes. Los cambios tan radicales en las varianzas pueden provocar una modificación del residuo y una interpretación errónea de las posiciones de las escuelas. Otra cuestión que puede llevar a descartar el M4_2 es que la puntuación en A1 tienen un mayor error de medida, por lo que volver a incluirla como predictor en el modelo aumenta el sesgo de las estimaciones. EL mismo fenómeno ocurre con el M5_2.

Los modelos M2_2 y M3_2 consiguen eliminar la covarianza negativa entre los residuos de las escuelas sin llevar a cabo grandes modificaciones en las varianzas de los distintos niveles. Las diferencias en la *deviance* de estos dos

modelos con el M1_2 no resulta significativa por lo que estos, al ser más simples, serían una mejor opción.

Por tanto, cambiar el punto inicial con una simple modificación de los valores de la función de tiempo. Incluir un predictor que ajusta por esas diferencias iniciales (M4_2) cambia radicalmente la varianza entre estudiantes y también la varianza de las pendientes de crecimiento entre escuelas y añadir el rendimiento previo como principal covariable no consigue eliminar la relación negativa entre estatus inicial y crecimiento entre escuelas, aunque si elimina esa parte de varianza entre estudiantes.

Otra opción es la que se lleva en el M6_2. Los coeficientes de este modelo son iguales que los del M1_2 y por eso no aparece en la tabla de coeficientes de regresión y medias estimadas. La diferencia se encuentra en el residuo de crecimiento asociado a las escuelas (u1). En el M6_2 se lleva a cabo una regresión lineal entre los residuos del estatus inicial y crecimiento para generar uno nuevo libre del efecto de los distintos puntos de partida de las escuelas en la primera aplicación. Esta nueva estimación del VA se ha denominado v. Al contar con los errores típicos de los residuos de las escuelas, la regresión se ha realizado con dos métodos distintos de estimación: con mínimos cuadrados y mínimos cuadrados ponderados por los errores típicos de los residuos del estatus inicial de las escuelas. Las tablas siguientes (Tabla VIII.15, Tabla VIII.16, Tabla VIII.17 y Tabla VIII.18) comparan los resultados de ambos modelos:

Modelo	R	R^2	R^2 corregida	ET de estimación
1	0,360	0,129	0,116	0,318
2_Ponderado	0,366	0,134	0,120	0,313

Tabla VIII.15. Ajuste de los modelos de regresión

El modelo que incluye los errores típicos tiene un mejor ajuste, aunque no es muy distinto del modelo que no los incorpora. La proporción de varianza explicada es de un 12% en el modelo ponderado y un 11,6% en el otro. Los análisis de varianza realizados para comprobar la significatividad de la regresión, que se incluyen en la Tabla VIII.15, muestran probabilidades de F similares. Aunque la media cuadrática de los residuales es algo inferior en el modelo ponderado.

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	0,946	1	0,946	9,369	0,003
	Residual	6,359	63	0,101		
	Total	7,305	64			
2	Regresión	0,953	1	0,953	9,759	0,003
	Residual	6,153	63	0,098		
	Total	7,106	64			

Tabla VIII.16. ANOVA de la regresión.

Como era de esperar las constantes (Tabla VIII.17) no difieren significativamente de cero ya que la media de los residuos es cero. La pendiente de regresión muestra ese efecto del estatus inicial sobre el residuo de crecimiento y toma valores de -0,008 en ambos modelos. Es decir, las escuelas con estatus iniciales superiores a la media tienen un crecimiento menor que aquellas que se encuentran por debajo.

Modelo		Coef.		Coef. tipificados	t	Sig.
		B	ET	Beta		
1	(Constante)	0,000	0,039		0,000	1,000
	u0_M1_2	-0,008	0,002	-0,360	-3,061	0,003
2	(Constante)	0,032	0,038		0,850	0,399
	u0_M1_2	-0,008	0,002	-0,366	-3,124	0,003

Tabla VIII.17. Coeficientes de regresión

Los nuevos residuos estimados están libres de ese efecto producido por la covarianza negativa entre estatus inicial y crecimiento.

Modelo		Mínimo	Máximo	Media	Desviación típica	N
1	Valor pronosticado	-0,281	0,253	0,00	0,122	65
	Residual	-0,826	0,659	0,00	0,315	65
2	Valor pronosticado	-0,254	0,290	0,032	0,124	65
	Residual	-0,862	0,622	-0,032	0,315	65

Tabla VIII.18. Estadísticos sobre los residuos.

La correlación entre los residuos de ambos modelos es igual a uno. No obstante, se utilizan los resultados producidos por el modelo ponderado por esa ligera disminución del error.

Una vez calculada la última estimación del VA en este problema (M6_2) se estudian las medias estimadas por cada modelo. Las medias de M1_2 y M6_2 coinciden por la cuestión explicada más arriba.

	A1	A2	A3	A4
Medias brutas	252,155	273,709	310,686	331,112
M1_2	247,170	280,706	301,666	331,010
M2_2	247,241	280,753	301,698	331,021
M3_2	247,196	280,724	301,679	331,016
M4_2	247,216	280,544	301,374	330,536
M5_2		278,264	302,069	335,396
M6_2	247,17	280,706	301,666	331,01

Tabla VIII.19. Medias brutas y medias estimadas con los seis modelos

El M5_2 no tiene estimación en la primera aplicación porque se utiliza como covariable en el modelo.

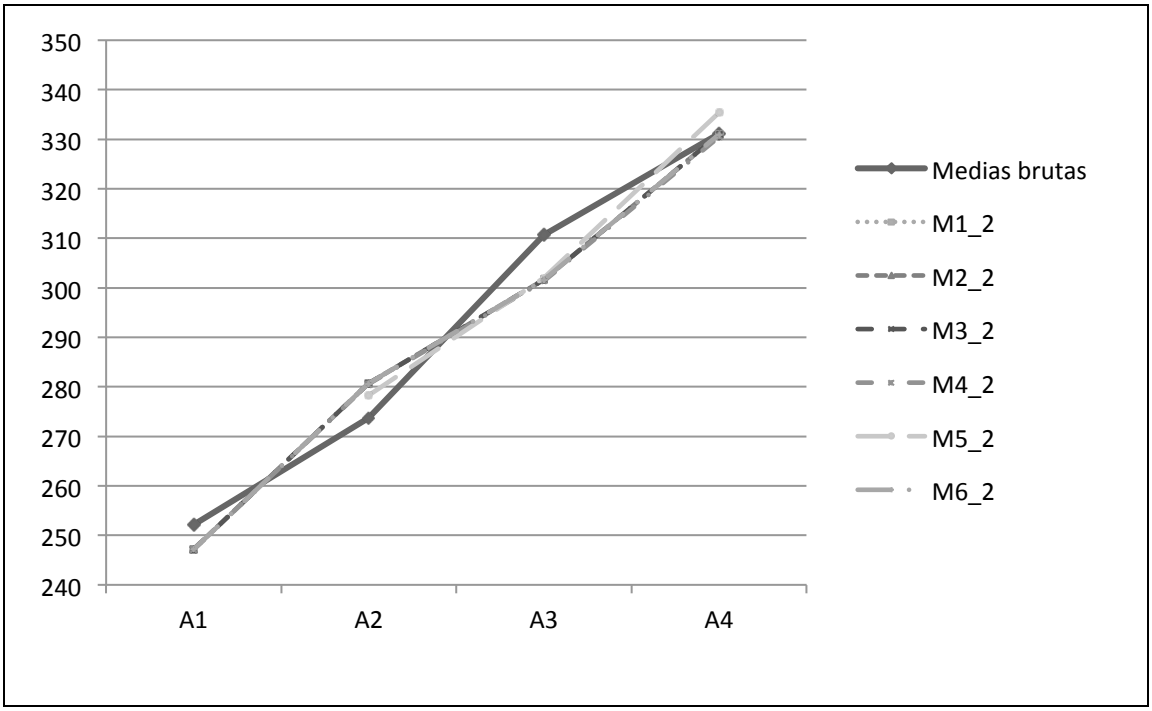


Gráfico VIII.3. Medias brutas y medias estimadas con los seis modelos.

Las medias estimadas con los cinco modelos son bastante similares entre sí. Se diferencian ligeramente de las medias brutas, aspecto que ya se explicó en el problema anterior y que está directamente relacionado con la utilización de los meses en la pendiente de crecimiento.

Para comprobar diferencias en las estimaciones de VA de las escuelas a continuación se presenta el análisis de los residuos de tercer nivel (escuelas)

VIII.3.2.2. Análisis de los residuos de las escuelas

En primer lugar, se presentan las correlaciones entre los residuos de los centros educativos estimados con los distintos modelos en la Tabla VIII.20 y, en segundo, las correlaciones de los rankings en la Tabla VIII.21.

		u1_M1_2	u1_M2_2	u1_M3_2	u1_M4_2	u1_M5_2	v_M6_2
u1_M1_2	**Pearson**	1	1,000	1,000	,877	,470	,933
	Sig.		,000	,000	,000	,000	,000
u1_M2_2	**Pearson**		1	1,000	,877	,470	,933
	Sig.			,000	,000	,000	,000
u1_M3_2	**Pearson**			1	,877	,470	,933
	Sig.				,000	,000	,000
u1_M4_2	**Pearson**				1	,305	,987
	Sig.					,011	,013
u1_M5_2	**Pearson**					1	,321
	Sig.						,009
v_M6_2	**Pearson**						1

Tabla VIII.20. Correlaciones de Pearson entre las estimaciones de VA en el Problema 2.

Los tres primeros modelos M1_2-M3_2 obtienen valores de correlación casi perfectos, también se aproxima a este nivel de relación el residuo del M6_2. Por tanto, llevar a cabo el ajuste posterior de los residuos produce nuevas estimaciones muy similares a las del modelo base. La correlación desciende cuando se incluye el índice ajuste en A1 como predictor (M4_2), aunque también puede considerarse elevada. Afirmación que no puede llevarse a cabo respecto al M5_2 que obtiene los valores de correlación inferiores, alrededor de 0,4.

Las correlaciones entre las distintas ordenaciones de escuelas siguen una tendencia similar a las anteriores como muestra la Tabla VIII.21.

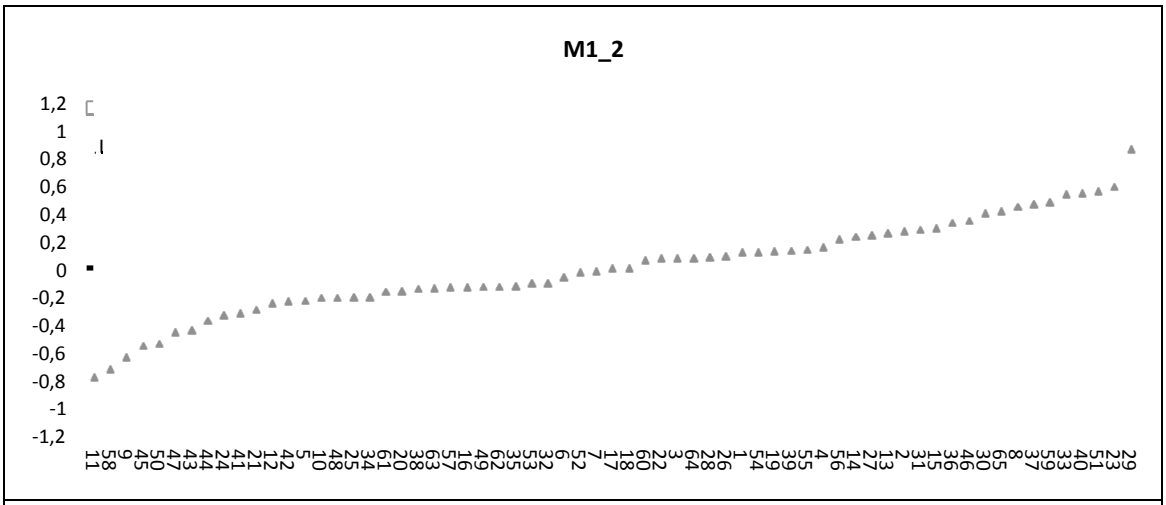
		R_M1_2	R_M2_2	R_M3_2	R_M4_2	R_M5_2	R_M6_2
R_M1_2	**Spearman**	1	1,000	1,000	,881	,483	,935
	Sig		,000	,000	,000	,000	,000
R_M2_2	**Spearman**		1	1,000	,881	,483	,935
	Sig			,000	,000	,000	,000
R_M3_2	**Spearman**			1	,881	,483	,935
	Sig				,000	,000	,000
R_M4_2	**Spearman**				1	,332	,983
	Sig					,007	,000
R_M5_2	**Spearman**					1	,350
	Sig						,004
R_M6_2	**Spearman**						1

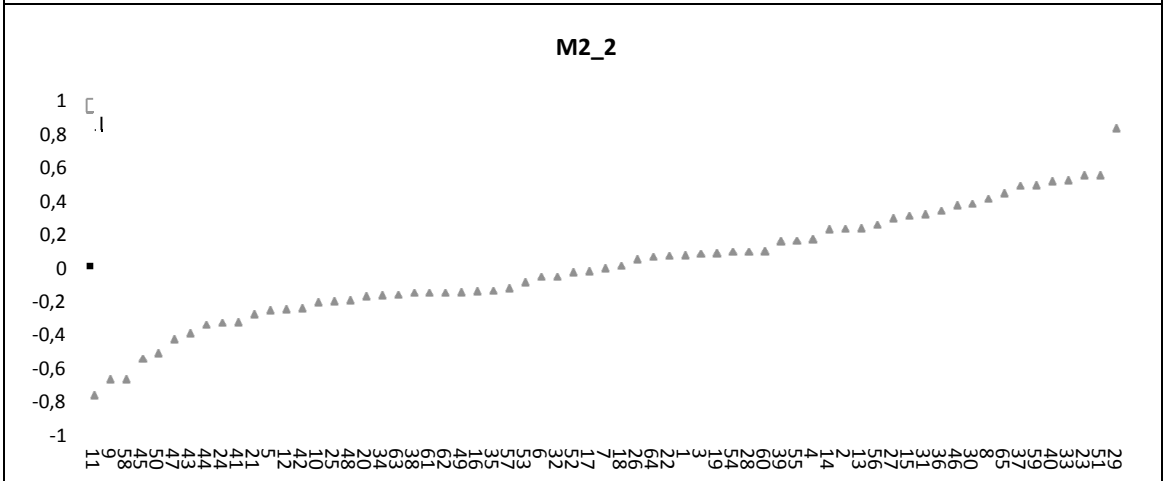
Tabla VIII.21. Correlaciones Rho de Spearman entre los rankings de escuelas del Problema 2.

Por tanto, tomando como referencia los resultados de las correlaciones, parece menos arriesgado utilizar los modelos que cambian la posición del punto de partida porque modifica la relación entre estatus inicial y crecimiento y produce

estimaciones del residuo muy similares a las del modelo de partida. Otra opción es eliminar el efecto de ese estatus inicial con los residuos estimados en la primera etapa, como lleva a cabo el M6_2. Y, por tanto, el sesgo será mayor si se vuelve a introducir la primera aplicación como covariable en el modelo debido a ese mayor error de medida.

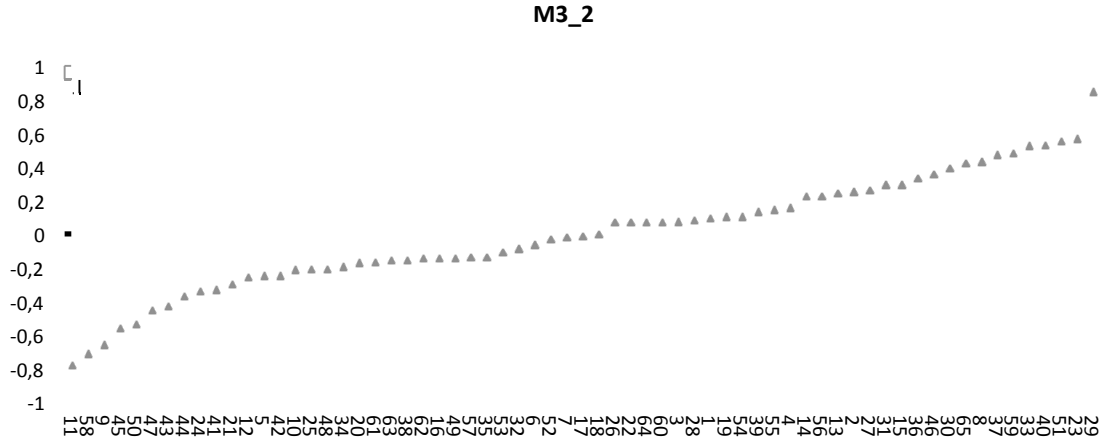
La Figura VIII.2 que se incluye a continuación informa sobre la ordenación de las escuelas y los posibles cambios en esa ordenación que producen los distintos modelos.





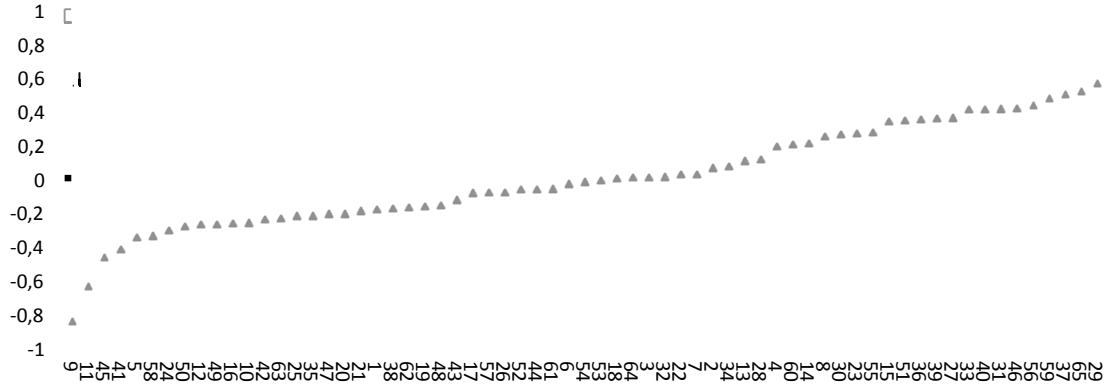
			R_cat_M2_2		
			igual a la media	**Superior a la media**	**Inferior a la media**
R_cat_M1_2	**igual a la media**	Recuento	11	0	0
		%	100%	,0%	,0%
	Superior a la media	Recuento	0	26	0
		%	,0%	100%	,0%
	Inferior a la media	Recuento	0	0	28
		%	,0%	,0%	100%

Chi-cuadrado de Pearson= 130,00 P=0,000



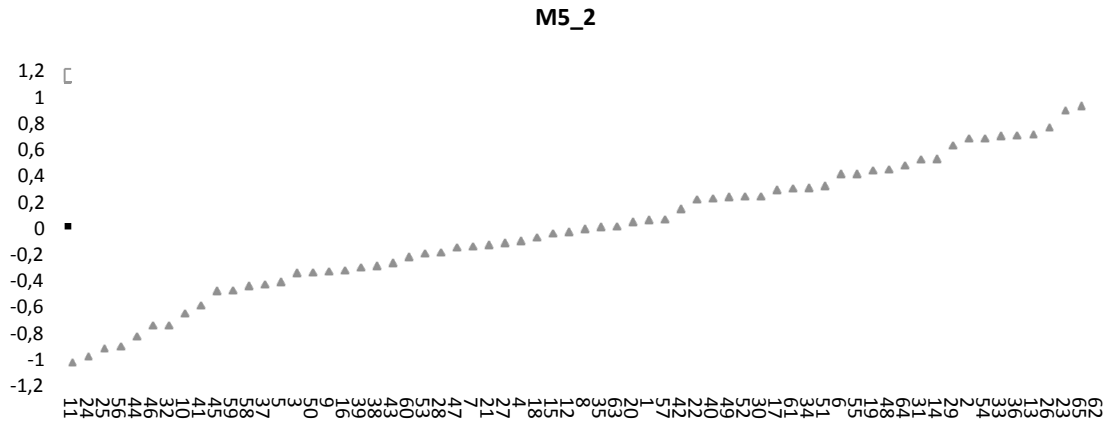
			R_cat_M3_2		
			igual a la media	**Superior a la media**	**Inferior a la media**
R_cat_M1_2	**igual a la media**	Recuento	11	0	0
		%	100%	,0%	,0%
	Superior a la media	Recuento	0	26	0
		%	,0%	100%	,0%
	Inferior a la media	Recuento	0	0	28
		%	,0%	,0%	100%

Chi-cuadrado de Pearson= 130,00 P=0,000



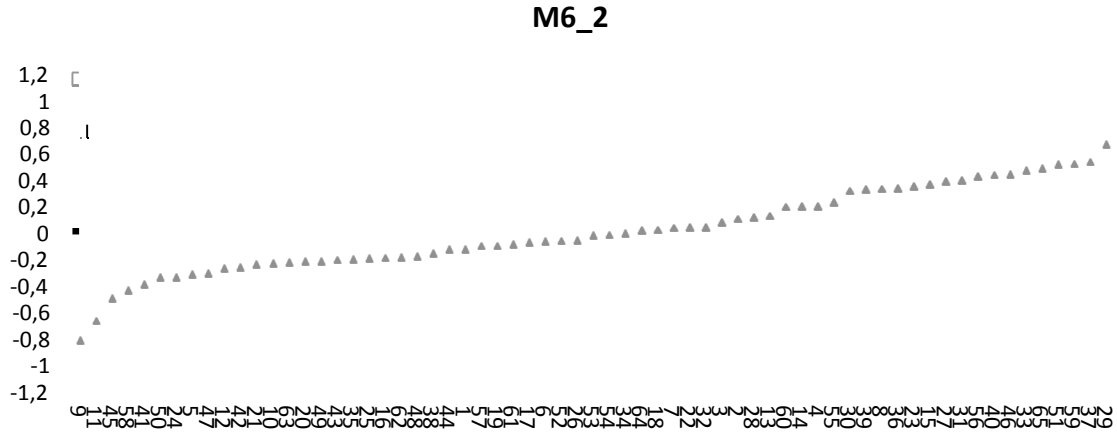
			R_cat_M4_2		
			igual a la media	**Superior a la media**	**Inferior a la media**
R_cat_M1_2	**igual a la media**	Recuento	8	1	2
		%	72,7%	9,1%	18,2%
	Superior a la media	Recuento	2	22	2
		%	7,7%	84,6%	7,7%
	Inferior a la media	Recuento	1	1	26
		%	3,6%	3,6%	92,9%

Chi-cuadrado de Pearson= 74,691 P=0,000



			R_cat_M5_2		
			igual a la media	**Superior a la media**	**Inferior a la media**
R_cat_M1_2	**igual a la media**	Recuento	1	5	5
		%	9,1%	45,5%	45,5%
	Superior a la media	Recuento	4	15	7
		%	15,4%	57,7%	26,9%
	Inferior a la media	Recuento	5	7	16
		%	17,9%	25,0%	57,1%

Chi-cuadrado de Pearson= 6,788 P=0,148



			R_cat_M6_2		
			igual a la media	**Superior a la media**	**Inferior a la media**
R_cat_M1_2	**igual a la media**	Recuento	7	1	3
		%	63,6%	9,1%	27,3%
	Superior a la media	Recuento	5	20	1
		%	19,2%	76,9%	3,8%
	Inferior a la media	Recuento	3	0	25
		%	10,7%	,0%	89,3%

Chi-cuadrado de Pearson= 59,714 P=0,000

Figura VIII.2. Grafico del Valor Añadido de las escuelas e Intervalo de Confianza al 95% y tablas de contingencia que reflejan los cambios en las posiciones respecto al modelo base en el problema 2

En los modelos que cambian el punto de partida modificando la función de tiempo (M1_2-M3_2) Hay el mismo número de escuelas clasificadas significativamente por encima y por debajo de la media global, 26 y 28 respectivamente. Y también son el mismo número las que no se diferencian significativamente de esa media, un total de 11 centros educativos. Estos modelos, por tanto, identifican 54 escuelas distintas de la media.

El resto de modelos producen algunos cambios es el ranking de centros educativos. Es más pronunciado en M5_2, el modelo que consta con solo tres mediciones de la variable dependiente y el rendimiento en A1 como principal covariable. En ese caso, de los 28 centros clasificados como inferiores a la media en M1_2, 7 pasan a estar por encima de la media y 5 a no diferenciarse estadísticamente de ella. El cambio que se produce es tan alto que la relación entre ambas variables deja de ser significativa como refleja el estadístico χ^2 (6,788) que acompaña a la tabla, con una probabilidad asociada por encima de 0,05. No obstante, este modelo identifica 55 centros distintos de la media.

Llevar a cabo el ajuste de los residuos de estatus inicial y crecimiento a posteriori (M6_2) también produce algunos cambios destacables en la clasificación de escuelas. Por ejemplo, aumenta el número de centros que no se diferencian de la media global de 11 a 15, pero con ligeros cambios respecto al modelo inicial, tres de ellos pasan a ser inferiores a la media y uno superior. El mayor cambio se produce en los centros clasificados como superiores a la media en el modelo inicial, cinco de ellos pasan a no diferenciarse de la media y uno es inferior a la media. Este modelo parece el más conservador, aunque como se pudo ver en las tablas Tabla VIII.20 Tabla VIII.21 la correlación entre los residuos de crecimiento de este modelo y el de base son elevadas, superiores a las obtenidas con el M4_2.

Para completar el análisis de los residuos se presenta a continuación los resultados referentes a la posición de las escuelas determinada por su residuo de crecimiento y también su estatus inicial (Tabla VIII.22 y Gráfico VI.4).

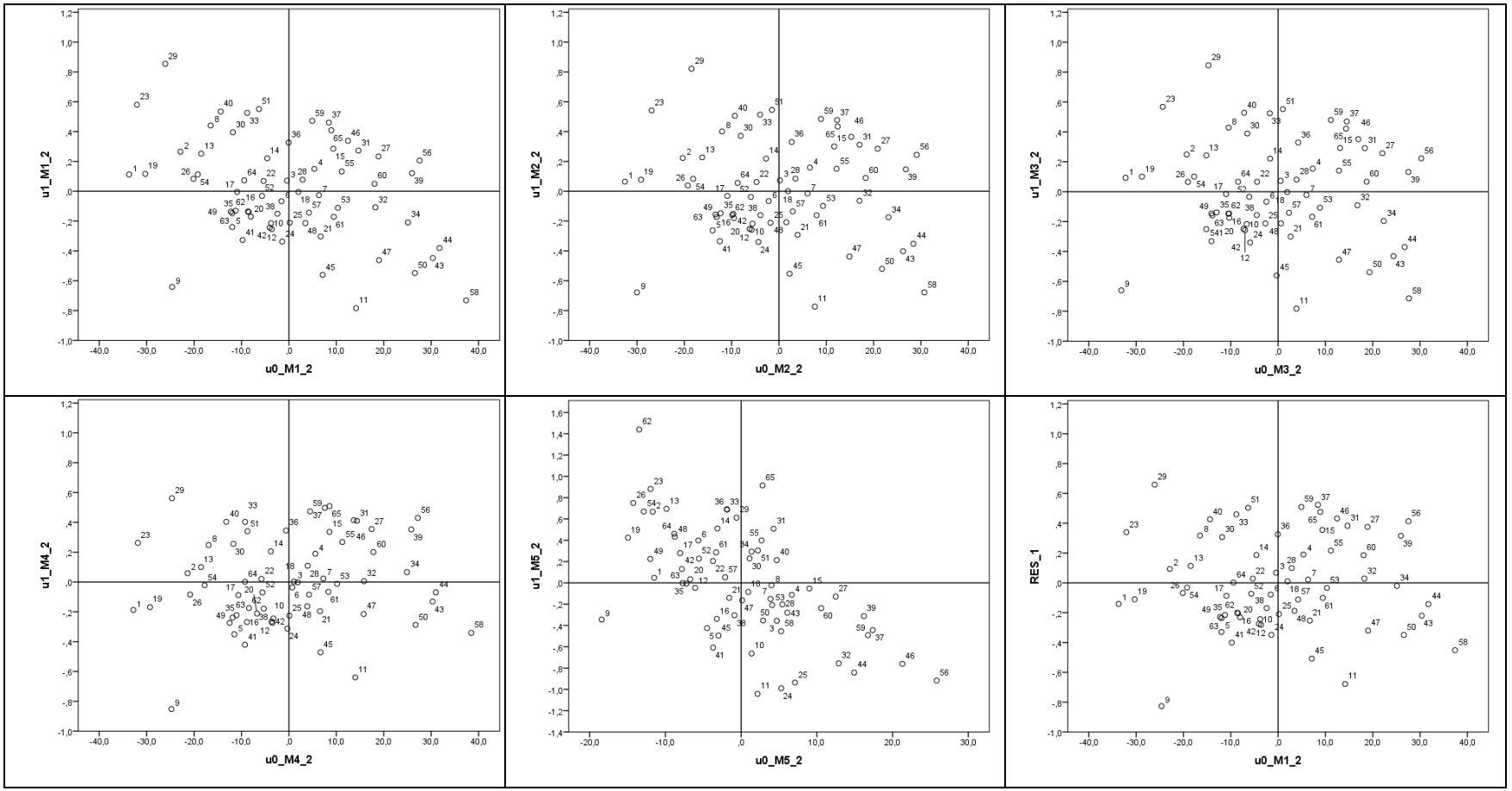


Gráfico VI.4. Gráficos de dispersión de los residuos de las escuelas (u0*u1) en los modelos del problema 2

			Disp_cat_M2_2			
			Bajo Estatus y Alto Crecimiento	**Bajo Estatus y Bajo Crecimiento**	**Alto Estatus y Alto Crecimiento**	**Alto Estatus y Bajo Crecimiento**
Disp_cat_M1_2	**Bajo Estatus y Alto Crecimiento**	Recuento	16	0	2	0
		%	88,9%	,0%	11,1%	,0%
	Bajo Estatus y Bajo Crecimiento	Recuento	0	17	0	0
		%	,0%	100%	,0%	,0%
	Alto Estatus y Alto Crecimiento	Recuento	0	0	13	0
		%	,0%	,0%	100%	,0%
	Alto Estatus y Bajo Crecimiento	Recuento	0	1	0	16
		%	,0%	5,9%	,0%	94,1%

Chi-cuadrado de Pearson= 172,852 P=0,000

			Disp_cat_M3_2			
			Bajo Estatus y Alto Crecimiento	**Bajo Estatus y Bajo Crecimiento**	**Alto Estatus y Alto Crecimiento**	**Alto Estatus y Bajo Crecimiento**
Disp_cat_M1_2	**Bajo Estatus y Alto Crecimiento**	Recuento	15	0	3	0
		%	83,3%	,0%	16,7%	,0%
	Bajo Estatus y Bajo Crecimiento	Recuento	0	17	0	0
		%	,0%	100%	,0%	,0%
	Alto Estatus y Alto Crecimiento	Recuento	0	0	13	0
		%	,0%	,0%	100%	,0%
	Alto Estatus y Bajo Crecimiento	Recuento	0	2	0	15
		%	,0%	11,8%	,0%	88,2%

Chi-cuadrado de Pearson= 160,326 P=0,000

			Disp_cat_M4_2			
			Bajo Estatus y Alto Crecimiento	**Bajo Estatus y Bajo Crecimiento**	**Alto Estatus y Alto Crecimiento**	**Alto Estatus y Bajo Crecimiento**
Disp_cat_M1_2	**Bajo Estatus y Alto Crecimiento**	Recuento	12	5	1	0
		%	66,7%	27,8%	5,6%	,0%
	Bajo Estatus y Bajo Crecimiento	Recuento	0	16	0	1
		%	,0%	94,1%	,0%	5,9%
	Alto Estatus y Alto Crecimiento	Recuento	0	0	13	0
		%	,0%	,0%	100%	,0%
	Alto Estatus y Bajo Crecimiento	Recuento	0	0	3	14
		%	,0%	,0%	17,6%	82,4%

Chi-cuadrado de Pearson= 131,401 P=0,00

			Disp_cat_M5_2			
			Bajo Estatus y Alto Crecimiento	**Bajo Estatus y Bajo Crecimiento**	**Alto Estatus y Alto Crecimiento**	**Alto Estatus y Bajo Crecimiento**
Disp_cat_M1_2	**Bajo Estatus y Alto Crecimiento**	Recuento	13	0	3	2
		%	72,2%	,0%	16,7%	11,1%
	Bajo Estatus y Bajo Crecimiento	Recuento	7	8	0	2
		%	41,2%	47,1%	,0%	11,8%
	Alto Estatus y Alto Crecimiento	Recuento	0	0	3	10
		%	,0%	,0%	23,1%	76,9%
	Alto Estatus y Bajo Crecimiento	Recuento	3	2	1	11
		%	17,6%	11,8%	5,9%	64,7%
Chi-cuadrado de Pearson= 48,489 P=0,000						

			Disp_cat_M6_2			
			Bajo Estatus y Alto Crecimiento	**Bajo Estatus y Bajo Crecimiento**	**Alto Estatus y Alto Crecimiento**	**Alto Estatus y Bajo Crecimiento**
Disp_cat_M1_2	**Bajo Estatus y Alto Crecimiento**	Recuento	12	6	0	0
		%	66,7%	33,3%	,0%	,0%
	Bajo Estatus y Bajo Crecimiento	Recuento	0	17	0	0
		%	,0%	100%	,0%	,0%
	Alto Estatus y Alto Crecimiento	Recuento	0	0	13	0
		%	,0%	,0%	100%	,0%
	Alto Estatus y Bajo Crecimiento	Recuento	0	0	0	17
		%	,0%	,0%	,0%	100%
Chi-cuadrado de Pearson= 162,029 P=0,00						

Tabla VIII.22. Tablas de contingencia y χ^2 para la relación. Cambios en los cuadrantes del gráfico de dispersión respecto al modelo base en el problema 2.

Con este último análisis se producen cambios en las posiciones de las escuelas en todos los modelos elaborados. No obstante, de la misma forma que ocurrían en el estudio de los rankings, el cambio es más pronunciado en M5_2. Cambian de cuadrante un mayor número de centro pero la relación entre ambas variables sigue siendo significativa. Por ejemplo, siete escuelas cambian de bajo estatus inicial y bajo crecimiento a tener un crecimiento por encima de la media.

El M4_2 cambia un total de 10 de escuelas de cuadrante. Por ejemplo, cinco de ellas pasan de estar por debajo de la media en estatus inicial pero con un crecimiento por encima a estar por debajo también en crecimiento.

El modelo que menos cambios produce es el M2_2, únicamente cambian tres escuelas. Dos pasan de tener un estatus inicial por debajo de la media y un crecimiento por encima a tener un estatus también superior a la media. La otra cambia de tener un estatus inicial alto y crecimiento bajo a estar por debajo de la media también es el punto de partida. El valor de χ^2 alcanza el mayor valor (172,852) entre estos dos modelos

En el M6_2 no cambian muchas escuelas de cuadrante y obtiene el segundo mayor valor de χ^2 (162,029), aunque muy cercano al obtenido con el M3_2 (160,326). Únicamente seis centros educativos pasan de tener estatus por debajo de la media y un crecimiento por encima, a un estatus y crecimiento por debajo de la media. Parece, por tanto, más conservador con las puntuaciones de crecimiento de esas seis escuelas.

En conclusión, es conveniente modificar el modelo de base para paliar los efectos que la primera aplicación puede tener en las estimaciones finales ya sea utilizando la segunda toma de datos como punto inicial, es decir, el final de 1ºde ESO o eliminando el efecto a posteriori a través de un análisis de regresión simple. Ambas opciones son los métodos más adecuados.

Optar por el M2_2 parece la opción adecuada. Este modelo descarta A1 como punto de partida del análisis lo que evita esos problemas asociados a la primera toma de datos que pueden afectar a las estimaciones de VA. Y en los MVA la relación entre el punto inicial y el crecimiento es un aspecto clave. Incluir la puntuación en A1 como predictor, de cualquiera de las dos formas mencionadas, puede aumentar del error de los modelos y también sesgar esas estimaciones.

En cualquier caso, llevar a cabo el ajuste de los residuos de las escuelas en una segunda fase como el M6_2, produce unos resultados más conservadores con un aumento de escuelas que no difieren significativamente de la media global.

VIII.3.3 Problema 3: Comparación de modelos de ganancia y crecimiento

En este tercer y último problema que se plantea en el estudio empírico se analizan las supuestas estimaciones de VA producidas por diferentes modelos que

varían en su concepción y medición del cambio en el rendimiento. La mayor parte utiliza los residuos asociados a las escuelas obtenidos a través del análisis lineal mixto o multinivel. Debido a que en algunos casos es necesario calcular la diferencia entre residuos (M2.1_3, M2.2_3, M2.3_3, M2.4_3, M2.5_3 y M3_3) se utiliza la letra P para denominar a esas estimaciones. Un resumen del cálculo de esas diferencias entre residuos y del resto de estimaciones que se utilizan en los modelos aparecen en la siguiente tabla (Tabla VII.23):

M1_3:	Modelo Multinivel de Curva de Crecimiento $P_{M1\_3} = u_{1j}$; Considerando A2 como estatus inicial $P_{M1.1\_3} = u_{1j}$; Con ajuste de los residuos a posteriori $P_{M1.2\_3} = v_{1j}$
M2_3:	Modelo Lineal Mixto $P_{M2\_3} = u_{3j}$; $P_{M2.1\_3} = u_{3j} - u_{1j}$; Curso 1 $P_{M2.2\_3} = u_{1j} - u_{0j}$; Curso 2 $P_{M2.3\_3} = u_{3j} - u_{2j}$;Verano $P_{M2.4\_3} = u_{2j} - u_{1j}$; Total $P_{M2.5\_3} = u_{3j} - u_{0j}$
M3_3:	Ganancia Estimada=$P_{M3\_3} = u_{1j} - u_{0j}$
M4_3:	Ganancia Residual $P_{M4\_3} = u_{0j}$; con dos predictores de rendimiento previo $P_{M4.1\_3} = u_{0j}$
M5_3:	Ganancia Bruta=$P_{M5\_3} = G_j - \overline{G}$; Modelo multinivel con la ganancia bruta como variable dependiente=$P_{M5.1\_3} = u_{0j}$; incluyendo la puntuación en A1 como covariable=$P_{M5.2\_3} = u_{0j}$
M6_3:	Modelo de Estatus $P_{M6\_3} = \frac{\sum_{i=1}^{i=n}(Y_{4ij} - \overline{Y}_4)}{n_j}$

Tabla VIII.23. Puntuaciones (P) utilizadas en los distintos modelos elaborados en el problema 3.

Los resultados de estos 16 modelos se presentan a continuación. En primer lugar, los coeficientes fijos y aleatorios estimados y, en segundo, el análisis pormenorizado de los residuos.

VIII.3.3.1. Coeficientes estimados

La Tabla VIII.24 muestra esos coeficientes de los distintos modelos. No a parecen aquellos que utilizan puntuaciones calculadas a partir de los residuos iniciales. Es el caso del M1.2_3 que calcula la puntuación mediante el ajuste a posteriori, mediante un análisis de regresión, de los residuos del estatus inicial y la pendiente de crecimiento obtenidos con el modelo base. También ocurre con las variaciones del modelo lineal mixto (M2.1_3-M2.5_3), que calculan el nuevo término residual a partir de la diferencia entre dos de los residuos estimados con el M2_3. Y, finalmente tampoco aparecen los modelos que utilizan términos brutos como la ganancia bruta (M5_3) o las puntuaciones brutas (M6_3).

	M1_3		M1.1_3		M2_3				M3_3		M4_3		M4.1_3		M5.1_3		M5.2_3	
EFECTOS FIJOS																		
	Coef.	**ET**	**Coef.**	**ET**	**Coef.**	**ET**			**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**
β0	247,170	2,102	280,753	2,083	250,121	2,206			272,689	2,455	331,611	0,979	331,942	0,980	58,182	0,997	58,386	0,945
β1	4,192	0,060	4,189	0,060	271,490	2,411			330,912	2,128	0,602	0,013	0,428	0,018			-0,141	0,014
β2					309,559	1,990							0,230	0,016				
β3					329,682	2,095												
EFECTOS ALEATORIOS																		
NIVEL 3	**u0**	**u1**			**u0**	**u1**	**u2**	**u3**	**u0**	**u1**	**u0**	**u1**	**u0**	**u1**	**u0**	**u1**	**u0**	**u1**
u0	291,774		254,159		272,525				339,52		42,812		40,752		39,053		31,803	
ETu0	58,088		49,269		5532				67,494		10,675		10,266		11,017		9,813	
u1	-2,676	0,163	0	0,159	275,13	340,053			277,018	253,408								
ETu1	1,189	0,041	0	0,04	56,5059	65,677			56,161	50,78								
u2					228,598	261,188	226,419											
ETu2					46,8057	51,8749	44,5687											
u3					228,884	276,39	227,178	255,37										
ETu3					48,0934	54,6476	45,1453	49,74										
NIVEL 2	**r0**	**r1**	**r0**	**r1**	**r0**	**r1**	**r2**	**r3**	**r0**	**r1**	**r0**	**r1**	**r0**	**r1**	**r0**	**r1**	**r0**	**r1**
r0	1154,862		919,988		1651,93				1365,592		617,751		567,768		848,655		817,736	
ETr0	41,048		28,030		44,1439				38,485		17,391		16,497		23,904		23,770	
r1	-16,443	0,434	-139	0,435	1046,92	1384,47			804,452	1090,00								
ETr0	1,526	0,088	1,088	0,088	34,6828	37,0574			29,121	30,720								
r2					903,07	873,244	1133,58											
ETr2					31,1426	28,9534	30,7576											
r3					817,155	816,373	771,183	1103,1										
ETr3					30,33	28,1603	25,9114	29,837										
NIVEL 1																		
e	479,956		479,944															
ET	9,359		9,358															
	Estu.	**Esc.**	**Estu.**	**Esc.**	**Estu.**		**Esc.**		**Estu.**	**Esc.**	**Estu.**	**Esc.**	**Estu.**	**Esc.**	**Estu.**	**Esc.**	**Estu.**	**Esc.**
N	2964	65	2964	65	2964		65		2581	63	2582	63	2427	63	2582	63	2427	63
Ajuste (Deviance)	106188,561		106190,461		80837,929				50083,992		24000,905		22358,4		24803,645		23219,367	
p_u	0,151		0,151		0,051				0,044		0,065		0,067		0,044		0,037	
p_r	0,605		0,563		0,949				0,956		0,935		0,933		0,956		0,963	
corr (estatus*cambio)	-0,388		0,00		0,661				-0,592		0,485		0,435		-0,504		-0,254	

Tabla VIII.24 Coeficientes, errores típicos, ajuste, correlación intraclase y correlación entre estatus inicial y cambio en los modelos del Problema 3.

La tabla anterior (Tabla VIII.24) permite identificar claramente los modelos de tres niveles. Es el caso del M1_3 y sus variantes que incorporan el termino residual (e) para el nivel vinculado al tiempo. Son los modelos que incluyen coeficientes fijos y aleatorios para el estatus inicial (β_0, μ_0 y r_0 respectivamente) y la pendiente de crecimiento entre las aplicaciones (β_1, μ_1 y r_1 respectivamente).

Los modelos M1_3 y M1.1_3 ya se compararon en el problema 2 y su única variación es la escala de la variable tiempo para situar la puntuación de rendimiento en la segunda aplicación como punto de partida. Este cambio tiene como consecuencia la eliminación de la covarianza negativa entre los residuos de las escuelas. Este modelo es más simple que el de base y evita el posible sesgo producido por el EFM.

El M2_3 es el modelo lineal mixto que estima un coeficiente fijo en cada una de las aplicaciones. Son las puntuaciones medias en cada una de ellas, β_0 es la media global en la primera aplicación, β_1 es la media global en la segunda y así respectivamente. Se observa un crecimiento entre las aplicaciones, de unos 20 puntos aproximadamente en cada curso, es decir, entre β_0 y β_1 y entre β_2 y β_3. También puede observarse ese mayor crecimiento en el periodo de verano, entre β_1 y β_2, uno 28 puntos. Una característica de este modelo es que permite analizar las ganancias en cada uno de los dos cursos analizados y obviar el periodo de verano.

Estos términos varían de forma aleatoria para estudiantes (r) y escuelas (μ). Un total de 16 términos aleatorios en cada nivel, tomando en cuenta la covarianza entre cada par de aplicaciones. Estas covarianzas indican la relación significativa entre aplicaciones que se reduce a medida que aumenta la distancia entre ellas.

Si analizamos los valores de la autocorrelación, se observa una reducción en la proporción de varianza que se debe a las escuelas respecto al M1_3. De un 15% se pasa al 5%. Esa disminución se mantiene el resto de modelos basados en la ganancia.

La relación entre estatus inicial y crecimiento dependerá de la puntuación utilizada como estimación del VA de la escuelas. El M2_3, al considerar el efecto persistente entre aplicaciones, utiliza el residuo en la última aplicación (u_3). Y, por

tanto, el residuo en la segunda aplicación (u_1) se considerara el estatus inicial. La correlación entre estos dos residuos es de 0,661, similar a la obtenida con las puntuaciones brutas estudiada en el problema 2 (Tabla VIII.12). Para las variaciones de este modelo que calculan las puntuaciones llevando a cabo diferencias entre pares de residuos también se han calculado estos valores de correlación. En el M2.1_3 el valor es de -0,566, es decir, cuando se correlaciona el residuo en A2 (u_2) con la diferencia entre los residuos de A2 y A4 la relación es negativa. También ocurre cuando se analiza el periodo de verano (M2.4_3), el valor de la correlación es de -0,702. La correlación es negativa, aunque moderada, en el M2.5_3, que calcula la diferencia entre los residuos de la primera y última aplicación, con un valor de -0,206. Cuando se analiza la ganancia intra-curso en los modelos M2.2_3 (curso 1) y M2.3_3 (curso2) los valores de correlación (0,122 y 0,102 respectivamente) no resultan significativos.

En el M3_3, la ganancia estimada, los coeficientes fijos también son medias globales en las dos aplicaciones analizadas. En este caso, β_0 es la media global en A2 y β_1 en A4. También incorpora términos de varianza aleatoria entre estudiantes y escuelas. Los modelos de ganancia cuentan con una muestra total de 63 centros porque hay dos de ellos a los que le falta una de sus puntuaciones en alguna de las dos aplicaciones utilizadas. No ocurre eso con los modelos de crecimiento, aunque utilicen como referencia A2 porque el software empleado para el análisis estima el residuo aunque tenga valor perdido y teniendo en cuenta las puntuaciones en el resto de aplicaciones. La correlación intraclase desciende aquí al 4,4% y la correlación entre estatus (u_0) y la ganancia ($u_1 - u_0$) es negativa (-0,592).

Los modelos de ganancia residual (4_3 y 4.1_3) se encuentran anidados entre sí. El primero incorpora la puntuación en A2 como única covariable para controlar su efecto sobre la puntuación en la aplicación final, el segundo modelo añade además la puntuación en la primera aplicación como covariable. Por tanto, β_0 es la puntuación ajustada en la última aplicación, una vez eliminados los efectos de las covariables. β_1 es el efecto de la segunda aplicación y β_1 el efecto de la primera. Las covariables se incluyen como coeficientes fijos en el modelo, por ese motivo solo hay varianza aleatoria entre estudiantes y escuelas de la puntuación ajustada en A4. Si se observan esos valores de las covariables (0,428 y 0,230)

puede afirmarse que a medida que mayor es la puntuación en A2 y A1, mayor es el rendimiento final en A4. El modelo 4.1_3 obtiene un mejor ajuste, con un menor valor de *deviance* y aunque incluye un parámetro más que el 4_3, la diferencia entre índices de ajuste es significativa. La correlación intraclase es similar en ambos modelos (6,5% y 6,7%). La correlación entre la puntuación en A2 que se considera el estatus inicial y el residuo ajustado es positiva, con valores ligeramente inferiores en el modelo que incluye dos covariables (M4.1_3).

Finalmente los modelos de ganancia bruta (M5.1_3 y M5.2_3) utilizan es puntuación como variable dependiente en el modelo. El segundo, incluye también la puntuación en A1 como covariable. En este caso, los coeficientes fijos hacen referencia a la media global en ganancia entre A2 y A4 (β_0). Y en el M5.2_3 el coeficiente β_1 es el efecto de la puntuación en la primera aplicación sobre esa ganancia que en este caso es negativo. Por tanto, a mayor puntuación en A1 menor ganancia. Son los modelos con los valores de autocorrelación más bajos, un 4,4% y3,7% respectivamente. La correlación entre el estatus inicial (A2) y la ganancia ajustada es negativa con un valor de -0,5 en el modelo multinivel sin predictores (M5.1_3). El valor de es relación desciende cuando se incluye la puntuación en la primera aplicación como covariable (-0,254).

A continuación se presenta el análisis de los resultados obtenidos mediante la comparación de los mismos. Se correlacionan 16 puntuaciones (P) distintas (Tabla VIII.25), algunas de ellas son residuos, otras diferencias entre residuos y el resto se obtienen a partir de las puntuaciones brutas. También se analiza la relación en las 16 clasificaciones de escuelas elaboradas a partir de la ordenación de sus puntuaciones (Tabla VIII.26).

VIII.3.3.2. Análisis de los residuos de las escuelas

		PM1_3	PM1.1_3	PM1.2_3	PM2_3	PM2.1_3	PM2.2_3	PM2.3_3	PM2.4_3	PM2.5_3	PM3_3	PM4_3	PM4.1_3	PM5_3	PM5.1_3	PM5.2_3	PM6_3
PM1_3	Pearson	1	1,000	,933	,050	,593	,534	,546	,166	,960	,534	,415	,623	,563	,569	,455	,089
	Sig		,000	,000	,694	,000	,000	,000	,187	,000	,000	,001	,000	,000	,000	,000	,486
PM1.1_3	Pearson		1	,933	,050	,593	,534	,546	,166	,960	,534	,415	,623	,563	,569	,455	,089
	Sig			,000	,694	,000	,000	,000	,187	,000	,000	,001	,000	,000	,000	,000	,486
PM1.2_3	Pearson			1	,411	,416	,658	,606	-,048	,943	,346	,644	,817	,433	,436	,430	,440
	Sig				,001	,001	,000	,000	,703	,000	,005	,000	,000	,000	,000	,000	,000
PM2_3	Pearson				1	-,305	,447	,364	-,560	,194	-,346	,752	,698	-,178	-,184	,081	,991
	Sig					,014	,000	,003	,000	,121	,005	,000	,000	,162	,148	,529	,000
PM2.1_3	Pearson					1	-,338	,320	,720	,435	,983	,388	,411	,966	,977	,888	-,220
	Sig						,006	,009	,000	,000	,000	,002	,001	,000	,000	,000	,080
PM2.2_3	Pearson						1	,468	-,665	,701	-,379	,172	,373	-,309	-,316	-,349	,418
	Sig							,000	,000	,000	,002	,177	,003	,014	,012	,005	,001
PM2.3_3	Pearson							1	-,427	,690	,314	,589	,665	,411	,397	,400	,426
	Sig								,000	,000	,012	,000	,000	,001	,001	,001	,000
PM2.4_3	Pearson								1	-,091	,715	-,063	-,096	,627	,648	,559	-,530
	Sig									,472	,000	,622	,454	,000	,000	,000	,000
PM2.5_3	Pearson									1	,389	,455	,663	,442	,443	,344	,225
	Sig										,002	,000	,000	,000	,000	,006	,073
PM3_3	Pearson										1	,347	,351	,971	,985	,880	-,278
	Sig											,005	,005	,000	,000	,000	,027
PM4_3	Pearson											1	,952	,501	,496	,696	,794
	Sig												,000	,000	,000	,000	,000
PM4.1_3	Pearson												1	,496	,490	,652	,742
	Sig													,000	,000	,000	,000
PM5_3	Pearson													1	,987	,924	-,100
	Sig														,000	,000	,435
PM5.1_3	Pearson														1	,935	-,112
	Sig															,000	,381
PM5.2_3	Pearson															1	,152
	Sig																,236
PM6_3	Pearson																1

Tabla VIII.25. Correlaciones de Pearson entre las estimaciones de las escuelas con los modelos del Problema 3.

En primer lugar, si analizamos los modelos multinivel de crecimiento puede verse en los dos primeros (PM1_3 y PM1.1_3), que únicamente se diferencian en el cambio del punto de partida mediante la modificación de la escala de la variable tiempo, y como ya se observó en el problema 2 (Tabla VIII.20), sus residuos tienen un correlación de uno aproximadamente por lo que sus coeficientes de Pearson son iguales con el resto de modelos. Ambos obtienen correlaciones positivas y significativos con el resto de modelos, excepto con PM2_3, PM2.4_3 y PM6_3 que obtienen valores de correlación que no se diferencian estadísticamente de cero.

Recordemos que el PM2_3 es el modelo lineal mixto utiliza el residuo de la última medición para calcular la puntuación de la escuela; el PM2.4_3 es la diferencia entre los residuos de A2 y A3, es decir, el periodo de varano; y el PM6_3 utiliza las medias brutas de esta última aplicación para elaborarla. Por tanto, las puntuaciones obtenidas con los dos primeros modelos multinivel de crecimiento parece que no tienen nada que ver con esos resultados. Curiosamente la correlación entre las puntuaciones de PM2_3 y PM6_3 es muy alta (0,991) lo que indica que utilizar únicamente el residuo de la última aplicación en el modelo lineal mixto no muy distinta a utilizar las medias brutas de las escuelas en es última toma de datos. Esto no quiere decir que haya que descartar totalmente el modelo PM2_6, si se utilizan las puntuaciones calculadas mediante la diferencia de residuos los resultados cambian.

La correlación con PM2.1_3 (calculada a través de la diferencia entre los residuos de A2 y A4) de los modelos de crecimiento es de 0,593. Obtienen valores similares con los modelos que calculan la ganancia intra-curso (PM2.2_3 y PM2.3_4). Con la puntuación PM2.5_3, que estima la diferencia entre A1 y A4, el valor de la correlación asciende a 0,960, un valor muy alto. Esta es una cuestión destacable porque se pueden obtener estimaciones del VA de las escuelas muy parecidas mediante ambas metodologías. Aunque, al menos con este tipo de datos, deben calcularse a través de la diferencia entre residuos.

El PM1.2_3, a diferencia de los otros dos modelos de crecimiento, obtiene una correlación positiva con el PM2_3 y PM6_3. El único valor no significativo es con la puntuación del periodo verano (PM2.4_3). Los patrones de correlación son similares a los que obtienen los dos primeros modelos, también muestra una

correlación muy alta con el PM2.5_3, cercana al 0,95. No obstante es algo mayor con los modelos de ganancia residual (PM4_3 y PM4.1_3) que la que obtenían PM1_3 y PM1.1_3 con valores que pasan de los 0,4 y 0,6 al 0,6 y 0,8. En cambio, es algo inferior respecto al modelo de ganancia estimada (PM3_3) donde disminuye de 0,5 a 0,3.

En segundo lugar, los valores de las correlaciones entre los seis modelos lineales mixtos (PM2_3-PM2.5_3) son otro aspecto reseñable. Las ganancias entre el primer curso (PM2.2_3 y PM2.3_3) son positivas pero de una intensidad media (0,468). Ambas correlacionan negativamente con el periodo de verano (PM2.4_3), con valores de -0,665 y -0,427 respectivamente. La ganancia en el primer curso (PM2.2_3) también obtiene un valor negativo (-0,338) en la relación con PM2.1_3, aunque es positivo para la ganancia en el segundo (PM2.3_3). Al contrario, obtienen valores de correlación medio-altos, alrededor de 0,7 con PM2.5_3. Algo mayor que la que obtienen con PM2_3 que se sitúa en 0,447 con el primer curso (PM2.2_3) y 0,364 con el segundo (PM2.3_3).

La relación es negativa, aunque moderada, entre el PM2_3 y el PM2.1_3, con un valor de -0,305. Esta aspecto señala que la consideración de la persistencia o no de los efectos cambia drásticamente la estimación de las puntuaciones de las escuelas. Los resultados de PM2_3 son prácticamente idénticos a los obtenidos con las puntuaciones brutas de la última aplicación (0,991). También tiene cierta similitud con las puntuaciones de ganancia residual (PM4_3 y PM4.1_3), aunque el valor de la correlación desciende hasta el 0,7 aproximadamente, con el resto de puntuaciones no tiene relación. El PM2.1_3, en cambio, está relacionado con el resto de modelos de forma positiva, incluso con el periodo de verano (PM2.4_3) con un valor de 0,720. Los valores superan el 0,9 con la puntuación de ganancia bruta (PM5_3) y la regresión multinivel de la ganancia bruta con una covariable (PM5.1_3). El valor de la correlación se sitúa en 0,888 cuando se incluye una segunda covariable (PM5.2_3).

Y, en tercer lugar, los patrones de correlación las puntuaciones obtenidas con los distintos tipos de ganancia (estimada, residual y bruta) y las puntuaciones brutas. La ganancia estimada (M3_3) correlaciona de forma negativa con la puntuación bruta en A4 (PM6_3), un coeficiente de -0,278. Este tipo de ganancia

estimada muestra resultados similares a los de la ganancia bruta y sus variaciones (M5_3, M5.1_3 y M5.2_3), con valores de correlación de sus puntuaciones en torno al 0,9.

De los modelos de ganancia, el que utiliza A2 y A1 como predictores de A4 para obtener la ganancia residual (PM4.1_3) es el que obtiene mayores valores de correlación con los modelos multinivel longitudinales (0,623). También obtiene valores similares con PM2.3_3 y PM2.5_3.

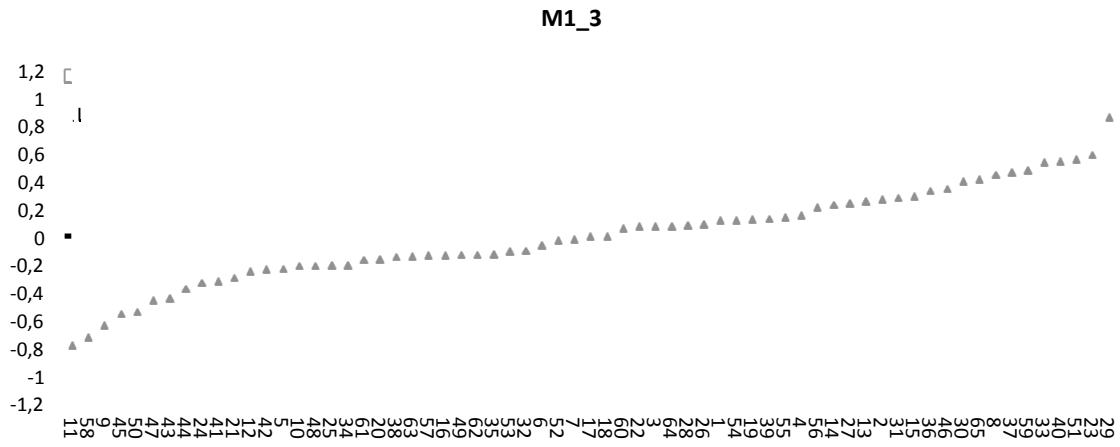
Los valores de correlación de las puntuaciones de ganancia residual (PM4_3 y PM4.1_3) con la puntuación bruta en la última aplicación (PM6_3) es positiva y con intensidad medio-alta (0,794 y 0,742 respectivamente).

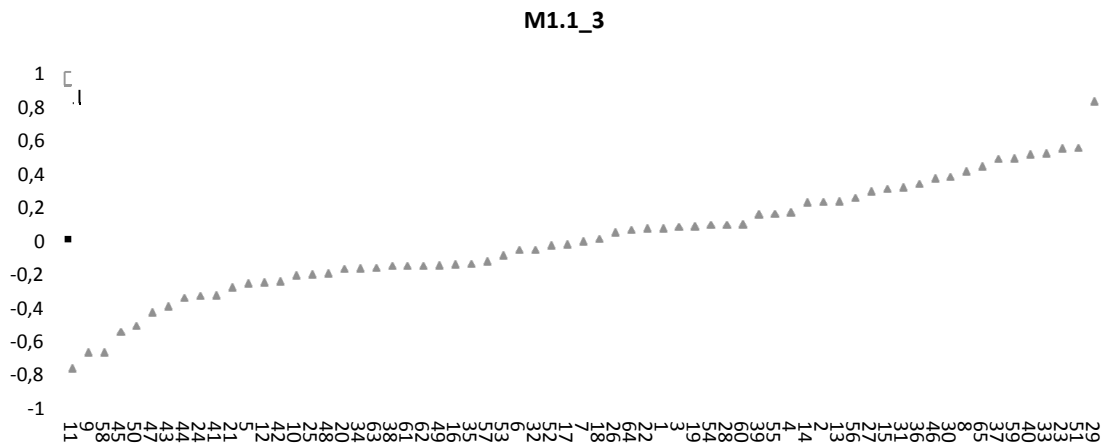
Las correlaciones entre las clasificaciones de las escuelas elaboradas con las distintas puntuaciones se presentan a continuación (Tabla VIII.26) pero no es necesario añadir más comentarios porque los patrones son similares a los que se obtienen del estudio esas puntuaciones.

		RM1_3	RM1.1_3	RM1.2_3	RM2_3	RM2.1_3	RM2.2_3	RM2.3_3	RM2.4_3	RM2.5_3	RM3_3	RM4_3	RM4.1_3	RM5_3	RM5.1_3	RM5.2_3	RM6_3
RM1_3	Spearman	1	1,000	,935	,081	,569	,546	,558	,142	,935	,508	,414	,661	,547	,547	,456	,113
	Sig		,000	,000	,523	,000	,000	,000	,260	,000	,000	,001	,000	,000	,000	,000	,375
RM1.1_3	Spearman		1	,935	,081	,569	,546	,558	,142	,935	,508	,414	,661	,547	,547	,456	,113
	Sig			,000	,523	,000	,000	,000	,260	,000	,000	,001	,000	,000	,000	,000	,375
RM1.2_3	Spearman			1	,375	,427	,641	,616	-,021	,933	,371	,611	,821	,454	,453	,432	,398
	Sig				,002	,000	,000	,000	,865	,000	,003	,000	,000	,000	,000	,000	,001
RM2_3	Spearman				1	-,308	,415	,364	-,538	,229	-,323	,766	,677	-,163	-,182	,049	,991
	Sig					,013	,001	,003	,000	,067	,010	,000	,000	,203	,154	,705	,000
RM2.1_3	Spearman					1	-,269	,302	,719	,407	,980	,318	,367	,963	,970	,885	-,230
	Sig						,030	,014	,000	,001	,000	,011	,003	,000	,000	,000	,067
RM2.2_3	Spearman						1	,532	-,616	,722	-,307	,227	,443	-,228	-,240	-,281	,384
	Sig							,000	,000	,000	,015	,074	,000	,072	,058	,026	,002
RM2.3_3	Spearman							1	-,375	,734	,300	,552	,656	,383	,366	,371	,415
	Sig								,002	,000	,017	,000	,000	,002	,003	,003	,001
RM2.4_3	Spearman								1	-,112	,722	-,072	-,100	,639	,658	,565	-,505
	Sig									,374	,000	,572	,436	,000	,000	,000	,000
RM2.5_3	Spearman									1	,375	,469	,708	,440	,433	,349	,253
	Sig										,002	,000	,000	,000	,000	,005	,044
RM3_3	Spearman										1	,296	,329	,975	,984	,891	-,262
	Sig											,019	,008	,000	,000	,000	,038
RM4_3	Spearman											1	,921	,441	,423	,603	,802
	Sig												,000	,000	,001	,000	,000
RM4.1_3	Spearman												1	,460	,448	,574	,714
	Sig													,000	,000	,000	,000
RM5_3	Spearman													1	,992	,937	-,095
	Sig														,000	,000	,461
RM5.1_3	Spearman														1	,936	-,119
	Sig															,000	,354
RM5.2_3	Spearman															1	,113
	Sig																,378
RM6_3	Spearman																1

Tabla VIII.26. Correlaciones de Pearson entre Los rankings de las escuelas con los modelos del Problema 3

Los gráficos y pruebas χ^2que aparecen a continuación (Figura VIII.3) son útiles para averiguar los cambios concretos en las clasificaciones de las escuelas que se llevan a cabo con las puntuaciones que se utilizarían para evaluarlas. También es posible saber qué modelo identifica mayor número de escuelas distintas de la media.

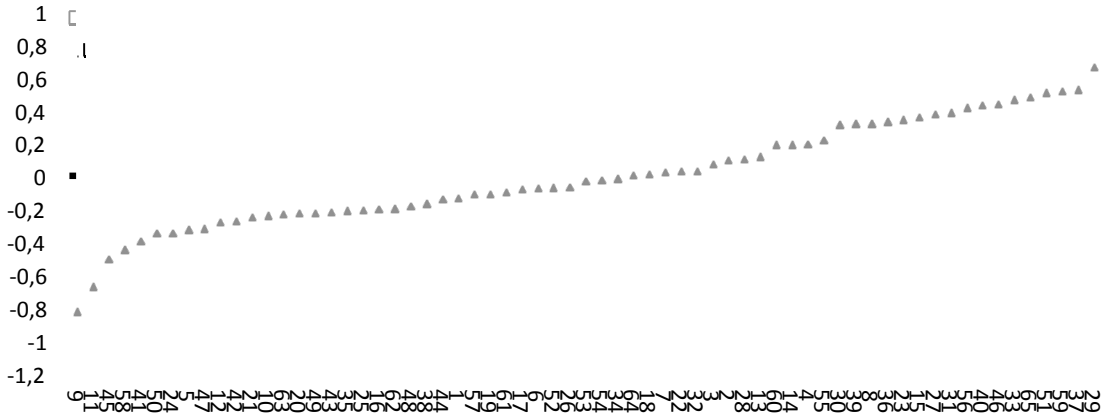




			R_cat_M1.1_3		
			igual a la media	**Superior a la media**	**Inferior a la media**
R_cat_M1_3	**igual a la media**	Recuento	11	0	0
		%	100%	0%	0%
	Superior a la media	Recuento	0	26	0
		%	0%	100%	0%
	Inferior a la media	Recuento	0	0	28
		%	0%	0%	100%

Chi-cuadrado de Pearson= 130,00 P=0,000

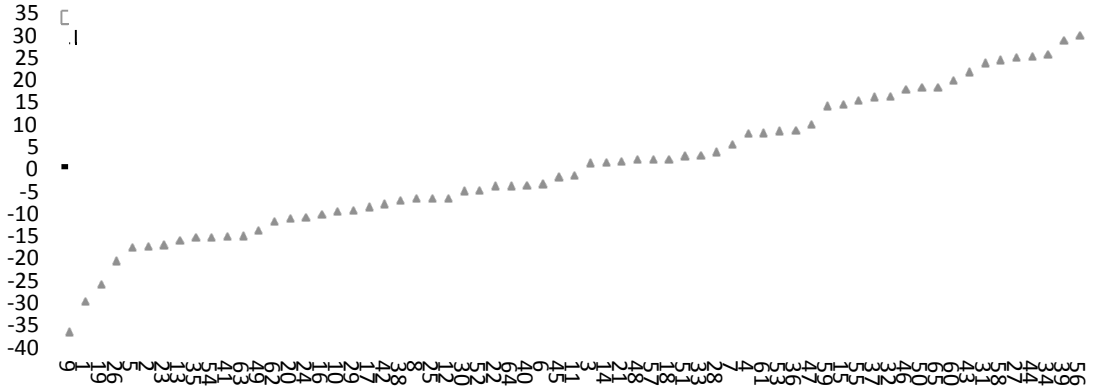
M1.2_3



			R_cat_M1.2_3		
			igual a la media	**Superior a la media**	**Inferior a la media**
R_cat_M1_3	**igual a la media**	Recuento	7	1	3
		%	63,6%	9,1%	27,3%
	Superior a la media	Recuento	5	20	1
		%	19,2%	76,9%	3,8%
	Inferior a la media	Recuento	3	0	25
		%	10,7%	0%	89,3%

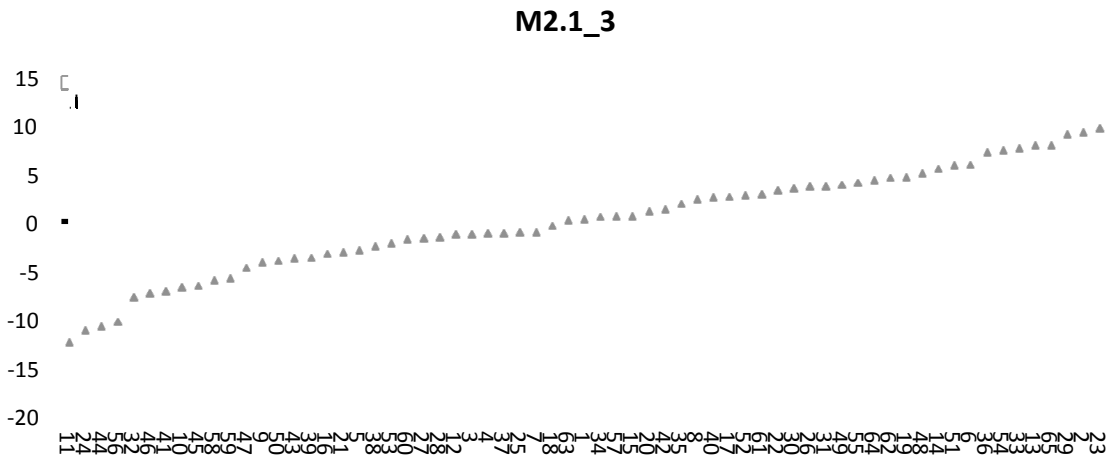
Chi-cuadrado de Pearson= 59,714 P=0,000

M2_3



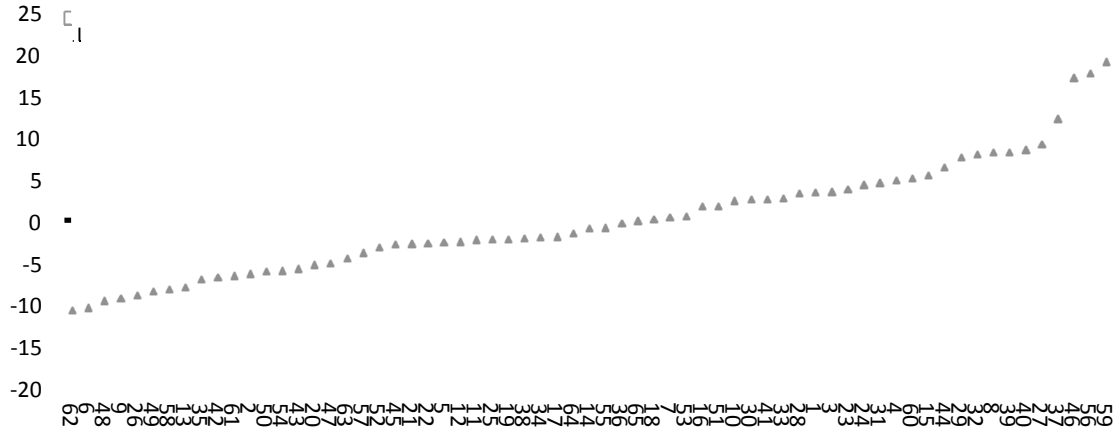
			R_cat_M2_3		
			igual a la media	**Superior a la media**	**Inferior a la media**
R_cat_M1_3	**igual a la media**	Recuento	1	4	6
		%	9,1%	36,4%	54,5%
	Superior a la media	Recuento	2	14	10
		%	7,7%	53,8%	38,5%
	Inferior a la media	Recuento	3	8	17
		%	10,7%	28,6%	60,7%

Chi-cuadrado de Pearson= 3,683P=0,451



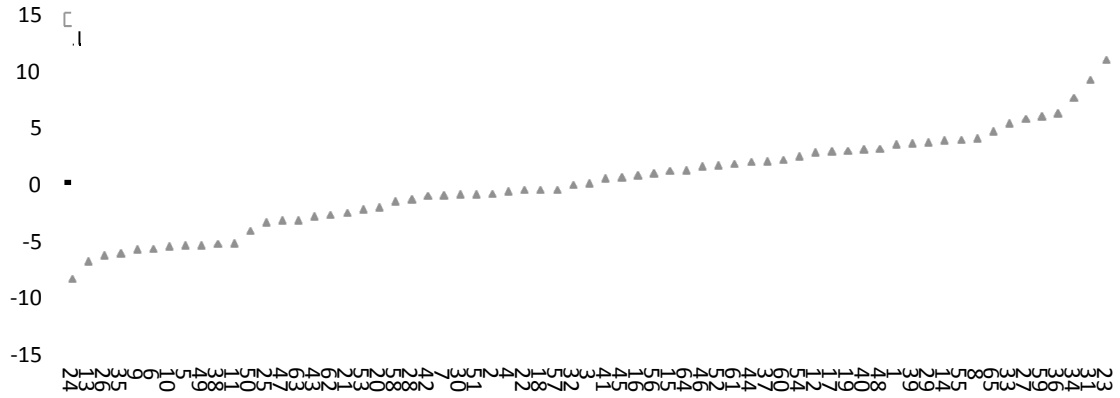
			R_cat_M2.1_3		
			igual a la media	**Superior a la media**	**Inferior a la media**
R_cat_M1_3	**igual a la media**	Recuento	8	2	1
		%	72,7%	18,2%	9,1%
	Superior a la media	Recuento	10	12	4
		%	38,5%	46,2%	15,4%
	Inferior a la media	Recuento	11	5	12
		%	39,3%	17,9%	42,9%

Chi-cuadrado de Pearson= 11,958 P=0,018



			R_cat_M2.2_3		
			igual a la media	**Superior a la media**	**Inferior a la media**
R_cat_M1_3	**igual a la media**	Recuento	9	1	1
		%	81,8%	9,1%	9,1%
	Superior a la media	Recuento	12	11	3
		%	46,2%	42,3%	11,5%
	Inferior a la media	Recuento	13	2	13
		%	46,4%	7,1%	46,4%

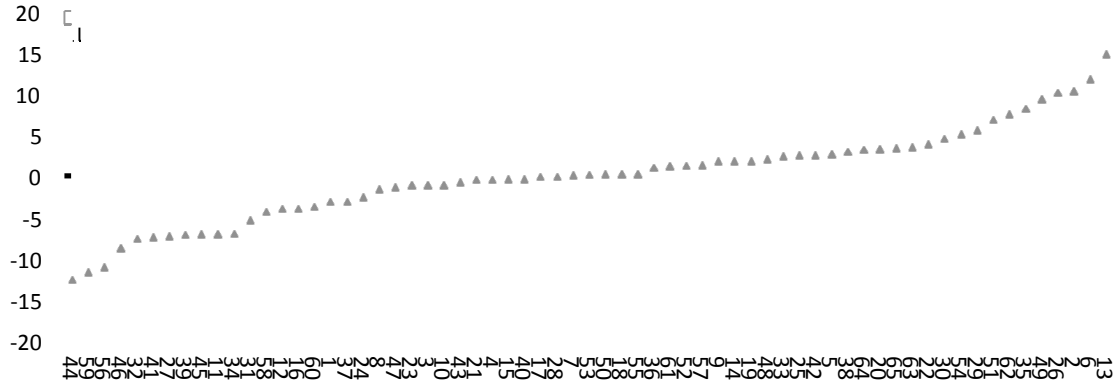
Chi-cuadrado de Pearson= 18,646 P=0,001



			R_cat_M2.3_3		
			igual a la media	**Superior a la media**	**Inferior a la media**
R_cat_M1_3	**igual a la media**	Recuento	9	0	2
		%	81,8%	0%	18,2%
	Superior a la media	Recuento	12	13	1
		%	46,2%	50,0%	3,8%
	Inferior a la media	Recuento	11	3	14
		%	39,3%	10,7%	50,0%

Chi-cuadrado de Pearson= 25,999 P=0,000

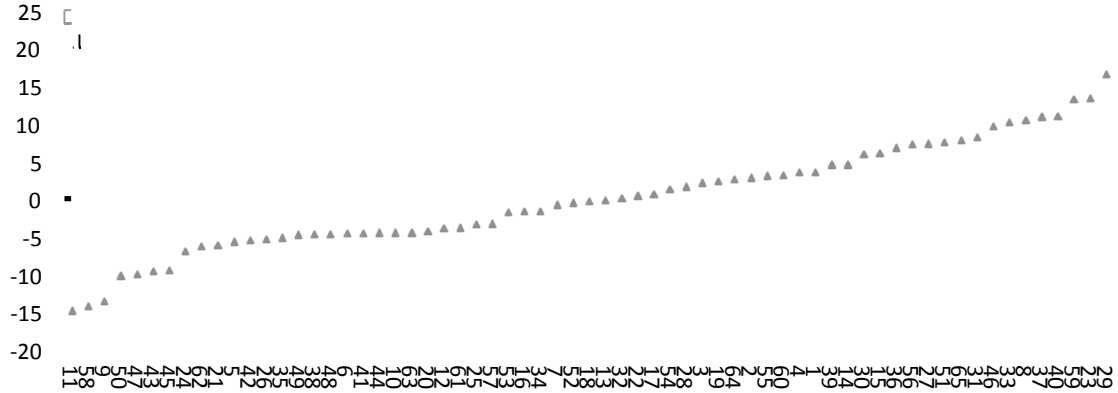
M2.4_3



			R_cat_M2.4_3		
			igual a la media	**Superior a la media**	**Inferior a la media**
R_cat_M1_3	**igual a la media**	Recuento	8	2	1
		%	72,7%	18,2%	9,1%
	Superior a la media	Recuento	14	6	6
		%	53,8%	23,1%	23,1%
	Inferior a la media	Recuento	16	4	8
		%	57,1%	14,3%	28,6%

Chi-cuadrado de Pearson= 2,350 P=0,672

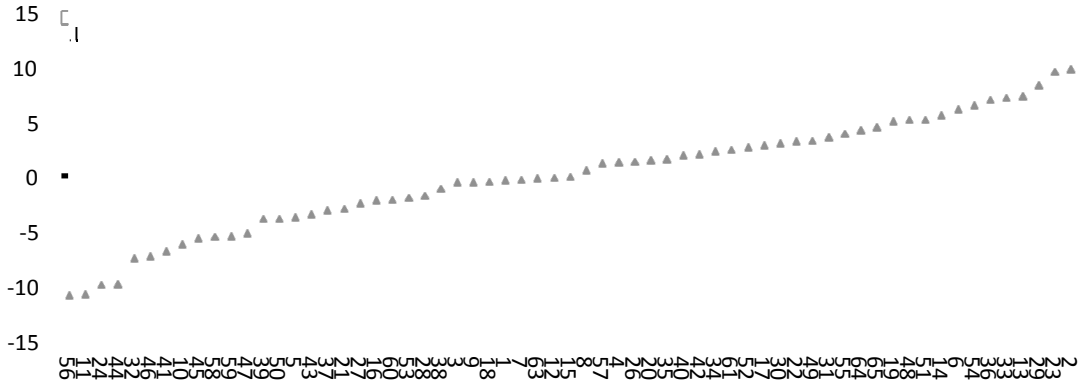
M2.5_3



			R_cat_M2.5_3		
			igual a la media	**Superior a la media**	**Inferior a la media**
R_cat_M1_3	**igual a la media**	Recuento	11	0	0
		%	100%	0%	0%
	Superior a la media	Recuento	10	16	0
		%	38,5%	61,5%	,0%
	Inferior a la media	Recuento	12	0	16
		%	42,9%	0%	57,1%

Chi-cuadrado de Pearson= 51,515 P=0,000

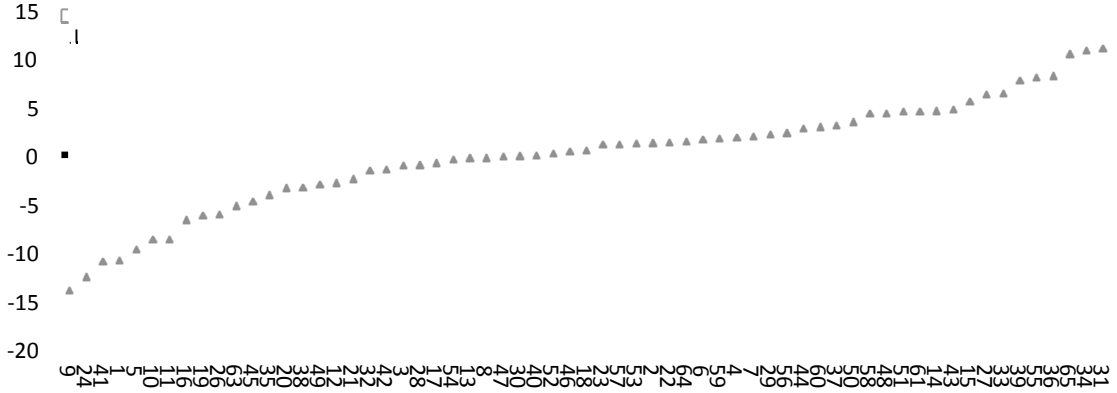
M3_3



			R_cat_M3_3		
			igual a la media	**Superior a la media**	**Inferior a la media**
R_cat_M1_3	**igual a la media**	Recuento	9	1	1
		%	81,8%	9,1%	9,1%
	Superior a la media	Recuento	10	12	4
		%	38,5%	46,2%	15,4%
	Inferior a la media	Recuento	14	2	12
		%	50,0%	7,1%	42,9%

Chi-cuadrado de Pearson= 18,245 P=0,001

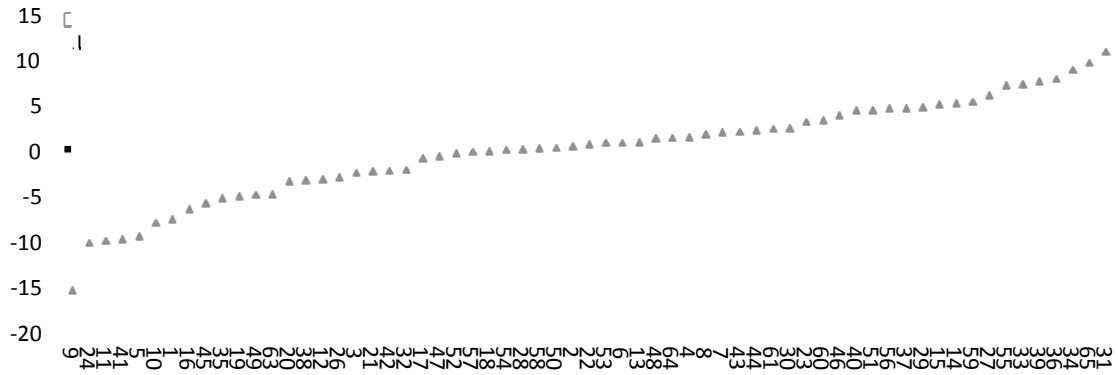
M4_3



			R_cat_M4_3		
			igual a la media	**Superior a la media**	**Inferior a la media**
R_cat_M1_3	**igual a la media**	Recuento	6	3	2
		%	54,5%	27,3%	18,2%
	Superior a la media	Recuento	7	16	3
		%	26,9%	61,5%	11,5%
	Inferior a la media	Recuento	3	9	16
		%	10,7%	32,1%	57,1%

Chi-cuadrado de Pearson= 19,226 P=0,001

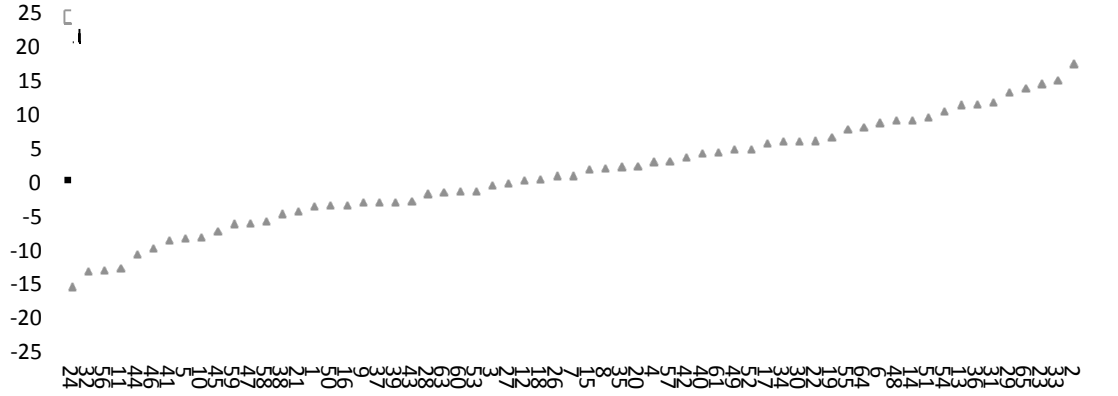
M4.1_3



			R_cat_M4.1_3		
			igual a la media	**Superior a la media**	**Inferior a la media**
R_cat_M1_3	**igual a la media**	Recuento	5	2	4
		%	45,5%	18,2%	36,4%
	Superior a la media	Recuento	5	19	2
		%	19,2%	73,1%	7,7%
	Inferior a la media	Recuento	6	6	16
		%	21,4%	21,4%	57,1%

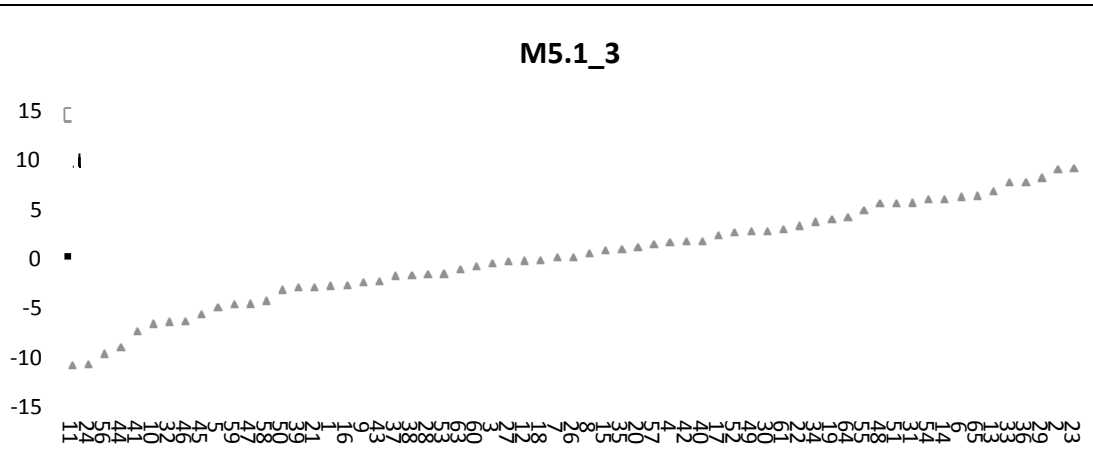
Chi-cuadrado de Pearson= 22,524 P=0,000

M5_3



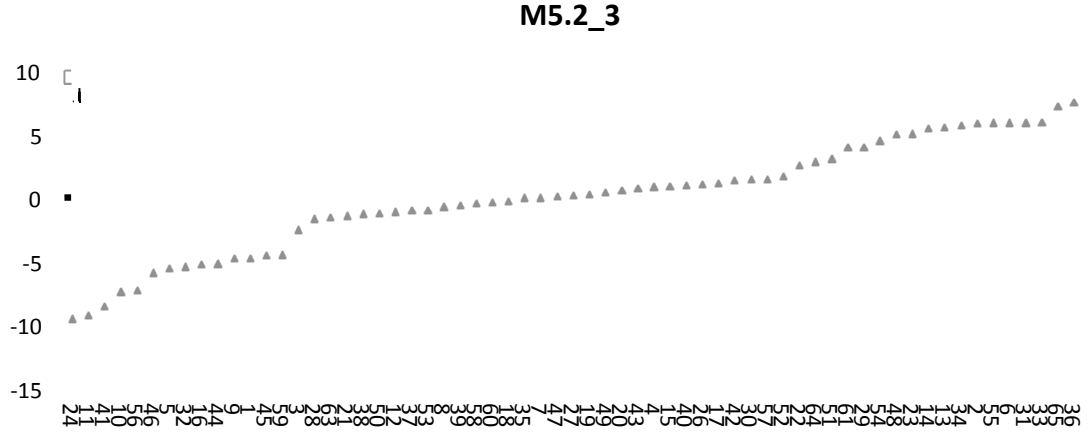
			R_cat_M5_3		
			igual a la media	**Superior a la media**	**Inferior a la media**
R_cat_M1_3	**igual a la media**	Recuento	9	1	1
		%	81,8%	9,1%	9,1%
	Superior a la media	Recuento	9	13	4
		%	34,6%	50,0%	15,4%
	Inferior a la media	Recuento	16	3	9
		%	57,1%	10,7%	32,1%

Chi-cuadrado de Pearson= 15,652 P=0,004



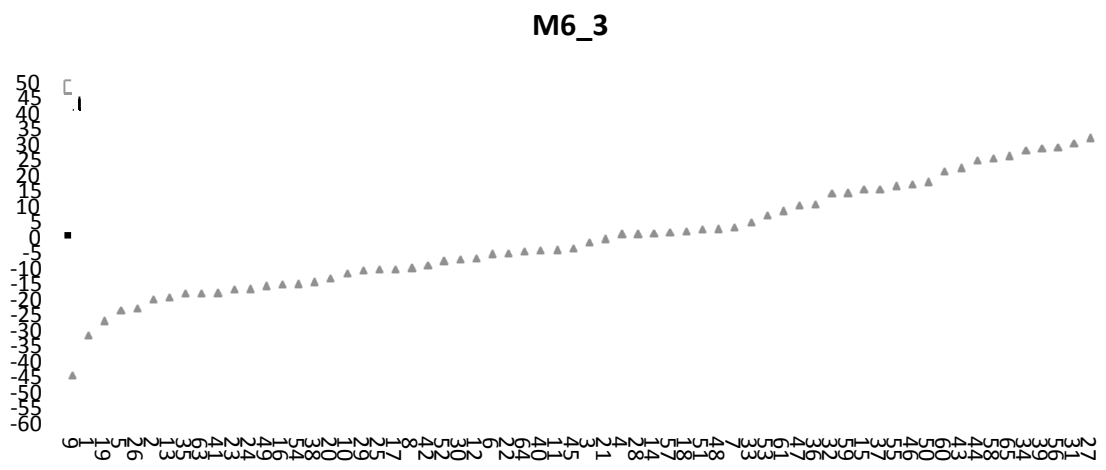
			R_cat_M5.1_3		
			igual a la media	**Superior a la media**	**Inferior a la media**
R_cat_M1_3	**igual a la media**	Recuento	7	3	1
		%	63,6%	27,3%	9,1%
	Superior a la media	Recuento	4	15	7
		%	15,4%	57,7%	26,9%
	Inferior a la media	Recuento	6	6	16
		%	21,4%	21,4%	57,1%

Chi-cuadrado de Pearson= 18,549 P=0,001



			R_cat_M5.2_3		
			igual a la media	**Superior a la media**	**Inferior a la media**
R_cat_M1_3	**igual a la media**	Recuento	6	3	2
		%	54,5%	27,3%	18,2%
	Superior a la media	Recuento	6	14	6
		%	23,1%	53,8%	23,1%
	Inferior a la media	Recuento	8	6	14
		%	28,6%	21,4%	50,0%

Chi-cuadrado de Pearson= 15,652 P=04



			R_cat_M6_3		
			igual a la media	**Superior a la media**	**Inferior a la media**
R_cat_M1_3	**igual a la media**	Recuento	9	1	1
		%	81,8%	9,1%	9,1%
	Superior a la media	Recuento	8	11	7
		%	30,8%	42,3%	26,9%
	Inferior a la media	Recuento	12	7	9
		%	42,9%	25,0%	32,1%

Chi-cuadrado de Pearson= 9,382 P=0,052

Figura VIII.3. Grafico de las puntuaciones de las escuelas e Intervalo de Confianza al 95% y tablas de contingencia que reflejan los cambios en las posiciones respecto al modelo base en el problema 3.

En la Figura VIII.3 se observa que hay variaciones entre las puntuaciones estimadas con los diferentes modelos de este problema respecto al modelo base (M1_3), excepto el M1.2_3 que no produce cambios entre los centros que fueron clasificados como significativamente distintos de la media y aquellos que no los son. Esto no quiere decir que algunos centros cambien su posición en el ranking. Por ejemplo, en el M1_3 el centro número 58 obtiene el segundo residuo más bajo, en cambio, en el M1.1_3 es el tercero por la cola.

Las diferencias del modelo base con el M1.2_3 son aspectos que ya se mencionaron en los resultados del problema 2[92] y hacen que este último sea más conservador con las puntuaciones de determinados centros. Principalmente los cambios son esas cinco escuelas que pasan de estar por encima de la media a no diferenciarse de ella y tres que pasan de no diferenciarse a estar por debajo. Además, otra de las escuelas sufre un cambio radical al cambiar de estar por encima de la media a estar por debajo, es la escuela número 19. La relación entre las categorías es significativa con un valor de χ^2=59,714 (P=0,000).

Los cambios que se producen entre las tres categorías con respecto al M2_3 anulan la relación, reduciendo el valor de χ^2 hasta que no resulta significativo (χ^2=3,683; P=0,451). Este modelo identifica un mayor número de centros distintos de la media debido a que las estimaciones, obtenidas mediante el modelo lineal mixto, del residuo de las escuelas en la última aplicación tienen asociados errores típicos más bajos. Únicamente seis centros, frente a los 11 del modelo base, no se diferencian de la media, aunque solo uno de ellos estaba en esa categoría en la clasificación del M1_3. Los otros 10 cambian de la siguiente forma: 6 pasan a estar por debajo de la media y 4 por encima. De los 26 clasificados como superiores a la media en el M1_3, diez pasan a estar por debajo. Y de los 28 que estaban inicialmente por debajo, ocho pasan a obtener una puntuación por encima.

Con el resto de modelos basados en el análisis lineal mixto (M2.1_3-M2.5_3) también hay cambios sustanciales. Estos modelos tienen mayores errores de estimación al estimarse a partir de la diferencia entre dos residuos y, por tanto, identifican un menor número de centros distintos de la media. El número de escuelas que no se diferencian de la media cambia de 11 en el modelo base a unos

[92]Para más detalle consultar el apartado VIII.3.2.2

30 aproximadamente en estos modelos. La mayor similitud en la clasificación se encuentra con el M2.5_5, con valores de χ^2 similares a los que se obtenían con el M1.2_3 (χ^2=51,515; P=0,000). Los únicos cambios que se producen son esos centros que pasan a no diferenciarse de la media, el resto de categorías mantiene a los centros, es decir, no hay escuelas que cambian de tener un residuo inferior a la media a tenerlo superior y viceversa. Si cambian ligeramente las posiciones de algunos centros, como el centro 51 que obtiene el tercer residuo más alto en el M1_3 y ocupa la undécima posición en el M2.5_3. No hay relación, en cambio, entre el M1_3 y el periodo de verano M2.4_3 (χ^2=2,350; P=0,672).

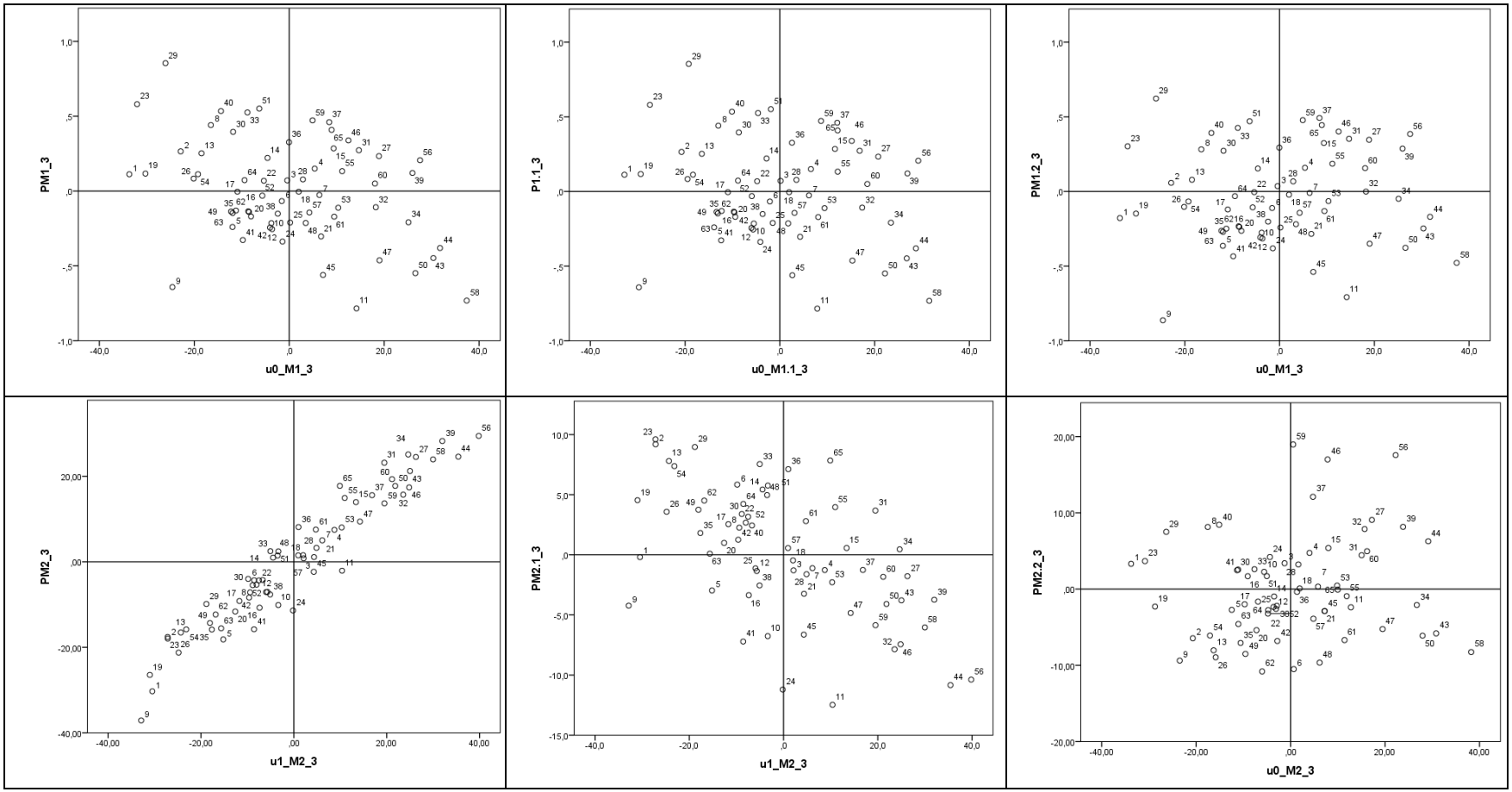
La mayor parte de los cambios que se producen entre el modelo base (M1_3) y la ganancia estimada (M3_3) se encuentran también en los centros que pasan a no diferenciarse de la media. Aunque también algunos centros pasan de tener valores superiores a la media en el modelo inicial a situarse por debajo en este modelo de ganancia estimada. No obstante, la relación es significativa (χ^2=18,245; P=0,001). Es debido a ese mayor tamaño de los errores típicos.

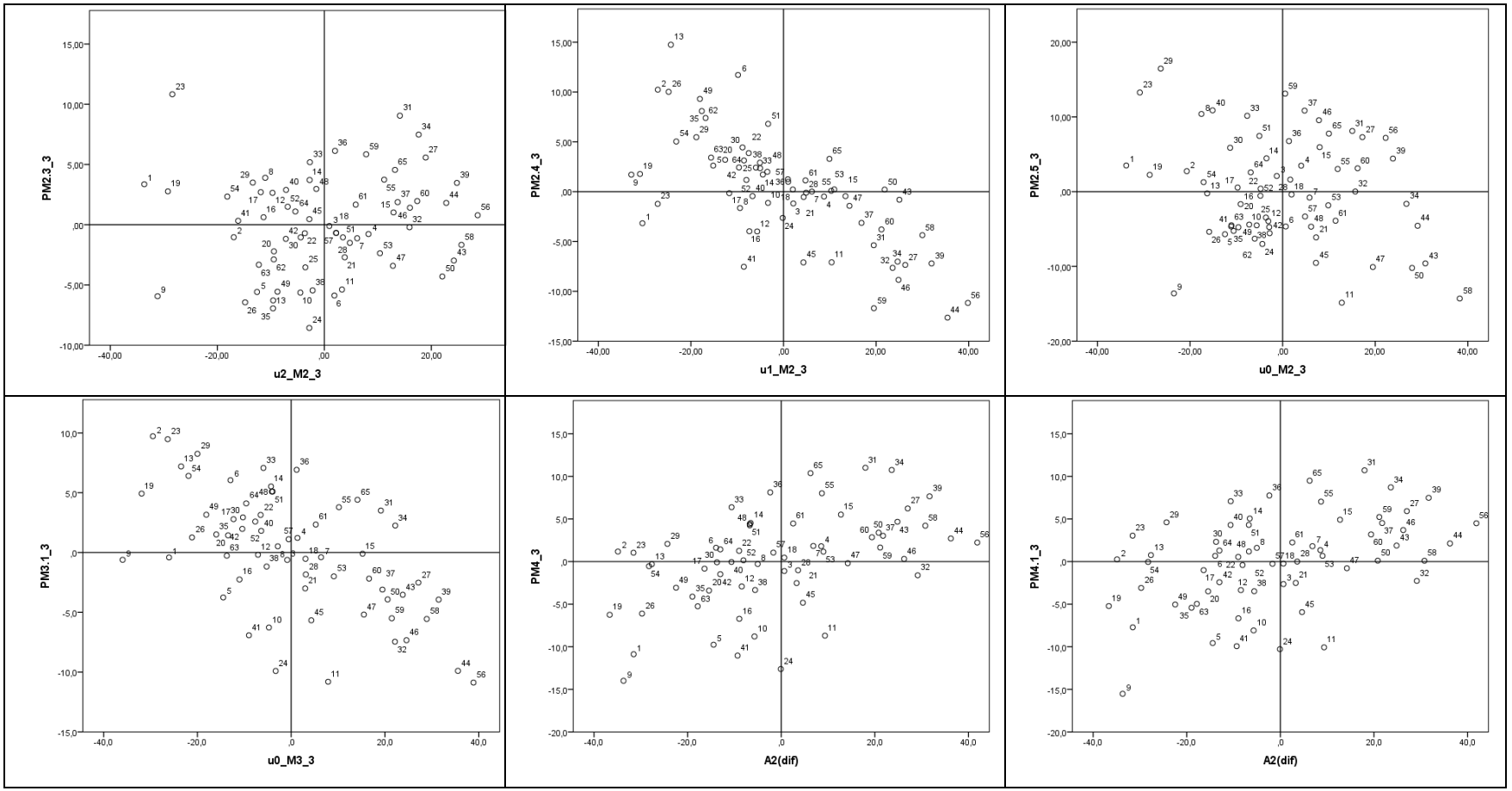
Los modelos de ganancia residual (M4_3 y M4.1_3) clasifican una mayor número de centros como iguales a la media, un total de 16. Cuatro más que en el modelo base M1_3. También hay cambios en las posiciones de algunos centros, por ejemplo, nueve centros cambian de estar por debajo de la media a estar por encima en M4_4. Son solo seis, en el caso del modelo con dos covariables (M4.1_3). El estadístico χ^2 resulta significativo en ambos casos, con un valor más alto en el caso de M4.1_3 (χ^2=22,254; P=0,000 frente χ^2=19,226; P=0,001).

La clasificación realizada mediante el modelo de ganancia bruta (M5_3) sigue estando relacionada con el modelo multinivel longitudinal (M1_3) (χ^2=15,652; P=0,004), aunque muchas de las escuelas pasan a no diferenciarse de la media, un total de 34, debido a la cuestión mencionada de los errores típicos. Los modelos multinivel realizados con esa puntuación de ganancia bruta (M5.1_3 y M5.2_3) reducen ese número de centros que no se diferencian de la media respecto, un total de 17 y 20 respectivamente. Aunque la relación también es significativa entre estos modelos y el de base, existe un número considerable de centros que cambian su posición respecto a la media, un total de 12 en ambos casos.

Finalmente, si los resultados de la clasificación de M1_3 se comparan con los resultados brutos de la escuelas en la cuarta toma de datos la relación deja de ser significativa (χ^2=9,382; P=0,052) .

El Gráfico VIII.5 y la Tabla VIII.27, que aparecen a continuación, analizan como la puntuación de los modelos se relaciona con el estatus inicial utilizado en cada uno de ellos y la comparación de esos resultados con el modelo base M1_3. No es posible calcular resultados en este apartado para el M6_4 que únicamente utiliza la puntuación bruta en la última aplicación de la evaluación.


PM1_3
u0_M1_3
P1.1_3
u0_M1.1_3
PM1.2_3
u0_M1_3
PM2_3
u1_M2_3
PM2.1_3
u1_M2_3
PM2.2_3
u0_M2_3


PM2.3_3
u2_M2_3
PM2.4_3
u1_M2_3
PM2.5_3
u0_M2_3
PM3.1_3
u0_M3_3
PM4_3
A2(dif)
PM4.1_3
A2(dif)

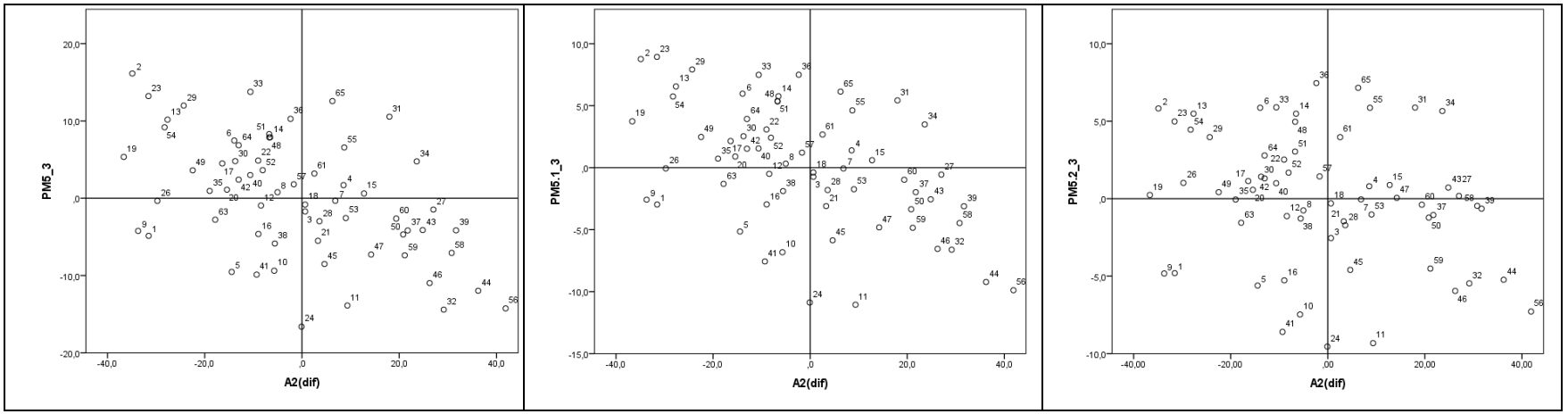


Gráfico VIII.5 Gráficos de dispersión de las puntuaciones de las escuelas y su correspondiente estatus inicial en los modelos del problema 3.

Los gráficos muestran una relación más intensa entre estatus inicial y la puntuación PM2_3. Es tan alta porque que se correlaciona el residuo de la segunda aplicación con el de la última toma de datos. Los resultados de este modelo son problemáticos debido a esa gran distancia con el modelo EVAAS, que originalmente asume la persistencia de los efectos, por lo que debe interpretarse con cautela.

			Disp_cat_M1.1_3			
			Bajo Estatus y Alto Crecimiento	**Bajo Estatus y Bajo Crecimiento**	**Alto Estatus y Alto Crecimiento**	**Alto Estatus y Bajo Crecimiento**
Disp_cat_M1_3	**Bajo Estatus y Alto Crecimiento**	Recuento	16	0	2	0
		%	88,9%	,0%	11,1%	,0%
	Bajo Estatus y Bajo Crecimiento	Recuento	0	17	0	0
		%	,0%	100%	,0%	,0%
	Alto Estatus y Alto Crecimiento	Recuento	0	0	13	0
		%	,0%	,0%	100%	,0%
	Alto Estatus y Bajo Crecimiento	Recuento	0	1	0	16
		%	,0%	5,9%	,0%	94,1%

Chi-cuadrado de Pearson= 172,852 P=0,000

			Disp_cat_M1.2_3			
			Bajo Estatus y Alto Crecimiento	**Bajo Estatus y Bajo Crecimiento**	**Alto Estatus y Alto Crecimiento**	**Alto Estatus y Bajo Crecimiento**
Disp_cat_M1_3	**Bajo Estatus y Alto Crecimiento**	Recuento	12	6	0	0
		%	66,7%	33,3%	,0%	,0%
	Bajo Estatus y Bajo Crecimiento	Recuento	0	17	0	0
		%	,0%	100%	,0%	,0%
	Alto Estatus y Alto Crecimiento	Recuento	0	0	13	0
		%	,0%	,0%	100%	,0%
	Alto Estatus y Bajo Crecimiento	Recuento	0	0	0	17
		%	,0%	,0%	,0%	100%

Chi-cuadrado de Pearson= 162,029 P=0,000

			Disp_cat_M2_3			
			Bajo Estatus y Alto Crecimiento	**Bajo Estatus y Bajo Crecimiento**	**Alto Estatus y Alto Crecimiento**	**Alto Estatus y Bajo Crecimiento**
Disp_cat_M1_3	**Bajo Estatus y Alto Crecimiento**	Recuento	3	13	2	0
		%	16,7%	72,2%	11,1%	,0%
	Bajo Estatus y Bajo Crecimiento	Recuento	0	17	0	0
		%	,0%	100%	,0%	,0%
	Alto Estatus y Alto Crecimiento	Recuento	0	0	13	0
		%	,0%	,0%	100%	,0%
	Alto Estatus y Bajo Crecimiento	Recuento	1	1	13	2
		%	5,9%	5,9%	76,5%	11,8%

Chi-cuadrado de Pearson= 60,955 P=0,000

			Disp_cat_M2.1_3			
			Bajo Estatus y Alto Crecimiento	**Bajo Estatus y Bajo Crecimiento**	**Alto Estatus y Alto Crecimiento**	**Alto Estatus y Bajo Crecimiento**
Disp_cat_M1_3	**Bajo Estatus y Alto Crecimiento**	Recuento	15	1	1	1
		%	83,3%	5,6%	5,6%	5,6%
	Bajo Estatus y Bajo Crecimiento	Recuento	9	8	0	0
		%	52,9%	47,1%	,0%	,0%
	Alto Estatus y Alto Crecimiento	Recuento	0	0	4	9
		%	,0%	,0%	30,8%	69,2%
	Alto Estatus y Bajo Crecimiento	Recuento	1	1	3	12
		%	5,9%	5,9%	17,6%	70,6%

Chi-cuadrado de Pearson= 63,608 P=0,000

			Disp_cat_M2.2_3			
			Bajo Estatus y Alto Crecimiento	**Bajo Estatus y Bajo Crecimiento**	**Alto Estatus y Alto Crecimiento**	**Alto Estatus y Bajo Crecimiento**
Disp_cat_M1_3	**Bajo Estatus y Alto Crecimiento**	Recuento	9	8	0	1
		%	50,0%	44,4%	,0%	5,6%
	Bajo Estatus y Bajo Crecimiento	Recuento	4	12	0	1
		%	23,5%	70,6%	,0%	5,9%
	Alto Estatus y Alto Crecimiento	Recuento	0	0	11	2
		%	,0%	,0%	84,6%	15,4%
	Alto Estatus y Bajo Crecimiento	Recuento	0	1	5	11
		%	,0%	5,9%	29,4%	64,7%

Chi-cuadrado de Pearson= 76,071 P=0,000

			Disp_cat_M2.3_3			
			Bajo Estatus y Alto Crecimiento	**Bajo Estatus y Bajo Crecimiento**	**Alto Estatus y Alto Crecimiento**	**Alto Estatus y Bajo Crecimiento**
Disp_cat_M1_3	**Bajo Estatus y Alto Crecimiento**	Recuento	10	5	1	2
		%	55,6%	27,8%	5,6%	11,1%
	Bajo Estatus y Bajo Crecimiento	Recuento	5	11	0	1
		%	29,4%	64,7%	,0%	5,9%
	Alto Estatus y Alto Crecimiento	Recuento	0	0	11	2
		%	,0%	,0%	84,6%	15,4%
	Alto Estatus y Bajo Crecimiento	Recuento	2	1	3	11
		%	11,8%	5,9%	17,6%	64,7%

Chi-cuadrado de Pearson= 69,690 P=0,000

			Disp_cat_M2.4_3			
			Bajo Estatus y Alto Crecimiento	**Bajo Estatus y Bajo Crecimiento**	**Alto Estatus y Alto Crecimiento**	**Alto Estatus y Bajo Crecimiento**
Disp_cat_M1_3	**Bajo Estatus y Alto Crecimiento**	Recuento	12	4	1	1
		%	66,7%	22,2%	5,6%	5,6%
	Bajo Estatus y Bajo Crecimiento	Recuento	11	6	0	0
		%	64,7%	35,3%	,0%	,0%
	Alto Estatus y Alto Crecimiento	Recuento	0	0	2	11
		%	,0%	,0%	15,4%	84,6%
	Alto Estatus y Bajo Crecimiento	Recuento	2	0	6	9
		%	11,8%	,0%	35,3%	52,9%

Chi-cuadrado de Pearson= 56,107 P=0,000

			Disp_cat_M2.5_3			
			Bajo Estatus y Alto Crecimiento	**Bajo Estatus y Bajo Crecimiento**	**Alto Estatus y Alto Crecimiento**	**Alto Estatus y Bajo Crecimiento**
Disp_cat_M1_3	**Bajo Estatus y Alto Crecimiento**	Recuento	15	2	1	0
		%	83,3%	11,1%	5,6%	,0%
	Bajo Estatus y Bajo Crecimiento	Recuento	1	15	0	1
		%	5,9%	88,2%	,0%	5,9%
	Alto Estatus y Alto Crecimiento	Recuento	0	0	13	0
		%	,0%	,0%	100%	,0%
	Alto Estatus y Bajo Crecimiento	Recuento	0	1	1	15
		%	,0%	5,9%	5,9%	88,2%

Chi-cuadrado de Pearson=145,666 P=0,000

			Disp_cat_M3_3			
			Bajo Estatus y Alto Crecimiento	**Bajo Estatus y Bajo Crecimiento**	**Alto Estatus y Alto Crecimiento**	**Alto Estatus y Bajo Crecimiento**
Disp_cat_M1_2	**Bajo Estatus y Alto Crecimiento**	Recuento	15	2	1	0
		%	83,3%	11,1%	5,6%	,0%
	Bajo Estatus y Bajo Crecimiento	Recuento	7	10	0	0
		%	41,2%	58,8%	,0%	,0%
	Alto Estatus y Alto Crecimiento	Recuento	0	0	4	9
		%	,0%	,0%	30,8%	69,2%
	Alto Estatus y Bajo Crecimiento	Recuento	2	1	2	12
		%	11,8%	5,9%	11,8%	70,6%

Chi-cuadrado de Pearson= 67,748 P=0,000

			Disp_cat_M4_3			
			Bajo Estatus y Alto Crecimiento	**Bajo Estatus y Bajo Crecimiento**	**Alto Estatus y Alto Crecimiento**	**Alto Estatus y Bajo Crecimiento**
Disp_cat_M1_2	**Bajo Estatus y Alto Crecimiento**	Recuento	9	8	0	1
		%	50,0%	44,4%	,0%	5,6%
	Bajo Estatus y Bajo Crecimiento	Recuento	2	14	0	0
		%	12,5%	87,5%	,0%	,0%
	Alto Estatus y Alto Crecimiento	Recuento	0	0	12	1
		%	,0%	,0%	92,3%	7,7%
	Alto Estatus y Bajo Crecimiento	Recuento	2	0	9	5
		%	12,5%	,0%	56,3%	31,3%

Chi-cuadrado de Pearson= 70,165 P=0,000

			Disp_cat_M4.1_3			
			Bajo Estatus y Alto Crecimiento	**Bajo Estatus y Bajo Crecimiento**	**Alto Estatus y Alto Crecimiento**	**Alto Estatus y Bajo Crecimiento**
Disp_cat_M1_2	**Bajo Estatus y Alto Crecimiento**	Recuento	13	4	0	1
		%	72,2%	22,2%	,0%	5,6%
	Bajo Estatus y Bajo Crecimiento	Recuento	1	15	0	0
		%	6,3%	93,8%	,0%	,0%
	Alto Estatus y Alto Crecimiento	Recuento	0	0	12	1
		%	,0%	,0%	92,3%	7,7%
	Alto Estatus y Bajo Crecimiento	Recuento	1	1	8	6
		%	6,3%	6,3%	50,0%	37,5%

Chi-cuadrado de Pearson= 90,506 P=0,000

			Disp_cat_M5_3			
			Bajo Estatus y Alto Crecimiento	**Bajo Estatus y Bajo Crecimiento**	**Alto Estatus y Alto Crecimiento**	**Alto Estatus y Bajo Crecimiento**
Disp_cat_M1_1	**Bajo Estatus y Alto Crecimiento**	Recuento	15	2	0	1
		%	83,3%	11,1%	,0%	5,6%
	Bajo Estatus y Bajo Crecimiento	Recuento	7	10	0	0
		%	41,2%	58,8%	,0%	,0%
	Alto Estatus y Alto Crecimiento	Recuento	0	0	5	8
		%	,0%	,0%	38,5%	61,5%
	Alto Estatus y Bajo Crecimiento	Recuento	2	1	2	12
		%	11,8%	5,9%	11,8%	70,6%
Chi-cuadrado de Pearson= 69,785 P=0,000						

			Disp_cat_M5.1_3			
			Bajo Estatus y Alto Crecimiento	**Bajo Estatus y Bajo Crecimiento**	**Alto Estatus y Alto Crecimiento**	**Alto Estatus y Bajo Crecimiento**
Disp_cat_M1_1	**Bajo Estatus y Alto Crecimiento**	Recuento	15	2	0	1
		%	83,3%	11,1%	,0%	5,6%
	Bajo Estatus y Bajo Crecimiento	Recuento	7	9	0	0
		%	43,8%	56,3%	,0%	,0%
	Alto Estatus y Alto Crecimiento	Recuento	0	0	5	8
		%	,0%	,0%	38,5%	61,5%
	Alto Estatus y Bajo Crecimiento	Recuento	2	0	2	12
		%	12,5%	,0%	12,5%	75,0%
Chi-cuadrado de Pearson= 70,268 P=0,000						

			Disp_cat_M5.2_3			
			Bajo Estatus y Alto Crecimiento	**Bajo Estatus y Bajo Crecimiento**	**Alto Estatus y Alto Crecimiento**	**Alto Estatus y Bajo Crecimiento**
Disp_at_M1_1	**Bajo Estatus y Alto Crecimiento**	Recuento	15	2	0	1
		%	83,3%	11,1%	,0%	5,6%
	Bajo Estatus y Bajo Crecimiento	Recuento	6	10	0	0
		%	37,5%	62,5%	,0%	,0%
	Alto Estatus y Alto Crecimiento	Recuento	0	0	6	7
		%	,0%	,0%	46,2%	53,8%
	Alto Estatus y Bajo Crecimiento	Recuento	2	0	4	10
		%	12,5%	,0%	25,0%	62,5%
Chi-cuadrado de Pearson= 71,074 P=0,000						

Tabla VIII.27 Tablas de contingencia y χ^2 para la relación. Cambios en los cuadrantes del gráfico de dispersión respecto al modelo base en el problema 3.

Las posiciones de los centros en los cuatro cuadrantes de los modelos multinivel longitudinales son muy similares. Con solo tres centros que cambian en M1.1_3 y seis en M1.2_3, respecto al modelo base (M1_3). Aunque son centros de clasificados en cuadrantes distintos, es decir, en el M1.1_3 son dos centros educativos que cambian de bajo estatus y alto crecimiento a tener también alto estatus inicial y otra que pasa de alto estatus y bajo crecimiento a estar por debajo también en ese estatus inicial; en cambo, en el M1.2_3 las seis escuelas pasan de tener bajo estatus y alto crecimiento a estar por debajo en crecimiento. Los valores de χ^2 elevados indican la alta relación entre ambas clasificaciones, alrededor de 170.

Un valor de correlación cercano obtiene la comparación con el M2.5_3 (χ^2=145,666; P=0,000). Un total de siete centros cambian de cuadrante respecto al modelo base. El modelo base también está relacionado, aunque en menor medida con las variantes del modelo lineal mixto que calculan las ganancias en cada curso evaluado, son M2.2_3 (curso1) y M2.3_4 (curso2) y los valores de χ^2 son 76,071 (P=0,000) y 69,690 (P=0,000) respectivamente. En ambos modelos, un total de 22 centros cambia de cuadrante pero los modelos que cambian el estatus inicial respecto al modelo base, es decir, que cambien de tener un nivel inicial por debajo de la media a estar por encima y viceversa, son solo tres en el M2.2_3 y siete en el M2.3_3. Recordemos que el M2.2_3 utiliza la tercera aplicación como punto de partida porque la puntuación en la ganancia en el segundo curso, entre la tercera y cuarta aplicación. Aun así la mayoría de la escuelas mantiene su estatus inicial con respecto a la media también en el segundo curso.

Con respecto a los modelos de ganancia, la comparación muestra una mayor relación con M4.1_3 (χ^2=90,506; P=0,000), mayor incluso que la relación con las ganancias de cada uno de los cursos (M2.2_3 y M2.3_3). El M4.1_3 estima el residuo ajustado de la puntuación en la última aplicación, utilizando dos predictores de rendimiento previo como covariables en el modelo multinivel. El mayor cambio de cuadrante se produce entre las escuelas clasificadas por el modelo base con alto estatus y bajo crecimiento. En el M4.1_3, ocho de esas escuelas superan a la media en crecimiento.

En resumen, una vez estudiados los residuos que los diferentes modelos de ganancia y crecimiento estiman para las escuelas, se ha hallado un gran parecido entre las puntuaciones de los distintos modelos multinivel longitudinales. También en las clasificaciones que realizan a partir de esas estimaciones.

De forma opuesta, utilizar las medias brutas de las escuelas produce resultados totalmente distintos a los conseguidos con el modelo multinivel longitudinal (no se ha encontrado correlación significativa entre estos modelos). No obstante, las correlaciones de los residuos muestran que está relacionado con las estimaciones obtenidas por el modelo de ganancia residual, con valores de Pearson por encima de 0,7.

Si se comparan los resultados del modelo multinivel longitudinal y los distintos tipos de ganancia (bruta, residual y estimada), hay una menor distancia. Se ha demostrado que tienen cierta similitud y pueden ser modelos válidos para estimar el VA cuando las evaluaciones cuentan con dos únicas mediciones del rendimiento. Las correlaciones muestran un gran parecido entre las estimaciones del modelo de ganancia estimada y los modelos de ganancia bruta, con valores de Pearson por encima de 0,9.

El modelo lineal mixto se ha revelado como un tipo de análisis muy útil si sus estimaciones se emplean en el cálculo de ganancias. Por ejemplo, calcular las ganancias en cada curso sin considerar el periodo de verano. Además, las estimaciones del modelo multinivel longitudinal y la diferencia entre los residuos de la primera y última aplicación del modelo lineal mixto es prácticamente idéntica. Por tanto, con este tipo de datos, el análisis lineal mixto parece adecuado.

Empleando los modelos lineales mixtos se consiguen dos puntuaciones de cambio para cada escuela, una por curso. Este característica puede ser una ventaja a la hora de elaborar las estimaciones de VA. Con esta información es posible saber si los centros educativos obtienen puntuaciones estables de crecimiento en cada curso, es decir, si su ritmo de cambio se mantiene o si ha sufrido alguna variación destacable. Por ejemplo, si los centros que tienen una ganancia superior a la media en el primer curso mantienen esa tendencia en el segundo.

A modo de ilustración se incluyen las Tabla VIII.28 y Tabla VIII.29. En la primera de ellas se compara el ranking realizado utilizando el residuo de crecimiento,

estimado mediante el modelo lineal mixto, para el primer curso y para el segundo. Se pone la atención sobre la cantidad de escuelas que se encuentran significativamente por encima o por debajo de la media, o escuelas que no se diferencian de esa media global. En la segunda tabla, se compara la clasificación de escuelas que se obtiene al relacionar el estatus inicial con la medida de ganancia.

			R_cat_M2.3_3		
			igual a la media	**Superior a la media**	**Inferior a la media**
R_cat_M2.1_3	**igual a la media**	Recuento	16	11	7
		%	47,1%	32,4%	20,6%
	Superior a la media	Recuento	9	4	1
		%	64,3%	28,6%	7,1%
	Inferior a la media	Recuento	7	1	9
		%	41,2%	5,9%	52,9%

Chi-cuadrado de Pearson= 11,242 P=0,024

Tabla VIII.28. Tabla de contingencia y χ2 para la relación entre las clasificaciones de las estimaciones de ganancia intra-curso en el modelo lineal mixto.

			Disp_cat_M2.3_3			
			Bajo Estatus y Alto Crecimiento	**Bajo Estatus y Bajo Crecimiento**	**Alto Estatus y Alto Crecimiento**	**Alto Estatus y Bajo Crecimiento**
Disp_cat_M2.1_1	**Bajo Estatus y Alto Crecimiento**	Recuento	8	3	0	2
		%	61,5%	23,1%	,0%	15,4%
	Bajo Estatus y Bajo Crecimiento	Recuento	7	14	0	0
		%	33,3%	66,7%	,0%	,0%
	Alto Estatus y Alto Crecimiento	Recuento	0	0	10	6
		%	,0%	,0%	62,5%	37,5%
	Alto Estatus y Bajo Crecimiento	Recuento	2	0	5	8
		%	13,3%	,0%	33,3%	53,3%

Chi-cuadrado de Pearson= 64,128 P=0,000

Tabla VIII.29. Tabla de contingencia y χ^2 para la relación entre las clasificaciones de las estimaciones de ganancia intra-curso y estatus inicial en el modelo lineal mixto.

Los resultados de la Tabla VIII.29 reflejan cambios entre las ganancias de las escuelas en cada curso, pero no en el estatus inicial. Únicamente dos escuelas que partían con bajo estatus en el primer curso pasan a tenerlo alto en el segundo y otras dos en sentido opuesto. Por tanto, los centros educativos pueden tener ganancias distintas en cada curso y conviene llevar un estudio por separado de cada uno de ellos.

La Tabla VIII.28 también muestra cambios en la posición de algunas escuelas respecto a la media si se comparan los resultados de cada curso por separado.

Capítulo IX: Conclusiones, limitaciones y prospectiva

El último capítulo de la tesis se dedica a la recopilación de los principales hallazgos encontrados a lo largo de todo el trabajo. Se extraen conclusiones de la revisión teórica que se ha llevado a cabo sobre el Valor Añadido en educación. Y también, por supuesto, un resumen de los resultados y las conclusiones de los dos estudios empíricos realizados.

Otras secciones de este capítulo son las limitaciones halladas en el desarrollo del trabajo y las posibles líneas de investigación futuras o estudios empíricos que pueden realizarse también con estos datos.

IX.1 Conclusiones

El auge de las evaluaciones generales en los sistemas educativos de muchos países es un hecho constatado. Pensar que es una cuestión pasajera o que carece de importancia dentro del sistema educativo es un error. La información obtenida con estas evaluaciones puede ser utilizada con fines distintos, aunque no todos ellos son recomendables. El diagnóstico de la situación educativa, la elección de escuela o la rendición de cuentas con alto o bajo impacto son algunos de los ejemplos de posible utilización de los resultados de las evaluaciones.

Es necesario contar con herramientas que proporcionen una información lo más fiable y ajustada a la realidad escolar. No se pueden tomar decisiones sobre las escuelas utilizando como justificación datos que pueden estar sesgados o no ser del todo fieles a lo que está pasando en un centro educativo. Y si la tendencia es

utilizar sistemas de evaluación de centros basados en la información que aportan las respuestas de sus estudiantes a pruebas estandarizadas, conviene contar con una medida fiable y precisa del rendimiento de los centros educativos y el VA es una posible solución.

El VA se entiende como la estimación de la contribución de la escuela o el docente al aprendizaje del estudiante, intentando aislarla de otros posibles factores que se escapan de su control. Se utiliza el término VA para hacer referencia a esa estimación, al dato concreto, pero también a la metodología de análisis de la información y, de forma global, al diseño de la evaluación.

Uno de los aspectos característicos de esta metodología es la consideración del aprendizaje como un proceso de cambio y no como un aspecto estático. El análisis del VA pone la atención en el estudio de la ganancia o crecimiento en rendimiento, es decir, trata de evaluar el cambio que se produce en el logro académico de los estudiantes dentro de un centro educativo. Poner el foco de atención en el estudio del cambio se adecúa en mayor medida a la concepción del aprendizaje que la que subyace en los modelos transversales que analizan el logro escolar en un punto temporal concreto.

Otro rasgo característico es el concepto de aportación. Esta metodología trata de averiguar cuál es la "verdadera" aportación de una escuela al aprendizaje. Debe evaluarse el proceso que ocurre dentro de la escuela independientemente de otros elementos que pueden influir en los resultados, pero que escapan al control de la institución educativa.

Sin embargo, averiguar ese efecto "verdadero", ese efecto causal de la escuela sobre el aprendizaje del estudiante es muy difícil, por no decir imposible. La propia operativización del VA, como residuo de un modelo estadístico, es imperfecta y dependerá de las variables que se incluyan en ese modelo.

Los sistemas de evaluación basados en la rendición de cuentas de alto impacto tratan de dotar con carácter causal a las estimaciones de VA. En otras palabras, la diferencia entre las aportaciones estimadas de dos escuelas se interpreta como un reflejo de las diferencias en su eficacia para hacer progresar el aprendizaje de sus estudiantes. Conseguir esto solo posible en diseños experimentales puros, donde los sujetos se distribuyan de forma aleatoria entre las

diferentes escuelas. En los sistemas educativos actuales es improbable que los estudiantes se distribuyan de forma aleatoria. Factores geográficos y de coste son los dos grandes determinantes de la elección de un centro educativo concreto por parte de las familias. Por tanto, las estimaciones de VA deben interpretarse como medidas descriptivas y, como señala Linn (2008), estos resultados siguen teniendo valor sin hacer inferencias causales respecto a la calidad de las escuelas y pueden utilizarse como indicadores para identificar escuelas que requieren una investigación más profunda.

El VA mejora otras formas de evaluación de las escuelas o los docentes como los estudios transversales o la simple comparación de medias en diferentes momentos temporales pero, como apunta Doran (2003), no son la panacea. En consecuencia, no debe ser la única información con la que cuenten los sistemas de rendición de cuentas, sobre todo aquellos que pretenden juzgar a las instituciones educativas con fines sancionadores o reforzadores.

La utilización de los resultados de VA genera un debate sobre si deben ser una herramientas informativa que sirva para identificar la situación de las escuelas, encontrar las mejores prácticas o diagnosticar situaciones problemáticas. O, en cambio, si deben tener un propósito más cercano a la rendición de cuentas de los servicios públicos, con la finalidad de proponer orientaciones sobre la asignación de recursos, proporcionar herramientas a los progenitores para poder elegir escuela o para premiar o sancionar a los centros en función de sus resultados.

No obstante, independientemente de los objetivos que persigan las evaluaciones, los resultados se utilizan para la toma de decisiones sobre los centros educativos, por tanto, es importante que la medida del rendimiento de las escuelas refleje realmente la aportación que dichos centros educativos hacen independientemente de los factores ajenos al control escolar. Se estaría cometiendo en un error sí aquellos sistemas educativos que utilizan una evaluación basada en incentivos o sanciones (alto impacto), no considera los efectos producidos por el contexto del estudiante y la escuela. Por ejemplo, centros que llevan a cabo una selección de estudiantes con buenas notas o de entornos socioeconómicos altos tenderán a obtener buenos resultados en las evaluaciones

transversales pero esto no quiere decir que sea la propia escuela la que produce esos resultados de los estudiantes.

Respecto a los **estudios empíricos** realizados, el **primero** tenía por objeto la **comparación empírica de metodologías de equiparación para la construcción de una escala vertical de rendimiento en matemáticas**. Y se dividió en dos problemas:

A. La comparación de procedimientos para la equiparación horizontal, que trata de averiguar la manera adecuada de garantizar la comparabilidad de las formas de los instrumentos de medida elaborados para un mismo curso.
B. La comparación de procedimientos para el anclaje vertical, donde se analizan los resultados producidos al emplear diferentes metodologías para elaborar una escala vertical de rendimiento en matemáticas.

A continuación se detallan las conclusiones extraídas de estos estudios.

A. Comparación de procedimientos para la equiparación horizontal

En resumen, los resultados del primer problema son los siguientes: en primer lugar, los análisis preliminares desde la TCT, de forma general, señalan una gran similitud entre los resultados medios de aciertos obtenidos por las distintas formas de cada aplicación. Sin embargo, existen algunas diferencias en la fiabilidad y discriminación de las distintas formas elaboradas y también en la longitud, que pueden ser indicadores de la necesidad de aplicar una metodología de calibración horizontal que lleve a cabo una transformación de la habilidad, aunque el diseño inicial se hizo para contar con formas paralelas. El estudio de los ítems comunes muestra un buen funcionamiento de los mismos y, por tanto, pueden ser utilizados en los procesos de calibración horizontal.

En segundo lugar, respecto a los análisis realizados desde la TRI, las diferencias en las medias estimadas en las dos formas de cada aplicación a través de las diferentes metodologías de calibración muestran resultados similares. No

obstante, conviene mencionar que el método de calibración que produce la diferencia de medias más pequeña es la CS sin transformación.

En tercer lugar, el estudio de las distancias horizontales proporciona información más precisa sobre las diferencias en las puntuaciones del rasgo estimadas con las siete metodologías de calibración horizontal, ya que analiza toda la distribución. Aunque las mayores distancias entre metodologías se encuentran en los extremos de la distribución, es destacable que los resultados de la primera aplicación cuando se emplea CS, CC y CF siguen un patrón parecido. En cambio, los resultados producidos por los métodos de CS con transformación se apartan de esa tendencia. En la segunda aplicación son la CS y CF las que apenas producen diferencias en el rasgo entre las formas. En esta aplicación el patrón de la CC es similar al de CSSL, CSMS y CSH. En la tercera aplicación los patrones de las distancias son similares entre las siete metodologías, es la aplicación con menor distancia entre las formas. En la cuarta aplicación la CS, CS, CSSL y CSSH producen resultados similares a lo largo de toda la distribución del rasgo. Por tanto, aunque las distancias calculadas son, en general, bastante parecidas, las metodologías CS, CC y CF son las que producen los resultados más semejantes.

Atendiendo a las medias de esas distancias, en valor absoluto, todas son inferiores a 0,1. De forma concreta, la CS sin transformación es la que menos distancias medias produce entre las formas en las cuatro aplicaciones, seguida de la CF y CC. No obstante, en la CF la distancia media en la última aplicación supera al resto de metodologías.

En conclusión, **a la luz de los resultados observados en la calibración horizontal, la CS sin transformación produce unas estimaciones del rasgo con mayor similitud entre formas** por lo que el diseño de grupos equivalentes cumple con su objetivo. No obstante, emplear CC y CF tiende a producir estimaciones del rasgo también similares entre formas, aunque con pequeñas diferencias entre aplicaciones. Por lo tanto, es recomendable emplear las puntuaciones obtenidas con la CS de la equiparación horizontal ya que son las que menores diferencias producen entre formas. No obstante, teniendo en cuenta que este proceso tiene la finalidad de ser el primer paso en la elaboración de una escala vertical de rendimiento, el proceso de CC puede facilitar la elaboración de esa

escala final, disminuyendo el número de ejecuciones del software empleado para la estimación. De esta forma, el anclaje vertical a través de CC se lleva a cabo en una única ejecución, estimando las puntuaciones de las dos formas en las cuatro aplicaciones al mismo tiempo.

Por tanto, para llevar a cabo el problema número dos, es decir, el anclaje vertical de las distintas puntuaciones de los sujetos, se utiliza la metodología de calibración conjunta para acometer la equiparación horizontal.

B. Comparación de procedimientos para el anclaje vertical

La equiparación **vertical que emplea la metodología de CC produce resultados mas estables que la CF o CS**. El procedimiento de Calibración Fija parece tener problemas a medida que se avanza en las aplicaciones, sin llegar a producir crecimiento entre la tercera y cuarta aplicación. Los cuatro tipos de calibración por separado muestran una tendencia a la disminución del crecimiento a medida que aumenta el rasgo, esta tendencia se suaviza en la CC.

Respecto al método de estimación del rasgo, el procedimiento bayesiano MAP produce crecimiento en todos los tramos de la distribución, es el único caso donde ocurre. Aunque los distintos tipos de calibración por separado crecen en los siete percentiles concretos que han sido analizados.

Por consiguiente, considerando la información extraída, **la CC como metodología de anclaje vertical es la más adecuada, sobre todo si se utiliza junto con el método de bayesiano MAP**[93] para la estimación de las puntuaciones del rasgo. Con estas características de anclaje, la dispersión de la escala es ligeramente inferior en comparación con el resto de metodologías. Otra de las características de la escala elaborada con esta metodología es que existe crecimiento en todos los tramos del rasgo, no hay un menor crecimiento en los extremos superiores respecto a los inferiores y el cambio entre las dos primeras aplicaciones y las dos últimas es similar.

Como apuntan Kolen y Brennan (2009), la CC muestra un mejor funcionamiento sobre todo si se mantiene la unidimensionalidad del constructo

[93]Esta combinación de CC y estimación MAP se aplica para construir la escala que se utiliza en el segundo estudio empírico.

evaluado. Con los datos de este trabajo, que únicamente abarcan dos cursos académicos consecutivos, el cambio en el constructo no es tan sustancial como podría serlo al comparar cursos más distanciados.

Los **métodos de calibración por separado no obtienen malos resultados interaccionando con los estimadores bayesianos, aunque sí existe una mayor tendencia a mostrar menor cambio a medida que se avanza hacia las mejores puntuaciones del rasgo evaluado**. Sin embargo, con el procedimiento MVL, los resultados son más inestables, sobre todo, en el cambio que se produce en los extremos del rasgo entre la primera y segunda aplicación. Se descarta la CF que tiende a suavizar el crecimiento a media que la aplicación se distancia del punto inicial establecido como base de la escala. Sin llegar a producir crecimiento entre las dos últimas aplicaciones.

En el **segundo estudio empírico** llevado a cabo en esta tesis y denominado **comparación empírica de modelos de Valor Añadido: tiempo, ocasiones de medida y relación entre estatus inicial y crecimiento**, se abordaron tres problemas distintos :

C. ¿Qué medida de tiempo es la más adecuada en el modelo de crecimiento?
D. ¿Cómo afecta la relación entre el punto inicial y el crecimiento a la estimación del VA de las escuelas?
E. ¿Cómo afecta la forma de medir el cambio en aprendizaje a las estimaciones de VA de las escuelas?

Y se extraen las siguientes conclusiones:

C. Selección de una medida adecuada de tiempo

El tipo de diseño específico de los datos de rendimiento de la evaluación que no cuenta con información longitudinal recogida en el mismo momento temporal y cuyos análisis iniciales revelaron un extraño cambio en el periodo de verano son factores que plantean la necesidad de estudiar el efecto de la variable utilizada como medida del tiempo en el modelo multinivel longitudinal.

Recordemos que la evaluación cuenta con cuatro medidas de rendimiento tomadas al inicio y final de dos cursos académicos y, por tanto, existe un periodo vacacional entre cursos que debe considerarse. Otro posible problema es que la tercera aplicación, que evalúa el inicio del segundo curso, se llevó a cabo dos meses después del comienzo oficial, en el mes de noviembre. Y en esos meses se suele dar un repaso a los contenidos del curso anterior. Este factor, junto con un cambio de facilidad en los ítems comunes que se utilizan para el anclaje entre la segunda y tercera aplicación, puede producir los resultados de crecimiento tan extraños en la tercera aplicación. Por tanto, utilizar el número de aplicaciones como variable tiempo no parece los más recomendable.

Utilizar una medida de tiempo en meses es lo más adecuado, pues al reducir la distancia entre la segunda y tercera aplicación se suaviza también el crecimiento entre aplicaciones. Los resultados del estudio indican que si se divide por el factor de corrección cuando se utiliza el número de aplicaciones como medida de tiempo, las estimaciones de las escuelas son prácticamente iguales a las producidas por el modelo que utiliza el número de meses. Por tanto, la corrección por el cambio en la facilidad de los ítems comunes parece compensada cuando se utiliza la distancia en meses entre aplicaciones como unidad de tiempo.

D. Relación entre estatus inicial y crecimiento y efecto de regresión hacia la media.

La relación entre estatus y cambio, medido a través del crecimiento o la ganancia, es el aspecto central de este problema. Una covarianza negativa entre ambas variables puede producir el denominado Efecto de Regresión hacia la Media (ERM) y sesgar las estimaciones de las escuelas. No obstante, en los modelos de crecimiento ese efecto puede confundirse con otros elementos como los estimadores bayesianos (BLUP) de los residuos de las escuelas o un mayor error de medida en la aplicación considerada como punto de partida. Este último aspecto ha sido confirmado en los resultados.

Se utilizaron dos metodologías para modificar la relación entre estatus inicial y cambio. La primera a través del cambio en la ocasión de medida utilizada como punto de partida, de A1 a A2 modificando la escala de la variable tiempo en

el modelo longitudinal como ya probó Rogosa (1995). Y la segunda, mediante la utilización de una variable de ajuste como las empleadas por Marsh y Hau (2002) y Castro, Ruíz y López (2009). Se incluye una nueva metodología para paliar la relación entre estatus y cambio mediante el ajuste a posteriori, a través del análisis de regresión simple, de los residuos de estatus inicial y crecimiento de las escuelas producidos por el modelo multinivel.

El modelo que cambia el punto de partida de A1 a A2, según los resultados, es el que produce un mejor ajuste. No hay diferencias en la *deviance* respecto al modelo que utiliza A1 como punto de partida, pero elimina la covarianza significativa entre la pendiente y el estatus de las escuelas y, por tanto, es el modelo más parsimonioso. Este cambio en el punto de partida elimina el posible ERM y confirma los argumentos de Rogosa (1995). Además, los residuos de las escuelas estimados por los modelos que solo modifican la posición del punto de partida y el modelo de base correlacionan de forma casi perfecta. Por tanto, el mayor error de medida con el que cuenta la primera aplicación puede ser el causante de esa covarianza negativa en el modelo que utiliza la primera aplicación como estatus inicial.

El ajuste a posteriori de los residuos de estatus y crecimiento mediante un nuevo análisis de regresión simple proporciona resultados muy similares a los del modelo que cambia el punto de partida. Y, por tanto, también puede ser una buena opción para eliminar el efecto que el estatus inicial tiene sobre el crecimiento. Sin embargo, los modelos que incluyen un predictor de ajuste en el modelo multinivel pueden cambiar drásticamente las estimaciones de las escuelas. El índice de ajuste que utilizan Castro, Ruíz y López (2009), elimina la covarianza entre estatus inicial y pendiente en las escuelas pero también elimina la varianza del crecimiento entre estudiantes, es decir, no existirían diferencias significativas en los ritmos de crecimiento de los estudiantes. Este fenómeno también ocurre en el modelo que utiliza la primera aplicación como covariable y solo tres medidas de rendimiento, como ya hicieron Marsh y Hau (2002), pero no consigue eliminar la covarianza negativa entre el estatus y el crecimiento de las escuelas.

E. Comparación de modelos de ganancia y crecimiento

Medir el crecimiento con dos únicas tomas de datos no es posible pero sí medir el VA. Willet (1989a) califica los modelos de ganancia como métodos tradicionales para medir el cambio y considera que los modelos longitudinales son una herramienta más potente para medir ese cambio en que se produce en el aprendizaje. Sin embargo, también es cierto que los modelos con dos únicas mediciones de logro académico siguen aplicándose actualmente en las evaluaciones de algunos países, como los modelos de ganancia residual en Inglaterra (Ray, 2006) o el modelo de Dallas (Webster & Mendro, 1997; Webster, 2005), y también hay publicaciones que prueban modelos basados en la ganancia para estimar el VA de las escuelas (Tekwe et al., 2004).

La comparación de las estimaciones producidas por los modelos de ganancia y de crecimiento muestran cierto grado de similitud, con coeficientes de correlación medios, en torno al 0,5. El modelo de ganancia estimada y el modelo de ganancia bruta producen residuos similares, alcanzando un coeficiente de correlación por encima de 0,9.

El modelo lineal mixto, aunque calcula sus puntuaciones como ganancias entre aplicaciones, utiliza las cuatro puntuaciones de logro como variable dependiente en el modelo y, por tanto, puede considerarse un modelo de crecimiento. Esta metodología de análisis del VA se ha revelado como una de los más interesantes. Su versatilidad permite analizar los cambios en cada uno de las dos cursos evaluados o estimar la ganancia completa entre la primera y última aplicación. Este indicador de la ganancia global tiene un gran parecido con el residuo de crecimiento obtenido a través del modelo multinivel longitudinal, con valores de Pearson por encima del 0,95.

En el modelo lineal mixto no hay correlación significativa entre estatus y la puntuación de ganancia de cada curso, por lo que se evita el posible efecto de regresión. Además puede omitirse el periodo de verano en los análisis, ya que ha resultado problemático.

Con los modelos lineales mixtos se consiguen dos puntuaciones de cambio para cada escuela, una por curso. Esta característica puede ser una ventaja a la hora de elaborar las estimaciones de VA. Con esta información es posible saber si

los centros educativos obtienen puntuaciones estables de crecimiento en cada curso, es decir, si su ritmo de cambio se mantiene o si ha sufrido alguna variación destacable. Por ejemplo, si los centros que tienen una ganancia superior a la media en el primer curso mantienen esa tendencia en el segundo.

La estimación de dos medidas de cambio, una por curso, para cada escuela también ofrece la posibilidad de calcular nuevas puntuaciones de su VA, por ejemplo, un promedio de ambas puntuaciones de ganancia.

Las características de los datos de esta evaluación y los distintos resultados obtenidos revelan que **el modelo lineal mixto puede ser el más adecuado en este tipo de situaciones**. Esta metodología, **además de permitir la estimación de la ganancia para cada curso por separado, se adapta a las características de la escala vertical que reduce la varianza de la puntuaciones a medida que se avanza en la escala**. Esta falta de homogeneidad puede ser un problema en el modelo multinivel longitudinal que considera una varianza común entre las distintas aplicaciones que forman la pendiente de crecimiento. Además, utilizando esta metodología sería posible incluir los grupos que fueron eliminados en la reducción muestral realizada entre la segunda y la tercera aplicación. Sería posible ofrecer la información de VA del primer curso académico evaluado para estos grupos.

IX.2 Limitaciones

En primer lugar, la utilización de datos empíricos tiene la ventaja de reflejar la realidad educativa que está siendo evaluada. Sin embargo, este tipo de datos también puede acarrear ciertos problemas que no se tendrían al trabajar con datos simulados, considerando que el objetivo principal de esta tesis doctoral es el estudio del comportamiento de distintas metodologías de análisis del Valor Añadido. Sin embargo, tiene la riqueza de las distribuciones y características de los datos que proceden de evaluaciones reales.

En relación con el uso de datos empíricos, una de las limitaciones es la mencionada perdida de muestra. Este fenómeno suele ser típico de los estudios longitudinales y se conoce como mortalidad experimental. En este trabajo esa

mortalidad experimental[94] está dentro de los valores aceptados, pero la situación problemática se encuentra en la reducción intencionada de casos que se produce entre la segunda y tercera aplicación. Eliminar grupos completos de algunos centros conlleva que sus puntuaciones de VA se estiman en función de la información de la que se dispones. Este aspecto no garantiza que el resultado sea representativo de la escuela. Le solución es vincular los resultados de VA con el grupo o aula de pertenencia y no con la escuela. Cada grupo puede tener profesores distintos lo que conllevaría metodologías de enseñanza distintas y, posiblemente, resultados de rendimiento distintos que, en parte, dependen del profesor.

Ligado a este problema aparece otro que tiene que ver con el número de estudiantes de alguno de los centros educativos evaluados. En ocasiones no supera los 20 casos, cantidad recomendada para llevar a cabo el análisis multinivel. Una escuela con pocos estudiantes, cuando se utilizan estimadores BLUP, tenderá a no diferenciarse de la media. De cualquier modo, sus estimaciones de VA deberán interpretarse con cautela.

En segundo lugar, los índices de ajuste TRI de alguno de los ítems empleados en las pruebas no son tan buenos como se esperaba. Aunque se han descartado aquellos con problemas de discriminación también debe considerarse la eliminación de esos ítems que no consiguen un buen ajuste.

Finalmente, en tercer lugar, la estimación de modelos lineales mixtos requiere equipos con mucha capacidad de cálculo. La estimación de modelos más complejos, con varias materias al mismo tiempo o con un tercer nivel de agregación, es muy dificil de ejecutar con los ordenadores habituales.

IX.3 Prospectiva

El análisis teórico llevado a cabo para este trabajo plantea campos de estudio que merecen ser tenidos en cuenta:

- El estudio en profundidad de los resultados de las evaluaciones que se desarrollan en la actualidad en España, sobre todo, el estudio PISA.

[94] Más información sobre los valores perdidos en el apartado VI.3.3 y en el Anexo I.

Los resultados de este estudio se asocian con el funcionamiento del sistema educativo español pero conviene mencionar que PISA evalúa competencias y no contenidos curriculares. El sistema educativo español ha comenzado, de forma reciente, el cambio hacia una enseñanza por competencias y, por tanto, los estudiantes españoles no están formados todavía para la resolución de este tipo de pruebas de evaluación. Las familias deberían ser consideradas como uno de los principales factores que, en parte, puede determinar los resultados de está evaluación, porque son ellos los encargados, junto con la escuela, de enseñar a sus hijos a desenvolverse en situaciones cotidianas.

- El tratamiento de los datos perdidos. Los modelos longitudinales, a medida que aumentan en número de tomas de datos, tienen una mayor probabilidad de perder sujetos en la muestra. El estudio de diferentes formas de abordar esta mortalidad experimental es otro aspecto que debe tratarse. Una opción posible es probar métodos distintos para la imputación de datos perdidos.

También los estudios empíricos realizados en esta tesis generan nuevos campos de trabajo para futuras investigaciones. En relación al primer estudio empírico sobre metodología para la elaboración de una escala vertical de rendimiento:

- El problema planteado sobre las distintas metodologías de equiparación horizontal podría vincularse con el de anclaje vertical, analizando cómo afecta la utilización de las distintas formas de calibración horizontal en la posterior elaboración de la escala vertical de rendimiento.
- Los resultados obtenidos mediante la calibración fija en el anclaje vertical que suaviza el crecimiento a medida que se avanza en las aplicaciones y que, en el presente trabajo, a penas estima crecimiento entre las dos últimas aplicaciones, es una cuestión que debe analizarse con mayor profundidad. Un análisis del error de equiparación y los posibles efectos que puede tener en esta metodología concreta de calibración es otro campo de estudio futuro.

- En este trabajó se optó por un modelo logístico de tres parámetros basado en la Teoría Respuesta al Ítem para la estimación de la habilidad de los sujetos pero puede ser relevante estudiar modelos distintos como los de uno y dos parámetros.

Y en relación con el segundo estudio empírico:

- Los modelos de crecimiento son menos vulnerables a los efectos de los factores del contexto. Sin embargo, los modelos basados en puntuaciones de ganancia pueden asumir un alto riesgo si no estudian esta cuestión. Por tanto, conviene llevar a cabo un estudio de los efectos de la incorporación de, por ejemplo, el nivel socioeconómico en los modelos basados en la ganancia y también en los de crecimiento.
- Un análisis en profundidad de las trayectorias de crecimiento de los estudiantes mediante la incorporación de predictores como el género o determinadas actitudes hacia los estudios. De esta forma se puede contrastar la existencia de diferencias en las trayectorias de grupos distintos. También puede ser interesante modelar la varianza del crecimiento entre escuelas, buscando, por ejemplo, diferencias entre centros con distinto tipo de titularidad (privada, concertada o pública).
- Se asume que los efectos de las escuelas son aleatorios porque los datos son solo una muestra representativa de la población. No todas las escuelas de interés se encuentran en la muestra. Sin embargo, comparar resultados de modelos con efectos aleatorios y fijos es otro aspecto que permite elaborar MVA distintos a los desarrollados en este estudio.

El Valor Añadido en Educación se presenta como una herramienta de evaluación muy potente y versátil si se realiza con rigurosidad metodológica. Además es un campo de investigación muy amplio que los investigadores educativos pueden aprovechar.

Bibliografía

Aitkin, M. & Longford, N. (1986). Statistical Modelling Issues in School Effectiveness Studies. *Journal of the Royal Statistical Society*, 149 (1), 1-43.

Alexander, J. C. (2008). *On the Fallacy of Value-Added Assessment.* Wesleyan College, Kentucky.

Armein-Beardsley, A. (2008). Methodological Concerns About the Education Value-Added Assessment System. *Educational Measurement, 37* (2), 65-75.

Astin, A. (1982). Excellence and Equity in American Education. *Paper presented at a Meeting of the National Commission on Excellence in Education.* Washington, DC: Department of Education.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Linn, R. L., Rothstein, R., Ladd, H. F., Ravitch, D., Shavelson, J. & Shepard, A. (2010). *Problems with the use of student test scores to evaluate teachers.* Economic Policy Institue. Washington, D.C.: Economic Policy Institue.

Ballou, D. (2009). Test Scaling and Value-Added Measurement. *Education Finance & Policy, 4* (4), 351-383.

Ballou, D., Sanders, W. & Wright, P. (2004). Controlling for student background in Value-Added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29* (1), 37-67.

Betebenner, D. (2004). *An Analysis of School District Data Using Value-added Methodology (CSE Report 622).* Los Angeles, CA: CRESST.

Betebenner, D. (2009). *Growth, Standards and Accountability.* Dover: The Center for Assessment.

Betebenner, D. & Linn, R. L. (2010). *Growth in student achievement: issues of measurement, longitudinal data analysis, and accountability*. Paper presentado en: Measurement challenges within the race to the top agenda ceter for k-12 assessment & performance management. Educational Testing Service (ETS)

Blanco, A., González, C. & Ordóñez, X. (2009). Patrones de correlación entre medidas de rendimiento escolar en evaluaciones longitudinales: un estudio de simulación desde un enfoque multinivel. *Revista de Educación, 348*, 195-215.

Bloom, B. J. (1972). *Taxonomía de los objetivos de la educación; la clasificación de metas educativas.* Buenos Aires: El Ateneo.

Braun, H. & Wainer, H. (2006). Value-Added Modeling. *Handbook of Statistics, 26*, 867-892.

Braun, H., Chudowsky, N. & Koenig, J. (. (2010). *Getting Value Out of Value-Added. Report of a Workshop.* Washington, D.C.National Research Council & National Academy of Education: The National Academies Press.

Briggs, D. C. (2008). *The Goals and Uses of Value-Added Models.* Washington: Paper prepared at Commitee on Value-Added Methodology for Instructional Improvemente. National Research Council and the National Academy of Education.

Briggs, D. C. & Weeks, J. (in review). Diagnosing volatility in measures of school-level status. *Educational Measurement: Issues & Practice* .

Briggs, D. & Betebenner, D. (2009). Is Growth is Student Achievement Scale Dependent? *Annual meeting of the National Council for Measurement in Education.* San Diego.

Briggs, D. & Weeks, J. (2009). The Impact of Vertical Scaling Decisions on Growth Interpretations. *Educational Meaurement: Issues and Practice, 28* (4), 3-14.

Briggs, D., Weeks, J. & Wiley, E. (2008). *Vertical Scaling in Value-Added Models for Student Learning.* Madison, WI.: Paper presented at the National Conference on Value-Added Modeling.

Bryk, A. S., & Raudenbush, S. W. (2002). *Hierarchical Linear Models. Aplications and data analysis methods.* California: Sage Publications.

Bryk, A. S. & Weisberg, H. I. (1976). Value-Added analysis: A dinamic approach to the estimation of treatment effects. *Journal of Educational Statistics, 1* (2), 127-155.

Bryk, A. S., Thum, Y. M., Easton, J. Q. & Luppescu, S. (1998). Assessing school academic productivity: The case of Chicago school reform. *Social Psychology of Education*, 103-142.

Bryk, A. & Woods, E. (1980). *An Introduction to the Value-Added Model and its Use in Short Term Impact Assessment.* Huron Institution, Cambridge, MA. Washtngton, D.C.: Department of Education.

Campbell, D. T. & Kenny, D. A. (1999). *A primer on regression artifacts.* New York: The Guildford Press.

Carabaña, J. (2011). Competencias y universidad, o un desajuste por mutua ignorancia. *Bodón. Resvista de Pedagogía , 63* (1), 15-31.

Castejon, J. L., Navas, L. & Sampascual, G. (1996). Un modelo estructural del rendimiento académico en matemáticas en la educación secundaria. *Revista de Psicología General y Aplicada, 49* (1), 27-43.

Castro, M. (2011). ¿Qué sabemos de la medida de las competencias? *Bordón. Revista de Pedagogía , 63* (1), 109-123.

Castro, M., Ruíz, C., & López, E. (2009). Forma básica del crecimiento en los modelos de valor añadido: vías para la supresión del efecto de regresión. *Revista de Educación, 348*, 111-136.

Cervini, R. (2004). Influencia de los factores institucionales sobre el logro en matemática de los estudiantes en el último año de la educación media de argentina. *Revista Electrónica Iberoamericana sobre Calidad, Eficacia y Cambio en Educación , 2* (1).

Chin, T., Kim, W. & Nering, M. (2006). *Five Statistical Factors That Influence IRT Vertical Scaling.* San Francisco: Paper presented at the annual meeting of National Council on Measurement in Education (NCME).

Choi, K., Goldschmidt, P. & Yamashiro, K. (2006). *Exploring models of school performance form theory to practice.* University of California, CRESST, CSE. Los Angeles: University of California.

Choi, K., Seltzer, M., Herman, J. & Yamashiro, K. (2007). Children Left Behind in AYP and Non-AYP Schools: Using Student Progress ande the Distribution of Student Gains to Validate AYP. *Educational Measurement: Issues and Practice, 26* (3), 21-32.

CIDE. (1990). *Hacia un modelo causal del rendimiento académico.* Madrid: Ministerio de Educación y Ciencia.

Coleman, J. S. (1975). Methods and results in the IEA studies of effects of school on learning. *Review of Educational Research*, 355-386.

Coleman, J. S., Campbell, E., Hobson, E., McPartland, J., Mood, A., Winfield, F. & York, R. L. (1966). *Eguality of educational opportunity.* Washington DC: National Center for Educational Statistics.

Creemers, B., Kyriakides, L. & Sammons, P. (2010). *Methodological Advances in Educational Effectiveness Research.* New York: Routledge.

Daniel, L. H. (2012). *Comparing cross-classified growth models with and without the cumulative effect of teachers to a hierarchical growth model on cross-classified data.* Pittsburgh: University of Pittsburgh. Recuperado el 09/09/2012.

Darmawan, I. G. & Keeves, J. P. (2006). Accountability of teachers and schools: A value-added approach. *International Education Journal, 7* (2), 174-188.

De la Orden, A. (2011). El problema de las competencias en la educación general. *Bordón. Revista de Pedagogía, 63* (1), 47-61.

De la Orden, A. (1985). Hacia una conceptualización del producto educativo. *Revista de Investigación Educativa, 3* (6), 271-283.

Demie, F. (2003). Using Value-added data for school self-evaluation: a case study of practice in inner-city schools. *School Leadership & Management, 23* (4), 445-467.

Dermitas, H. (2004). Modeling Incomplete Longitudinal Data. *Journal of Modern Applied Statistical Methods, 3* (2), 305-321.

Doran, H. C. (2003). Value-Added Analysis: A Review of Related Issues. *Annual Meeting of the American Educational Research Association.* Chicago: AERA.

Doran, H. C. & Izumi, L. T. (2004). *Putting Education to the Test: A Value-Added Model for California.* Consultado el 13/07/2009 en: http://www.heartland.org/custom/semod_policybot/pdf/15626.pdf.

Dossett, D. & Muñoz, M. A. (2003). *Classroom Accountability: A Value-Added Methodology. Paper presented at the annual meeting of the American Educational Research Association.* Chicago, IL: AERA.

Edmonds, R. R. (1979). *Search for effective school: The identification and analysis of city school that are instructionally effective for poor children.* Michigan State University: East Lansing.

Fernández, M. J. & Gonzalez, A. (1997). Desarrollo y situación actual de los estudios de eficacia escolar. *Revista Electrónica de Investigación y Evaluación Educativa, 3*.

Fernández, T. & Blanco, E. (2004). ¿Cuánto importa la escuela? El caso de México en el contexto de Ámerica Latina. *Revista Electrónica Iberoamericana sobre Calidad, Eficacia y Cambio en Educación, 2* (1).

Ferrão, M. E. (2009). Sensibilidad de las especificaciones del modelo de valor añadido: midiendo el estatus socioeconómico. *Revista de Educación* (348), 137-152.

Ferrão, M. E. & Goldstein, H. (2009). Adjusting form measurement error in the value added model: evidence from Portugal. *Cuality & Quantity, 43* (6), 951-963.

Fitz-Gibbon, C. T. (1992). School effects at A level: genesis of an information system. En D. eds Reynolds, & P. Cuttance, *School Effectiveness Research Policy and Practice.* London: Cassel.

Fitz-Gibbon, C. T. (2001). *Value-Added for those in Despair: research matter.* Londres: British Psychological Society.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland , 15*, 246–263.

Gaviria, J. L. & Castro, M. (2005). *Modelos jerárquicos lineales.* Madrid: La Muralla.

Gaviria, J. L., Biencinto, C. & Navarro, E. (2009). Invarianza de la estructura de covarianzas de las medidas de rendimiento académico en estudios longitudinales en la transición de Educación Primaria a Secundaria. *Revista de Educación,* 348, 153-173.

Gaviria, J. L., Martínez, R. & Castro, M. (2004). Un Estudio Multinivel Sobre los Factores de Eficacia Escolar en Países en desarrollo: El Caso de los Recursos en Brasil. *Education Policy Analysis Archives, 12* (20), Consultado el 20/04/2010 en: http://epaa.asu.edu/epaa/v12n20.

Goldschmidt, P., Choi, K. & Martinez, F. (2004). *Using Hierarchical Growth Models to Monitor School Performance Over Time: Comparing NCE to Scale Score Results-.* Los Angeles, CA: CSE Report 61ado8. Consultado el 12/03/2009 en: http://www.google.es/url?sa=t&source=web&cd=1&ved=0CCMQFjAA&url=http%3A%2F%2Fwww.cse.ucla.edu%2Fproducts%2Freports%2Fr618.pdf&rct=j&q=cse%20report%20618%20goldschmidt&ei=hQIzTpDiG46SswbfrqzpBg&usg=AFQjCNGhTPv-s.

Goldschmidt, P., Roschewski, P., Choi, K., Auty, W., Hebbler, S., Blank, R. & Williams, A. (2005). *Policymakers' Guide to Growth Models for School Accountabillity: How do Accountability Moels Differ.* Washington: CCSSO.

Goldstein, H. (1986). Efficient statiscal modelling of longitudinal data. (129-141, Ed.) *Annals of human biology, 13* (2).

Goldstein, H. (1987). *Multilevel Models in Educational and Social Research.* London: C. Griffin and Co.

Goldstein, H. (1997). Methods in School Effectiveness Research. *School Effectiveness and School Improvement, 8* (4), 369-395.

Goldstein, H. (1999). *Multilevel Statistical Models.* Institute of Education, London.

Goldstein, H. & Spiegelhalter, D. J. (1996). League tables and tehir limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society, 159* (3), 385-443.

Goldstein, H. & Woodhouse, G. (2001). Modelling Repeated Measurements. En A. H. Leyland, & H. Goldstein, *Multilevel Modelling of Health Statistics* (págs. 13-26). Chichester: John Willey and Sons.

Goldstein, H., Kounali, D. & Robinson, A. (2008). Modelling masurement errors and category misclassifications in multilevel models. *Statistical Modelling, 8* (3), 243-261.

González, D. (1975). Procesos escolares inexplicables. *Aula Abierta, 11* (12).

Grenn, J. L. (2010). *Estimating Teacher Effects Using Value-Added Models* . Lincoln: Dissertations and Theses in Statistics. University of Nebraska. Consultado el 16/09/2011 de: http://digitalcommons.unl.edu/statisticsdiss/6.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squared method. *Japanese Psychological Research, 22*, 144-149.

Haegeland, T. & Kirkeboen, L. (2008). *School performance and Value-Added indicators - What is the importance of controlling for socioeconomic background? A simple empirical illustration using Norwegian data.* Report 2008/8. Statistics Norway.

Hanushek, E. (1971). Teacher Characteristics and Gains in Student Achievement: Estimation usig micro data. *The American. Economic Review, 61* (2), 280-288.

Hanushek, E. (1972). *Education and Race.* Lexington, MA: D.C. Heath and Company.

Hanushek, E. (1979). Conceptual and Empirical Issues in the Estimation of Educational Production Functions. *The Journal of Human Resources, 14* (3), 351-388.

Hanushek, E. (2003). The Failure of Input-Based Schooling Policies. *The Economic Journal , 113*, 64-98.

Healy, M. J. & Goldstein, H. (1978). Regression to the mean. *Annals of Human Biology, 5* (3), 277-280.

Henderson, C. R. (1975). Best Linear Unbiased Estimation and Prediction under a selection model. *Biometrics*, 31 (2), 423-447

Hibpshman, T. (2004). A review of Value-Added Models. *Kentacky education Professional Standards Board.* Consultado el 28/03/2010 en: http://www.kyepsb.net/documents/Stats/Journals/Heterogeneity%20of%20regression.pdf .

Hill, P. W. & Goldstein, H. (1998). Multilevel Modeling of Educational Data With Cross-Classification and Missing Identification for Units . *Journal of Educational and Behavioral Statistics, 23* (2), 117-128 .

Holland, P. (2002). Two Measures of Change in the Gaps between the CDFs of Test-Score Distributions. *ournal of Educational and Behavioral Statistics, 27* (1), 3-17.

Hox, J. (1995). *Applied Multilevel Analysis.* Amsterdam: TT- Publikaties.

Hutchison, D. (2004). The Effect of Measurement Errors on Apparent Group Level Effects in Educational Progress. *Quality & Quantity, 38* (4), 407–424.

Hutchison, D. & Misfud, C. M. (2005). The Malta Primary Literacy Value-Added Proyect: a template for value-added in small island states? *Research Papers in Education, 20* (3), 303-345.

INECSE. (1998). *Diagnóstico general del sistema educativo. Avance de resultados.* Madrid: Ministerio de Educación y Ciencia.

INCE. (1999). *PISA: La medida de los conocimientos y destrezas de los alumnos.* Madrid: MECD.

INCE (2002). *Marcos teóricos y especificaciones de evaluación de TIMSS 2003.* Madrid: Ministerio de Educación Cultura y Deporte. Consultado el 06/05/2010 en: http://www.educacion.gob.es/dctm/ievaluacion/internacional/marcosteoricostimss2003.pdf?documentId=0901e72b8011071e

INECSE (2000). *Evaluación de la educación secundaria obligatoria. Datos básicos.* Madrid: Ministerio de Educación y Ciencia.

INECSE (2002). *Marco teórico y especificaciones de evaluación de TIMMS 2003.* Madrid: Ministerio de Educación y Ciencia.

INECSE (2002). *Conocimientos y destrezas para la vida. Primeros resultados del proyecto PISA 2000.* Madrid: Ministerio de Educación Cultura y Deporte. Consultado el 10/07/2011 en: http://www.educacion.gob.es/dctm/ievaluacion/internacional/pisa2000-int.pdf?documentId=0901e72b80110720

INECSE (2003). *Evaluación de la Educación Secundaria Obligatoria 2000.* Madrid: Ministerio de Educación Cultura y Deporte. Consultado el 15/07/2011 en: http://www.educacion.gob.es/dctm/ievaluacion/nacional/12evaluacion-de-la-educacion-secundaria-obligatoria-2000.pdf?documentId=0901e72b80110dcb

INECSE (2004). *Evaluación PISA 2003. Resumen de los resultados en España.* Madrid: Ministerio de Educación y Ciencia.

INECSE (2004). *Marcos teóricos de PISA 2003. Conocimientos y destrezas en Matemáticas, Lectura, Ciencias y Resolución de Problemas .* Madrid: Ministerio de Educación Cultura y Deporte. Consultado el 23/04/2008 en: http://www.educacion.gob.es/dctm/ievaluacion/internacional/marcoteoricopisa2003.pdf?documentId=0901e72b801106cd

INECSE (2006). *PIRLS 2006 - Marcos teóricos y especificaciones de evaluación (Traducción al Español).* Madrid: Ministerio de Educación y Ciencia. Consultado el 21/09/2009 en: http://www.educacion.gob.es/dctm/ievaluacion/internacional/pirlsmarcos2006.pdf?documentId=0901e72b80110474

Instituto de Evaluación (2007). *Educación primaria 2007. Evaluación General del Sistema Educativo.* Madrid: Ministerio de Educación. Consultado el 14/07/2011 en: http://www.educacion.gob.es/dctm/ievaluacion/nacional/educacion-primaria-2007.-evaluacion-del-sistema-educativo-espanol.pdf?documentId=0901e72b8046dc96

Instituto de Evaluación (2007b). *PISA 2006 Programa para la Evaluación Internacional de Alumnos de la OCDE. Informe español.* Madrid: Ministerio de Educación y Ciencia. Consultado el 20/03/2009 en: http://www.educacion.gob.es/dctm/ievaluacion/internacional/pisainforme2006.pdf?documentId=0901e72b8010c472

Instituto de Evaluación (2009). *Evaluación General de Diagnóstico 2009. Marco de la evaluación.* Madrid: Ministerio de Educación. Consultado el 14/07/2011 de: http://www.mecd.gob.es/dctm/ievaluacion/evaluaciongeneraldiagnostico/egd-2009-marco-evaluacion.pdf?documentId=0901e72b8044a2e5

Instituto de Evaluación (2010). *Evaluación General de Diagnóstico 2009. Educación primaria. Cuarto curso.* Madrid: Ministerio de Educación y Ciencia. Consultado el 14/07/2011 en: http://www.educacion.gob.es/dctm/ievaluacion/evaluaciongeneraldiagnostico/pdf-completo-informe-egd-2009.pdf?documentId=0901e72b8015e34e

Instituto de Evaluación (2010b). *PISA 2009 Programa para la Evaluación Internacional de los Alumnos OCDE. Informe Español.* Madrid: Ministerio de Educación. Consultado el 15/07/2011 en: http://www.educacion.gob.es/dctm/ievaluacion/internacional/pisa-2009-con-escudo.pdf?documentId=0901e72b808ee4fd

Instituto de Evaluación (2011). *Evaluación General de Diagnóstico 2010. Educación Secundaria Obligatoria. Segundo Curso. Informe de resultados.* Madrid: Ministerio de Educación. Consultado el 14/07/2011 en: http://www.educacion.gob.es/dctm/ievaluacion/informe-egd-2010.pdf?documentId=0901e72b80d5ad3e

Ito, K., Sykes, R. & Yao, L. (2008). Concurrent and Separate Grade-Goups Linging Procedures for Vertical Scaling. *Applied Measurement in Education, 21*, 187-206.

Jakubowski, M. (2008). *Implementing Value-Added Models of School Assesment.* San Domenico di Fiesole: RSCAS.

Jencks, C. (1971). *Inequality.* Londres: Allen Lane.

Jungnam, K. (2007). *A Comparison Of Calibration Methods And Proficiency Stimators for Creating IRT Vertical Scales (Tesis Doctoral).* Consultado el 3/10/10 en: http://ir.uiowa.edu/etd/163.

Kane, T. J. & Staiger, D. O. (2008). *Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. (NBER Working paper 14607).*

Cambridge, MA: NATIONAL BUREAU OF ECONOMIC RESEARCH. Consultado 10/11/2009 en: http://www.dartmouth.edu/~dstaiger/Papers/w14607.pdf.

Kang, T. & Petersen, N. (2009). *Linking Item Parameters to a Base Scale.* San Diego, CA: Paper presented at the National Council on Measurement in Education.

Keeves, J. P., Hungi, N. & Afrassa, T. (2005). Measuring Value Added effects across schools: Should schools be compared in performance? *Studies in Educational Evaluation, 31*, 247-266.

Koedel, C. & Betts, J. R. (2009). *Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique.* Working Papers 0902. Department of Economics, University of Missouri. Consultado el 6/11/2009 en: http://economics.missouri.edu/working-papers/2009/wp0902_koedel.pdf.

Koenker, R. & Basset, G (1978). Regression quantiles. *Econometrica*, 46 (1), 33-50

Kolen, M. J. & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and Practices. (2nd ed.).* New York: Springer.

Kreft, I. G. & De Leeuw, J. (1998). *Introducing multilevel modeling.* Newbury Park, CA: Sage.

Kupermintz, H. (2002). Value added assessment of teachers: the empirical evidence. En A. Molnar, *School Reform Proposals: The Research Evidence* (págs. 217-234). Charlotte: Information Age Publishing.

Kupermintz, H., Shepard, L. & Linn, R. (2001). Teacher Effects as a Measure of Teacher Effectiveness: Construct Validity Considerations in TVAAS. *Annual Meeting of the National Council on Measurement in Eduation.* Seattle: NCME.

Ladd, H. F. & Walsh, R. P. (2002). Implementing value-added measures of school effectiveness: getting the incentives right. *Economics of Education Review , 21*, 1-17.

Lee, W-C. & Ban, J.-C. (2010). A Comparison of IRT Linking Procedures. *Applied Measurement in Education , 23* (1), 23-48.

Leyland, A. (2004). A review of multilevel modelling in SPSS. Consultado el 23/07/2012 en: www.bristol.ac.uk/cmm/learning/mmsoftware/reviewspss.pdf

Linn, R. L. (1993). Linking Results of Distinct Assessments. *Applied Measurement in Education, 6* (2), 83-102.

Linn, R. L. (2008). Educational Accountability systems. En K. E. Ryan, & L. A. Shepard, *The future of test-based educational accountability* (págs. 3-24). New York: Taylor & Francis.

Linn, R. L. (2008). *Measurement Issues Associated with Value-Added Methods.* Washington, DC: Paper prepared for a Workshop Held by the Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation and Educational Accountability sponsored by the National Research Council and the National Academy of Education.

Lissitz, R. W., Doran, H., Schafer, W. D. & Willhoft, J. (2006). Growth Modeling, Value Added Modeling and Linking: An introduction. En R. W. Lissitz, *Longitudinal and Value Added Models of Student Performance* (págs. 1-46). Minnesota: JAM Press.

Lizasoain, L. & Joaristi, L. (2009). Análisis de la dimensionalidad en modelos de valor añadido: estudio de las pruebas de matemáticas empleando métodos no paramétricos basados en TRI. *Revista de Educación, 348*, 175-194.

Lockwood, J. R., Louis, T. A. & McCaffrey, D. F. (2003). Uncertainty in Rank Estimation: Implications for Value-Added Modeling Accountability. *Journal of Educational and Behavioral Statistics, 27* (3), 255-270.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V-N. & Martínez, J. F. (2007). The Sensitivity of Value-Added Teacher Effect Estimates to Different Mathematics Achievement Measures. *Journal of Educational Measurement, 44* (1), 47–67.

LOE. (2006). Ley Orgánica 2/2006, de 3 de Mayo, de Educación. *BOE nº 106*, 17158-17207.

Loyd, B., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement* (17), 179-193.

López, E. (2011). *Evaluación de la eficiencia técnica a partir del valor añadido en educación.* Tesis doctoral sin publicar, Universidad Complutense de Madrid.

López, E., Navarro, E., Ordoñez, X. G. & Romero, S. J. (2009). Estudio de variables determinantes de eficiencia a través de los modelos jerárquicos lineales en la evaluación PISA 2006: el caso de España. *Archivos Analíticos de Políticas Educativa, 17* (16).

Marchesi, A. & Martínez, R. (2006). *Escuelas de éxito en España. Sugerencias e interrogantes a partir del informe PISA 2003.* Madrid: Fundación Santillana.

Marchesi, A., Martínez, R. & Martín, E. (2004). Estudio longitudinal sobre la influencia del nivel sociocultural en el aprendizaje de los alumnos en la Educación Secundaria Obligatoria. *Infancia y Aprendizaje, 27* (3), 307-323.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement,* (14), 139-160.

Marnineau, J. A. (2006). Distorting value added: the use of longitudinal, vertically scaled studient achievement data for value-added accountability. *Journal of Educational and Behavioral Statistics, 31*, 35-62.

Martineau, J. A. (2009). Measuring Student Achievement Growth at the Hisgh School Level. En L. M. Pinkus, *Meaningful Measurement: The Role of Assessments in Improving High School Education in the Twenty-Frist Century* (págs. 119-142). Washington: Alliance for Excellent Education.

Marsh, H. W. & Hau, K. T. (2002). Multilevel Modeling of Longitudinal Growth and Change: Substantive Effects or Regression toward the Mean Artifacts? *Multivariate Behavioral Research, 37* (2), 245-282.

Martinez, M. R. & Hernández, M. J. (2006). *Psicometría.* Madrid: Alianza Editorial.

Martinez, R., Gaviria, J. L. & Castro, M. (coord) (2009). El valor añadido en educación (Monográfico). *Revista de Educación,* 348.

Martínez, R., Gaviria, J. L. & Castro, M. (2009). Concepto y evolución de los modelos de valor añadido en educación. *Revista de Educación,* 348, 15-34.

McCaffrey, D. F. & Hamilton, L. S. (2007). *Value-Added Assessment in Practice. Lessons from the Pennsylvania Value-Added Assessment System.* Pittsburgh, PA: RAND Corporation.

McCaffrey, D. F. & Lockwood, J. R. (2008). Value-Added Models: Analytic Issues. *National Research Council and National Academy of Edcuation, Board on Testing and Accountability.* Washington: Paper presented in Workshop on Value-Added Modeling.

McCaffrey, D. F., Koretz, D., Louis, T. A. & Hamilton, L. (2004). Models for Value-Added Modeling of Teacher Effects. *Journal of Educational and Behavioral Statistics, 29* (1), 67-101.

McCaffrey, D. F., Lockwood, J. R., Doretz, D. M. & Hamilton, L. S. (2003). *Evaluating Value-Added Models for Teacher Accountability.* Santa Monica: RAND Corporation.

Meyer, R. (1997). Value-Added Indicators of School Performance: A primer. *Economics of Education Review, 16* (3), 283-301.

Meyers, J. L. & Beretvas, S. N. (2006). The Impact of Inappropriate Modeling of Cross-Classified Data Structures. *Multivariate Behavioral Research, 41* (4), 473–497.

Muñiz, J. (1994). *Teoría clásica de los test.* Madrid: Ediciones Pirámide.

Muñiz, J. (1990). *Teoría de Respuesta a los Ítems.* Madrid: Ediciones Pirámide.

Muñiz, J. & Fidalgo, A. M. (2005). *Análisis de los ítems.* Madrid: La Muralla.

Murillo, F. J. (2005). *La investigación sobre eficacia escolar.* Barcelona: Octaedro.

Murillo, F. J. (2005). ¿Importa la escuela?. Una estimación de los efectos escolares en España. *Tnedencias Pedagógicas, 10*, 29-45.

Murnane, R. J. (1975). *The impact of shool resources on the learning of inner city children.* Cambridge, MA: Ballilnger.

Myslevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects.* Princeton, NJ: ETS Policy Information Center.

Navarro, E. & Redondo, S. (2006). Estudio sobre el Rendimiento en Matemáticas en España a partir de los Datos del Informe PISA 2003. Un Modelo Jerárquico de Dos Niveles. *Revista Electrónica Iberoamericana sobre Calidad, Eficacia y Cambio en Educación, 5* (3), 118-136.

Nesselroade, J. R., Stigler, S. M. & Baltes, P. B. (1980). Regression Toward the Mean and the Study of Chage. *Psychological Bulletin, 88* (3), 622-637.

OCDE. (2006). *Demand Sensitive Schooling? Evidence and Issues.* París: OCDE.

OCDE. (2006). *PISA 2006. Marco de la evaluación. Conocimientos y habilidades en Ciencias, Matemáticas y Lectura.* OCDE.

OCDE. (2008). *Measuring Improvements in Learning Outcomes.* Paris: OCDE.

Patz, R. J. (2007). *Vertical Scaling in Standards-Based Educational Assessment and Accountability Systems.* Washington, DC: CCSSO.

Pérez, G. (1981). *Origen social y rendimiento escolar.* Madrid: Centro de Investigaciones Sociológicas.

Ponisciak, S. M. & Bryk, A. S. (2005). Value Added Analysis of the Chicago Public Schools: An application of Hierarchical Models. En L. R, *Value Added Models in Education: Theory and Applications* (págs. 40-79). Maple Grove, MN: JAM Press.

Potamites, L., Chaplin, D. B. & Isenberg, E. (2009). Informe: Measuring School effectveness in Memphis. *Mathematica Policy Research Inc. Consultado el 10/02/2011 en: http://www.policyarchive.org/handle/10207/bitstreams/22107.pdf*

Rasbash, J., Steele, F., Browne, W. & Goldstein, H. (2009). A User's Guide to MLwiN version 2.10. Londres: Center for Multilevel Modelling. Consultado el 21/03/2010 de: http://www.bristol.ac.uk/cmm/software/mlwin/download/manual-print.pdf

Raudenbush, S. W. (2004). *Schooling, Statistics, and Poverty: Can we measure school improvement?* Princeton: Educational Testing Service.

Raudenbush, S. W. (2004). What are Value-Added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics , 29* (1), 121-129.

Raudenbush, S. W. & Bryk, A. S. (1986). A Hierarchical Model for Studying School Effects. *Sociology of Education, 59* (1), 1-17.

Raudenbush, S. W. & Willms, J. D. (1995). The Estimation of School Effects. *Journal of Educational and Behavioral Statistics, 20* (4), 307-335.

Ray, A. (2006). *School Value Added Measures in England: A Paper for the OECD Project on the Development of Value-Added Models in Education System.* London: Department for Education and Skills. Consultado el 17/4/2009: http://www.dcsf.gov.uk/research/data/uploadfiles/RW85.pdf

Ray, A., Evans, H., & McCormack, T. (2009). El uso de los modelos nacionales de valor añadido para la mejora de las escuelas británicas. *Revista de Educación* (348), 47-66.

Ray, A., McCormack, T. & Evans, H. (2009). Value-Added in English Schools. *Education Finance and Policy, 4* (4), 415-438.

Reardon, S. F. & Raudenbush, S. W. (2008). Assumptions of Value-Added Models for Estimating School Effects. *National Conference on Value-Added Modeling.* Madison: University of Wisconsin.

Reckase, M. D. (2008). *Measurement Issues Associated with Value-added Methods.* Artículo escrito por petición del: National Research Council Committee on Value-added Methodology. Washington, DC.

Reckase, M. D. (2010). *Study of best practices for vertical scaling and standard seting with recommendations for FCAT 2.0*. Florida Department of Educativon. Consultado el 13/08/2012 en: www.fldoe.org/asp/k12memo/pdf/StudyBestPracticesVerticalScalingStandardSetting.pdf

Roberts, J. S. & Ma, Q. (2006). IRT models for hte assessment of change across repeated measurements. En R. Lissitz, *Longitudinal and Value Added Models of Student Performance* (págs. 100-129). Maple Grove: JAM Press.

Rocconi, L. M. & Ethington, C. A. (2006). Assessing Longitudinal Change: Adjustment for Regression to the Mean Effects. *Research in Higher Education, 50* (4), 368-376.

Rogosa, D. (1995). Myths and methods: "Myths about longitudinal research," plus supplemental questions. En J. M. Gottman, *The analysis of change* (págs. 3-65). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Rogosa, D. R. & Willett, J. B. (1983). Demostrating the reliability of the difference score in the measurement of change. *Journal of Educational Measurement, 20* (4), 335-343.

Rothstein, J. (2009). *Student sorting and bias in value added estimation: Selection on observables and unobservables.* NBER Working Papers 14666 National Bureau of Economic Research. Consultado el 7/11/2009 en: http://www.princeton.edu/~ceps/workingpapers/170rothstein.pdf.

Rovine, M. J. & Molenaar, C. P. (2002). A Structural Equations Modeling Approach to the General Lineear Mixed Model. En L. M. Collins, & A. G. Sayer, *New Methods for the Analysis of Change* (págs. 67-104). Washington, DC: American Psychological Association.

Rubin, D. B., Stuart, E. A. & Zanutto, E. L. (2004). A Potential Outcomes View of Value-Added Assessment in Education. *Journal of Eduational and Behavioral Statistics , 29* (1), 103-116.

Ruiz, C. (2009). Las escuelas eficaces: un estudio multinivel de factores explicativos del rendimiento escolar en el área de matemáticas. *Revista de Educación,* 348, 355-376.

Ruiz, C. & Castro, M. (2006). Un estudio multinivel basado en PISA 2003: factores de eficacia escolar en el área de matemáticas. *Archivos Analíticos de Políticas Educativas, 14* (29), Consultado el 03-08-2009 en: http://epaa.asu.edu/epaa/v14n.29.

Sanders, W. L. (2006). *Comparisons among varius educational assessment value-added models.* Columbus: Presented at: The power of two national value-added conference.

Sanders, W. L. & Horn, S. P. (1994). The Tenessee Value Added Assessment System (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Educatio , 8*, 299-311.

Sanders, W. L. & Rivers, J. C. (1996). *Cumulative and Residual Effects of Teachers on Future Student Academic Achievement.* Knoxville: University of Tennessee Value-Added Research and Assessment Center.

Sanders, W. L. & Wright, S. P. (2008). *A Response to Amrein-Beardsley (2008) "Methodological Concerns about the Education Value-Added Assessment System.* Conultado el 05/08/2009, en: http://www.sas.com/govedu/edu/services/Sanders_Wright_response_to_Amrein-Beardsley_4_14_2008.pdf

Sanders, W. L., Saxton, A. M. & Horn, S. P. (1997). The Tennessee Value-Added Accountability System: A Quantitative, Outcomes-Based Approach to Educational Assessment. En J. Millman, G*rading Teachers, Grading Schools: Is Student Achievement a Valid Evaluation Measure?* (págs. 137-162). Thousands Oaks: Corwin Press.

Schafer, W. D. (2006). Growth Scales as an Alternative to Vertical Scales. *Practical Assessment, Research & Evaluation, 11* (4).

Schmidt, W. H., Houang, R. T. & McKnight, C. C. (2005). Value-Added Research: Right Idea but Wrong Solution? En R. W. Lissitz, *Value Added Models in Education* (págs. 145-165). Maple Grove: JAM Press.

Sean, K., & Monczunski, L. (2007). Overcoming the volatility in school-level gain socres: a new approach to identifying value added with cross-sectional data. *Educational Researcher, 36* (5), 279-289.

Seltzer, M., Choi, K. & Thum, Y. M. (2002). *Examining relationships between where students start and how rapidly they progress: Implications for constructing indicators that help illuminate the distribution of achievement within schools.* Los Angeles CA: CRESST. Consultado el 24/08/2010 en: http://www.eric.ed.gov/ERICWebPortal/search/detailmini.jsp?_nfpb=true&_&ERICExtSearch_SearchValue_0=ED465804&ERICExtSearch_SearchType_0=no&accno=ED465804

Singer, J. D. & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and even occurrence.* New York: Oxford University.

Stevens, J. (2005). The Study of School Effectiveness as a Problem in Research Design. En R. W. Lissitz, *Value Added Models in Education. Theory and Applications* (págs. 167-208). Maaple Grove, MN: JAM Press.

Stevens, J. & Zvoch, K. (2006). Issues in the implementation of longitudinal growth models for studen achievement. En R. W. Lissitz, *Longitudinal and Value Added Models of Student Performance* (págs. 170-209). Maple Grove: JAM Press.

Stocking, M. L. & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement* (7), 201-210.

Strand, S. (1998). A "value-added" analysis of the 1996 primary school performance tables. *Educational Research , 40* (2), 123-137.

Tekwe, C. D., Carter, R. L., Ma, C-X., Lucas, M. E., Roth, J., Ariet, M., Fisher, T. & Resnick, M. B. (2004). An Empirical Comparison of Statistical Models for Value-Added Assessment of School Performance. *Jounal of Educational and Behavioral Statistics, 29* (1), 11-36.

Thum, Y. M. (2002). *Measuring Student and School Progress With the California API (CSE Technical Report 578).* Los Angeles, CA: CRESST.

Thum, Y. M. (2003). Measuring Progress Toward a Goal: Estimating Teacher Productivity Using a Multivariate Multilevel Model for Value-Added Analysis. *Sociological Methods & Research, 32* (2), 153-207.

Thum, Y. M. (2006). Designing Gross Productiviy Indicators: A Proposal for Connecting Accountability Goals, Data, and Analysis. En R. W. Lissitz, *Longitudinal and Value Added Models of Student Performance* (págs. 436-479). Maple Grove: JAM Press.

Thum, Y. M. (2009). No Child Left Behind: Retos metodológicos y recomendaciones para la medida del progreso anual adecuado. *Revista de Educación, 348*, 67-90.

Tiana, A. (2011). Análisis de las competencias básicas como núcleo curricular en la educación obligatoria española. *Bordón. Revista de Pedagogía , 63* (1), 63-75.

Tong, Y. & Kolen, M. J. (2007). Coparisons of Methodologiaes and Results in Vertical Scaling for Educational Achievement Test. *Applied Measurement in Education, 20* (2), 227-253.

Touron, J. (1985). La predicción del rendimiento académico: procedimientos, resultados e implicaciones. *Revista Española de Pedagogía, 45* (175), 103-124.

Vasquez, J. & Darling-Hammond, L. (2008). Accountability Texas-Style: The progress and learning of urban minority students in a high-stakes testing context. *Educational Evaluation and Policy Analysis, 30* (2), 75-110.

Webster, W. J. (2005). The Dallas School-Level Accountability Model: The Marriage os Status and Value-Added Approches. En R. W. Lissitz, *Value-Added Models in Education. Theory and Applications* (págs. 233-271). Maple Grove: JAM Press.

Webster, W. J. & Mendro, R. L. (1997). The Dallas Value-Added Accountability System. En J. (. Millman, *Grading teachers, grading schools. Is student achievement a valid evaluation measure?* (págs. 81-99). Corwin: Thousand Oaks.

West, B. T., Welch, K. B. & Gallecki, A. T. (2007). *Linear mixed models. A Practical Guide Using Statistical Software.* Boca Raton, FL: Taylor & Francis Group.

Wiley, E. W. (2006). *A Practitioner's Guide to Value Added Assessment.* Tempe: Educational Policy Studies Laboratory, Arizona State University.

Willett, J. B. (1989). Questions and answers in the measurement of change. *Review of Research in Education, 15*, 345-422.

Willett, J. B. (1989b). Some results on reliability for the longitudinal measurement of change: implications for the design of studies of individual growth. *Educational and Psychological Measurement, 49*, 587-601.

Willett, J. B. (1994). Measurement of change. En T. Husen, & T. N. Postlethwaite, *The International Encyclopedia of Education* (págs. 671-678). Oxford, UK: Pergamon Press.

Willett, J. B. (1997). Measuring change: what individual growth modelling buys you. En E. Amsel, & K. A. Renninge, *Change and Development: Issues of Theory, Method, and Application.* Mahwah, NJ: Lawrence Erlbaum Associates.

Willms, J. D. (2008). *Seven Key Issues for Assessing 'Value Added' in Education.* Washington D.C: Paper prepared for a workshop sponsored by the US National Research Council and the US National Academy of Education on the use of value-added methods for instructional improvement, program evaluation and accountability.

Willms, J. D. & Raudenbush, S. W. (1989). A longitudinal Hierarchical Linear Model For Estimating School Effects and Their Stability. *Journal Oof Educational Measurement, 26* (3), 209-232.

Wright, P. S., Sanders, W. L. & Rivers, J. C. (2006). Measurement of Academic Growth of Individual Students toward Variable and Meaningful Academic Standards. En R. W. Lissitz, *Longitudinal and Value Added Models of Student Performance.* Minnesota: JAM Press .

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement, 23*, 299–326.

Young, D. J. (1999). *The Usufulness of Value-Added research in Identifying Effective Schools. Paper Presented at the Joing Conference of The Australian Association of Research in Education an the New Zeland Association of Research in Education.* Melbourne: Australian Research Council.

Younk, D. J. (1999). *Studen Progress in Australian Schools: A Multilevel Multivariate Model.* Montreal: Paper presented to the American Educational Research Association Annual Meeting.

Zaidman-Zait, A. & Zumbo, B. D. (2005). *Multilevel (HLM) models for modeling change with incomplete data: demonstrating the effects of missing data and level-1 model mis-specifications.* Montreal: Paper presented at the Hierarchical Leneral Modeling (SIG) of the American Educational Research Association conference.

Zvoch, K. & Stevens, J. J. (2003). A multilevel, longitudinal analysis of middle school math and language achievement. *Education Policy Analysis Archives, 11* (20).

Zvoch, K. & Stevens, J. J. (2006). Successive Student Cohorts and Longitudinal Growth Models: An Investigation of Elementary School Mathematics Performance. *Education Policy Analysis Archives, 14* (2), Consultado el 14/07/2009 en: http://epaa.asu.edu/epaa/v14n2/.

Anexo I: Datos perdidos, características de la escala vertical y análisis de ítems

En este anexo se incluye información complementaria del estudio de valores perdidos realizado en el capítulo VI (apartado VI.3.2). Una caracterización de la escala vertical de rendimiento elaborada y la comprobación de los supuestos de las puntuaciones estimadas par los estudiantes. Además se incluye un estudio pormenorizado de cada uno de los ítems que formaron parte de los instrumentos de medida.

Anexo I.1 Datos perdidos por centro y estadísticos descriptivos

La siguiente tabla (Tabla AI.1) presenta los resultados de rendimiento en matemáticas obtenidos por todos los centros de la muestra original. Además se incorpora en la misma tabla el número de estudiantes evaluados en cada ocasión de medida para comprobar la pérdida de datos y la reducción muestral llevada a cabo entre la segunda y tercera aplicación

CENTRO	N	M	DT	N	M	DT	N	M	DT	N	M	DT	% N reducido A2 A3
	A1			A2			A3			A4			
1	115	-0,803	0,815	111	-0,266	0,642	52	0,537	0,706	47	0,976	0,782	53
2	40	-0,432	0,762	38	-0,031	0,584	27	1,030	0,446	28	1,332	0,473	29
3	39	-0,004	0,878	43	0,371	0,751	24	1,240	0,782	22	1,662	0,571	44
4	156	0,011	0,802	145	0,598	0,782	69	1,499	0,588	62	1,682	0,656	52
5	81	-0,353	0,906	63	0,173	0,712	35	1,081	0,671	38	1,243	0,581	44
6	203	-0,203	0,835	203	0,127	0,791	83	1,379	0,682	84	1,578	0,711	59
7	138	0,187	0,939	123	0,625	0,795	50	1,402	0,618	50	1,634	0,716	59
8	37	-0,769	0,658	17	0,127	0,641	16	1,005	0,557	18	1,428	0,640	6

9	156	-0,327	0,831	142	0,039	0,751	68	0,700	0,676	50	0,712	0,816	52
10	128	-0,190	0,801	128	0,340	0,708	66	1,243	0,653	67	1,461	0,666	48
11	171	0,272	0,877	168	0,695	0,773	104	1,365	0,658	98	1,598	0,779	38
12	113	0,015	0,867	104	0,242	0,762	59	1,087	0,553	57	1,569	0,616	43
13	114	-0,422	0,813	111	-0,004	0,643	47	1,194	0,555	42	1,266	0,667	58
14	83	0,077	0,744	72	0,538	0,725	55	1,304	0,645	65	1,634	0,626	24
15	88	0,173	0,702	83	0,664	0,735	64	1,512	0,596	61	1,901	0,619	23
16	119	-0,224	0,783	124	0,204	0,761	53	1,031	0,721	49	1,384	0,710	57
17	38	-0,346	0,821	32	0,316	0,752	16	1,014	0,459	21	1,360	0,633	50
18	98	-0,021	0,753	96	0,348	0,763	52	1,279	0,636	49	1,631	0,634	46
19	91	-0,523	0,760	92	-0,272	0,596	45	0,704	0,732	49	1,095	0,748	51
20	106	-0,212	0,831	94	0,188	0,615	51	1,107	0,645	54	1,350	0,680	46
21	84	-0,121	0,807	84	0,278	0,824	27	1,323	0,646	41	1,430	0,807	68
22	68	-0,061	0,782	57	0,429	0,723	34	1,248	0,544	37	1,574	0,675	40
23	107	-0,317	0,718	109	0,040	0,708	49	0,680	0,578	49	1,311	0,669	55
24	115	-0,018	0,812	107	0,457	0,742	49	1,271	0,635	52	1,273	0,686	54
25	116	-0,141	0,909				71	1,037	0,747	25	1,385	0,604	39
26	60	-0,188	0,706	98	0,017	0,619	43	0,979	0,591	31	1,202	0,547	56
27	39	0,278	0,938	42	0,895	0,900	23	1,627	0,704	23	2,240	0,543	45
28	71	0,290	0,819	69	0,752	0,732	47	1,387	0,613	47	1,687	0,579	32
29	51	-0,542	0,681	53	0,066	0,685	38	1,029	0,476	40	1,488	0,545	28
30	21	-0,179	0,474	29	0,382	0,661	21	1,237	0,422	22	1,554	0,622	28
31	46	0,138	0,908	38	0,691	0,725	12	1,510	0,595	23	2,247	0,576	68
32	27	0,358	0,750	26	1,002	0,945	21	1,639	0,514	20	1,955	0,716	19
33	25	-0,109	0,705	31	0,297	0,821	22	1,301	0,627	25	1,780	0,636	29
34	80	0,651	0,770	79	1,047	0,655	1	-0,537	.	52	2,182	0,517	99/34
35	48	-0,109	0,838	45	0,196	0,661	39	1,167	0,601	38	1,342	0,575	13
36	62	0,249	0,846	64	0,589	0,791	55	1,322	0,641	54	1,890	0,676	14
37	82	-0,038	0,715	57	0,743	0,689	46	1,590	0,471	46	1,971	0,478	19
38	31	0,129	0,858	34	0,268	0,847	16	1,216	0,704	14	1,314	0,818	53
39	117	0,671	0,775	123	1,166	0,659	85	1,772	0,599	85	2,228	0,513	31
40	22	-0,272	0,743	28	0,440	0,553	24	1,103	0,683	21	1,609	0,542	14
41	107	0,012	0,823	108	0,457	0,823	82	0,984	0,712	81	1,349	0,739	24
42	44	0,124	0,671	45	0,166	0,772	38	1,131	0,556	36	1,506	0,569	16
43	48	0,725	0,838	74	1,000	0,702	47	1,734	0,568	48	2,055	0,530	36
44	147	0,690	0,724	149	1,205	0,626	86	1,721	0,518	83	2,154	0,652	42
45	89	0,441	0,791	87	0,817	0,842	51	1,252	0,626	57	1,569	0,739	41
46	59	0,305	0,870	62	0,884	0,730	46	1,406	0,714	45	1,831	0,842	26
47	50	0,577	0,611	56	0,869	0,662	52	1,571	0,488	53	1,816	0,607	7
48	52	0,306	0,697	48	0,493	0,622	44	1,259	0,474	44	1,747	0,518	8
49	52	-0,166	0,815	52	0,219	0,706	23	1,394	0,702	45	1,299	0,832	56
50	158	0,381	0,875	156	0,697	0,800	25	1,966	0,640	73	1,994	0,578	84
51	28	-0,149	0,619	48	0,451	0,563	42	1,416	0,597	43	1,682	0,648	13
52	61	0,160	0,860	63	0,445	0,774	40	1,176	0,635	38	1,548	0,523	37
53	109	0,397	0,772	104	0,802	0,797	52	1,559	0,593	72	1,786	0,672	50
54	52	-0,288	0,733	49	-0,006	0,628	38	0,983	0,571	39	1,404	0,577	22

55	58	0,395	0,842	57	0,764	0,615	53	1,498	0,495	51	1,965	0,565	7
56	55	0,617	0,641	55	1,402	0,665	55	1,864	0,579	55	2,221	0,623	0
57	47	0,185	0,913	46	0,426	0,874	34	1,355	0,757	33	1,716	0,884	26
58	45	1,104	0,681	47	1,096	0,758	42	1,779	0,592	42	2,170	0,521	11
59	80	0,091	0,727	75	0,999	0,637	50	1,441	0,518	50	1,945	0,601	33
60	62	0,400	0,565	22	0,853	0,836	19	1,647	0,590	38	2,106	0,585	14
61	114	0,311	0,787	116	0,648	0,644	85	1,417	0,578	80	1,850	0,572	27
62	57	-0,004	0,829	53	0,258	0,844	42	1,154	0,550				21
63	47	0,065	0,889	46	0,188	0,803	24	1,064	0,855	27	1,346	0,826	48
64	29	0,109	0,755	30	0,454	0,652				30	1,680	0,621	100/0
65	70	0,261	0,649	72	0,761	0,642	41	1,574	0,495	20	2,188	0,420	43

Tabla AI.1. Nº de estudiantes, medias y desviaciones típicas y reducción del tamaño muestral en A2-A3 por centro educativo.

Anexo I.2 Características de la escala vertical

A la luz de los resultados del estudio empírico realizado en el capítulo VII, la calibración tanto horizontal como vertical se llevó a cabo de forma conjunta y la estimación del rasgo (rendimiento en matemáticas) se realizó con metodología bayesiana, concretamente con estimación Máxima a Posteriori (MAP). Se fijo la métrica de la escala mediante una transformación lineal para facilitar la interpretación, con una media de 250 puntos y desviación típica de 50 en la A1. Este cambio se hizo directamente desde el software BILOGMG incluyendo en la sección de calificación de la sintaxis los siguientes parámetros:

RSCTYPE = 1, LOCATION = (250.0000), SCALE = (50.0000)

El total de ítems a calibrar fue de 162 y se contaba inicialmente con 170. Los ocho ítems que faltan fueron eliminados por motivos de ajuste. Algunos tenían una correlación biserial puntual negativa lo que impedía la ejecución del programa de estimación y el resto por cuestiones de redacción como no tener alternativa correcta.

La calibración se lleva a cabo con una única estimación, calculando al mismo tiempo los parámetros de los ítems y la estimación de las puntuaciones de rendimiento en las cuatro ocasiones de medida. Esta metodología utiliza los parámetros de los ítems comunes entre aplicaciones para llevar a cabo el proceso y se desarrolla con la opción multigrupo de BILOGMG. En el comando INPUT se incluye el parámetro "NGROUPS=n", donde n es el número de mediciones. Después se define cada grupo añadiendo comandos "GROUP" para especificar las

características (tamaño del test e ítems comunes). Por último en el comando CALIB, debe incorporarse el parámetro "REFERENCE=n" para indicar que medición va a ser la base de la escala, el punto de partida.

El Gráfico AI.6 nos proporciona información sobre la fiabilidad de las estimaciones a lo largo de todas las puntuaciones del rasgo. La curva continua representa la función de información del test que equivale a la suma de las funciones de información de cada uno de los ítems y la discontinua es el error típico, es decir, la inversa de la raíz cuadrada de la función de información del test.

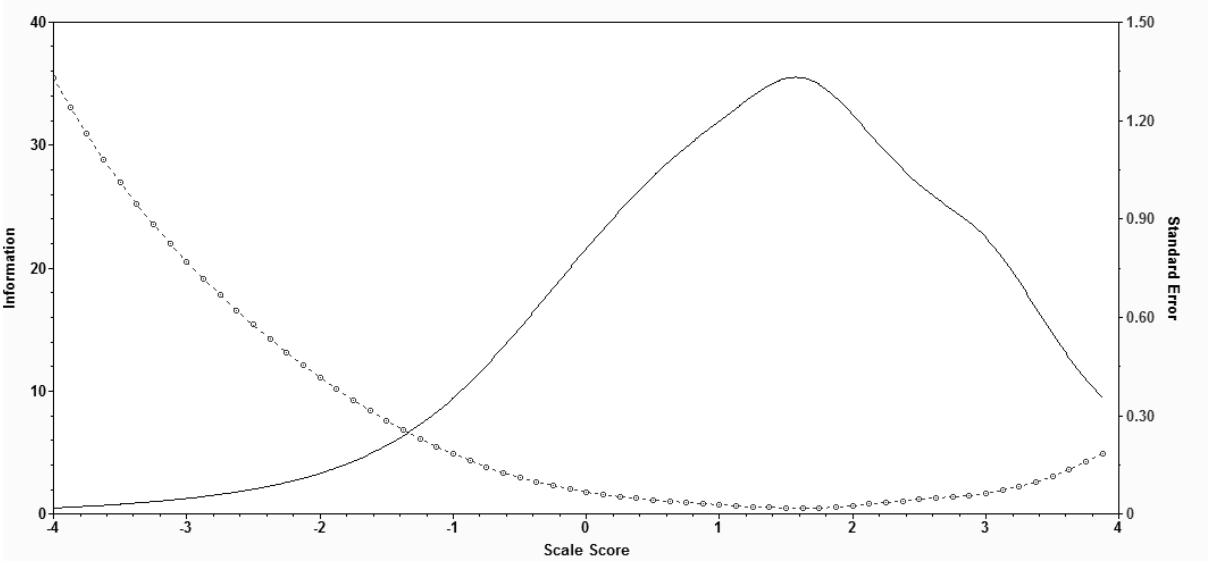


Gráfico AI.6. Función de Información y error típico de la escala.

Las estimaciones son más fiables en los valores positivos del rasgo, por encima de la media inicial. Por debajo de -1,5 los errores superan el 0,3. Los estadísticos descriptivos de la escala vertical son los siguientes:

Escala Vertical		A1	A2	A3	A4
N	Válidos	2809	2801	2695	2745
	Perdidos	155	163	269	219
Media		252,155	273,709	310,686	331,112
Desv. típ.		44,032	41,633	36,909	36,880
Varianza		1938,790	1733,286	1362,285	1360,116
Percentiles	5	181,943	202,938	249,141	266,764
	10	194,782	218,961	262,615	283,042
	25	220,427	246,632	285,500	307,741
	50	251,008	273,108	311,311	332,643
	75	282,489	302,170	336,589	356,251
	90	312,431	326,570	357,830	377,229
	95	325,410	343,722	370,112	388,420

Tabla AI.2. Medias, desviaciones típicas y tamaño muestral en cada aplicación

La media de rendimiento en matemáticas en la primera toma de datos, en Octubre de 2005, es de 252,1 puntos y se mantiene ese mayor crecimiento entre A2 y A3, en comparación con el resto, como ocurría con la muestra completa. La dispersión de los datos también es similar, con menor desviación típica en las dos últimas aplicaciones.

Los tamaños muestrales entre aplicaciones son similares entre aplicaciones y número total de sujetos de los que se posee información de rendimiento en las cuatro es de 2158. El siguiente diagrama de caja y bigote muestra las características concretas de las cuatro distribuciones

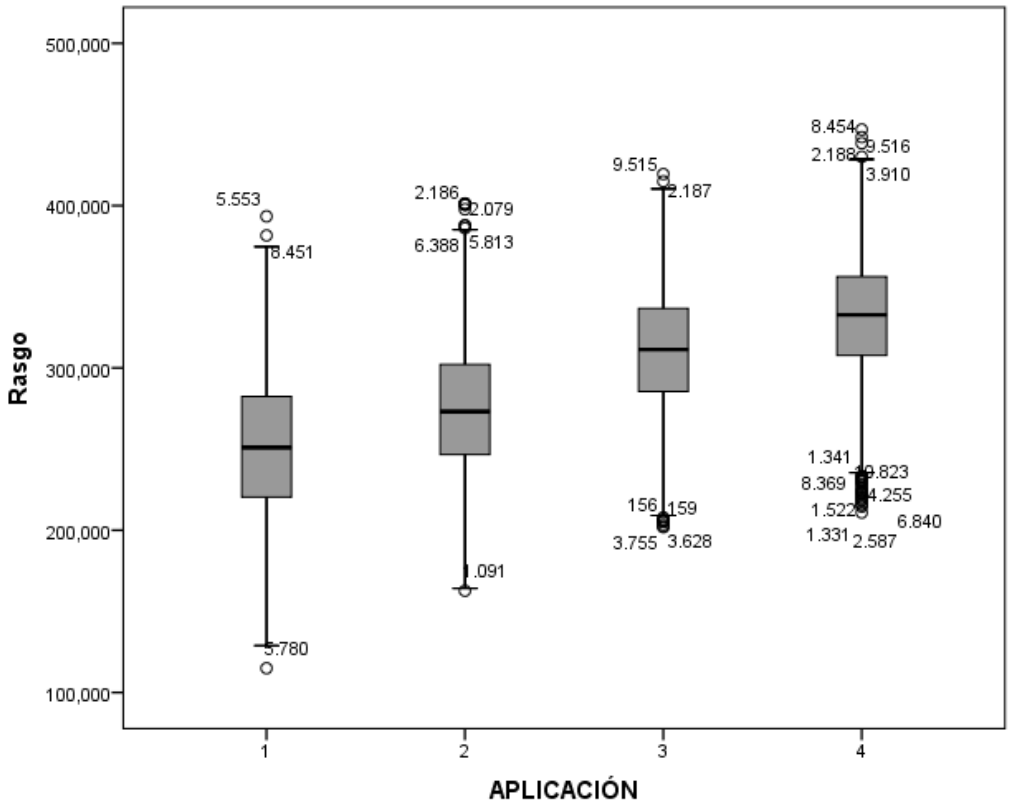


Gráfico AI.7. Diagramas de caja y bigote para cada en cada ocasión de medida.

El Gráfico AI.7 plasma la tendencia creciente, lógica y ya anunciada, del rendimiento entre aplicaciones de medida. Al contrario de lo que ocurre con la dispersión que parece ser decreciente. Este último aspecto es una características propia de las escalas verticales elaborada bajo los supuestos de la TRI y también parece lógica. A medida que se avanza hacia el extremo superior de la escala se hace más difícil conseguir un crecimiento alto, es decir, los estudiantes con niveles

altos de rendimiento tienden a crecer menos porque se encuentran cerca del techo de la escala, de su límite superior.

Finalmente, como último aspecto de esta sección de descriptivos de la escala, se incluyen las medias y desviaciones típicas agregadas por centro educativo, una vez eliminados de las dos primeas aplicaciones los sujetos que formaron parte de la reducción muestral intencional realizada en la tercera aplicación.

CENTRO	N	M	DT	N	M	DT	N	M	DT	N	M	DT
	A1			A2			A3			A4		
1	43	212,671	212,671	45	242,178	37,193	46	271,919	38,031	42	298,868	38,944
2	27	227,057	34,112	27	238,820	27,250	26	292,433	23,729	26	310,363	24,242
3	22	248,349	47,171	23	274,358	38,403	22	311,788	39,707	22	328,804	30,380
4	79	253,896	39,320	77	282,249	37,358	67	321,517	31,968	59	331,482	33,630
5	32	236,532	41,710	29	259,252	38,818	31	298,116	32,645	32	306,883	24,925
6	80	253,517	42,716	78	259,780	45,526	80	313,621	38,142	81	325,061	36,139
7	46	257,345	37,836	45	280,558	39,068	47	316,800	33,121	44	333,683	34,285
8	13	218,283	29,268	10	268,675	18,817	13	291,652	31,472	13	320,585	27,117
9	56	226,163	44,281	54	240,014	37,948	52	277,991	36,182	43	285,769	41,692
10	65	243,785	40,454	66	267,993	33,560	66	305,991	36,164	66	318,709	35,114
11	100	264,451	43,280	99	283,039	40,529	102	312,774	36,654	96	326,461	40,702
12	60	247,996	43,278	56	265,271	36,570	59	296,844	30,698	55	323,582	33,010
13	41	230,990	40,000	44	246,063	29,658	44	304,456	30,941	39	311,001	33,816
14	56	247,077	40,290	55	267,172	39,434	52	308,583	36,944	56	331,778	30,457
15	57	256,762	37,309	56	286,459	36,609	58	323,952	32,718	55	346,046	31,351
16	46	239,523	43,944	49	264,745	40,928	49	295,684	39,731	46	315,122	35,832
17	18	239,529	33,753	18	257,312	38,468	16	292,981	25,969	18	320,241	27,595
18	42	251,859	36,952	45	274,345	39,284	43	313,285	34,275	43	332,431	29,046
19	49	219,493	37,241	49	237,057	28,751	43	276,961	40,338	45	303,480	36,250
20	46	244,391	42,774	44	258,275	32,932	48	299,642	35,972	48	317,128	30,156
21	32	259,704	35,871	34	276,992	38,454	25	314,280	32,989	31	329,864	36,961
22	34	244,277	40,561	33	264,704	39,860	34	306,318	30,277	35	325,403	35,532
23	48	215,124	32,121	46	242,144	31,544	47	275,535	31,944	46	313,551	33,671
24	48	245,922	41,358	48	273,614	36,779	49	307,501	35,110	46	313,801	33,338
25	22	246,322	44,869				22	307,330	33,365	22	320,140	25,802
26	20	228,699	44,800	37	243,963	37,143	36	296,876	29,901	26	307,485	26,397
27	19	266,891	48,490	19	300,720	51,618	19	330,282	41,536	19	362,554	29,660
28	45	252,168	42,204	44	277,302	39,278	45	316,381	32,481	44	331,610	30,738
29	38	217,601	35,720	40	249,382	34,284	38	293,734	26,369	40	319,789	28,643
30	14	229,239	21,968	22	259,977	32,474	21	305,135	23,150	22	323,198	32,709
31	22	267,552	52,848	22	291,687	35,007	12	320,302	33,573	22	360,659	31,294
32	20	264,767	37,317	20	302,858	46,220	21	327,937	28,946	20	344,611	38,276
33	21	237,161	37,973	25	263,100	40,786	22	308,797	35,010	25	335,284	33,354

34	51	281,090	42,541	51	297,306	39,712	1	207,866	.	50	358,416	26,375
35	39	239,058	41,894	36	254,721	33,659	39	301,504	33,253	37	312,232	30,985
36	55	250,990	46,095	56	271,349	41,688	55	310,238	35,690	54	341,129	36,079
37	44	251,502	34,036	22	295,452	34,872	45	325,059	26,684	45	346,045	25,501
38	11	242,326	47,836	12	268,121	45,460	13	312,135	37,745	12	315,984	40,707
39	80	274,905	45,132	84	305,387	33,603	85	335,205	33,571	85	359,168	27,514
40	15	224,230	36,769	21	263,070	33,921	21	300,389	36,575	20	326,071	29,659
41	80	239,039	41,110	80	264,380	45,651	82	291,417	39,577	81	312,426	38,832
42	36	249,638	33,886	34	260,685	33,133	36	301,072	29,963	33	321,436	29,735
43	40	286,980	43,482	42	298,527	39,658	42	337,182	30,669	42	352,832	26,668
44	85	280,915	37,768	85	309,937	27,128	86	332,119	28,968	83	355,238	34,718
45	52	259,479	38,057	53	278,332	40,980	51	306,016	34,770	54	326,691	38,079
46	34	256,107	53,339	35	299,954	37,722	37	327,188	30,751	36	347,477	36,961
47	45	270,349	33,149	49	287,923	34,980	52	323,825	27,222	50	340,677	23,646
48	43	259,870	37,146	42	266,981	33,928	44	306,748	25,976	44	333,149	27,776
49	38	239,699	42,886	40	251,239	39,152	23	314,191	38,912	39	314,802	41,696
50	66	281,678	36,856	67	294,510	35,308	25	345,888	35,928	67	348,249	30,616
51	19	234,648	33,134	37	266,963	28,235	38	316,669	30,020	39	332,905	30,510
52	40	245,652	45,556	38	265,629	40,401	39	301,107	35,641	38	322,821	27,617
53	72	260,514	40,418	69	282,717	39,427	52	323,367	33,161	68	337,610	34,615
54	37	233,275	36,710	35	245,448	31,008	36	289,818	31,499	37	315,400	30,074
55	51	263,644	45,829	50	282,392	31,225	51	321,686	26,960	48	346,993	30,278
56	53	271,725	35,147	53	315,585	36,116	55	340,430	32,476	54	359,417	32,982
57	34	256,948	49,348	33	272,027	49,592	34	312,247	41,859	33	331,887	46,551
58	40	298,023	37,710	41	304,487	34,414	41	337,471	30,964	42	355,964	28,033
59	51	247,280	39,357	49	294,835	33,595	50	316,813	28,825	49	344,697	32,009
60	27	267,682	30,456	15	293,081	30,011	18	328,831	33,942	27	351,805	35,366
61	81	262,048	42,430	82	276,254	33,990	85	315,462	32,117	80	338,893	30,462
62	39	245,698	39,576	39	252,429	43,265	39	300,204	30,830			
63	23	237,953	50,472	24	255,863	45,687	24	295,999	47,018	24	312,300	46,016
64	27	241,873	41,464	27	260,686	35,399				27	325,963	31,839
65	40	260,384	34,225	41	279,963	28,905	41	324,405	27,717	20	356,677	22,634

Tabla AI.3. Nº de estudiantes, medias y desviaciones típicas por centro educativo.

Las medias brutas de las escuelas proporcionan información poco fiable sobre su rendimiento en matemáticas debido a que se encuentra sesgada por otros factores de contexto ajenos a su control. También la ganancia bruta tiene sus problemas como ya se ha detallado. A pesar de estos inconvenientes puede utilizarse como base de comparación.

El tamaño medio de los centros educativo es de 43 estudiantes aproximadamente. Más de la ratio media por aula en la Comunidad de Madrid. Se recomienda que para llevar a cabo el análisis multinivel las unidades de análisis tengan entre 20 y 25 sujetos. Sin embargo, en esta muestra, hay centros con un

tamaño inferior al recomendado, un total de 4 (con id 8, 17, 27 y 38). Con las estimaciones bayesianas (BLUP), los centros de este tipo tienden a no diferenciarse de la media por lo que sus resultados deben interpretarse con cautela.

Anexo I.3 Supuestos de las puntuaciones de rendimiento

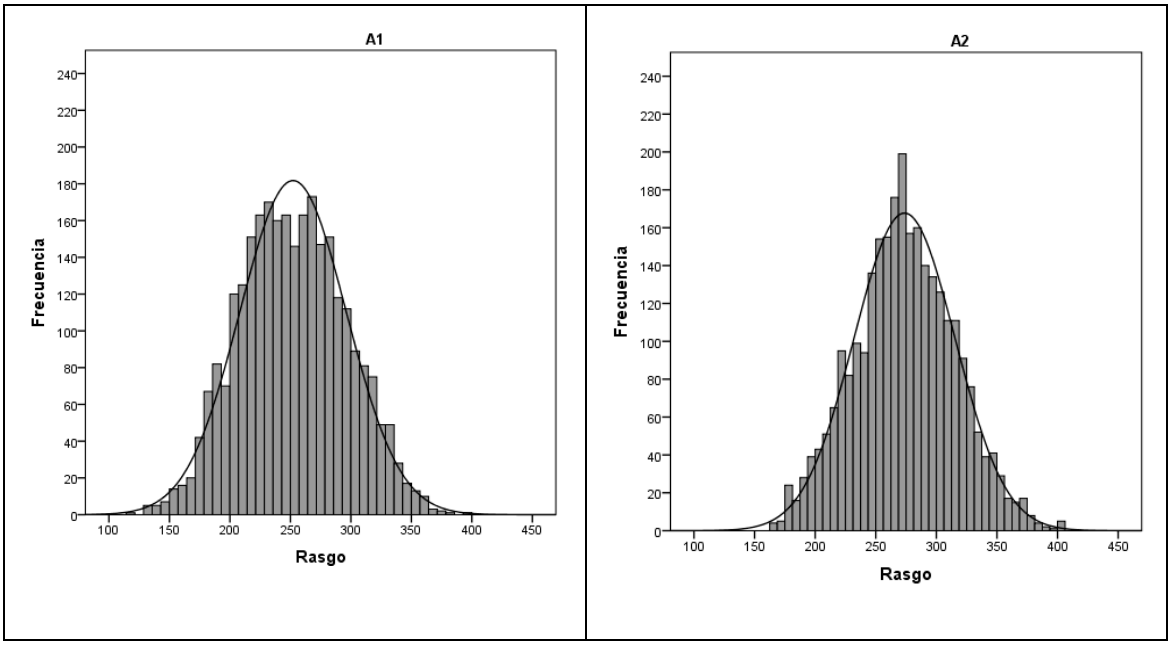
Anexo I.3.1 Estudio de normalidad

Para comprobar la normalidad de las puntuaciones de rendimiento en matemáticas obtenidas con la escala vertical, en cada una de las ocasiones de medida, se contrasta la siguiente hipótesis nula:

La distribución de puntuaciones de rendimiento es igual a la normal.

Se han elaborado histogramas con curva normal y calculado los estadísticos de Kolmogorv-Smirnov y Shapiro-Wilk y los gráficos de normalidad Q-Q para contrastar esta hipótesis.

En primer lugar, los histogramas de la distribución de frecuencias en las distintas aplicaciones parecen mostrar esa normalidad esperada. Los distintos intervalos de puntuaciones se sitúan dentro de la curva normal estimada para la distribución.



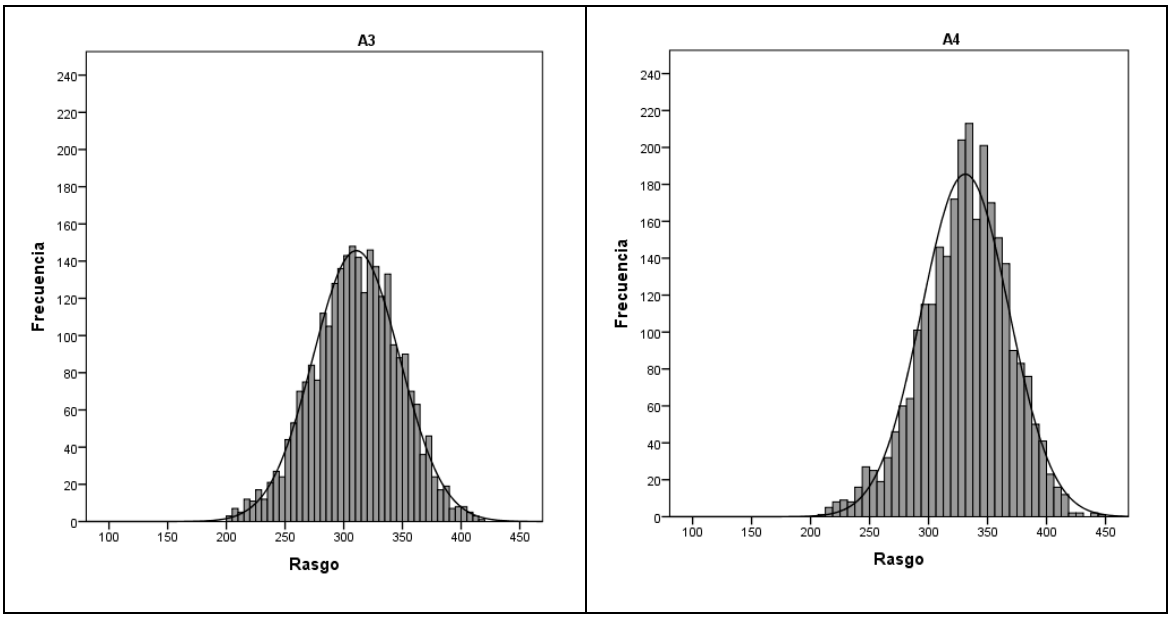


Gráfico AI.8. Histogramas con curva normal.

No obstante, los resultados de los estadísticos de normalidad no siguen la misma dirección que los histogramas.

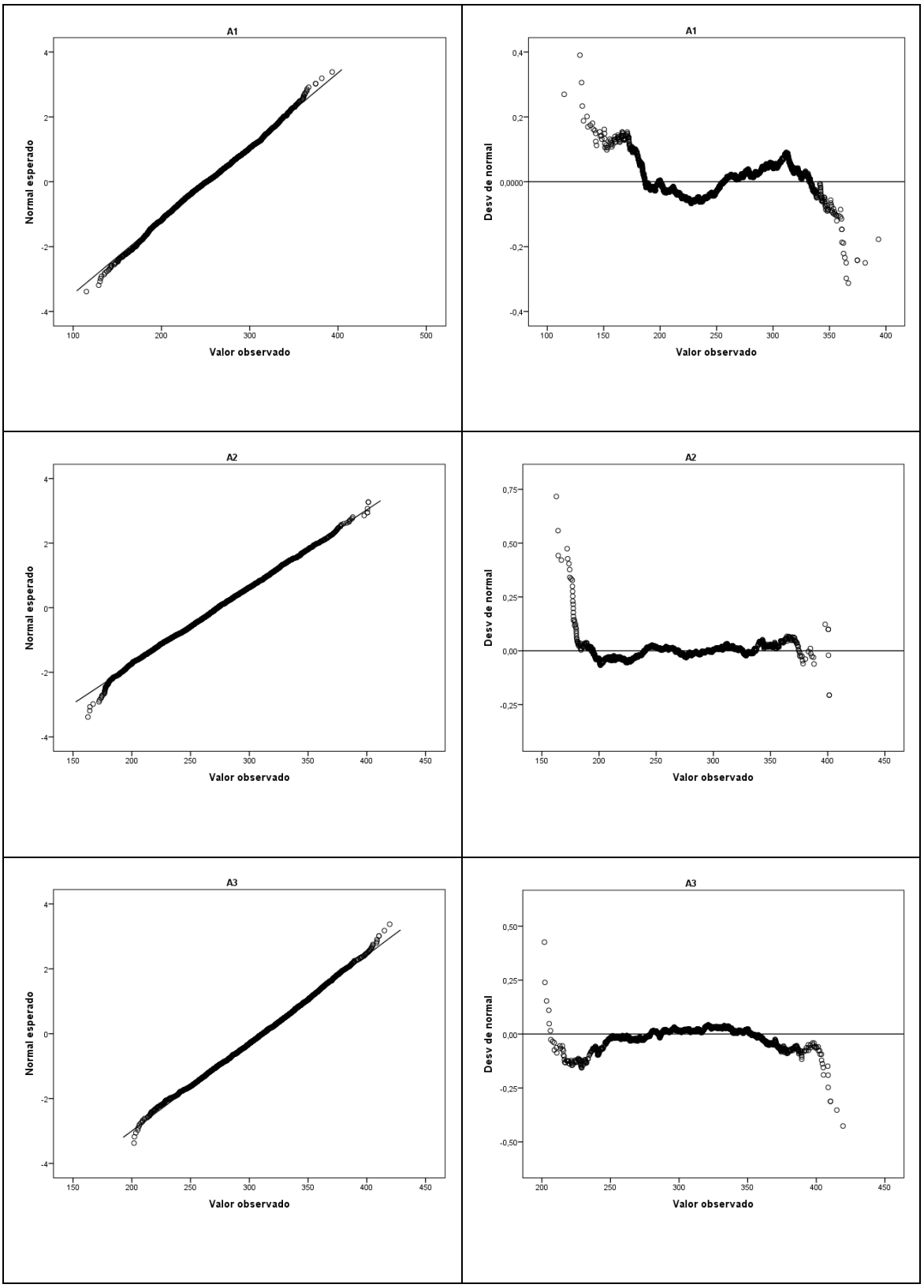
	TIEMPO	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Estadístico	gl	Sig.	Estadístico	gl	Sig.
Rasgo	1	0,024	2809	0,001	0,997	2809	0,000
	2	0,013	2801	0,200*	0,998	2801	0,001
	3	0,017	2695	0,081	0,998	2695	0,003
	4	0,031	2745	0,000	0,994	2745	0,000

a. Corrección de la significación de Lilliefors
*. Este es un límite inferior de la significación verdadera.

Tabla AI.4. Pruebas de normalidad de la distribución de las puntuaciones de rendimiento.

Los estadísticos para comprobar el supuesto de normalidad de la distribución indican que la hipótesis nula debe ser rechazada.

El conjunto de gráficos siguiente es otro elemento de información sobre la normalidad de la variable de rendimiento estudiada. La primera columna son los denominados Q-Q normales que comparan cada valor observado de rendimiento en matemáticas con la puntuación típica que le correspondería en una escala normal estandarizada. Si los puntos se sitúan sobre la línea continua la distribución de las puntuaciones es normal. La segunda columna incluye los gráficos Q-Q normal sin tendencias que representan las distancias entre los distintos puntos y la línea recta trazada en el primer grupo de gráficos. Para asumir la normalidad las diferencias no deben seguir ninguna tendencia.


A1
Normal esperado
Valor observado
A1
Desv de normal
Valor observado
A2
Normal esperado
Valor observado
A2
Desv de normal
Valor observado
A3
Normal esperado
Valor observado
A3
Desv de normal
Valor observado

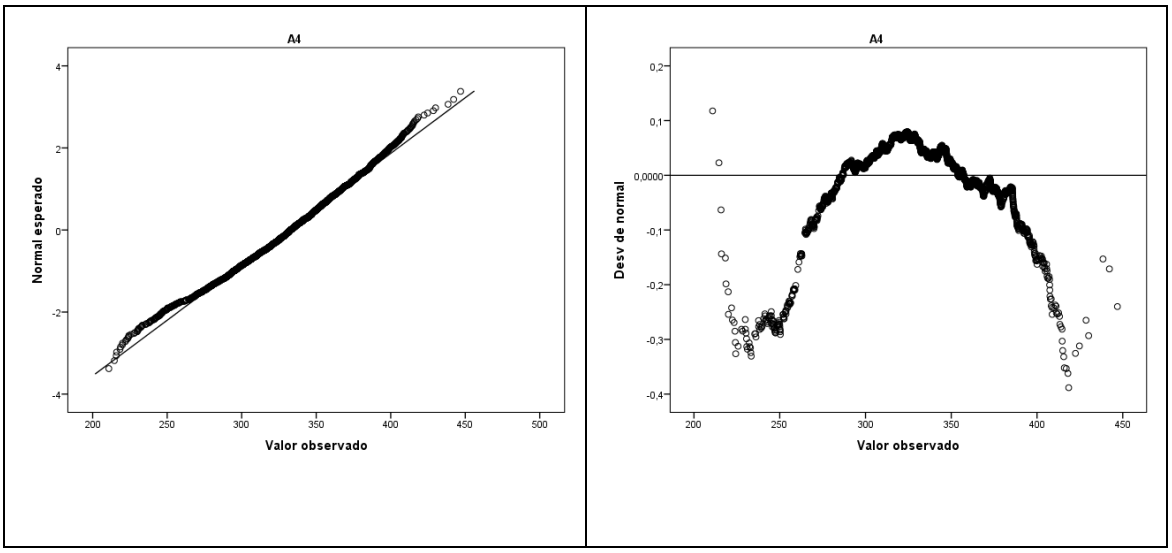


Gráfico AI.9. Gráficos de normalidad Q-Q

Si se analizan los gráficos Q-Q normales, excepto en los extremos del rasgo, la distribución parece tener las características de normalidad estadística. En la A4, esa distancia entre la distribución normal teórica y la empírica se hace más pronunciada en los extremos.

Los gráficos sin tendencias estudian en profundidad esas diferencias entre las puntuaciones observadas normalizadas y las que teóricamente les corresponden en una escala normal. Igual que los anteriores, excepto en los extremos, las distancias parecen no seguir ningún patrón. Sin embargo, en A1 y A2 las puntuaciones observadas del extremo inferior del rasgo tienen una puntuación por encima de lo que les correspondería en la escala normal. En A3 ambos extremos parecen tener puntuaciones observadas más bajas de lo esperados si se comparan con las normales teóricas y esta tendencia es más fuerte en A4.

En conclusión, los gráficos muestran, en términos generales, una normalidad en la distribución. Excepto en la A4 donde los gráficos de normalidad Q-Q sin tendencias ponen de manifiesto ese valor más bajo de las puntuaciones observadas en los extremos en comparación con la escala normal. De forma opuesta, los estadísticos no aseguran esa normalidad. Por estos motivos, no puede confirmarse esa normalidad estadística.

Anexo I.3.2 Estudio de homocedasticidad

La homocedasticidad o igualdad de varianza entre grupos se contrasta utilizando dos tipos de información. Por un lado, la prueba de Levene para contrastar la siguiente hipótesis nula:

La varianza de las puntuaciones del rasgo en las distintas ocasiones de medida es homogénea

Y, por otro lado, el gráfico que muestra la dispersión de la variable de rendimiento en cada nivel definido por el factor.

		Estadístico de Levene	gl1	gl2	Sig.
Rasgo	**Basándose en la media**	51,840	3	11046,000	0,000
	Basándose en la mediana.	51,940	3	11046,000	0,000
	Basándose en la mediana y con gl corregido	51,940	3	10891,651	0,000
	Basándose en la media recortada	52,043	3	11046,000	0,000

Tabla AI.5. Prueba de Levene para la Homogeneidad de varianzas.

La prueba de Levene utiliza como variable dependiente las puntuaciones diferenciales, es decir, el valor absoluto de la diferencia entre la puntuación observada y la media, la mediana o la media recortada del grupo. Todos los contrastes han resultados significativos, por tanto debe rechazarse la hipótesis de igualdad de varianzas entre mediciones.

Otro indicador de esa falta de igualdad de varianza se observa en el gráfico de dispersión (Gráfico AI.10). Para que la variable muestre homocedasticidad los puntos del gráfico deben distribuirse de forma horizontal y, como puede verse, no ocurre así.

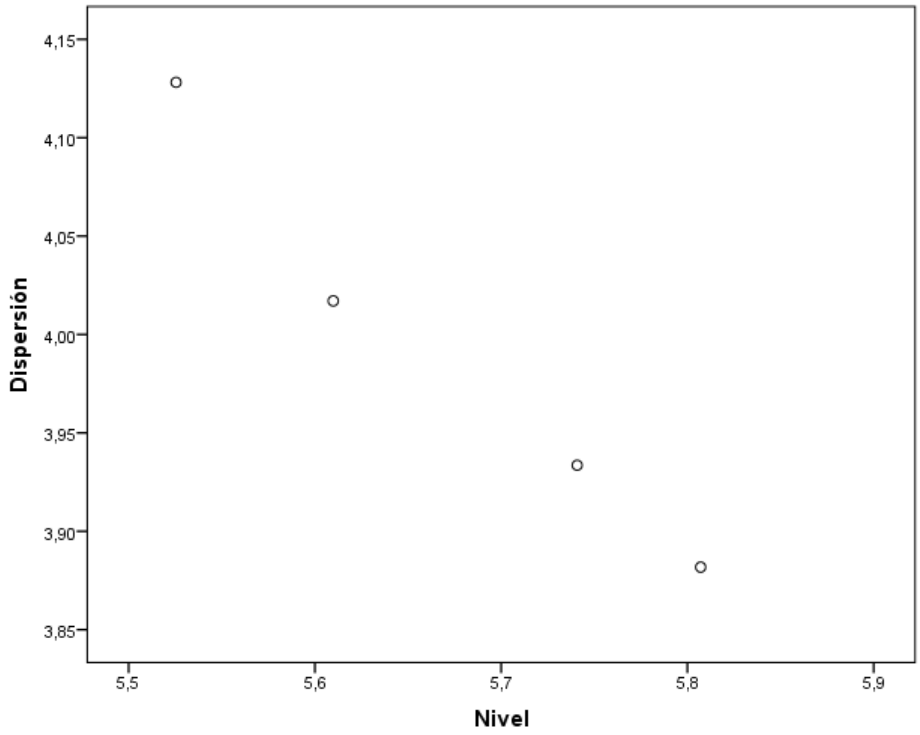


Gráfico AI.10. Dispersión en cada una de las aplicaciones.

En conclusión y tomando como referencia los resultados no se puede asumir la igualdad de varianzas entre aplicaciones de medida. Existe una tendencia decreciente, ya anunciada, que parece inherente a las escalas verticales que se desarrollan con modelos psicométricos TRI.

Anexo I.3.3 ¿Es posible asumir la propiedad de intervalo de la escala vertical?

Reckase (2008) considera que las escalas de rendimiento construidas bajo los supuestos de la TRI poseen la propiedad de intervalo por dos motivos:

A. En primer lugar, si el modelo TRI empleado ajusta con los datos puede considerarse una escala de intervalo debido a que la forma de la función dentro de esta teoría no está definida al menos que las escala de rendimiento tenga dicha propiedad. En este trabajo, el modelo de tres parámetros consigue alcanzar los criterios de convergencia establecidos (ciclo EM: 50; nº máximo de ciclos Newton: 25; y criterio de convergencia: 0,01)

B. En segundo lugar, poniendo la atención en la forma específica de la distribución de las puntuaciones verdaderas en el test. Si la distribución observada coincide con la distribución asumida, entonces puede argumentarse que los resultados tienen esa propiedad. Por tanto, si el número de respuestas correctas sigue la misma distribución que el rasgo podrá asumirse dicha la propiedad de intervalo de la escala.

De la misma forma que ocurre con el rasgo, el número de respuestas correctas y la proporción de respuestas correctas tampoco se distribuyen de forma normal como muestra la Tabla AI.6.

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	APLICACION	Estadístico	gl	Sig.	Estadístico	gl	Sig.
Nº Correctas	1	,052	2809	,000	,993	2809	,000
	2	,044	2801	,000	,993	2801	,000
	3	,054	2695	,000	,990	2695	,000
	4	,052	2745	,000	,992	2745	,000
% Correctas	1	,039	2809	,000	,994	2809	,000
	2	,040	2801	,000	,994	2801	,000
	3	,046	2695	,000	,992	2695	,000
	4	,040	2745	,000	,994	2745	,000
Rasgo	1	,024	2809	,001	,997	2809	,000
	2	,013	2801	,200*	,998	2801	,001
	3	,017	2695	,081	,998	2695	,003
	4	,031	2745	,000	,994	2745	,000

a. Corrección de la significación de Lilliefors
*. Este es un límite inferior de la significación verdadera.

Tabla AI.6. Pruebas de normalidad de las variables: Rasgo, nº y % de respuestas correctas.

Es necesario un análisis con mayor detalle de la forma de la distribución de las variables rasgo, número y porcentaje de respuestas correctas para comprobar si se asemejan. Los resultados del estudio de la simetría y curtosis se presentan en la Tabla AI.7.

APLICACION		Nº Correctas	% Correctas	Rasgo
1	N	2809	2809	2809
	Asimetría	-,030	-,013	,064
	ET. de asimetría	,046	,046	,046
	Curtosis	-,495	-,496	-,387
	ET. de curtosis	,092	,092	,092
2	N	2801	2801	2801
	Asimetría	-,022	-,027	,051
	ET. de asimetría	,046	,046	,046
	Curtosis	-,517	-,513	-,196
	ET. de curtosis	,092	,092	,092
3	N	2695	2695	2695
	Asimetría	-,186	-,181	-,121
	ET. de asimetría	,047	,047	,047
	Curtosis	-,505	-,504	-,138
	ET. de curtosis	,094	,094	,094
4	N	2745	2745	2745
	Asimetría	-,259	-,258	-,303
	ET. de asimetría	,047	,047	,047
	Curtosis	-,116	-,115	,115
	ET. de curtosis	,093	,093	,093

Tabla AI.7. Simetría, Curtosis y errores típicos de las variables Rasgo, nº y % de respuestas correctas.

El estudio de la simetría revela que la distribución es simétrica en las tres variables en las dos primeras aplicaciones. Los coeficientes de simetría no resultan significativos. En cambio, A3 y A4 poseen una distribución asimétrica negativa.

La curtosis negativa, es decir, una distribución platicúrtica es característica de las tres variables analizadas en las dos primeras aplicaciones. En la tercera aplicación, la curtosis no es significativamente distinta de cero y, por tanto, la distribución es mesocúrtica. Esta misma forma se da en las tres variables durante la cuarta aplicación.

Teniendo en cuenta los resultados anteriores la distribución de las tres variables parece similar y puede asumirse, por tanto, la propiedad de intervalo de la escala vertical de rendimiento en matemáticas elaborada.

Anexo I.4 Análisis de ítems

Anexo I.4.1 Análisis desde la Teoría Clásica de los Test

	Item	N Total	N Correctas	% Correctas	R_Item*Test	Rbp
1	**MO5AB1**	2809	2674	0,952	0,177	0,378
2	**MO5AB2**	2809	2586	0,921	0,156	0,285
3	**MO5AB3**	2809	1558	0,555	0,307	0,386
4	**MO5AB4**	2809	1612	0,574	0,138	0,174
5	**MO5AB5**	2809	1510	0,538	0,334	0,419
6	**MO5AB6**	2809	1161	0,413	0,122	0,154
7	**MO5AB7**	2809	1473	0,524	0,354	0,444
8	**MO5AB8**	2809	2207	0,786	0,308	0,433
9	**MO5AB9**	2809	843	0,300	0,030	0,040
10	**MO5AB10**	2809	1246	0,444	0,216	0,272
11	**MO5AB11**	2809	1982	0,706	0,244	0,323
12	**MO5AB12**	2809	797	0,284	0,323	0,430
13	**MO5AB13**	2809	811	0,289	0,234	0,310
14	**MO5AB14**	2809	2126	0,757	0,370	0,508
15	**MO5AB15**	2809	1289	0,459	0,234	0,293
16	**MO5AB16**	2809	2643	0,941	0,054	0,109
17	**MO5AB17**	2809	1821	0,648	0,202	0,259
18	**MO5AB18**	2809	2537	0,903	0,237	0,409
19	**MO5AB19**	2809	2240	0,797	0,203	0,289
20	**MO5A28**	1390	817	0,588	0,363	0,459
21	**MO5A29**	1390	890	0,640	0,280	0,360
22	**MO5A30**	1390	565	0,406	0,251	0,318
23	**MO5A31**	1390	688	0,495	0,348	0,436
24	**MO5A32**	1390	1148	0,826	0,213	0,315
25	**MO5A33**	1390	1019	0,733	0,371	0,499
26	**MO5A34**	1390	1122	0,807	0,357	0,515
27	**MO5A35**	1390	1144	0,823	0,251	0,369
28	**MO5A36**	1390	790	0,568	0,433	0,545
29	**MO5A37**	1390	950	0,683	0,256	0,334
30	**MO5B30**	1419	1034	0,729	0,261	0,351
31	**MO5B31**	1419	833	0,587	0,200	0,252
32	**MO5B32**	1419	593	0,418	0,241	0,305
33	**MO5B33**	1419	1222	0,861	0,299	0,467
34	**MO5B34**	1419	821	0,579	0,406	0,513
35	**MO5B35**	1419	755	0,532	0,255	0,32
36	**MO5B36**	1419	339	0,239	0,210	0,288
37	**MO5B37**	1419	1193	0,841	0,025	0,038
38	**MO5B38**	1419	1159	0,817	0,257	0,375
39	**MO5A20_J6B19**	2818	2526	0,896	0,141	0,239
40	**MO5A21_J6B20**	2818	1048	0,372	0,352	0,45
41	**MO5A22_J6B21**	2818	945	0,335	0,203	0,263

42	**MO5A23_J6B22**	2818	2186	0,776	0,308	0,429
43	**MO5A24_J6B23**	2818	1986	0,705	0,328	0,434
44	**MO5A25_J6B24**	2818	2264	0,803	0,280	0,402
45	**MO5A26_J6B25**	2818	899	0,319	0,219	0,286
46	**MO5A27_J6B26**	2818	1651	0,586	0,344	0,435
47	**MO5B20_J6A21**	2792	2369	0,848	0,235	0,359
48	**MO5B21_J6A22**	2792	1429	0,512	0,329	0,413
49	**MO5B22_J6A23**	2792	1797	0,644	0,309	0,397
50	**MO5B23_J6A24**	2792	1468	0,526	0,328	0,411
51	**MO5B24_J6A25**	2792	2107	0,755	0,205	0,281
52	**MO5B25_J6A26**	2792	2087	0,747	0,314	0,427
53	**MO5B26_J6A27**	2792	729	0,261	0,226	0,306
54	**MO5B27_J6A28**	2792	1435	0,514	0,192	0,241
55	**MO5B28_J6A29**	2792	1683	0,603	0,228	0,289
56	**MO5B29_J6A30**	2792	419	0,150	0,033	0,050
57	**MJ6AB1**	2801	2159	0,771	0,423	0,586
58	**MJ6AB2**	2801	1593	0,569	0,225	0,284
59	**MJ6AB3**	2801	2081	0,743	0,395	0,535
60	**MJ6AB4**	2801	1845	0,659	0,289	0,373
61	**MJ6AB5**	2801	2121	0,757	0,327	0,448
62	**MJ6AB6**	2801	1194	0,426	0,125	0,157
63	**MJ6AB7**	2801	2263	0,808	0,279	0,402
64	**MJ6AB8**	2801	2222	0,793	0,300	0,425
65	**MJ6AB9**	2801	1596	0,570	0,380	0,479
66	**MJ6AB10**	2801	1256	0,448	0,260	0,327
67	**MJ6AB11**	2801	616	0,220	0,186	0,261
68	**MJ6AB12**	2801	2084	0,744	0,405	0,549
69	**MJ6AB13**	2801	1820	0,650	0,288	0,370
70	**MJ6AB14**	2801	1387	0,495	0,345	0,432
71	**MJ6AB15**	2801	645	0,230	0,128	0,178
72	**MJ6AB16**	2801	844	0,301	0,237	0,312
73	**MJ6AB17**	2801	1697	0,606	0,331	0,421
74	**MJ6AB18**	2801	1526	0,545	0,272	0,341
75	**MJ6A29_N6B20**	2734	2067	0,756	0,327	0,448
76	**MJ6A30_N6B21**	2734	1810	0,662	0,196	0,253
77	**MJ6A31_N6B22**	2734	1408	0,515	0,426	0,534
78	**MJ6A32_N6B23**	2734	1809	0,662	0,335	0,434
79	**MJ6A33_N6B24**	2734	1352	0,495	0,363	0,455
80	**MJ6A34_N6B25**	2734	973	0,356	0,148	0,190
81	**MJ6A35_N6B26**	2734	1498	0,548	0,377	0,474
82	**MJ6A36_N6B27**	2734	2047	0,749	0,319	0,434
83	**MJ6A37_N6B28**	2734	1884	0,689	0,285	0,373
84	**MJ6A38_N6B29**	2734	2187	0,800	0,338	0,484
85	**MJ6B27_N6A20**	2762	1835	0,664	0,259	0,336
86	**MJ6B28_N6A21**	2762	1991	0,721	0,345	0,461
87	**MJ6B29_N6A22**	2762	1943	0,703	0,377	0,498

88	**MJ6B30_N6A23**	2762	1675	0,606	0,347	0,441
89	**MJ6B31_N6A24**	2762	939	0,340	0,271	0,351
90	**MJ6B32_N6A25**	2762	1858	0,673	0,316	0,410
91	**MJ6B33_N6A26**	2762	1305	0,472	0,364	0,456
92	**MJ6B34_N6A27**	2762	1695	0,614	0,399	0,508
93	**MJ6B35_N6A28**	2762	1474	0,534	0,271	0,341
94	**MJ6B36_N6A29**	2762	1704	0,617	0,353	0,449
95	**MN6AB1**	2695	1387	0,515	0,380	0,476
96	**MN6AB2**	2695	1149	0,426	0,179	0,225
97	**MN6AB3**	2695	1952	0,724	0,394	0,526
98	**MN6AB4**	2695	1831	0,679	0,332	0,434
99	**MN6AB5**	2695	2445	0,907	0,356	0,622
100	**MN6AB6**	2695	2276	0,845	0,344	0,523
101	**MN6AB7**	2695	503	0,187	0,128	0,186
102	**MN6AB8**	2695	1651	0,613	0,295	0,375
103	**MN6AB9**	2695	1526	0,566	0,369	0,465
104	**MN6AB10**	2695	2080	0,772	0,238	0,33
105	**MN6AB11**	2695	617	0,229	0,103	0,143
106	**MN6AB12**	2695	1727	0,641	0,283	0,363
107	**MN6AB13**	2695	878	0,326	0,347	0,451
108	**MN6AB14**	2695	1112	0,413	0,424	0,536
109	**MN6AB16**	2695	1453	0,539	0,186	0,234
110	**MN6AB17**	2695	2033	0,754	0,347	0,474
111	**MN6AB18**	2695	1679	0,623	0,339	0,432
112	**MN6A31**	1334	676	0,507	0,327	0,410
113	**MN6A35**	1334	1121	0,840	0,332	0,501
114	**MN6A40**	1334	428	0,321	0,203	0,264
115	**MN6AB19_J7B1**	4080	1491	0,365	0,207	0,265
116	**MN6A30_J7B2**	2719	2152	0,791	0,313	0,443
117	**MN6A32_J7B4**	2719	2001	0,736	0,358	0,482
118	**MN6A33_J7B5**	2719	1551	0,570	0,311	0,392
119	**MN6A34_J7B6**	2719	1911	0,703	0,340	0,449
120	**MN6A37_J7B9**	2719	1382	0,508	0,290	0,363
121	**MN6A38_J7B10**	2719	2328	0,856	0,344	0,532
122	**MN6B30_J7A1**	2721	1992	0,732	0,318	0,427
123	**MN6B31_J7A2**	2721	2558	0,940	0,305	0,608
124	**MN6B32_J7A3**	2721	2187	0,804	0,355	0,509
125	**MN6B33_J7A4**	2721	812	0,298	0,158	0,208
126	**MN6B34_J7A5**	2721	1021	0,375	0,184	0,235
127	**MN6B35_J7A6**	2721	2340	0,860	0,353	0,550
128	**MN6B36_J7A7**	2721	1323	0,486	0,401	0,503
129	**MN6B37_J7A8**	2721	1448	0,532	0,307	0,385
130	**MN6B38_J7A9**	2721	798	0,293	0,126	0,167
131	**MN6B39_J7A10**	2721	2335	0,858	0,332	0,516
132	**MJ7AB11**	2745	729	0,266	0,268	0,361
133	**MJ7AB13**	2745	1375	0,501	0,310	0,389

134	**MJ7AB14**	2745	950	0,346	0,093	0,120
135	**MJ7AB15**	2745	658	0,240	0,225	0,309
136	**MJ7AB16**	2745	2108	0,768	0,376	0,520
137	**MJ7AB17**	2745	1608	0,586	0,327	0,413
138	**MJ7AB18**	2745	2410	0,878	0,222	0,360
139	**MJ7AB19**	2745	1813	0,660	0,231	0,298
140	**MJ7AB20**	2745	843	0,307	0,041	0,053
141	**MJ7AB21**	2745	1105	0,403	0,174	0,220
142	**MJ7AB22**	2745	2286	0,833	0,222	0,331
143	**MJ7AB23**	2745	1601	0,583	0,247	0,312
144	**MJ7AB24**	2745	847	0,309	0,059	0,078
145	**MJ7AB25**	2745	1688	0,615	0,074	0,095
146	**MJ7AB26**	2745	1311	0,478	0,258	0,324
147	**MJ7AB27**	2745	2315	0,843	0,386	0,585
148	**MJ7AB28**	2745	1537	0,560	0,340	0,427
149	**MJ7AB29**	2745	781	0,285	0,082	0,110
150	**MJ7AB30**	2745	1141	0,416	0,238	0,301
151	**MJ7AB31**	2745	1395	0,508	0,431	0,540
152	**MJ7AB32**	2745	1683	0,613	0,420	0,535
153	**MJ7AB33**	2745	2317	0,844	0,421	0,638
154	**MJ7AB34**	2745	1953	0,711	0,297	0,393
155	**MJ7AB35**	2745	2066	0,753	0,349	0,477
156	**MJ7AB36**	2745	1735	0,632	0,335	0,429
157	**MJ7AB37**	2745	1461	0,532	0,208	0,261
158	**MJ7AB38**	2745	2152	0,784	0,294	0,412
159	**MJ7AB39**	2745	1011	0,368	0,141	0,180
160	**MJ7AB40**	2745	1577	0,574	0,277	0,350
161	**MJ7B3**	1385	524	0,378	0,238	0,304
162	**MJ7B7**	1385	1151	0,831	0,345	0,513

Tabla AI.8. Análisis de TCT dificultad (% respuestas correctas) y discriminación (correlación biserial puntual) de los ítems.

Anexo I.4.2 Análisis desde la Teoría Respuesta al Ítem

Anexo I.4.2.1 Curvas características de los ítems

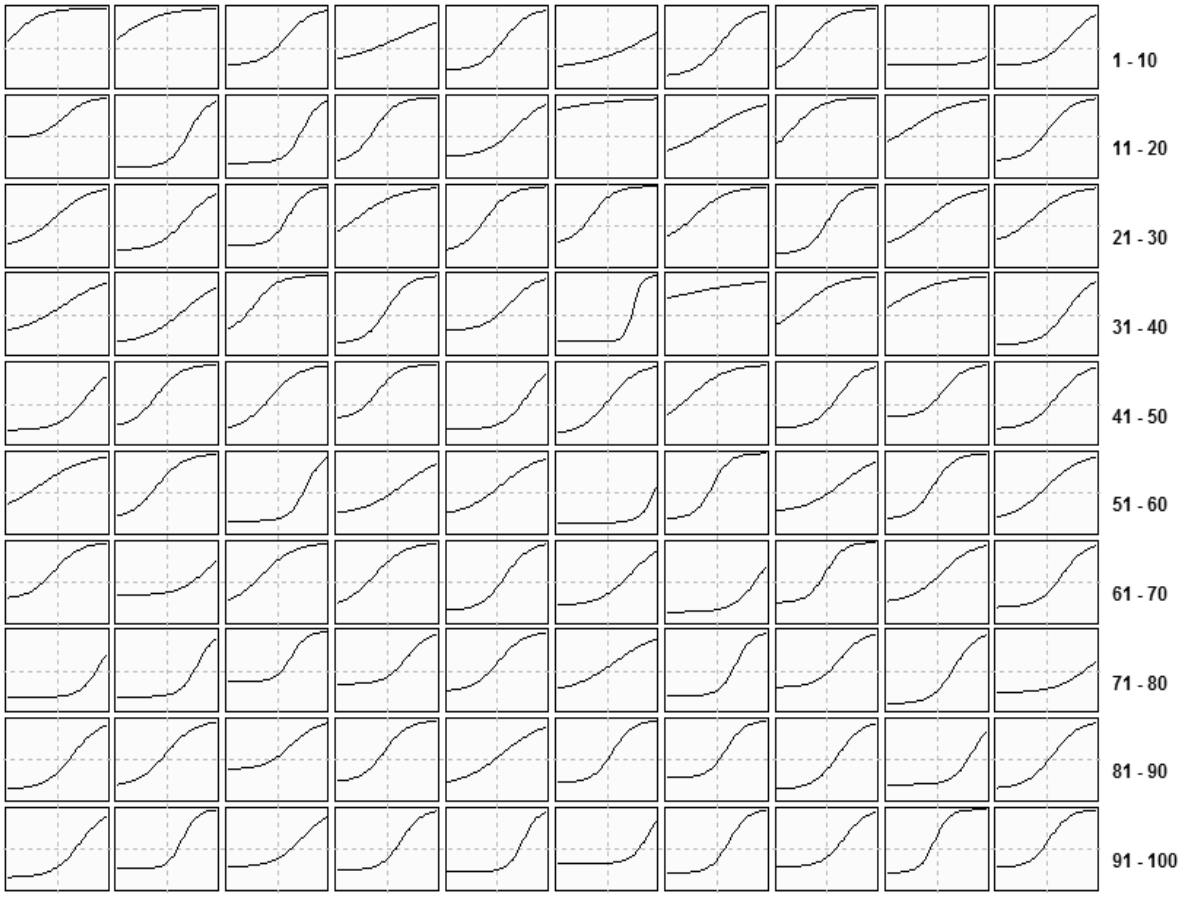


Gráfico AI.11. Curvas Características de los Ítems 1 a 100

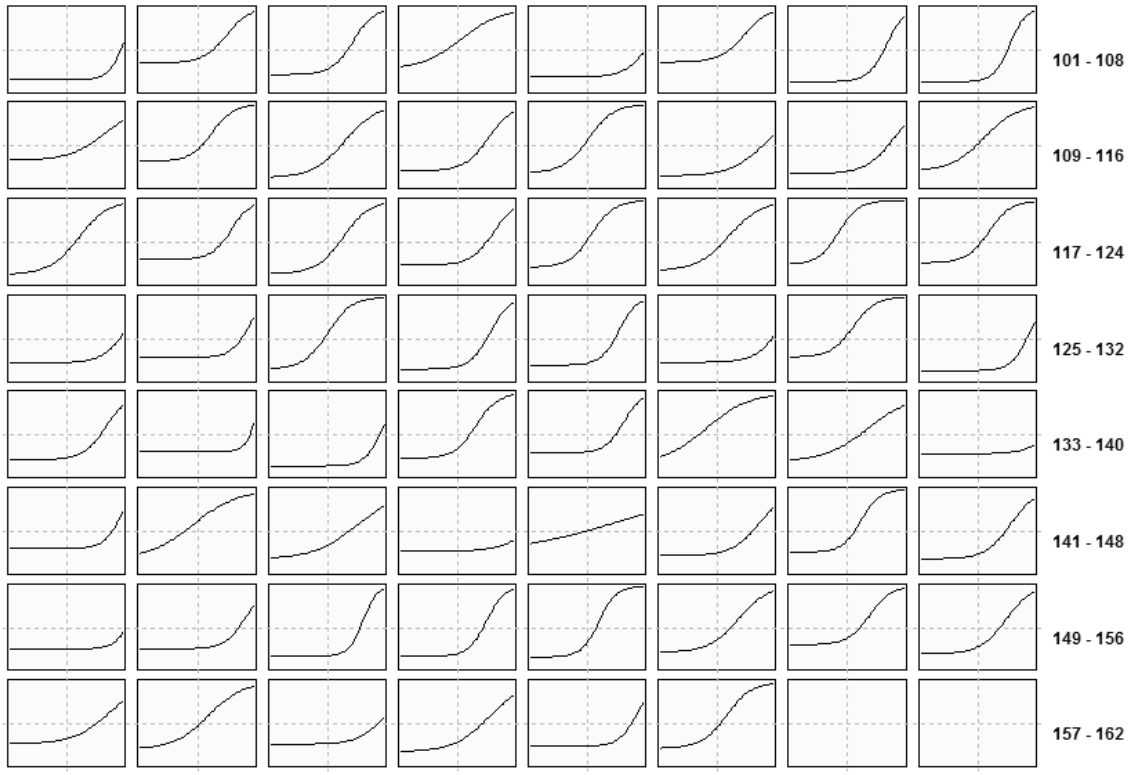


Gráfico AI.12. Curvas Características de los Ítems 101 a 162

Anexo I.4.2.2 Parámetros de los ítems e índice de ajuste

		a	b	c	chi2 (prob)	gl
MO5AB1	Coef	0,657	-2.900	0,218	12,7	6
	ET	0,069	0,28	0,091	(-0,0474)	
MO5AB2	Coef	0,459	-3.069	0,23	12	7
	ET	0,053	0,388	0,095	(-0,1006)	
MO5AB3	Coef	0,778	0,533	0,29	305,5	8
	ET	0,111	0,131	0,043	(0)	
MO5AB4	Coef	0,301	0,8	0,283	13,6	9
	ET	0,061	0,509	0,085	(-0,1367)	
MO5AB5	Coef	0,776	0,382	0,219	215,4	8
	ET	0,093	0,121	0,043	(0)	
MO5AB6	Coef	0,343	2.346	0,249	28,6	9
	ET	0,091	0,332	0,056	(-0,0008)	
MO5AB7	Coef	0,697	0,255	0,148	178,1	8
	ET	0,075	0,127	0,047	(0)	
MO5AB8	Coef	0,69	-1.113	0,185	33,3	7
	ET	0,057	0,175	0,073	(0)	
MO5AB9	Coef	0,712	4.541	0,294	423,3	9
	ET	0,317	1.475	0,012	(0)	
MO5AB10	Coef	0,732	1.330	0,289	22,7	8
	ET	0,139	0,11	0,032	(-0,0038)	
MO5AB11	Coef	0,862	0,367	0,5	34,1	8
	ET	0,154	0,166	0,042	(0)	
MO5AB12	Coef	1.005	1.301	0,125	764,5	4
	ET	0,134	0,062	0,018	(0)	
MO5AB13	Coef	1.136	1.522	0,182	63	8
	ET	0,181	0,072	0,015	(0)	
MO5AB14	Coef	0,859	-0,824	0,18	36,2	7
	ET	0,073	0,137	0,065	(0)	
MO5AB15	Coef	0,638	1.131	0,258	172,1	8
	ET	0,113	0,133	0,041	(0)	
MO5AB16	Coef	0,223	-6.732	0,224	8,1	8
	ET	0,042	1.306	0,094	(-0,4241)	
MO5AB17	Coef	0,37	-0,364	0,214	22,3	8
	ET	0,047	0,365	0,082	(-0,0044)	
MO5AB18	Coef	0,654	-2.174	0,192	23,7	7
	ET	0,058	0,213	0,082	(-0,0013)	
MO5AB19	Coef	0,415	-1.525	0,261	10,9	8
	ET	0,046	0,384	0,098	(-0,2101)	
MO5A28	Coef	0,779	0,084	0,202	90,1	7
	ET	0,108	0,161	0,059	(0)	
MO5A29	Coef	0,561	-0,115	0,241	31,2	8
	ET	0,083	0,269	0,079	(-0,0001)	
MO5A30	Coef	0,669	1.166	0,198	66,8	8

	ET	0,128	0,143	0,043	(	0	)	
MO5A31	Coef	1.122	0,672	0,26		18,1		7
	ET	0,202	0,095	0,036	(	-0,0115	)	
MO5A32	Coef	0,443	-1.890	0,2		46,9		7
	ET	0,057	0,338	0,085	(	0	)	
MO5A33	Coef	0,79	-0,728	0,176		56,1		7
	ET	0,084	0,165	0,07	(	0	)	
MO5A34	Coef	0,896	-0,915	0,282		25,4		7
	ET	0,112	0,199	0,087	(	-0,0007	)	
MO5A35	Coef	0,59	-1.350	0,268		33,2		7
	ET	0,077	0,301	0,098	(	0	)	
MO5A36	Coef	0,947	0,057	0,157		51,8		6
	ET	0,125	0,12	0,052	(	0	)	
MO5A37	Coef	0,48	-0,565	0,206		41,6		8
	ET	0,064	0,293	0,082	(	0	)	
MO5B30	Coef	0,555	-0,686	0,263		5,7		8
	ET	0,076	0,306	0,093	(	-0,678	)	
MO5B31	Coef	0,431	0,371	0,264		21,7		8
	ET	0,082	0,369	0,084	(	-0,0055	)	
MO5B32	Coef	0,482	1.031	0,143		73,2		7
	ET	0,084	0,205	0,054	(	0	)	
MO5B33	Coef	0,824	-1.395	0,268		54,4		6
	ET	0,106	0,224	0,097	(	0	)	
MO5B34	Coef	0,886	-0,027	0,152		31,6		7
	ET	0,1	0,123	0,05	(	0	)	
MO5B35	Coef	0,731	0,768	0,307		39,5		8
	ET	0,148	0,178	0,053	(	0	)	
MO5B36	Coef	2.195	1.559	0,175		104,5		6
	ET	0,754	0,081	0,014	(	0	)	
MO5B37	Coef	0,164	-4.856	0,237		47,1		8
	ET	0,035	1.260	0,097	(	0	)	
MO5B38	Coef	0,527	-1.500	0,238		24,3		7
	ET	0,068	0,309	0,094	(	-0,001	)	
MO5A20_J	Coef	0,377	-2.927	0,224		8,4		8
	ET	0,042	0,432	0,093	(	-0,3939	)	
MO5A21_J	Coef	0,76	1.210	0,134		12,5		9
	ET	0,094	0,077	0,028	(	-0,1855	)	
MO5A22_J	Coef	0,718	1.731	0,18		181,5		8
	ET	0,114	0,1	0,026	(	0	)	
MO5A23_J	Coef	0,8	-0,674	0,215		70,8		7
	ET	0,073	0,166	0,07	(	0	)	
MO5A24_J	Coef	0,655	-0,447	0,165		26,2		8
	ET	0,058	0,164	0,062	(	-0,001	)	
MO5A25_J	Coef	0,885	-0,605	0,319		12,9		7
	ET	0,096	0,187	0,075	(	-0,0751	)	
MO5A26_J	Coef	0,849	1.773	0,19		252,4		9

	ET	0,137	0,094	0,021	(	0	)
MO5A27_J	Coef	0,686	0,115	0,125		24,8	9
	ET	0,06	0,118	0,044	(	-0,0032	)
MO5B20_J	Coef	0,527	-1.719	0,177		9,6	8
	ET	0,045	0,235	0,077	(	-0,295	)
MO5B21_J	Coef	0,819	0,67	0,204		12,8	9
	ET	0,092	0,098	0,036	(	-0,1698	)
MO5B22_J	Coef	0,902	0,384	0,343		139,3	8
	ET	0,119	0,123	0,042	(	0	)
MO5B23_J	Coef	0,723	0,592	0,193		67,6	9
	ET	0,086	0,12	0,043	(	0	)
MO5B24_J	Coef	0,406	-1.070	0,209		34,1	9
	ET	0,043	0,334	0,085	(	-0,0001	)
MO5B25_J	Coef	0,707	-0,662	0,175		5,5	8
	ET	0,061	0,165	0,067	(	-0,7078	)
MO5B26_J	Coef	1.174	1.733	0,156		184,5	9
	ET	0,163	0,069	0,014	(	0	)
MO5B27_J	Coef	0,469	0,999	0,234		80,5	9
	ET	0,081	0,233	0,061	(	0	)
MO5B28_J	Coef	0,509	0,252	0,217		31,9	9
	ET	0,067	0,241	0,068	(	-0,0002	)
MO5B29_J	Coef	1.245	2.960	0,135		35,3	9
	ET	0,395	0,292	0,009	(	-0,0001	)
MJ6AB1	Coef	1.133	-0,293	0,187		38,4	7
	ET	0,099	0,099	0,05	(	0	)
MJ6AB2	Coef	0,517	0,903	0,266		48,3	9
	ET	0,083	0,223	0,061	(	0	)
MJ6AB3	Coef	0,923	-0,245	0,182		15,1	8
	ET	0,081	0,119	0,054	(	-0,0568	)
MJ6AB4	Coef	0,554	-0,032	0,161		11,4	9
	ET	0,055	0,187	0,062	(	-0,2482	)
MJ6AB5	Coef	0,767	-0,19	0,295		12,9	8
	ET	0,087	0,19	0,069	(	-0,1153	)
MJ6AB6	Coef	0,693	2.428	0,343		87,7	9
	ET	0,172	0,199	0,027	(	0	)
MJ6AB7	Coef	0,594	-0,936	0,189		14	8
	ET	0,052	0,215	0,077	(	-0,0828	)
MJ6AB8	Coef	0,629	-0,796	0,176		24,2	8
	ET	0,054	0,192	0,072	(	-0,0021	)
MJ6AB9	Coef	0,819	0,464	0,148		85,4	9
	ET	0,074	0,092	0,038	(	0	)
MJ6AB10	Coef	0,634	1.373	0,209		13,7	9
	ET	0,092	0,12	0,039	(	-0,1329	)
MJ6AB11	Coef	0,778	2.484	0,132		35	9
	ET	0,134	0,147	0,017	(	-0,0001	)
MJ6AB12	Coef	1.073	-0,088	0,249		11,9	8

	ET	0,107	0,108	0,05	(	-0,1547	)	
MJ6AB13	Coef	0,62	0,292	0,25		20,1		9
	ET	0,077	0,197	0,063	(	-0,0175	)	
MJ6AB14	Coef	0,818	0,924	0,191		32,3		9
	ET	0,092	0,089	0,035	(	-0,0002	)	
MJ6AB15	Coef	0,992	2.595	0,179		71,2		9
	ET	0,209	0,165	0,014	(	0	)	
MJ6AB16	Coef	1.114	1.825	0,184		287,8		9
	ET	0,157	0,069	0,017	(	0	)	
MJ6AB17	Coef	1.276	0,846	0,371		51,4		8
	ET	0,149	0,072	0,027	(	0	)	
MJ6AB18	Coef	0,904	1.173	0,343		139,9		9
	ET	0,133	0,096	0,032	(	0	)	
MJ6A29_N	Coef	0,772	0,086	0,252		15,2		9
	ET	0,077	0,164	0,063	(	-0,0845	)	
MJ6A30_N	Coef	0,447	0,459	0,237		32,2		9
	ET	0,058	0,299	0,075	(	-0,0002	)	
MJ6A31_N	Coef	1.174	1.147	0,193		79,6		8
	ET	0,113	0,059	0,027	(	0	)	
MJ6A32_N	Coef	0,8	0,751	0,3		20,7		9
	ET	0,089	0,132	0,046	(	-0,014	)	
MJ6A33_N	Coef	0,869	1.036	0,098		111,5		7
	ET	0,068	0,063	0,026	(	0	)	
MJ6A34_N	Coef	0,531	2.895	0,234		47,6		9
	ET	0,123	0,213	0,033	(	0	)	
MJ6A35_N	Coef	0,716	0,917	0,141		38,1		9
	ET	0,068	0,104	0,039	(	0	)	
MJ6A36_N	Coef	0,687	-0,082	0,176		70		9
	ET	0,061	0,163	0,062	(	0	)	
MJ6A37_N	Coef	0,754	0,864	0,392		31,9		9
	ET	0,103	0,164	0,049	(	-0,0002	)	
MJ6A38_N	Coef	0,846	-0,176	0,231		31,2		9
	ET	0,073	0,149	0,063	(	-0,0003	)	
MJ6B27_N	Coef	0,477	0,319	0,183		22,2		9
	ET	0,052	0,238	0,068	(	-0,0082	)	
MJ6B28_N	Coef	0,993	0,345	0,225		22,1		9
	ET	0,086	0,1	0,045	(	-0,0087	)	
MJ6B29_N	Coef	1.072	0,563	0,287		34,3		9
	ET	0,107	0,096	0,042	(	-0,0001	)	
MJ6B30_N	Coef	0,8	0,673	0,141		107,2		8
	ET	0,067	0,097	0,04	(	0	)	
MJ6B31_N	Coef	1.004	2.119	0,201		36,4		9
	ET	0,146	0,071	0,02	(	0	)	
MJ6B32_N	Coef	0,739	0,353	0,15		24,4		9
	ET	0,061	0,121	0,048	(	-0,0037	)	
MJ6B33_N	Coef	0,748	1.339	0,153		50,8		9

	ET	0,08	0,09	0,035	(	0	)
MJ6B34_N	Coef	1.335	0,917	0,26		67,9	9
	ET	0,123	0,061	0,029	(	0	)
MJ6B35_N	Coef	0,703	1.484	0,282		21,3	9
	ET	0,1	0,125	0,041	(	-0,0114	)
MJ6B36_N	Coef	0,977	0,864	0,243		30,7	9
	ET	0,091	0,087	0,036	(	-0,0003	)
MN6AB1	Coef	1.300	1.559	0,222		10	9
	ET	0,137	0,056	0,028	(	-0,3504	)
MN6AB2	Coef	1.188	2.413	0,325		15,1	9
	ET	0,239	0,084	0,021	(	-0,0876	)
MN6AB3	Coef	1.066	0,721	0,206		18,6	9
	ET	0,094	0,096	0,051	(	-0,0291	)
MN6AB4	Coef	0,915	1.034	0,286		7,2	9
	ET	0,104	0,125	0,053	(	-0,6163	)
MN6AB5	Coef	1.319	-0,159	0,209		29,3	8
	ET	0,113	0,104	0,061	(	-0,0003	)
MN6AB6	Coef	0,99	0,175	0,277		13	9
	ET	0,092	0,14	0,068	(	-0,1607	)
MN6AB7	Coef	1.476	2.948	0,146		50,5	8
	ET	0,338	0,13	0,011	(	0	)
MN6AB8	Coef	0,979	1.517	0,348		40,6	9
	ET	0,136	0,106	0,041	(	0	)
MN6AB9	Coef	1.050	1.354	0,207		64,9	9
	ET	0,107	0,078	0,037	(	0	)
MN6AB10	Coef	0,547	0,232	0,267		6,1	9
	ET	0,063	0,272	0,082	(	-0,733	)
MN6AB11	Coef	0,888	3.359	0,178		18,3	9
	ET	0,24	0,262	0,018	(	-0,0317	)
MN6AB12	Coef	0,93	1.404	0,356		5,3	9
	ET	0,12	0,117	0,043	(	-0,811	)
MN6AB13	Coef	1.238	2.040	0,124		45,9	8
	ET	0,141	0,046	0,019	(	0	)
MN6AB14	Coef	1.391	1.678	0,122		107,8	8
	ET	0,126	0,039	0,02	(	0	)
MN6AB16	Coef	0,608	2.089	0,332		13,8	9
	ET	0,126	0,173	0,051	(	-0,1284	)
MN6AB17	Coef	1.053	0,773	0,322		37,2	9
	ET	0,109	0,118	0,054	(	0	)
MN6AB18	Coef	0,7	0,922	0,128		15,6	9
	ET	0,063	0,117	0,048	(	-0,0752	)
MN6A31	Coef	0,936	1.647	0,207		11,3	9
	ET	0,152	0,113	0,05	(	-0,2549	)
MN6A35	Coef	0,892	0,052	0,185		5,9	7
	ET	0,107	0,167	0,075	(	-0,5476	)
MN6A40	Coef	0,629	2.674	0,147		12,3	8

	ET	0,143	0,179	0,043	(	-0,1378	)
MN6AB19_	Coef	0,818	2.410	0,176		242	9
	ET	0,103	0,067	0,024	(	0	)
MN6A30_J	Coef	0,678	0,346	0,206		4,6	9
	ET	0,064	0,182	0,068	(	-0,8685	)
MN6A32_J	Coef	0,758	0,606	0,121		33,3	7
	ET	0,062	0,114	0,048	(	0	)
MN6A33_J	Coef	1.133	1.773	0,299		72,5	8
	ET	0,139	0,076	0,033	(	0	)
MN6A34_J	Coef	0,84	0,834	0,13		22	9
	ET	0,068	0,101	0,046	(	-0,0089	)
MN6A37_J	Coef	1.024	1.887	0,235		14,1	9
	ET	0,126	0,073	0,033	(	-0,1192	)
MN6A38_J	Coef	1.009	0,226	0,208		20,6	9
	ET	0,089	0,128	0,064	(	-0,0147	)
MN6B30_J	Coef	0,689	0,622	0,165		22,6	9
	ET	0,06	0,143	0,055	(	-0,0072	)
MN6B31_J	Coef	1.105	-0,48	0,242		19,8	8
	ET	0,119	0,176	0,076	(	-0,0113	)
MN6B32_J	Coef	1.008	0,586	0,259		29,5	9
	ET	0,092	0,119	0,056	(	-0,0005	)
MN6B33_J	Coef	0,909	3.123	0,216		29,4	9
	ET	0,205	0,154	0,021	(	-0,0006	)
MN6B34_J	Coef	1.287	2.686	0,283		18,7	9
	ET	0,24	0,078	0,018	(	-0,028	)
MN6B35_J	Coef	0,922	0,009	0,142		8,2	9
	ET	0,074	0,122	0,054	(	-0,5183	)
MN6B36_J	Coef	1.123	1.716	0,142		75,9	7
	ET	0,112	0,055	0,028	(	0	)
MN6B37_J	Coef	1.239	1.639	0,188		85,3	9
	ET	0,111	0,053	0,026	(	0	)
MN6B38_J	Coef	0,912	3.248	0,223		17,7	9
	ET	0,229	0,187	0,02	(	-0,0385	)
MN6B39_J	Coef	1.018	0,289	0,285		8,7	9
	ET	0,091	0,133	0,063	(	-0,4614	)
MJ7AB11	Coef	1.277	2.689	0,123		29,5	8
	ET	0,163	0,054	0,016	(	-0,0003	)
MJ7AB13	Coef	0,974	2.049	0,209		30	9
	ET	0,115	0,073	0,033	(	-0,0004	)
MJ7AB14	Coef	2.269	3.000	0,309		100,6	9
	ET	0,846	0,094	0,013	(	0	)
MJ7AB15	Coef	1.360	2.858	0,137		35,9	9
	ET	0,21	0,066	0,015	(	0	)
MJ7AB16	Coef	1.038	0,973	0,226		32,3	9
	ET	0,09	0,098	0,048	(	-0,0002	)
MJ7AB17	Coef	1.237	1.871	0,291		19,5	9

	ET	0,142	0,067	0,032	(	-0,0214	)
MJ7AB18	Coef	0,54	-0,566	0,172		31,9	9
	ET	0,053	0,265	0,072	(	-0,0002	)
MJ7AB19	Coef	0,499	1.159	0,185		17,9	9
	ET	0,058	0,21	0,061	(	-0,0363	)
MJ7AB20	Coef	0,783	4.347	0,279		51	9
	ET	0,302	0,594	0,018	(	0	)
MJ7AB21	Coef	1.361	2.734	0,3		43,7	9
	ET	0,268	0,07	0,02	(	0	)
MJ7AB22	Coef	0,496	-0,225	0,169		21,2	9
	ET	0,049	0,258	0,069	(	-0,012	)
MJ7AB23	Coef	0,515	1.632	0,176		28,9	9
	ET	0,064	0,177	0,054	(	-0,0007	)
MJ7AB24	Coef	0,66	4.309	0,263		10,8	9
	ET	0,244	0,571	0,023	(	-0,2912	)
MJ7AB25	Coef	0,207	1.614	0,234		26,1	9
	ET	0,041	0,677	0,087	(	-0,002	)
MJ7AB26	Coef	0,803	2.229	0,214		17,9	9
	ET	0,114	0,094	0,039	(	-0,0362	)
MJ7AB27	Coef	1.211	0,692	0,255		24,3	9
	ET	0,108	0,099	0,052	(	-0,0039	)
MJ7AB28	Coef	0,9	1.740	0,178		28,1	8
	ET	0,093	0,082	0,038	(	-0,0005	)
MJ7AB29	Coef	1.317	3.428	0,249		78,7	9
	ET	0,484	0,245	0,016	(	0	)
MJ7AB30	Coef	1.087	2.543	0,251		102,6	9
	ET	0,167	0,068	0,026	(	0	)
MJ7AB31	Coef	1.610	1.860	0,166		38,6	8
	ET	0,156	0,039	0,024	(	0	)
MJ7AB32	Coef	1.232	1.530	0,164		16,7	8
	ET	0,108	0,059	0,033	(	-0,0334	)
MJ7AB33	Coef	1.279	0,599	0,154		18	8
	ET	0,098	0,08	0,044	(	-0,0211	)
MJ7AB34	Coef	0,768	1.102	0,212		15,7	9
	ET	0,073	0,13	0,052	(	-0,0723	)
MJ7AB35	Coef	1.020	1.160	0,3		27,9	9
	ET	0,102	0,111	0,049	(	-0,001	)
MJ7AB36	Coef	0,88	1.472	0,193		9,5	9
	ET	0,084	0,094	0,042	(	-0,3923	)
MJ7AB37	Coef	0,637	2.223	0,269		32,5	8
	ET	0,104	0,142	0,047	(	-0,0001	)
MJ7AB38	Coef	0,664	0,584	0,208		23,1	8
	ET	0,062	0,176	0,062	(	-0,0032	)
MJ7AB39	Coef	0,81	3.163	0,263		63,2	9
	ET	0,198	0,155	0,028	(	0	)
MJ7AB40	Coef	0,647	1.675	0,174		17,2	9

	ET	0,076	0,133	0,049	(	-0,0459	)
MJ7B3	Coef	1.241	2.601	0,238		10,5	9
	ET	0,264	0,086	0,029	(	-0,308	)
MJ7B7	Coef	0,887	0,487	0,215		4,8	9
	ET	0,109	0,173	0,07	(	-0,8538	)

Tabla AI.9. Parámetros a (discriminación), b (dificultad) y c (azar), errores típicos, chi-cuadrado y probabilidad asociada y grados de libertad de los ítems

En la Tabla AI.9 las probabilidades asociadas a chi-cuadrado por encima de 0,05 indican un buen ajuste del ítem, es decir, valores no significativos. Indicador de que la curvas características del ítem teóricas y empíricas coinciden. Los ítems con valores de probabilidad por encima de 0,05 no mostrarían un buen ajuste. Además, el estadístico chi-cuadado tiende a ser significativo cuando los tamaños muestrales son altos (Gaviria, Biencinto & Navarro, 2009). Si se observan las curvas características de los ítems en el Gráfico AI.11 y Gráfico AI.12, en la mayor parte, las probabilidades de acierto a lo largo del rasgo no son planas y cumplen con su propósito.

No se eliminó ningún ítem más a pesar del pobre ajuste alguno de ellos bajo el modelo TRI. Descartar tantos ítems aumentaría el sesgo de las estimaciones del rasgo y también de los MVA, por tanto, con fines empíricos se mantienen los 162 ítems incluidos en el análisis.

Con un modelo logístico de tres parámetros, la probabilidad de respuesta correcta de un sujeto a un determinado ítem en función de su nivel en el rasgo se calcula a través de la siguiente formula:

$$P(\theta|a,b,c) = c + (1-c)\frac{e^{Da(\theta-b)}}{1+e^{Da(\theta-b)}} \qquad \textbf{Ec. AI.1}$$

Tomando como base esta ecuación es posible llevar a cabo una transformación de los parámetros para situarlos en la misma escala que el rasgo, una media de 250 y desviación típica de 50 puntos. Se lleva a cabo mediante una transformación lineal de la puntuación de una escala X en una la nueva escala Y, de la siguiente manera:

$$\theta_Y = A\theta_X + B \qquad \textbf{Ec. AI.2}$$

A y B son las constantes en esa transformación (la desviación típica y la media respectivamente) y θ_Y y θ_X son las puntuaciones de un sujeto en las dos escalas. Los ítems se transforman de la siguiente forma:

$$a_Y = \frac{a_X}{A}$$
$$b_Y = Ab_X + B \qquad \textbf{Ec. AI.3}$$
$$c_Y = c_X$$

a_X, b_X y c_X son los parámetros de un ítem en la escala X; a_Y, b_Y y c_Y son los de un ítem en la escala Y. Es conveniente mencionar que el parámetro c (azar) es independiente de la transformación. Por tanto, si un modelo TRI ajusta con los datos llevando a cabo una transformación lineal de la escala también ajustará con lados siempre que se transformen los parámetros de los ítems (Kolen & Brennan, 2004). Los resultados se muestran en la Tabla AI.10:

ITEM	NOMBRE	Parámetros TRI					
		a	ET (a)	b)	ET (b)	c	ET (c)
1	MO5AB1	0,013	0,001	104,989	14,006	0,217	0,091
2	MO5AB2	0,009	0,001	96,571	19,396	0,23	0,095
3	MO5AB3	0,015	0,002	276,668	6,531	0,289	0,042
4	MO5AB4	0,006	0,001	289,975	25,442	0,283	0,085
5	MO5AB5	0,015	0,001	269,109	6,052	0,218	0,043
6	MO5AB6	0,006	0,001	367,315	16,615	0,249	0,055
7	MO5AB7	0,013	0,001	262,741	6,346	0,148	0,046
8	MO5AB8	0,013	0,001	194,368	8,735	0,184	0,073
9	MO5AB9	0,014	0,006	477,041	73,734	0,293	0,011
10	MO5AB10	0,014	0,002	316,507	5,496	0,288	0,031
11	MO5AB11	0,017	0,003	268,333	8,3	0,5	0,041
12	MO5AB12	0,02	0,002	315,043	3,088	0,124	0,017
13	MO5AB13	0,022	0,003	326,096	3,594	0,181	0,015
14	MO5AB14	0,017	0,001	208,778	6,826	0,18	0,064
15	MO5AB15	0,012	0,002	306,526	6,651	0,257	0,04
16	MO5AB16	0,004	0,000	-86,623	65,294	0,224	0,093
17	MO5AB17	0,007	0,000	231,802	18,254	0,214	0,082
18	MO5AB18	0,013	0,001	141,309	10,648	0,192	0,081
19	MO5AB19	0,008	0,00	173,772	19,214	0,26	0,098
20	MO5A28	0,015	0,002	254,178	8,052	0,202	0,059
21	MO5A29	0,011	0,001	244,264	13,467	0,24	0,079
22	MO5A30	0,013	0,002	308,297	7,15	0,197	0,042
23	MO5A31	0,022	0,004	283,623	4,771	0,26	0,035
24	MO5A32	0,008	0,001	155,488	16,889	0,199	0,084
25	MO5A33	0,015	0,001	213,577	8,243	0,175	0,07

26	MO5A34	0,017	0,002	204,269	9,967	0,281	0,086
27	MO5A35	0,011	0,001	182,492	15,033	0,267	0,098
28	MO5A36	0,018	0,002	252,862	5,997	0,157	0,051
29	MO5A37	0,009	0,001	221,768	14,643	0,206	0,081
30	MO5B30	0,011	0,001	215,691	15,323	0,262	0,092
31	MO5B31	0,008	0,001	268,545	18,452	0,263	0,084
32	MO5B32	0,009	0,001	301,574	10,233	0,143	0,054
33	MO5B33	0,016	0,002	180,254	11,219	0,267	0,097
34	MO5B34	0,017	0,001	248,644	6,125	0,151	0,05
35	MO5B35	0,014	0,002	288,405	8,906	0,307	0,052
36	MO5B36	0,043	0,015	327,936	4,052	0,174	0,014
37	MO5B37	0,003	0,000	7,176	62,985	0,237	0,097
38	MO5B38	0,01	0,001	175,02	15,457	0,237	0,094
39	MO5A20_J6B19	0,007	0,000	103,664	21,597	0,223	0,093
40	MO5A21_J6B20	0,015	0,001	310,514	3,85	0,133	0,027
41	MO5A22_J6B21	0,014	0,002	336,571	5,019	0,18	0,025
42	MO5A23_J6B22	0,016	0,001	216,289	8,306	0,214	0,07
43	MO5A24_J6B23	0,013	0,001	227,668	8,213	0,165	0,061
44	MO5A25_J6B24	0,017	0,001	219,755	9,334	0,319	0,074
45	MO5A26_J6B25	0,016	0,002	338,649	4,719	0,19	0,02
46	MO5A27_J6B26	0,013	0,001	255,758	5,924	0,125	0,044
47	MO5B20_J6A21	0,01	0,000	164,025	11,73	0,176	0,076
48	MO5B21_J6A22	0,016	0,001	283,516	4,88	0,203	0,036
49	MO5B22_J6A23	0,018	0,002	269,222	6,166	0,342	0,042
50	MO5B23_J6A24	0,014	0,001	279,599	6	0,193	0,043
51	MO5B24_J6A25	0,008	0,000	196,492	16,678	0,209	0,085
52	MO5B25_J6A26	0,014	0,001	216,914	8,25	0,174	0,066
53	MO5B26_J6A27	0,023	0,003	336,642	3,431	0,156	0,013
54	MO5B27_J6A28	0,009	0,001	299,939	11,628	0,233	0,061
55	MO5B28_J6A29	0,01	0,001	262,609	12,045	0,216	0,068
56	MO5B29_J6A30	0,024	0,007	398,019	14,576	0,135	0,008
57	MJ6AB1	0,022	0,001	235,332	4,972	0,187	0,05
58	MJ6AB2	0,01	0,001	295,17	11,168	0,265	0,061
59	MJ6AB3	0,018	0,001	237,737	5,925	0,181	0,054
60	MJ6AB4	0,011	0,001	248,405	9,368	0,16	0,061
61	MJ6AB5	0,015	0,001	240,476	9,479	0,294	0,069
62	MJ6AB6	0,013	0,003	371,377	9,944	0,343	0,026
63	MJ6AB7	0,011	0,001	203,175	10,763	0,188	0,076
64	MJ6AB8	0,012	0,001	210,208	9,622	0,176	0,071
65	MJ6AB9	0,016	0,001	273,188	4,594	0,148	0,038
66	MJ6AB10	0,012	0,001	318,67	6,003	0,209	0,039
67	MJ6AB11	0,015	0,002	374,205	7,332	0,131	0,017
68	MJ6AB12	0,021	0,002	245,623	5,392	0,248	0,049
69	MJ6AB13	0,012	0,001	264,616	9,865	0,25	0,063
70	MJ6AB14	0,016	0,001	296,178	4,439	0,19	0,034
71	MJ6AB15	0,019	0,004	379,773	8,244	0,179	0,014

72	MJ6AB16	0,022	0,003	341,226	3,471	0,184	0,016
73	MJ6AB17	0,025	0,002	292,276	3,583	0,371	0,027
74	MJ6AB18	0,018	0,002	308,643	4,816	0,342	0,032
75	MJ6A29_N6B20	0,015	0,001	254,279	8,204	0,251	0,062
76	MJ6A30_N6B21	0,008	0,001	272,945	14,969	0,236	0,075
77	MJ6A31_N6B22	0,023	0,002	307,348	2,96	0,193	0,026
78	MJ6A32_N6B23	0,016	0,001	287,567	6,62	0,3	0,046
79	MJ6A33_N6B24	0,017	0,001	301,817	3,161	0,098	0,026
80	MJ6A34_N6B25	0,01	0,002	394,752	10,653	0,234	0,032
81	MJ6A35_N6B26	0,014	0,001	295,846	5,182	0,141	0,039
82	MJ6A36_N6B27	0,013	0,001	245,88	8,131	0,175	0,062
83	MJ6A37_N6B28	0,015	0,002	293,198	8,214	0,391	0,048
84	MJ6A38_N6B29	0,016	0,001	241,185	7,432	0,231	0,063
85	MJ6B27_N6A20	0,009	0,001	265,942	11,918	0,183	0,067
86	MJ6B28_N6A21	0,019	0,001	267,25	5,011	0,224	0,045
87	MJ6B29_N6A22	0,021	0,002	278,144	4,779	0,286	0,042
88	MJ6B30_N6A23	0,016	0,001	283,635	4,87	0,141	0,04
89	MJ6B31_N6A24	0,02	0,002	355,972	3,528	0,2	0,019
90	MJ6B32_N6A25	0,014	0,001	267,629	6,048	0,149	0,048
91	MJ6B33_N6A26	0,014	0,001	316,937	4,499	0,152	0,034
92	MJ6B34_N6A27	0,026	0,002	295,834	3,075	0,26	0,028
93	MJ6B35_N6A28	0,014	0,002	324,187	6,272	0,282	0,04
94	MJ6B36_N6A29	0,019	0,001	293,192	4,36	0,242	0,036
95	MN6AB1	0,026	0,002	327,967	2,815	0,221	0,028
96	MN6AB2	0,023	0,004	370,639	4,212	0,325	0,021
97	MN6AB3	0,021	0,001	286,031	4,794	0,205	0,05
98	MN6AB4	0,018	0,002	301,718	6,243	0,285	0,052
99	MN6AB5	0,026	0,002	242,032	5,225	0,209	0,06
100	MN6AB6	0,019	0,001	258,764	7,013	0,277	0,068
101	MN6AB7	0,029	0,006	397,401	6,484	0,145	0,011
102	MN6AB8	0,019	0,002	325,864	5,282	0,348	0,04
103	MN6AB9	0,021	0,002	317,703	3,887	0,207	0,037
104	MN6AB10	0,01	0,001	261,623	13,579	0,267	0,082
105	MN6AB11	0,017	0,004	417,974	13,105	0,177	0,017
106	MN6AB12	0,018	0,002	320,186	5,841	0,356	0,043
107	MN6AB13	0,024	0,002	351,979	2,282	0,124	0,018
108	MN6AB14	0,027	0,002	333,889	1,945	0,122	0,019
109	MN6AB16	0,012	0,002	354,424	8,65	0,331	0,051
110	MN6AB17	0,021	0,002	288,64	5,92	0,321	0,053
111	MN6AB18	0,013	0,001	296,113	5,842	0,128	0,047
112	MN6A31	0,018	0,003	332,346	5,637	0,207	0,049
113	MN6A35	0,017	0,002	252,588	8,344	0,184	0,074
114	MN6A40	0,012	0,002	383,689	8,94	0,147	0,042
115	MN6AB19_J7B1	0,016	0,002	370,5	3,351	0,176	0,024
116	MN6A30_J7B2	0,013	0,001	267,299	9,122	0,206	0,068
117	MN6A32_J7B4	0,015	0,001	280,295	5,694	0,12	0,048

118	MN6A33_J7B5	0,022	0,002	338,627	3,797	0,299	0,033
119	MN6A34_J7B6	0,016	0,001	291,682	5,042	0,13	0,046
120	MN6A37_J7B9	0,02	0,002	344,355	3,672	0,235	0,032
121	MN6A38_J7B10	0,02	0,001	261,324	6,411	0,208	0,064
122	MN6B30_J7A1	0,013	0,001	281,09	7,16	0,164	0,054
123	MN6B31_J7A2	0,022	0,002	225,98	8,793	0,241	0,076
124	MN6B32_J7A3	0,02	0,001	279,277	5,953	0,259	0,055
125	MN6B33_J7A4	0,018	0,004	406,134	7,714	0,215	0,02
126	MN6B34_J7A5	0,025	0,004	384,32	3,882	0,283	0,018
127	MN6B35_J7A6	0,018	0,001	250,444	6,085	0,142	0,054
128	MN6B36_J7A7	0,022	0,002	335,777	2,769	0,141	0,028
129	MN6B37_J7A8	0,024	0,002	331,933	2,655	0,187	0,026
130	MN6B38_J7A9	0,018	0,004	412,414	9,341	0,223	0,02
131	MN6B39_J7A10	0,02	0,001	264,452	6,65	0,285	0,062
132	MJ7AB11	0,025	0,003	384,44	2,695	0,123	0,015
133	MJ7AB13	0,019	0,002	352,453	3,626	0,209	0,033
134	MJ7AB14	0,045	0,016	400,019	4,724	0,308	0,013
135	MJ7AB15	0,027	0,004	392,911	3,315	0,136	0,015
136	MJ7AB16	0,02	0,001	298,641	4,919	0,226	0,048
137	MJ7AB17	0,024	0,002	343,538	3,368	0,291	0,032
138	MJ7AB18	0,01	0,001	221,705	13,228	0,171	0,071
139	MJ7AB19	0,009	0,001	307,945	10,517	0,184	0,061
140	MJ7AB20	0,015	0,006	467,36	29,705	0,278	0,017
141	MJ7AB21	0,027	0,005	386,689	3,513	0,3	0,02
142	MJ7AB22	0,009	0,000	238,759	12,903	0,168	0,069
143	MJ7AB23	0,01	0,001	331,584	8,85	0,176	0,054
144	MJ7AB24	0,013	0,004	465,449	28,57	0,263	0,022
145	MJ7AB25	0,004	0,000	330,69	33,87	0,234	0,087
146	MJ7AB26	0,016	0,002	361,43	4,696	0,213	0,038
147	MJ7AB27	0,024	0,002	284,597	4,974	0,255	0,051
148	MJ7AB28	0,018	0,001	337,015	4,102	0,177	0,037
149	MJ7AB29	0,026	0,009	421,406	12,242	0,249	0,015
150	MJ7AB30	0,021	0,003	377,138	3,375	0,251	0,025
151	MJ7AB31	0,032	0,003	343,011	1,968	0,165	0,023
152	MJ7AB32	0,024	0,002	326,491	2,958	0,164	0,032
153	MJ7AB33	0,025	0,001	279,963	3,991	0,153	0,044
154	MJ7AB34	0,015	0,001	305,09	6,506	0,211	0,051
155	MJ7AB35	0,02	0,002	307,995	5,525	0,3	0,049
156	MJ7AB36	0,017	0,001	323,593	4,712	0,193	0,041
157	MJ7AB37	0,012	0,002	361,173	7,1	0,268	0,047
158	MJ7AB38	0,013	0,001	279,186	8,778	0,208	0,061
159	MJ7AB39	0,016	0,003	408,14	7,771	0,263	0,028
160	MJ7AB40	0,012	0,001	333,752	6,665	0,174	0,049
161	MJ7B3	0,024	0,005	380,043	4,284	0,237	0,028
162	MJ7B7	0,017	0,002	274,326	8,674	0,214	0,07

Tabla AI.10. Parámetros TRI: a (discriminación), b (dificultad) y c (azar) transformados

Anexo II: Marcas de clase de los intervalos entre proporciones acumuladas de la distribución en el rasgo como alternativa a los percentiles para el cálculo de las distancias horizontales

Para analizar las diferencias existentes entre la distribución de puntuaciones de un mismo test en grupos distintos, una opción es utilizar las Curvas de Distribución Acumulada (CDA en adelante) como hace Holland (2002). Estas curvas se estiman a partir de las proporciones acumuladas de esa distribución. Este autor utiliza los percentiles como puntos de referencia para calcular distancias horizontales entre dos CDA distantes pero, en este anexo, se especifica otra medida alternativa a dichos percentiles que tiene una finalidad similar. Es una medida de posición construida a partir de las marcas de clase de los intervalos de puntuaciones que determinan los mencionados percentiles de la distribución. En lugar de tomar como referencia un único punto se emplea el conjunto de puntuaciones de todos los sujetos que se encuentran en ese intervalo.

Anexo II.1 Cálculo de percentiles y distancias horizontales

Para calcular las CDA, en primer lugar se asigna la posición para todos los sujetos de la muestra en función de su valor estimado en el rasgo. Desde el menor valor del rasgo (Ranking (R)=1) hasta el mayor (Ranking (R)=N). Donde N es igual al número de sujetos de la muestra

En segundo lugar se calcula la proporción acumulada (de 0 a 100) de cada uno de los sujetos de la muestra. Por tanto, se parte la distribución en tantas partes como sujetos haya en la muestra

$$\mathrm{P}_\theta = \frac{\mathrm{R}_{n=1}^{N} * 100}{\mathrm{N}} \qquad n = 1 \dots N \qquad \textbf{Ec. AII.1}$$

Donde P es la proporción calculada para un valor del rasgo (θ) que está determinada por la posición que ocupa en el ranking (R) ordenado y el total de sujetos de la muestra (N). Por tanto, los valores de R alcanzan el valor máximo del tamaño muestral.

Por ejemplo, la puntuación del sujeto en el rasgo igual a -0,35 se encuentra en la posición 1500 (R= 1500) del ranking del rasgo, de un total de 5000 estudiantes. Por tanto, ese valor del rasgo tiene una proporción acumulada igual a 30. Esto quiere decir que el 30% de los estudiantes 30% tienen una puntuación igual o inferior, y este valor equivale al percentil 30.

$$P_\theta = \frac{1500 * 100}{5000} \qquad P_\theta = 30 \qquad \textbf{Ec. AII.2}$$

De esta forma, cada puntuación del rasgo de cada sujeto se encuentra ordenada en un continuo en función de estas proporciones acumuladas desde 0% hasta 100%.

Con esta información se pueden elaborar las curvas de distribución acumulada de puntuaciones en los test como define Holland (2002):

$$F(x) = proporción\ de\ sujetos\ con\ puntuaciones\ iguales\ o\ inferiores\ a\ x.$$

El siguiente gráfico muestra dos curvas de distribución acumuladas en dos grupos que han respondido a un mismo test o uno equiparado. Como se puede observar el grupo A2 tiene puntuaciones superiores al grupo A1

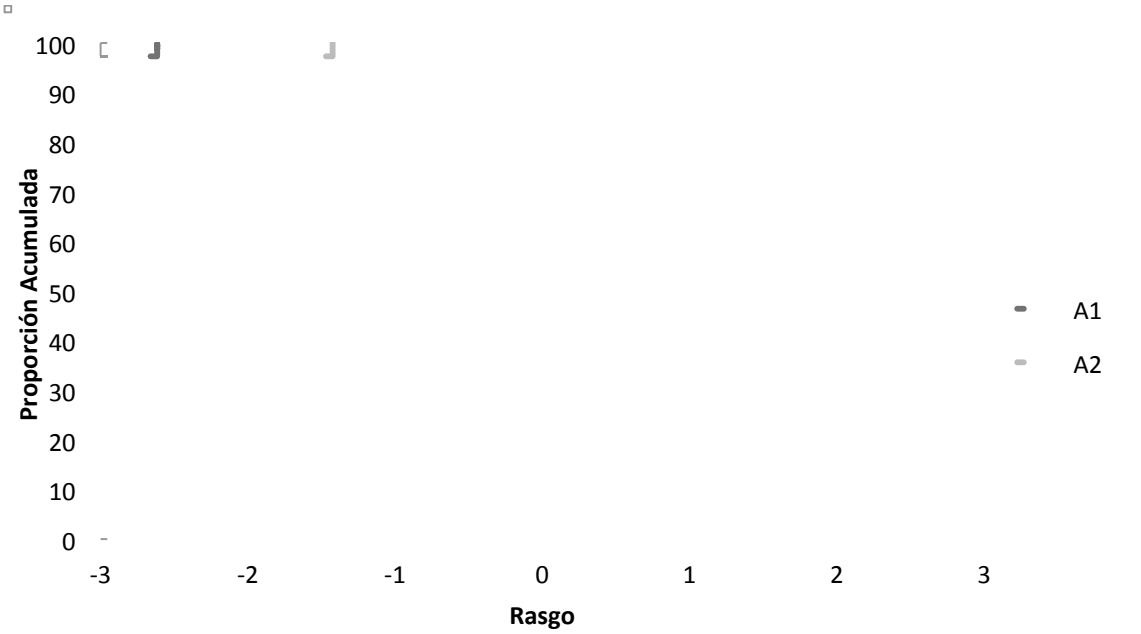


Gráfico AII.1. CDA de las puntuaciones de dos grupos al mismo test de rendimiento

Holland (2002) propone dos formas de analizar las diferencias entre las CDA. En primer lugar, la distancia vertical (denominada por Holland como D(x)) que se define como la diferencia existente entre las proporciones de dos grupos en una misma puntuación (x). Empleando el ejemplo del gráfico anterior, si F(x) es la CDA del grupo A1 y G(x) es la del grupo A2, entonces:

$$D(x) = F(x) - G(x) \qquad \textbf{Ec. AII.3}$$

El autor asume que F(x) es estocásticamente menor que G(x) y, por eso D(x) siempre toma valores positivos. En segundo lugar, las distancias horizontales (denominadas D*(p) por Holland y en este trabajo se han nombrado como delta (Δ)). Para calcularlas es necesario estimar los percentiles de la distribución de puntuaciones, ya que se definen como la diferencia entre las puntuaciones del rasgo que tienen el mismo percentil en ambos grupos.

Los percentiles parten la distribución del rasgo en cien partes iguales. Por ejemplo, el percentil 90 deja por debajo el 90% de la distribución y por encima el otro 10%. Utilizando las proporciones acumuladas se pueden extraer los 99 percentiles (p=1,2,…,99) que indican el valor del rasgo que deja por debajo de sí una determinada proporción estudiantes. En ocasiones, dependiendo del tamaño de la muestra, los percentiles pueden corresponder a una puntuación que no ha

sido estimada para ningún sujeto de la muestra. Por ejemplo, en nuestra distribución se puede encontrar la puntuación del sujeto con una proporción acumulada del 4,99% y del 5,01 pero no exactamente el percentil 5. No obstante, es posible calcular esos valores del rasgo asociados a los percentiles utilizando la información disponible mediante una interpolación lineal.

Cuando los tamaños muestrales son adecuados sí es posible encontrar puntuaciones de los sujetos de la muestra que equivalen a un percentil determinado. Por ejemplo, en una muestra de 5000 sujetos, el percentil 5 es igual a la puntuación del sujeto situado en la posición 250, el 10 en la posición 500 y el 90 en la posición 4500.

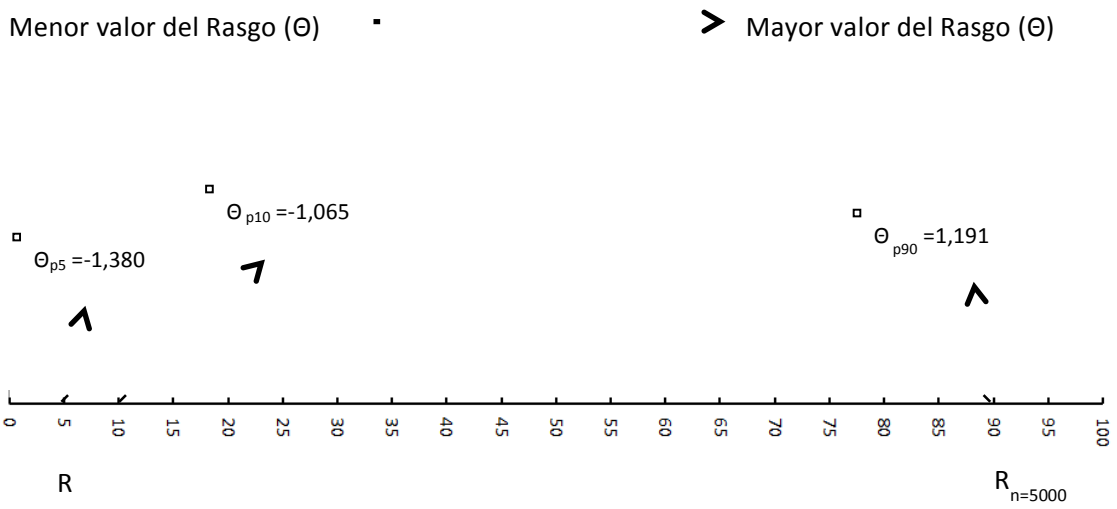


Figura AII.1. Representación de percentiles en una muestra de 5000 sujetos.
Fuente: elaboración propia

Por tanto, para calcular la distancia horizontal (Δ), si seguimos con el ejemplo del Gráfico AII.1:

$$\theta_X(p) = puntuación\ del\ test\ en\ el\ percentil\ p\ para\ el\ grupo\ A1$$

$$\theta_Y(p) = punt\square ación\ del\ test\ en\ el\ percentil\ p\ para\ el\ grupo\ A2$$

Ec. AII.4

$$p = 1 \ldots 99$$

$$\Delta(x) = Y(p) - X(p)$$

En el supuesto de que las distribuciones estén estocásticamente ordenadas, por ejemplo cuando los test que se aplican a los mismos estudiantes durante dos cursos consecutivos se encuentran equiparados verticalmente, se asume que debe haber un crecimiento al incluir en primer lugar la puntuación del curso superior (grupo A2). De esta forma se obtendrán siempre valores positivos en la distancia horizontal.

En cambio, si no podemos cumplir este supuesto, como por ejemplo en los diseños de equiparación horizontal entre dos formas de un mismo test, cada una de ellas aplicada a un grupo de la misma población. Lo esperado es que las puntuaciones en los diferentes percentiles sean similares pero es probable que cada grupo obtenga puntuaciones superiores en diferentes puntos de la distribución. Por tanto, utilizar como primer término de la diferencia la puntuación de un grupo u otro solo varía en la interpretación de los valores positivos o negativos de la distancia horizontal.

Anexo II.2 Cálculo de marcas de clase, distancias horizontales y comparación con percentiles.

El método propuesto consiste en intercambiar las puntuaciones de los 99 percentiles, por la marca de clase de los intervalos de puntuaciones determinados por los percentiles de la CDA. Se construyen 100 intervalos de puntuaciones que agrupa a todos los sujetos que se encuentran entre dos proporciones acumuladas determinadas, se toman los percentiles como referencia ($0<P\leq1$ $1<P\leq2$,…,$99<P\leq100$). Por tanto, se cuenta con 100 submuestras para describir diferentes tramos de la distribución en el rasgo. Una vez agrupadas se calculan las medias agregadas de cada uno de los 100 intervalos para determinar esa Marca de Clase (MC). Mientras que los percentiles muestran puntos concretos del rasgo, estas MC pueden reflejan mejor la distribución de puntuaciones al completo.

$$MC_{i-j} = \frac{\sum_{P>i}^{P\leq j} \theta}{n_{i-j}} \qquad \textbf{Ec. AII.5}$$

Donde P son las proporciones acumuladas que establecen los puntos de corte del intervalo i-j. Por tanto se utilizan todas las puntuaciones (θ) incluidas entre esas dos proporciones. Y n_{i-j} es el tamaño muestral del intervalo.

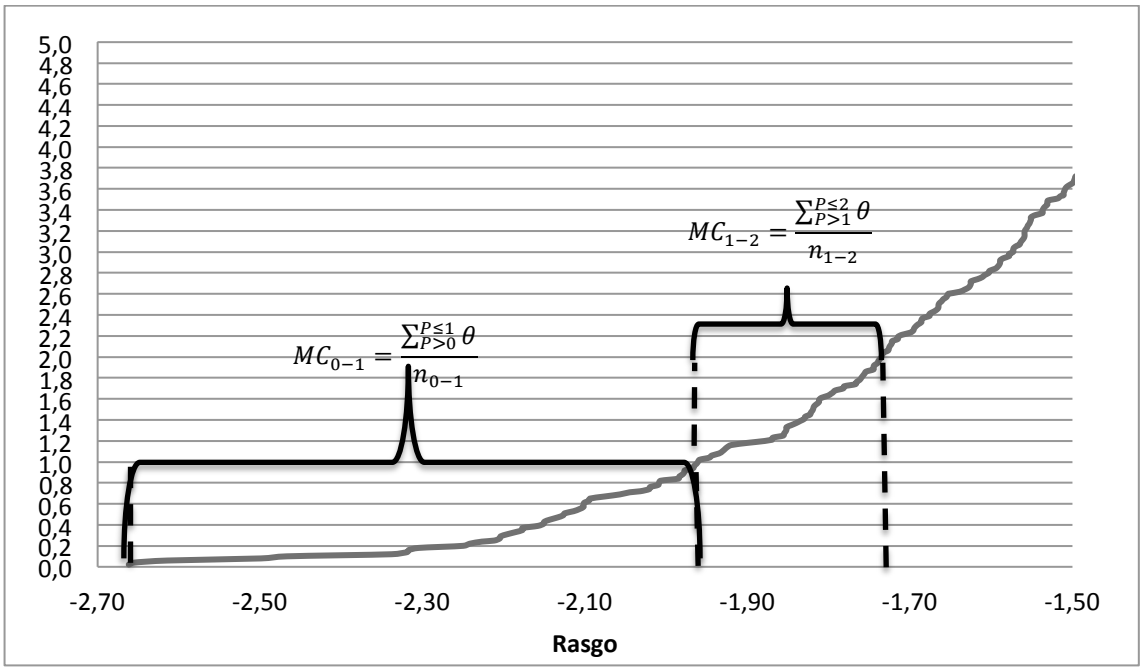


Gráfico AII.2. Representación de las Marcas de clase de los intervalos.

Por tanto, estas Marcas de Clase, cobran un significado ligeramente distinto a los percentiles. Son las puntuaciones que un sujeto obtendría como media si se sitúa en entre las proporciones acumuladas determinadas. Además, pueden acompañarse de desviaciones típicas para comprobar la dispersión dentro de los intervalos. Esta información resulta útil sobre todo en los intervalos situados en los extremos de la distribución.

La utilización de estas Marcas de Clase de los intervalos de puntuaciones construidos a partir de las CDA permite llevar a cabo una caracterización de la distribución más precisa. Mientras que los percentiles identifican puntos concretos de la distribución, estas Marcas de Clase utilizan todas las puntuaciones del rasgo. No obstante, ambos procedimientos conducen a resultados similares.

El procedimiento para calcular las distancias horizontales utilizando las Marcas de Clase es el mismo que ha sido comentado en el apartado de metodología del estudio empírico A. La única diferencia es que se emplean 100 intervalos, es decir, 100 Marcas de Clase en lugar de 99 percentiles. Por tanto, para comparar las puntuaciones de las dos formas de las pruebas producidas por los distintos procedimientos de equiparación horizontal:

$$\Delta_{mc} = \theta_A - \theta_B \qquad mc = 1,2,3 \ldots 100 \qquad \textbf{Ec. AII.6}$$

Y la media:

$$|\overline{\Delta MC}| = \frac{\sum_{mc=1}^{mc=100} |\Delta_{mc}|}{100} \qquad \textbf{Ec. AII.7}$$

Como ya se ha descrito en ese mismo apartado, las distancias horizontales empleadas para comparar los resultados del anclaje vertical varían ligeramente respecto a las de la horizontal y con las Marcas de Clase quedarían de la siguiente manera:

$$\Delta_{mc} = \theta_{superior} - \theta_{inferior} \qquad p = 1,2,3 \ldots 100 \qquad \textbf{Ec. AII.8}$$

Y la distancia media:

$$\overline{\Delta MC} = \frac{\sum_{mc=1}^{mc=100} \Delta_{mc}}{100} \qquad \textbf{Ec. AII.9}$$

Anexo 2.2.1 Sintaxis de SPSS 19 para calcular las marcas de clase

En este apartado se incluye la programación en lenguaje del software SPSS 19 para llevar a cabo el cálculo de las Marcas de Clase desde la hoja de sintaxis de dicho programa.

---

*CÁCLULO DE MARCAS DE CLASE de la variable "ABILITY". La sintaxis crea un archivo separado con los datos sin valores perdidos para llevar a cabo el cálculo. Este archivo puede conservarse quitando * en la parte indicada y poniendo * en la parte final de la sintaxis (también está indicado).
Debe sustituirse la variable ABILITY por el nombre de la variable que se desea utilizar.

DATASET NAME datos.

**Aquí ordena (ascendente) por la variable de interés, en este caso ABILITY

SORT CASES BY ABILITY(A).

*Aquí filtra para descartar los valores perdidos y crea un archivo nuevo con esos datos.

DATASET COPY MC.
DATASET ACTIVATE MC.
FILTER OFF.
USE ALL.
SELECT IF (~ SYSMIS(ABILITY)).

*Aquí asigna un orden a los casos de menor a mayor (desde n=1 a n=N)

COMPUTE RANKING=$CASENUM.

*Aquí calcula el número total de casos de la variable.

AGGREGATE
/OUTFILE=* MODE=ADDVARIABLES
/BREAK=
/N_Total=NU(ABILITY).

*Aquí calcula la proporción acumulada de cada caso.

COMPUTE Proporcion_Acumulada=(100 * (RANKING))/N_Total.

*Aquí recodifica la proporción acumulada para crear los 100 Intervalos.

RECODE Proporcion_Acumulada (0 thru 1=1) (1 thru 2=2) (2 thru 3=3) (3 thru 4=4) (4 thru 5=5) (5 thru 6=6) (6 thru 7=7) (7 thru 8=8) (8 thru 9=9) (9 thru 10=10) (10 thru 11=11) (11 thru 12=12) (12 thru 13=13) (13 thru 14=14) (14 thru 15=15) (15 thru 16=16) (16 thru 17=17) (17 thru 18=18) (18 thru 19=19) (19 thru 20=20) (20 thru 21=21) (21 thru 22=22) (22 thru 23=23) (23 thru 24=24) (24 thru 25=25) (25 thru 26=26) (26 thru 27=27) (27 thru 28=28) (28 thru 29=29) (29 thru 30=30) (30 thru 31=31) (31 thru 32=32) (32 thru 33=33) (33 thru 34=34) (34 thru 35=35) (35 thru 36=36) (36 thru 37=37) (37 thru 38=38) (38 thru 39=39) (39 thru 40=40) (40 thru 41=41) (41 thru 42=42) (42 thru 43=43) (43 thru 44=44) (44 thru 45=45) (45 thru 46=46) (46 thru 47=47) (47 thru 48=48) (48 thru 49=49) (49 thru 50=50) (50 thru 51=51) (51 thru 52=52) (52 thru 53=53) (53 thru 54=54) (54 thru 55=55) (55 thru 56=56) (56 thru 57=57) (57 thru 58=58) (58 thru 59=59) (59 thru 60=60) (60 thru 61=61) (61 thru 62=62) (62 thru 63=63) (63 thru 64=64) (64 thru 65=65) (65 thru 66=66) (66 thru 67=67) (67 thru 68=68) (68 thru 69=69) (69 thru 70=70) (70 thru 71=71) (71 thru 72=72) (72 thru 73=73) (73 thru 74=74) (74 thru 75=75) (75 thru 76=76) (76 thru 77=77) (77 thru 78=78) (78 thru 79=79) (79 thru 80=80) (80 thru 81=81) (81 thru 82=82) (82 thru 83=83) (83 thru 84=84) (84 thru 85=85) (85 thru 86=86) (86 thru 87=87) (87 thru 88=88) (88 thru 89=89) (89 thru 90=90) (90 thru 91=91) (91 thru 92=92) (92 thru 93=93) (93 thru 94=94) (94 thru 95=95) (95 thru 96=96) (96 thru 97=97) (97 thru 98=98) (98 thru 99=99) (99 thru 100=100) INTO MC.

*Aquí crea las Marcas de Clase (media), tamaño del intervalo y desviación típica (esta parte no es necesaria si no necesitamos que esa información esté en el archivo (quitar el * para conservar el archivo de cálculos.

***AGGREGATE**
/OUTFILE=* MODE=ADDVARIABLES
/BREAK=MC
/MC_Media=MEAN(ABILITY)
/MC_sd=SD(ABILITY)
/MC_n=NU(ABILITY).

*Aquí crea el archivo con los resultados (media y sd).

SORT CASES BY MC.

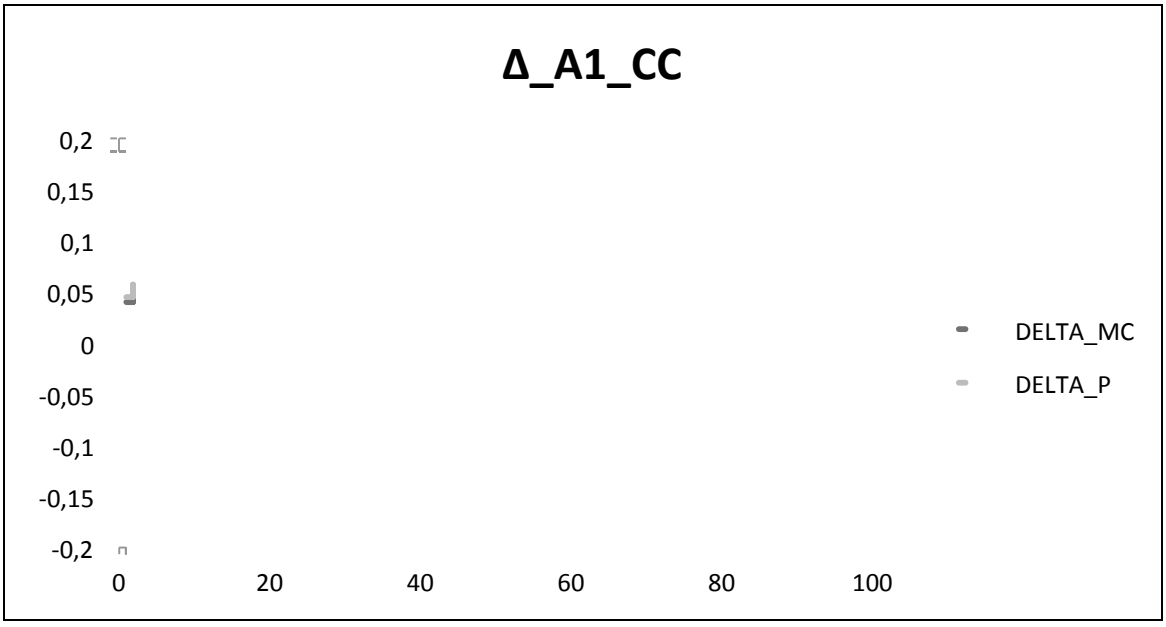
```
SPLIT FILE LAYERED BY MC.
DESCRIPTIVES VARIABLES=ABILITY
  /STATISTICS=MEAN STDDEV.
```

*Aquí cierra el archivo creado para los cálculos (poner * para conservar el archivo de cálculo).

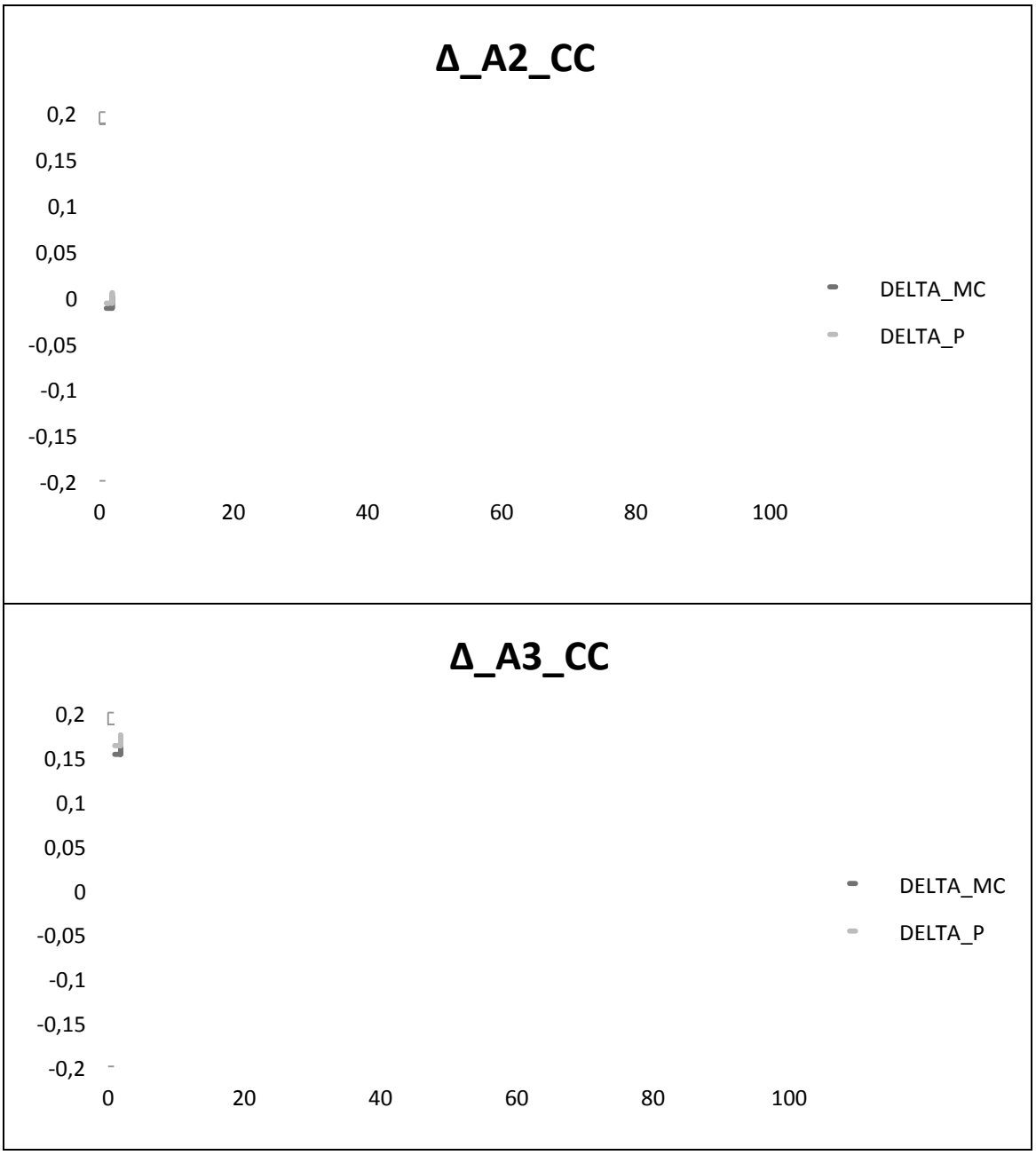
```
DATASET ACTIVATE DATOS.
DATASET CLOSE MC.

EXECUTE.
```

Anexo 2.2.2 Comparación percentiles y marcas de clase en la equiparación horizontal

A continuación, a modo de ejemplo, se presenta de forma gráfica las distancias horizontales calculadas para el problema 1[95] del primer estudio empírico empleando los percentiles de la distribución (DELTA_P o ΔP) y mediante estas marcas de clase de los intervalos (DELTA_MC o ΔMC) de las cuatro aplicaciones.



[95]Conviene recordar que el primer objetivo del estudio empírico compara distintos procedimientos de equiparación horizontal.


Δ_A2_CC
0,2
0,15
0,1
0,05
0
-0,05
-0,1
-0,15
-0,2
0
20
40
60
80
100
DELTA_MC
DELTA_P
Δ_A3_CC
0,2
0,15
0,1
0,05
0
-0,05
-0,1
-0,15
-0,2
0
20
40
60
80
100
DELTA_MC
DELTA_P

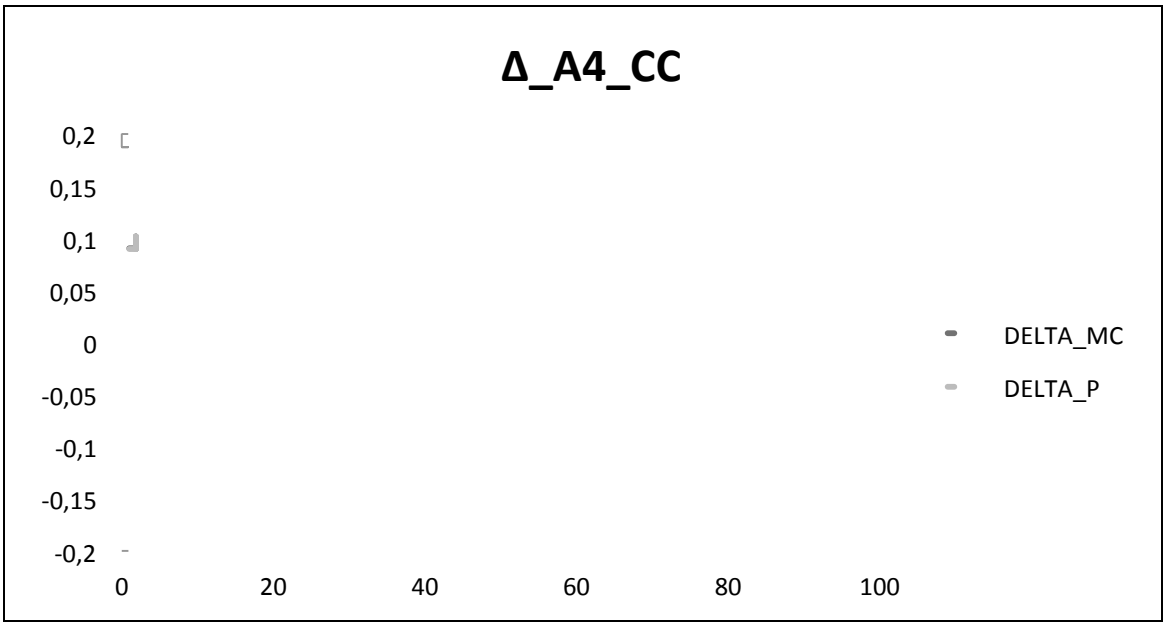


Gráfico AII.3. Comparación de distancias horizontales construidas con percentiles y marcas de clase utilizando los datos producidos por la Calibración Conjunta en la equiparación horizontal.

Las líneas, en las cuatro aplicaciones, indican que ambos estadísticos producen distancias horizontales con una misma tendencia a lo largo de todo el rasgo evaluado. Se han calculado las correlaciones entre las distancias calculadas por ambos procedimientos para cuantificar esa similitud:

	CS	CC	CF	CSMM	CSMS	CSSL	CSH
Aplicación 1	0,964	0,970	0,963	0,967	0,966	0,990	0,986
Aplicación 2	0,920	0,966	0,932	0,989	0,935	0,903	0,907
Aplicación 3	0,870	0,906	0,919	0,895	0,920	0,869	0,930
Aplicación 4	0,950	0,962	0,969	0,952	0,980	0,949	0,956

Tabla AII.1. Correlaciones (Pearson) entre las distancias horizontales calculadas con percentiles y con marcas de clase, en función del método de calibración horizontal y la aplicación.

La mayor parte de los valores correlaciónales encontrados se encuentran por encima del 0,9, lo que indica el alto grado de parecido entre ambas distancias horizontales. Solo en la tercera aplicación estos valores descienden ligeramente en la calibración por separado sin transformación (CS) y con los procedimientos media/media (CSMM) y media/sigma (CSMS), pero sin bajar del 0,85.

Otra prueba más del parecido se encuentra el superponer las curvas construidas utilizando los 99 percentiles y las 100 marcas de clase, como muestran los siguientes gráficos.

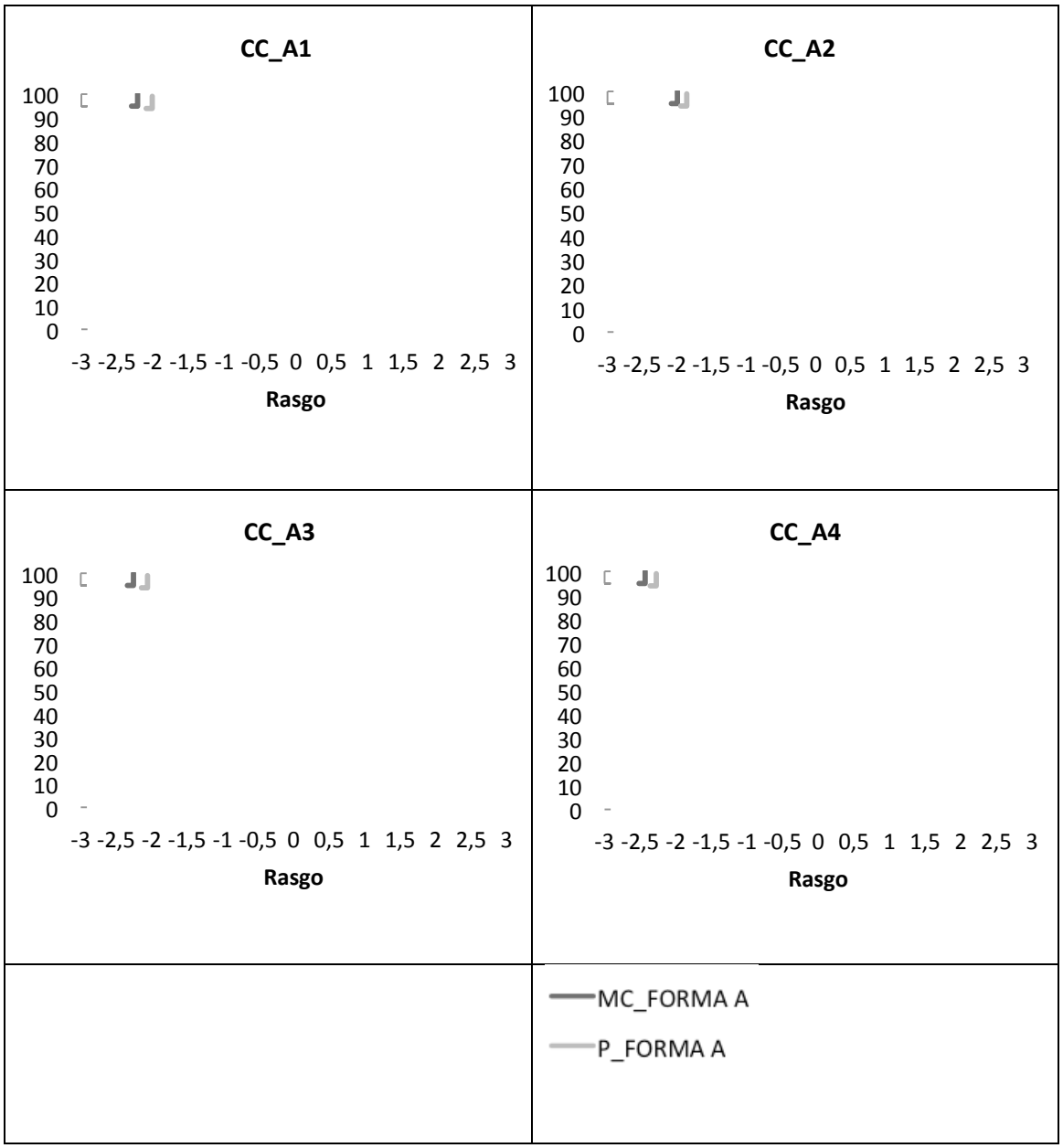


Gráfico AII.4. Curvas de distribución acumuladas construidas empleando solo los 99 percentiles y las 100 marcas de clase, utilizando los datos producidos por la Calibración Conjunta en la equiparación horizontal.

Las curvas, en las cuatro aplicaciones, se encuentran prácticamente superpuestas. Si ampliamos una de las zonas, por ejemplo entre los percentiles 45 y 55 de la tercera aplicación se pueden observar las ligeras diferencias entre los valores determinados por los percentiles y por las marcas de clase

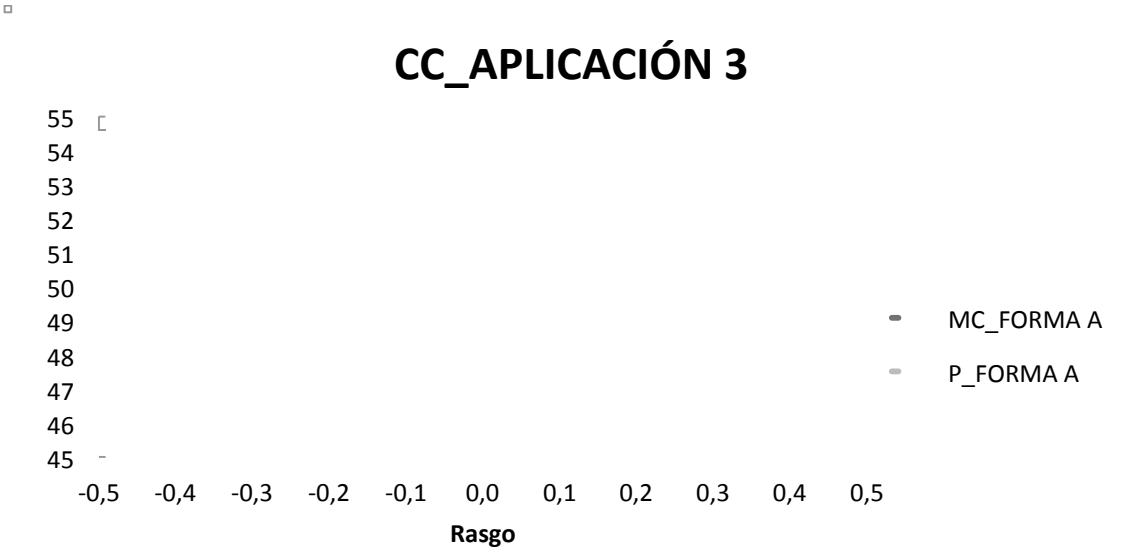


Gráfico AII.5. Ampliación de la sección 45-55 de las curvas de distribución acumulada.

La tabla siguiente muestra as correlaciones entre los percentiles y las marcas de clase y también confirma este parecido:

	A1	A2	A3	A4
CS_A	0,9995	0,9997	0,9995	0,9996
CS_B	0,9995	0,9997	0,9995	0,9996
CC_A	0,9995	0,9996	0,9995	0,9996
CC_B	0,9995	0,9997	0,9995	0,9996
CF_A	0,9995	0,9997	0,9995	0,9996
CF_B	0,9995	0,9997	0,9995	0,9996
CSMM_B	0,9995	0,9997	0,9995	0,9996
CSMS_B	0,9995	0,9997	0,9995	0,9996
CSSL_B	0,9995	0,9997	0,9994	0,9996
CSH_B	0,9995	0,9997	0,9994	0,9996

Tabla AII.2. Correlaciones de Pearson entre percentiles y marcas de clase calculadas en la equiparación horizontal.

Todos los valores de las correlaciones calculadas entre las puntuaciones del rasgo que determinan los percentiles y las marcas de clase se aproximan a valores perfectos.

Con el propósito de profundizar más en el estudio de esas posibles variaciones que pueden producirse en el cálculo de las distancias horizontales por ambos procedimientos, se incluye, a continuación, una tabla con las diferencias:

$$diferencia = |\Delta MC_i - \Delta P_i| \qquad \textbf{Ec. AII.10}$$

Esta diferencia se calcula para comprobar cuánta distancia, en valores de la escala del rasgo, existe entre la distancia horizontales empleando percentiles y las calculadas utilizando las marcas de clase.

	Aplicación 1	**CS**	**CC**	**CF**	**CSMM**	**CSMS**	**CSSL**	**CSH**
Dif	5	0,02	0,01	0,01	0,02	0,02	0,02	0,02
	10	0	0,01	0,01	0,01	0	0	0
	25	0,01	0	0,01	0	0,01	0,01	0,01
	50	0	0	0,01	0	0	0	0
	75	0	0	0	0,01	0,01	0	0
	90	0	0,01	0	0	0	0	0
	95	0,01	0	0,01	0,01	0,01	0,01	0
	Aplicación 2	CS	CC	CF	CSMM	CSMS	CSSL	CSH
Dif	5	0,01	0	0	0,01	0	0,01	0
	10	0	0,02	0,01	0	0,01	0	0,01
	25	0,01	0	0,01	0,02	0,01	0,01	0,02
	50	0	0	0	0	0,01	0	0
	75	0	0	0	0	0,01	0	0
	90	0	0	0	0	0	0,01	0,01
	95	0,01	0,01	0	0	0	0,01	0
	Aplicación 3	**CS**	**CC**	**CF**	**CSMM**	**CSMS**	**CSSL**	**CSH**
Dif	5	0,02	0,03	0,00	0,02	0,02	0,02	0,02
	10	0,01	0,01	0	0,02	0,02	0,01	0,01
	25	0	0	0,01	0	0,01	0	0,01
	50	0	0	0,01	0	0	0,01	0,01
	75	0	0	0	0	0	0	0
	90	0,01	0	0	0	0,01	0,01	0,01
	95	0,01	0,04	0	0,01	0,01	0,01	0,02
	Aplicación 4	**CS**	**CC**	**CF**	**CSMM**	**CSMS**	**CSSL**	**CSH**
Dif	5	0,01	0,02	0,07	0,02	0,01	0,02	0,01
	10	0,02	0,01	0	0,02	0,01	0,02	0,02
	25	0,01	0,01	0,01	0,01	0,01	0,01	0,01
	50	0,01	0,01	0	0,01	0	0,01	0
	75	0	0,01	0,01	0	0	0	0
	90	0	0,01	0,02	0	0,01	0,01	0
	95	0,01	0	0,04	0,02	0,02	0,02	0,01

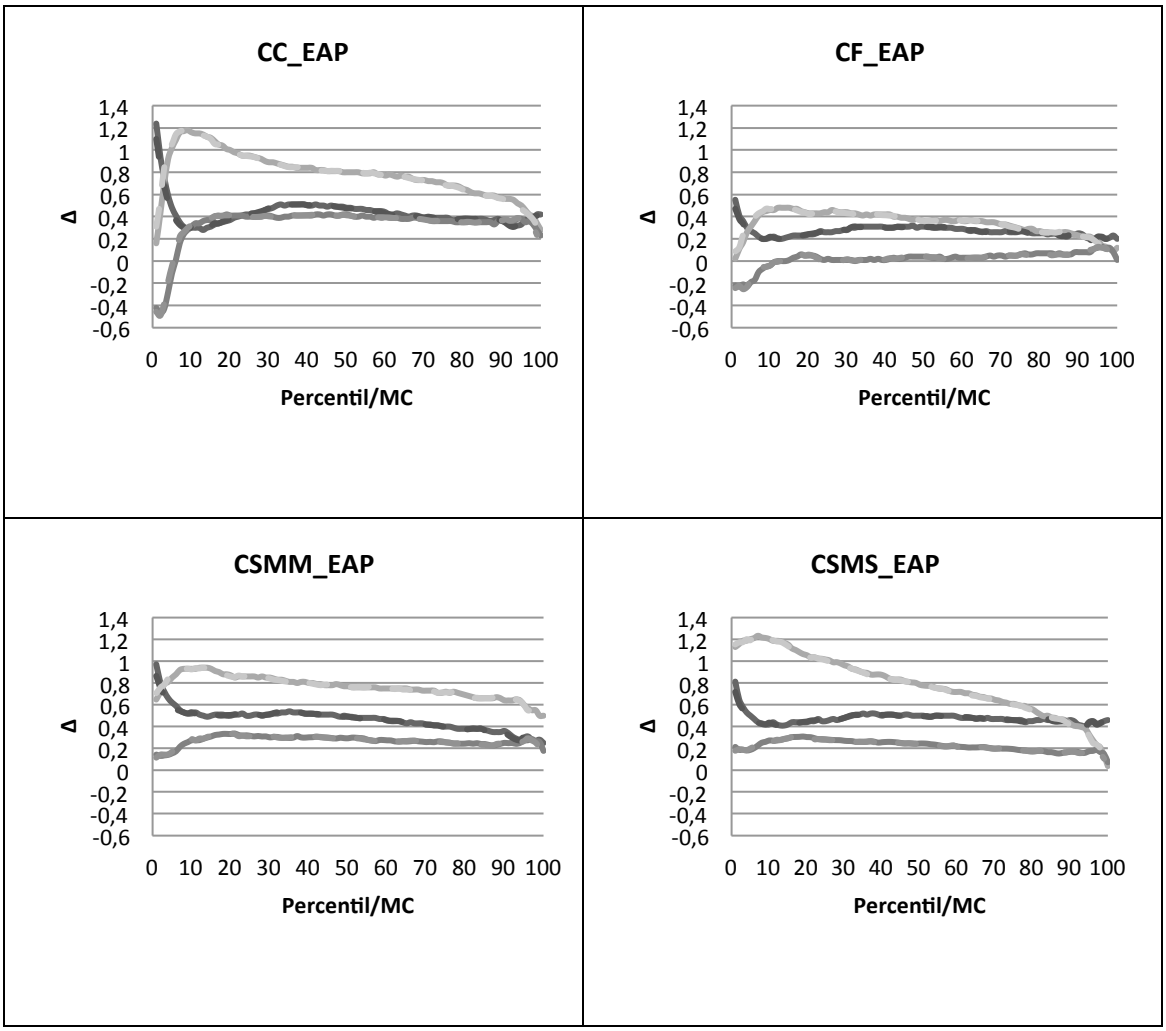
Tabla AII.3. Diferencias entre Distancias Horizontales calculadas con Marcas de Clase y con Percentiles en los 7 puntos de la distribución (calibración horizontal).

Estas diferencias calculadas entre las distancias horizontales (Δ) de ambas metodologías indican el gran parecido. Únicamente en la cuarta aplicación, con la metodología de calibración fija, las Δ calculadas se diferencian por encima de 0,05.

Todas las pruebas realizadas señalan que las Marcas de Clase de los intervalos de puntuaciones delimitados por los percentiles pueden ser una alternativa para el estudio de las distribuciones de variables de intervalo, así como para emplearlas en el cálculo de las distancias horizontales propuestas por Holland (2002).

Anexo 2.2.3 Comparación percentiles y marcas de clase en la equiparación vertical

A modo de ejemplo se presentan las distancias horizontales, calculadas a través de los dos procesos, de la metodología bayesiana de estimación del rasgo EAP (Gráfico AII.6).



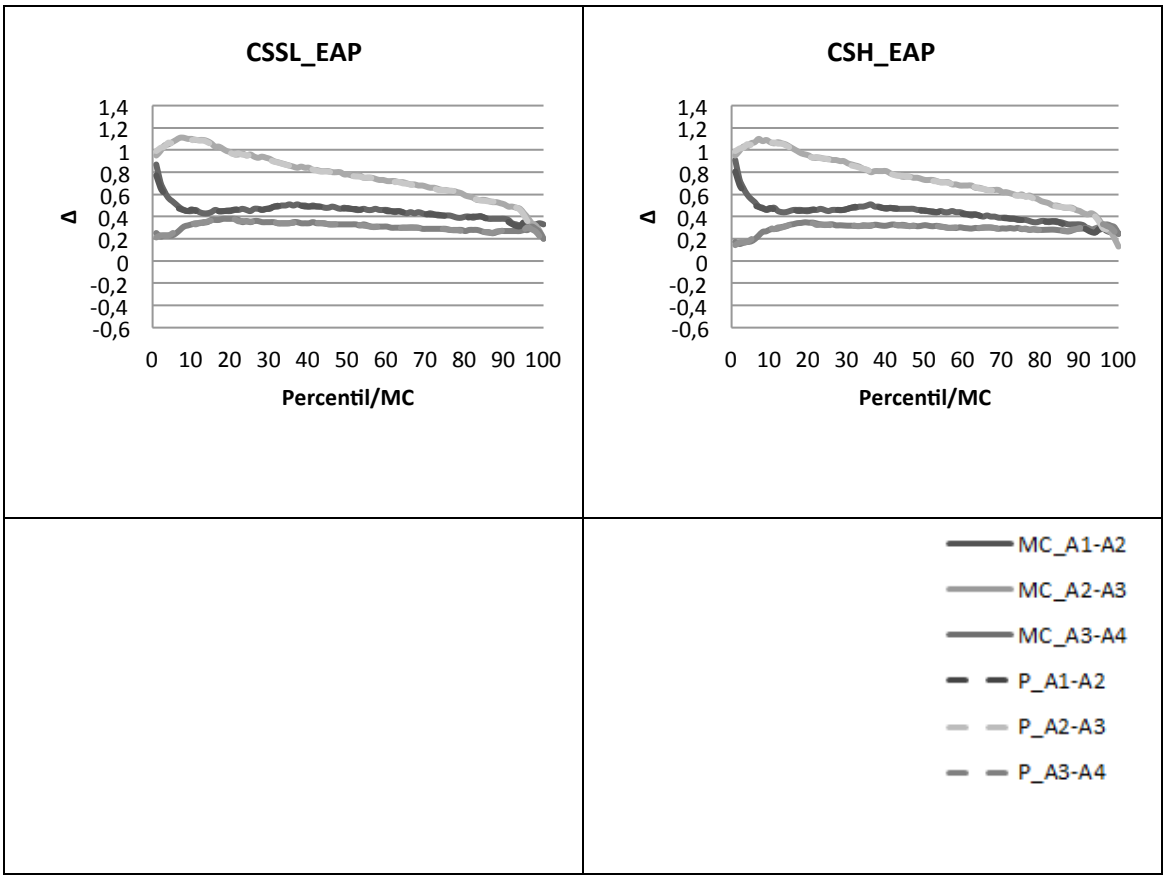


Gráfico AII.6. Comparación de las Distancias Horizontales calculadas con percentiles y Marcas de Clase por metodología de calibración. Método de estimación EAP

Como se puede ver en los gráficos anteriores, las distancias son prácticamente iguales. Además, se han calculado correlaciones de Pearson con el propósito de comprobar la relación existente entre las distancias horizontales calculadas con ambas metodologías:

		CC	CF	CSMM	CSMS	CSSL	CSH
EAP	A1-A2	0,986	0,978	0,993	0,977	0,988	0,991
	A2-A3	0,993	0,987	0,995	1,000	0,999	0,999
	A3-A4	0,996	0,987	0,978	0,988	0,973	0,978
MAP	A1-A2	0,985	0,976	0,996	0,986	0,994	0,995
	A2-A3	0,998	0,987	0,994	1,000	0,999	0,999
	A3-A4	0,996	0,985	0,947	0,974	0,944	0,949
MVL	A1-A2	0,925	0,947	0,964	0,943	0,939	0,958
	A2-A3	0,993	0,955	0,989	0,998	0,997	0,998
	A3-A4	0,986	0,946	0,961	0,971	0,951	0,946

Tabla AII.4. Correlaciones (Pearson) entre las distancias horizontales calculadas con percentiles y con marcas de clase, en función del método de calibración vertical.

Las diferencias entre las distancias horizontales calculadas por ambos procedimientos, de la misma forma que se hizo en la equiparación horizontal (diferencia $= |\Delta MC_i - \Delta P_i|$), se presentan en la siguiente tabla

		A1-A2	**CC**	**CF**	**CSMM**	**CSMS**	**CSSL**	**CSH**
EAP	Dif	5	0,04	0,01	0,01	0,01	0,01	0,00
		10	0,00	0,01	0,01	0,02	0,01	0,01
		25	0,01	0,00	0,01	0,01	0,01	0,01
		50	0,01	0,00	0,00	0,00	0,01	0,01
		75	0,00	0,00	0,00	0,00	0,00	0,00
		90	0,00	0,00	0,01	0,00	0,00	0,00
		95	0,03	0,02	0,01	0,01	0,01	0,01
		A2-A3	CC	CF	CSMM	CSMS	CSSL	CSH
	Dif	5	0,03	0,02	0,00	0,01	0,01	0,01
		10	0,01	0,01	0,01	0,01	0,01	0,01
		25	0,01	0,01	0,00	0,01	0,01	0,00
		50	0,00	0,00	0,01	0,01	0,01	0,00
		75	0,01	0,00	0,01	0,01	0,00	0,01
		90	0,01	0,01	0,01	0,01	0,02	0,02
		95	0,03	0,03	0,00	0,01	0,02	0,02
		A3-A4	CC	CF	CSMM	CSMS	CSSL	CSH
	Dif	5	0,07	0,03	0,01	0,01	0,02	0,02
		10	0,01	0,00	0,01	0,01	0,01	0,01
		25	0,00	0,01	0,00	0,00	0,01	0,00
		50	0,01	0,00	0,00	0,00	0,00	0,00
		75	0,00	0,00	0,00	0,00	0,00	0,00
		90	0,00	0,00	0,01	0,00	0,01	0,01
		95	0,01	0,01	0,01	0,00	0,00	0,00
		A1-A2	CC	CF	CSMM	CSMS	CSSL	CSH
MAP	Dif	5	0,00	0,01	0,01	0,00	0,01	0,01
		10	0,00	0,00	0,00	0,00	0,00	0,00
		25	0,00	0,01	0,00	0,00	0,01	0,01
		50	0,00	0,01	0,00	0,00	0,00	0,00
		75	0,00	0,01	0,00	0,00	0,00	0,00
		90	0,00	0,01	0,01	0,00	0,00	0,01
		95	0,02	0,00	0,01	0,00	0,00	0,00
		A2-A3						
	Dif	5	0,00	0,01	0,00	0,01	0,00	0,00
		10	0,01	0,00	0,00	0,00	0,00	0,01
		25	0,01	0,00	0,00	0,00	0,01	0,00
		50	0,00	0,00	0,01	0,01	0,00	0,01
		75	0,00	0,01	0,00	0,00	0,01	0,00
		90	0,00	0,00	0,01	0,01	0,02	0,01
		95	0,01	0,01	0,01	0,01	0,01	0,02

		A3-A4						
	Dif	5	0,03	0,02	0,01	0,01	0,01	0,01
		10	0,01	0,00	0,00	0,00	0,00	0,01
		25	0,00	0,00	0,00	0,01	0,01	0,00
		50	0,01	0,01	0,01	0,00	0,00	0,01
		75	0,01	0,00	0,00	0,00	0,00	0,00
		90	0,01	0,01	0,01	0,01	0,01	0,01
		95	0,01	0,01	0,00	0,00	0,00	0,01
		A1-A2	CC	CF	CSMM	CSMS	CSSL	CSH
	Dif	5	0,06	0,07	0,05	0,06	0,06	0,05
		10	0,01	0,01	0,01	0,01	0,02	0,01
		25	0,00	0,00	0,00	0,01	0,00	0,00
		50	0,00	0,00	0,00	0,00	0,00	0,00
		75	0,00	0,01	0,01	0,00	0,01	0,01
		90	0,00	0,00	0,00	0,01	0,01	0,01
		95	0,03	0,01	0,02	0,03	0,01	0,02
		A2-A3						
MVL	Dif	5	0,10	0,10	0,09	0,13	0,10	0,11
		10	0,03	0,02	0,02	0,03	0,02	0,03
		25	0,00	0,00	0,00	0,00	0,01	0,01
		50	0,00	0,00	0,00	0,00	0,01	0,01
		75	0,01	0,00	0,01	0,00	0,00	0,00
		90	0,01	0,01	0,01	0,01	0,00	0,01
		95	0,02	0,02	0,03	0,03	0,03	0,03
		A3-A4						
	Dif	5	0,03	0,03	0,00	0,00	0,00	0,00
		10	0,01	0,01	0,01	0,00	0,00	0,00
		25	0,00	0,00	0,01	0,00	0,01	0,00
		50	0,00	0,01	0,00	0,00	0,00	0,00
		75	0,00	0,01	0,00	0,00	0,00	0,00
		90	0,00	0,00	0,00	0,01	0,01	0,00
		95	0,01	0,01	0,01	0,00	0,00	0,00

Tabla AII.5. Diferencias entre Distancias Horizontales calculadas Marcas de Clase y con Percentiles en los 7 puntos de la distribución (anclaje vertical).

Aunque las diferencias son mínimas, la mayor separación se produce en la parte baja de la distribución (percentil o marca de clase 5) con una diferencia máxima de 0,13 puntos en el rasgo en el procedimiento de estimación de máxima verosimilitud (VML) entre la segunda y tercera aplicación.

El siguiente gráfico muestra los valores de los 99 percentiles y las 100 marcas de clase para las cuatro aplicaciones empleando la metodología de calibración por separado y la calificación EAP.

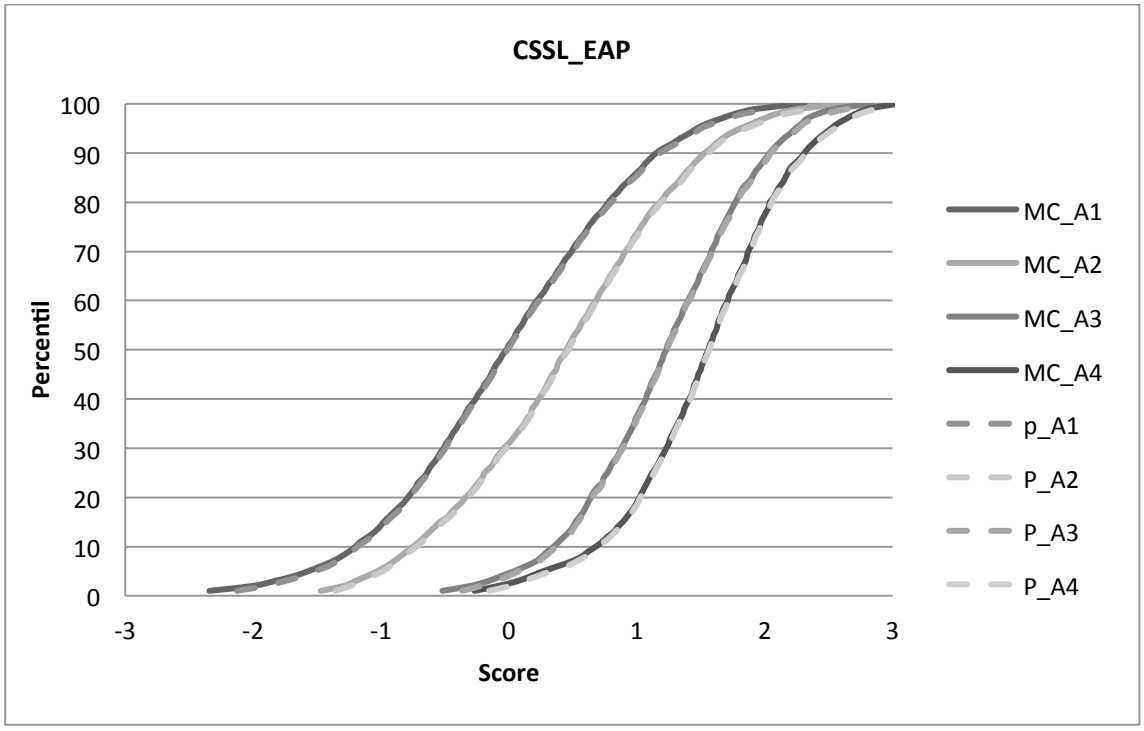


Gráfico AII.7. Curvas de distribución acumuladas construidas empleando solo los 99 percentiles y las 100 marcas de clase, utilizando los datos producidos por la Calibración Separada (Stocking y Lord) en la equiparación horizontal.

Los puntos parecen casi coincidentes. Si se amplía la sección entre 30 y 40 (percentil o marca de clase) no se aprecian diferencias:

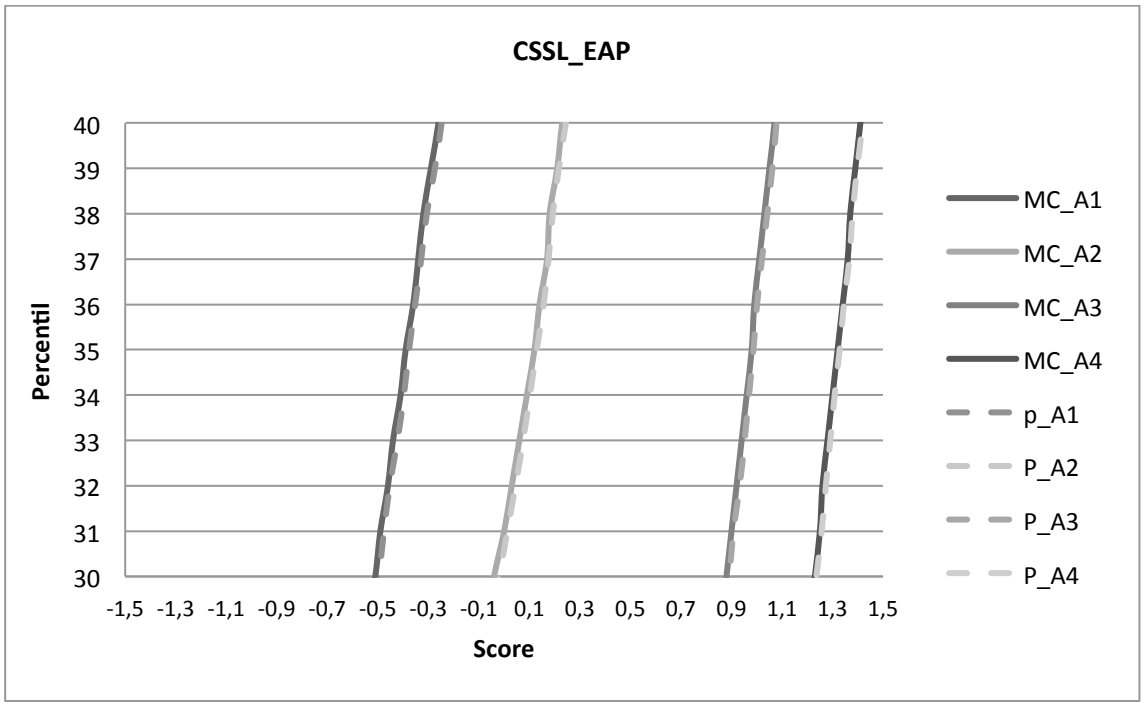


Gráfico AII.8. Ampliación del tramo entre el percentil 30 y 40.

Si ampliamos aún más, por ejemplo, la curva de la aplicación número dos (A2) vemos que las diferencias siguen la misma tendencia que las encontradas en la comparación de procesos durante la equiparación horizontal.

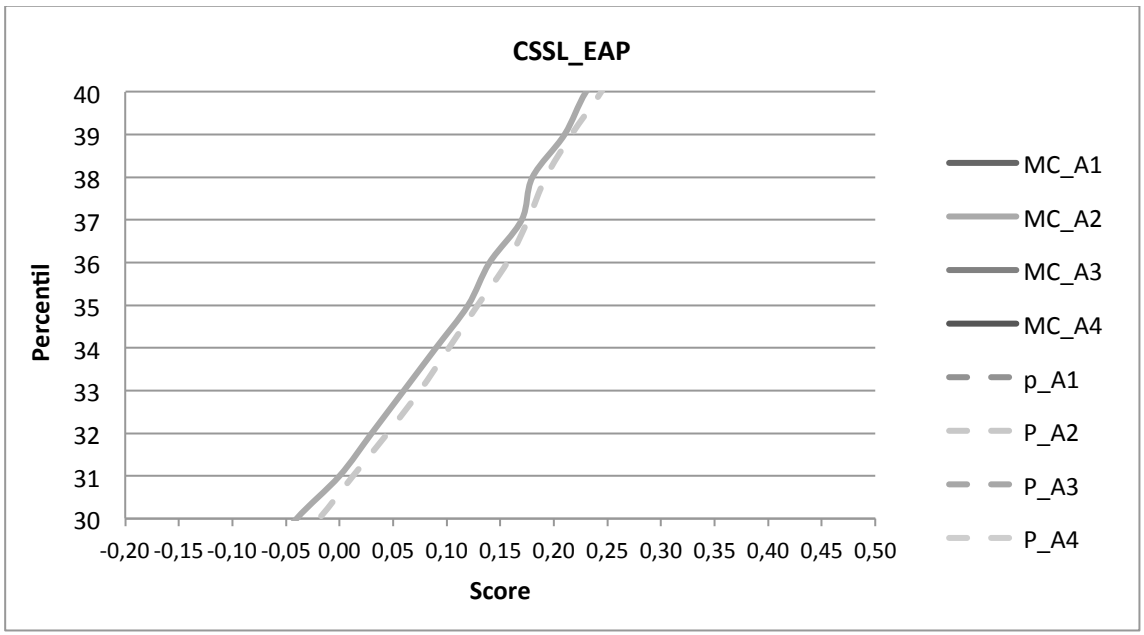


Gráfico AII.9. Ampliación del tramo entre el percentil 30 y 40 y sección del rasgo entre -0,2 y 0,5 (Aplicación 2).

Las correlaciones entre los percentiles y las marcas de clase son casi perfectas, como se puede ver en la siguiente tabla.

		CC	CF	CSMM	CSMS	CSSL	CSH
EAP	A1	0,999	1,000	1,000	1,000	1,000	1,000
	A2	1,000	1,000	1,000	1,000	1,000	1,000
	A3	0,999	0,999	1,000	1,000	0,999	1,000
	A4	0,999	1,000	1,000	1,000	1,000	1,000
MAP	A1	1,000	1,000	1,000	1,000	1,000	1,000
	A2	1,000	1,000	1,000	1,000	1,000	1,000
	A3	1,000	0,999	1,000	0,999	0,999	1,000
	A4	1,000	1,000	1,000	1,000	1,000	1,000
MLV	A1	0,999	0,999	0,999	0,999	0,999	0,999
	A2	0,999	0,999	0,999	0,999	0,999	0,999
	A3	0,998	0,999	0,999	0,999	0,999	0,999
	A4	0,999	0,999	0,999	0,999	0,999	0,999

Tabla AII.6. Correlaciones de Pearson entre percentiles y marcas de clase calculadas en la equiparación vertical.

A la luz de los resultados mostrados, las marcas de clase de los intervalos de puntuaciones pueden ser una alternativa para caracterizar la distribución. Y, por supuesto, su utilización en el cálculo de las distancias horizontales propuestas por Holland.

Anexo II.3 Resultados del estudio empírico 1 empleando las marcas de clase:

Las tablas y gráficos que se incluyen a continuación hacen referencia a esas distancias horizontales calculadas con la finalidad de comparar los resultados en el estudio empírico 1. Recordemos que este estudio analiza procedimientos distintos para la elaboración de una escala vertical del rendimiento académico como producto final.

Anexo II.3.1 Problema 1: Comparación de procedimientos de equiparación horizontal

	Aplicación 1	**CS**	**CC**	**CF**	**CSMM**	**CSMS**	**CSSL**	**CSH**
Marcas de Clase	5	-0,06	-0,08	-0,08	-0,03	-0,03	**-0,18**	**-0,14**
	10	0	-0,02	0	0,03	0,02	**-0,1**	-0,07
	25	-0,02	-0,04	-0,03	-0,01	-0,02	-0,09	-0,07
	50	0,01	-0,01	0,01	0	0	-0,03	-0,01
	75	0,02	0,02	0,02	-0,01	-0,01	0,01	0,02
	90	0,03	0,02	0,03	-0,03	-0,03	0,05	0,05
	95	0,05	0,05	0,04	-0,01	-0,01	0,08	0,08
	Aplicación 2	**CS**	**CC**	**CF**	**CSMM**	**CSMS**	**CSSL**	**CSH**
Marcas de Clase	5	-0,01	-0,04	-0,02	**0,1**	**-0,1**	**-0,12**	-0,08
	10	0,05	-0,01	0,04	**0,14**	-0,04	-0,06	-0,03
	25	0,02	-0,02	0,02	0,07	-0,05	-0,08	-0,05
	50	-0,03	**-0,14**	-0,04	-0,03	-0,08	**-0,12**	**-0,1**
	75	0,03	-0,04	0,01	-0,01	-0,01	-0,06	-0,04
	90	-0,01	-0,05	-0,01	**-0,1**	-0,02	-0,08	-0,07
	95	-0,01	-0,04	-0,01	**-0,11**	-0,01	-0,08	-0,07
	Aplicación 3	**CS**	**CC**	**CF**	**CSMM**	**CSMS**	**CSSL**	**CSH**
Marcas de Clase	5	0,04	0,08	0,05	0,05	0,07	**0,1**	**0,13**
	10	-0,02	0,02	0	-0,01	0	0,04	0,06
	25	0	0,05	0,01	0	0,01	0,06	0,07
	50	-0,01	0,06	0,02	-0,02	-0,02	0,04	0,03
	75	-0,02	0,02	-0,01	-0,04	-0,04	0,04	0,01
	90	0,01	0,03	0	-0,03	-0,03	0,06	0,02
	95	0,05	0,03	0,03	0,01	0,01	**0,1**	0,05
	Aplicación 4	**CS**	**CC**	**CF**	**CSMM**	**CSMS**	**CSSL**	**CSH**
Marcas de Clase	5	-0,03	0	-0,06	-0,02	**0,1**	0,01	0
	10	0,08	0,09	0,08	**0,1**	0,2	**0,13**	**0,11**
	25	0	0,01	0,09	0,02	0,07	0,04	0,03
	50	-0,02	0,01	0,07	0	0,02	0,02	0,01
	75	-0,04	-0,01	0,03	-0,01	-0,02	0	0
	90	0,05	0,08	**0,13**	0,08	0,05	0,1	0,09
	95	0,05	0,09	0,09	0,08	0,04	0,1	0,09

Tabla AII.7. Distancias Horizontales en siete puntos de la distribución (Marcas de Clase) en función de la metodología de calibración horizontal empleada, en cada una de las aplicaciones.

		CS	**CC**	**CF**	**CSMM**	**CSMS**	**CSH**	**CSSL**
$\lvert\overline{\Delta MC}\rvert$	**Aplicación 1**	0,027	0,035	0,029	0,027	0,027	0,060	0,051
	Aplicación 2	0,021	0,062	0,023	0,061	0,051	0,093	0,071
	Aplicación 3	0,018	0,043	0,017	0,025	0,025	0,056	0,050
	Aplicación 4	0,034	0,036	0,078	0,036	0,058	0,047	0,042

Tabla AII.8. Distancias Horizontales Medias (Marcas de Clase) en función de la metodología de calibración horizontal y la aplicación.

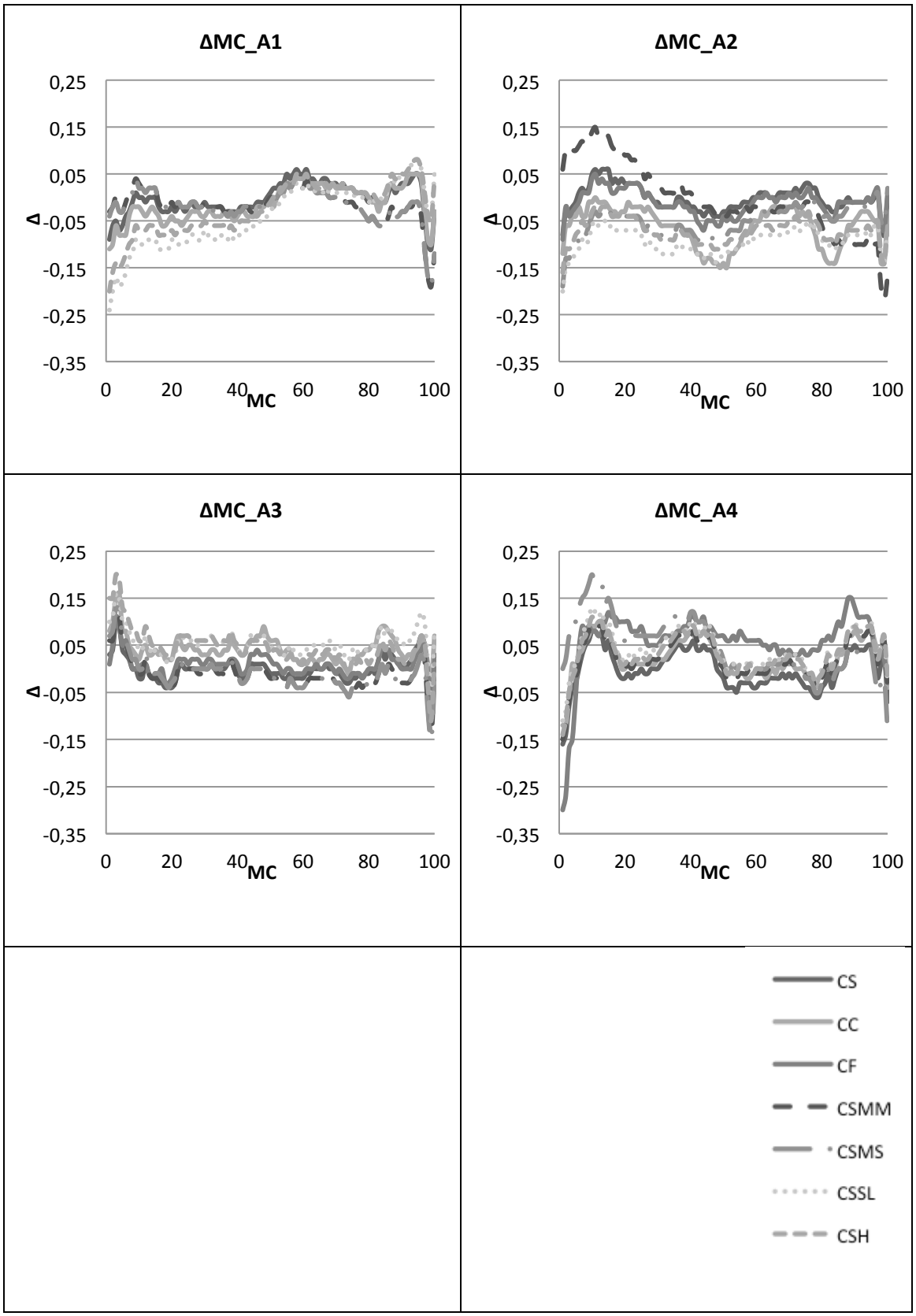


Gráfico AII.10. Distancias horizontales en las 100 Marcas de Clase, en función de la metodología de calibración horizontal, en cada una de las aplicaciones por separado.

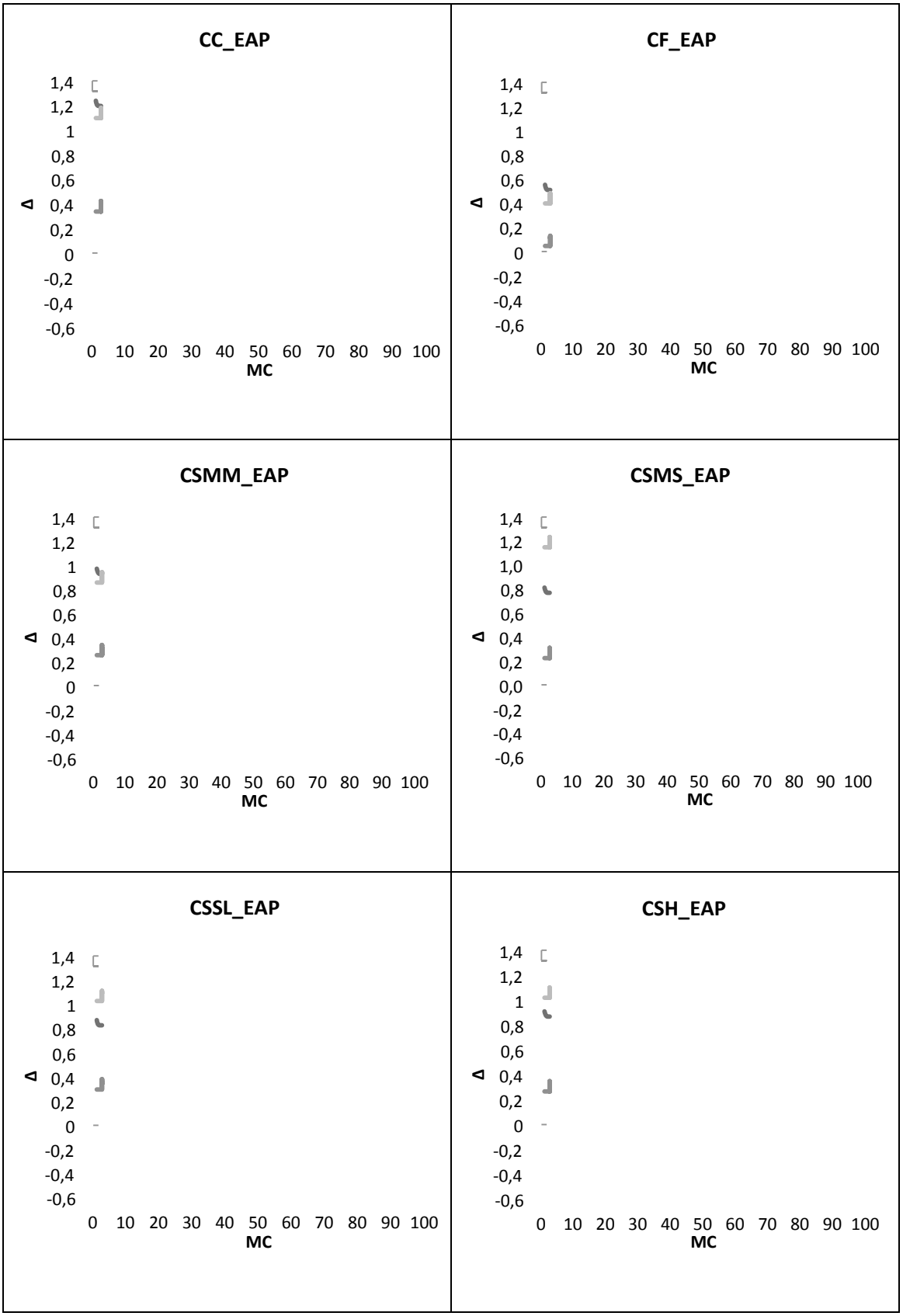
Anexo II.3.2 Problema 2: Comparación de procedimientos para el anclaje vertical

A continuación se incluyen los resultados referentes a las distancias horizontales pero calculadas utilizando las marcas de clase de los intervalos, en lugar de los percentiles.

		A1-A2	CC	CF	CSMM	CSMS	CSSL	CSH
EAP	Marca de Clase	5	0,48	0,27	0,50	0,62	0,54	0,56
		10	0,30	0,21	0,42	0,52	0,45	0,47
		25	0,42	0,26	0,45	0,50	0,45	0,45
		50	0,48	0,31	0,50	0,49	0,47	0,45
		75	0,39	0,27	0,46	0,40	0,41	0,37
		90	0,36	0,25	0,46	0,36	0,38	0,33
		95	0,33	0,20	0,43	0,30	0,34	0,28
		A2-A3	**CC**	**CF**	**CSMM**	**CSMS**	**CSSL**	**CSH**
	Marca de Clase	5	1,02	0,31	0,85	1,20	1,07	1,06
		10	1,16	0,46	0,93	1,20	1,10	1,07
		25	0,95	0,44	0,86	1,02	0,96	0,92
		50	0,80	0,36	0,77	0,78	0,78	0,73
		75	0,70	0,29	0,72	0,61	0,64	0,60
		90	0,57	0,23	0,64	0,43	0,51	0,45
		95	0,50	0,19	0,60	0,33	0,43	0,37
		A3-A4	**CC**	**CF**	**CSMM**	**CSMS**	**CSSL**	**CSH**
	Marca de Clase	5	-0,11	-0,19	0,16	0,19	0,23	0,18
		10	0,31	-0,04	0,28	0,27	0,33	0,29
		25	0,40	0,02	0,32	0,28	0,35	0,33
		50	0,42	0,04	0,29	0,25	0,33	0,32
		75	0,36	0,06	0,26	0,19	0,29	0,29
		90	0,37	0,08	0,25	0,17	0,27	0,29
		95	0,38	0,12	0,27	0,18	0,28	0,31
		A1-A2	**CC**	**CF**	**CSMM**	**CSMS**	**CSSL**	**CSH**
MAP	Marca de Clase	5	0,52	0,25	0,64	0,53	0,57	0,60
		10	0,46	0,21	0,56	0,48	0,50	0,52
		25	0,44	0,22	0,50	0,46	0,46	0,46
		50	0,45	0,27	0,48	0,49	0,47	0,45
		75	0,40	0,23	0,40	0,46	0,41	0,37
		90	0,36	0,22	0,34	0,44	0,36	0,31
		95	0,33	0,18	0,29	0,42	0,33	0,26

	A2-A3	CC	CF	CSMM	CSMS	CSSL	CSH
Marca de Clase	5	0,97	0,24	0,82	1,14	1,01	1,00
	10	0,99	0,36	0,86	1,11	1,01	1,00
	25	0,91	0,37	0,82	0,96	0,91	0,87
	50	0,79	0,33	0,76	0,77	0,75	0,71
	75	0,68	0,28	0,70	0,59	0,62	0,56
	90	0,57	0,20	0,64	0,42	0,51	0,42
	95	0,50	0,18	0,61	0,33	0,43	0,35
	A3-A4	**CC**	**CF**	**CSMM**	**CSMS**	**CSSL**	**CSH**
Marca de Clase	5	0,13	-0,14	0,18	0,20	0,25	0,25
	10	0,28	-0,06	0,27	0,26	0,32	0,31
	25	0,36	-0,02	0,29	0,27	0,33	0,34
	50	0,40	0,00	0,28	0,24	0,33	0,32
	75	0,38	0,03	0,26	0,19	0,30	0,30
	90	0,40	0,09	0,26	0,18	0,29	0,30
	95	0,41	0,11	0,28	0,18	0,31	0,31
	A1-A2	**CC**	**CF**	**CSMM**	**CSMS**	**CSSL**	**CSH**
Marca de Clase	5	0,25	0,11	0,49	0,22	0,38	0,43
	10	0,42	0,26	0,58	0,41	0,50	0,53
	25	0,47	0,30	0,54	0,47	0,49	0,49
	50	0,45	0,30	0,48	0,49	0,46	0,44
	75	0,37	0,24	0,37	0,45	0,38	0,34
	90	0,31	0,19	0,28	0,42	0,32	0,26
	95	0,30	0,16	0,24	0,43	0,29	0,22
MVL	**A2-A3**	**CC**	**CF**	**CSMM**	**CSMS**	**CSSL**	**CSH**
Marca de Clase	5	1,54	0,78	1,31	1,79	1,60	1,62
	10	1,19	0,58	1,02	1,34	1,21	1,22
	25	0,95	0,45	0,85	1,01	0,95	0,94
	50	0,79	0,36	0,75	0,76	0,75	0,72
	75	0,67	0,29	0,7	0,57	0,61	0,55
	90	0,55	0,23	0,65	0,39	0,46	0,40
	95	0,48	0,21	0,62	0,29	0,39	0,31
	A3-A4	**CC**	**CF**	**CSMM**	**CSMS**	**CSSL**	**CSH**
Marca de Clase	5	0,06	-0,21	0,18	0,22	0,25	0,24
	10	0,29	-0,07	0,27	0,28	0,33	0,32
	25	0,37	-0,02	0,29	0,27	0,33	0,33
	50	0,40	0,03	0,29	0,24	0,33	0,32
	75	0,38	0,09	0,27	0,20	0,30	0,30
	90	0,38	0,14	0,28	0,19	0,31	0,30
	95	0,37	0,15	0,28	0,18	0,31	0,31

Tabla AII.9. Distancias Horizontales en 7 puntos específicos (Marcas de Clase) de la distribución, en función de la metodología de calibración y el método de calificación.



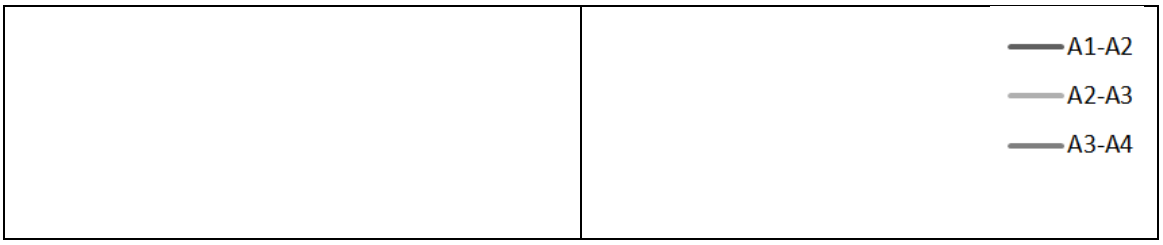
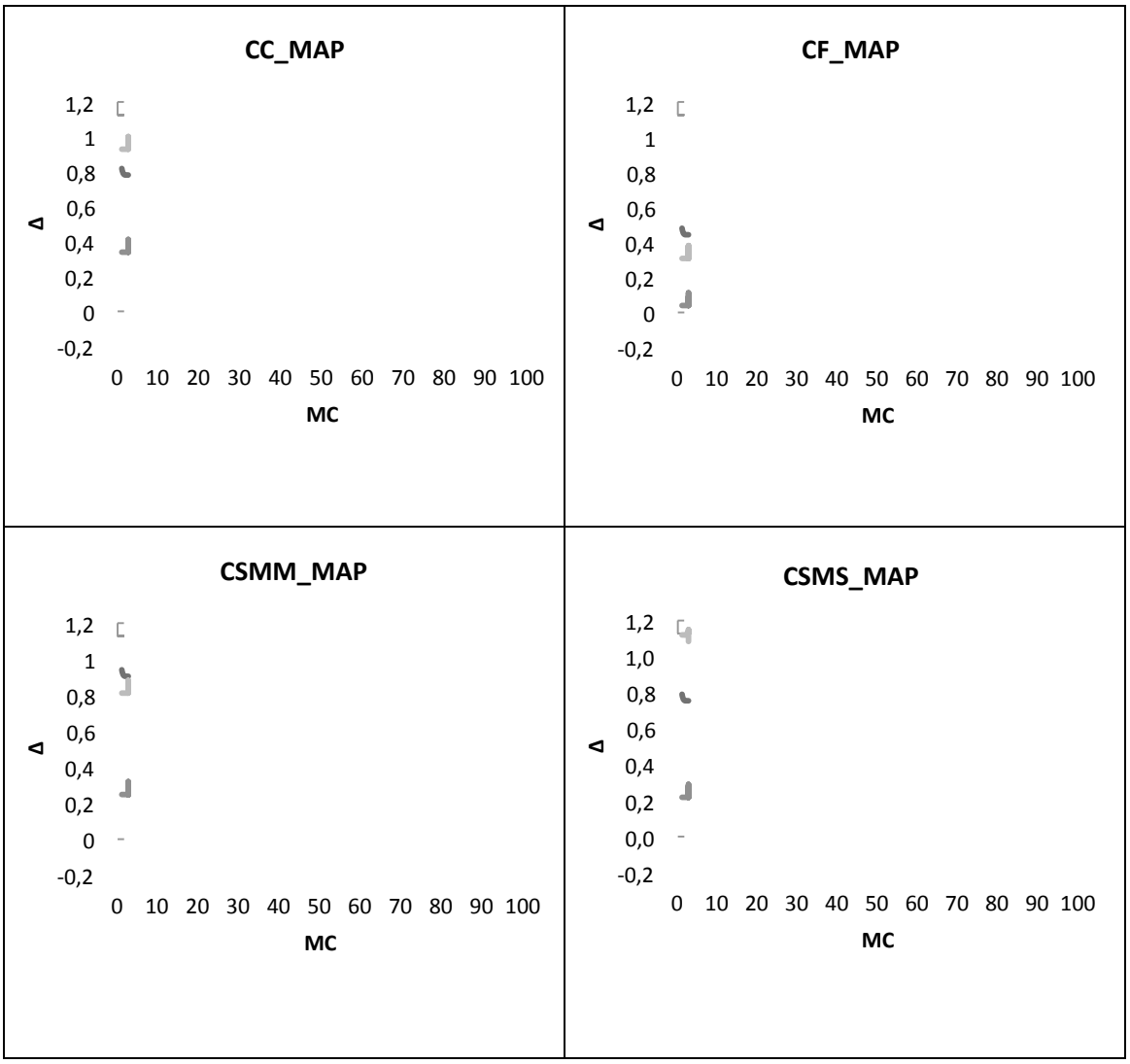


Gráfico AII.11. Distancias horizontales en las 100 Marcas de Clase, en función de la metodología de calibración horizontal, en cada una de las aplicaciones por separado. Método de estimación EAP



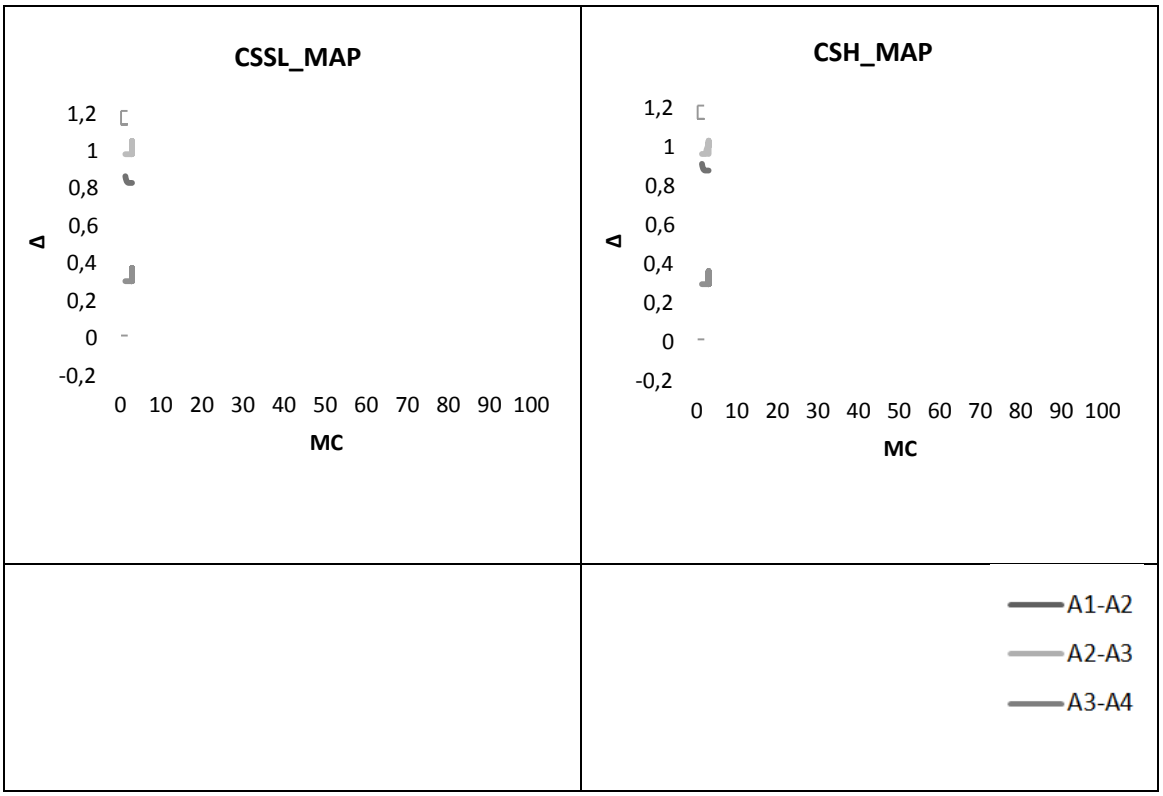
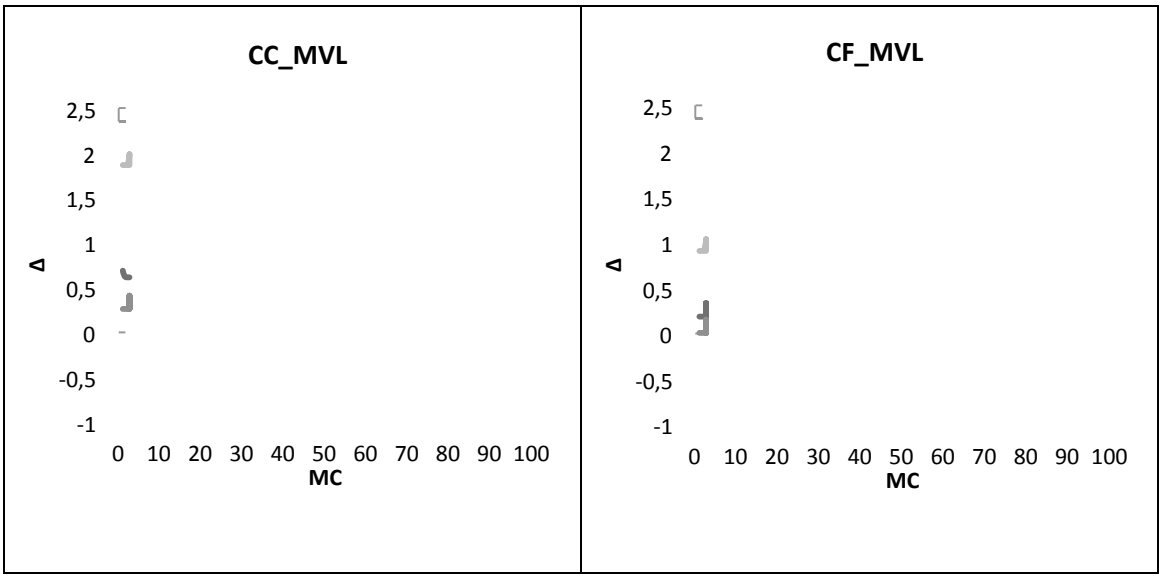


Gráfico AII.12. Distancias horizontales en las 100 Marcas de Clase, en función de la metodología de calibración horizontal, en cada una de las aplicaciones por separado. Método de estimación MAP



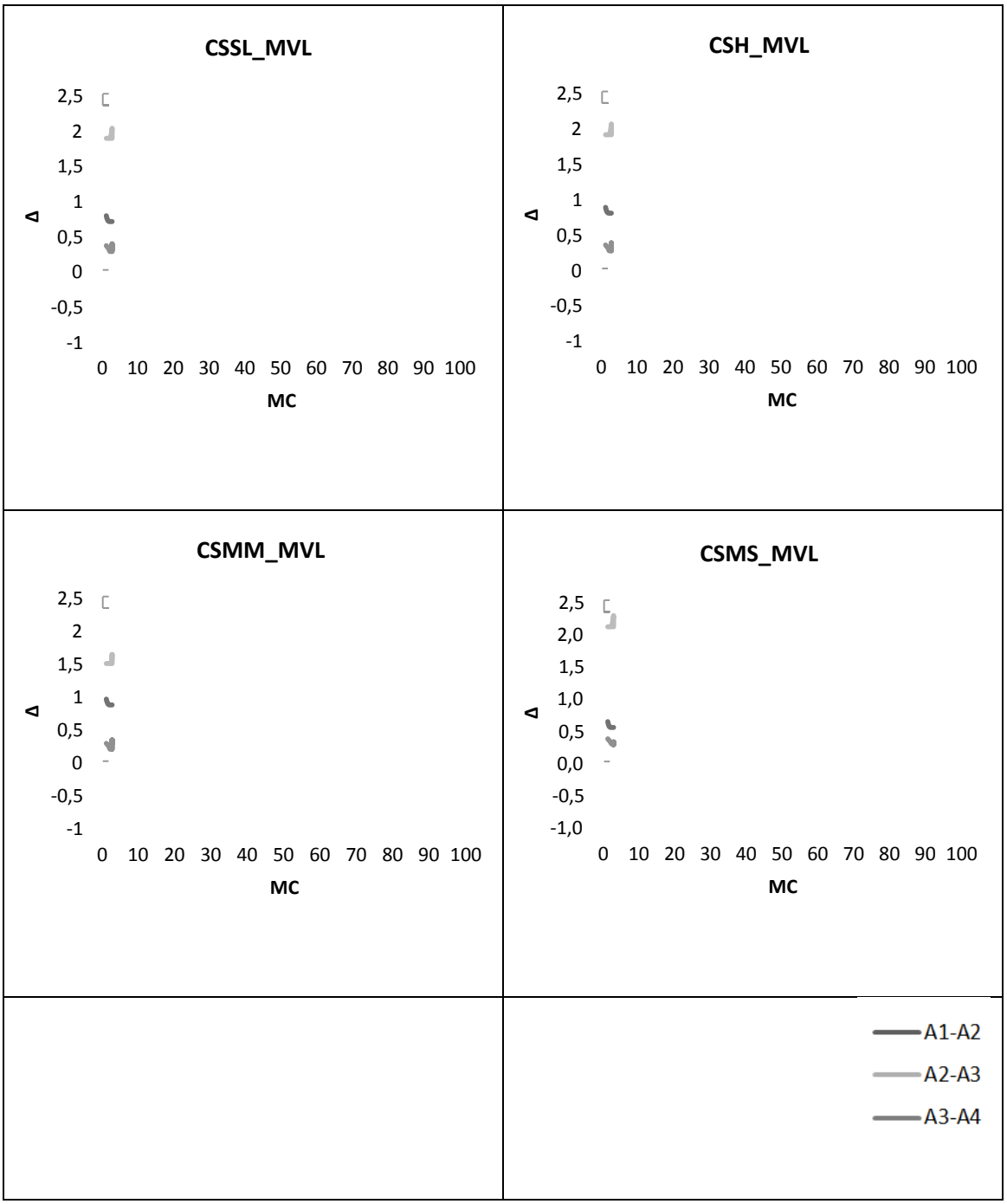


Gráfico AII.13. Distancias horizontales en las 100 Marcas de Clase, en función de la metodología de calibración horizontal, en cada una de las aplicaciones por separado. Método de estimación MVL

		CC	CF	CSMM	CSMS	CSSL	CSH
EAP	A1-A2	0,431	0,267	0,466	0,475	0,449	0,429
	A2-A3	0,789	0,344	0,767	0,785	0,773	0,732
	A3-A4	0,337	0,022	0,271	0,226	0,310	0,297
MAP	A1-A2	0,431	0,237	0,462	0,474	0,448	0,428
	A2-A3	0,774	0,297	0,747	0,758	0,748	0,700
	A3-A4	0,359	0,003	0,269	0,223	0,312	0,312
MVL	A1-A2	0,402	0,243	0,444	0,446	0,424	0,406
	A2-A3	0,834	0,377	0,797	0,823	0,802	0,771
	A3-A4	0,342	0,020	0,275	0,228	0,313	0,309

Tabla AII.10. Distancias Horizontales (Marcas de Clase) medias en función de la metodología de anclaje vertical y de estimación del rasgo.

Anexo III: Sintaxis

Anexo III.1 Sintaxis BILOG MG para la elaboración de la escala vertical de rendimiento

```
>COMMENT
Escala Vertical de Rendimiento en Matemáticas
Estimación: Calibración Conjunta y MAP
ELIMINADOS EL MN6AB15 Y MJ7AB12 Y MN6A36_J7B8

>GLOBAL DFName = 'F:\BBDD_Dep.dat', NPArm = 3, SAVe;
>SAVE PARm = 'F:\M_1-2_2.PAR', SCOre = 'F:\M_1-2_2.SCO';
>LENGTH NITems = (162);
>INPUT NTOtal = 162, NALt = 4, NGROUPS=4, TYPE = 1, NIDchar = 8, KFName = 'F:\BBDD_Dep.dat',
NFName = 'F:\BBDD_Dep.dat', OFName = 'F:\BBDD_Dep.dat';
>ITEMS INUM = (1(1)162),
INAmes = (MO5AB1 MO5AB2 MO5AB3 MO5AB4 MO5AB5 MO5AB6 MO5AB7 MO5AB8 MO5AB9  MO5AB10
MO5AB11 MO5AB12 MO5AB13 MO5AB14 MO5AB15 MO5AB16 MO5AB17 MO5AB18 MO5AB19 MO5A28
MO5A29 MO5A30 MO5A31 MO5A32 MO5A33 MO5A34 MO5A35 MO5A36 MO5A37 MO5B30 MO5B31 MO5B32
MO5B33 MO5B34 MO5B35 MO5B36 MO5B37 MO5B38 MO5A20_J6B19 MO5A21_J6B20 MO5A22_J6B21
MO5A23_J6B22 MO5A24_J6B23 MO5A25_J6B24 MO5A26_J6B25 MO5A27_J6B26 MO5B20_J6A21
MO5B21_J6A22 MO5B22_J6A23 MO5B23_J6A24 MO5B24_J6A25 MO5B25_J6A26 MO5B26_J6A27
MO5B27_J6A28 MO5B28_J6A29 MO5B29_J6A30 MJ6AB1 MJ6AB2 MJ6AB3 MJ6AB4 MJ6AB5 MJ6AB6 MJ6AB7
MJ6AB8 MJ6AB9 MJ6AB10 MJ6AB11 MJ6AB12 MJ6AB13 MJ6AB14 MJ6AB15 MJ6AB16 MJ6AB17 MJ6AB18
MJ6A29_N6B20 MJ6A30_N6B21 MJ6A31_N6B22 MJ6A32_N6B23 MJ6A33_N6B24 MJ6A34_N6B25 MJ6A35_N6B26
MJ6A36_N6B27 MJ6A37_N6B28 MJ6A38_N6B29 MJ6B27_N6A20 MJ6B28_N6A21 MJ6B29_N6A22 MJ6B30_N6A23
MJ6B31_N6A24 MJ6B32_N6A25 MJ6B33_N6A26 MJ6B34_N6A27 MJ6B35_N6A28 MJ6B36_N6A29 MN6AB1
MN6AB2 MN6AB3 MN6AB4 MN6AB5 MN6AB6 MN6AB7 MN6AB8 MN6AB9 MN6AB10 MN6AB11 MN6AB12
MN6AB13 MN6AB14 MN6AB16 MN6AB17 MN6AB18 MN6A31 MN6A35 MN6A40 MN6AB19_J7B1 MN6A30_J7B2
MN6A32_J7B4 MN6A33_J7B5 MN6A34_J7B6 MN6A37_J7B9 MN6A38_J7B10 MN6B30_J7A1 MN6B31_J7A2
MN6B32_J7A3 MN6B33_J7A4 MN6B34_J7A5 MN6B35_J7A6 MN6B36_J7A7 MN6B37_J7A8 MN6B38_J7A9
MN6B39_J7A10 MJ7AB11MJ7AB13 MJ7AB14 MJ7AB15 MJ7AB16 MJ7AB17 MJ7AB18 MJ7AB19 MJ7AB20 MJ7AB21
MJ7AB22 MJ7AB23 MJ7AB24 MJ7AB25 MJ7AB26 MJ7AB27 MJ7AB28 MJ7AB29 MJ7AB30 MJ7AB31 MJ7AB32
MJ7AB33 MJ7AB34 MJ7AB35 MJ7AB36 MJ7AB37 MJ7AB38 MJ7AB39 MJ7AB40 MJ7B3 MJ7B7);
>TEST TNAme = M_1-2ESO;
>GROUP1  GNAME='TIME 1', LENGTH=56, INUM=(1(1)56);
>GROUP2  GNAME='TIME 2', LENGTH=56, INUM=(39(1)94);
>GROUP3  GNAME='TIME 3', LENGTH=57, INUM=(75(1)131);
>GROUP4  GNAME='TIME 4', LENGTH=48, INUM=(115(1)162);
(8A1, I1, 108A1, 1X, 11A1, 1X, 13A1, 1X, 30A1)
>CALIB CYCles = 50, NEWton = 25, PLOt = 1, REFERENCE=1;
>SCORE INFo = 1, method= 3, RSCTYPE = 1, LOCATION = (250.0000), SCALE = (50.0000), NOPrint, POP;
```

Figura AIII.1. Sintaxis BILOG MG para estimar la escala vertical de rendimiento en matemáticas con calibración conjunta.

Anexo III.2 Sintaxis SPSS 19.0 para los Modelos de Valor Añadido

Los modelos se estimaron con el software MLWin pero es posible realizar el mismo proceso utilizando los modelos lineales mixtos de SPSS. A continuación se incluye la sintaxis.

***Modelo multinivel longitudinal con pendiente de crecimiento (M1_3)**

```
MIXED Rasgo with T
/CRITERIA=CIN(95) MXITER(100) MXSTEP(5) SCORING(1) SINGULAR(0.000000000001) HCONVERGE(0,
ABSOLUTE) LCONVERGE(0, ABSOLUTE) PCONVERGE(0.000001, ABSOLUTE)
/FIXED=  T |  SSTYPE(3)
/METHOD=ML
/PRINT=G  R SOLUTION TESTCOV
/RANDOM INTERCEPT T   | SUBJECT(ID_CENTRO) COVTYPE(UN)
/RANDOM INTERCEPT T   | SUBJECT (ID_CENTRO*ID_ALUMNO) COVTYPE(un)
/SAVE=PRED FIXPRED RESID.
```

***Modelo lineal mixto multinivel longitudinal (M2_3)**

```
MIXED Rasgo by T1 T2 T3 T4
/CRITERIA=CIN(95) MXITER(100) MXSTEP(5) SCORING(1) SINGULAR(0.000000000001) HCONVERGE(0,
ABSOLUTE) LCONVERGE(0, ABSOLUTE) PCONVERGE(0.000001, ABSOLUTE)
/FIXED=  T1 T2 T3 T4 | NOINT SSTYPE(3)
/METHOD=ML
/PRINT=G  R SOLUTION TESTCOV
/RANDOM T1 T2 T3 T4   | SUBJECT (ID_CENTRO) COVTYPE(UN)
/REPEATED TIEMPO   | SUBJECT (ID_CENTRO*ID_ALUMNO) COVTYPE(UN)
/SAVE=PRED FIXPRED RESID.
```

***Ganancia estimada (M3_3)**

```
MIXED Rasgo  by T2 T4
/CRITERIA=CIN(95) MXITER(100) MXSTEP(5) SCORING(1) SINGULAR(0.000000000001) HCONVERGE(0,
ABSOLUTE) LCONVERGE(0, ABSOLUTE) PCONVERGE(0.000001, ABSOLUTE)
/FIXED=  T2 T4 | NOINT SSTYPE(3)
/METHOD=ML
/PRINT=G  R SOLUTION TESTCOV
/RANDOM T2 T4   | SUBJECT(ID_CENTRO) COVTYPE(UN)
/REPEATED  T2 T4  | SUBJECT (ID_CENTRO*ID_ALUMNO) COVTYPE(un)
/SAVE= PRED FIXPRED RESID.
```

***Ganancia residual (M4_3)**

```
MIXED RasgoA4 with RasgoA2
/CRITERIA=CIN(95) MXITER(100) MXSTEP(5) SCORING(1) SINGULAR(0.000000000001) HCONVERGE(0,
ABSOLUTE) LCONVERGE(0, ABSOLUTE) PCONVERGE(0.000001, ABSOLUTE)
/FIXED=  RasgoA2| SSTYPE(3)
/METHOD=ML
/PRINT=G  R SOLUTION TESTCOV
/RANDOM RasgoA2   | SUBJECT(ID_CENTRO) COVTYPE(UN)
/SAVE= PRED FIXPRED RESID.
```

Figura AIII.2. Sintaxis de SPSS para estimar Modelos Lineales Mixtos

Anexo III.3 Resultados proporcionados por el software SPSS

A modo de ejemplo se incluyen los resultados de M1_3, el modelo multinivel de crecimiento, obtenidos con SPSS:

		Número de niveles	Estructura de covarianza	Número de parámetros	Variables del sujeto
Efectos fijos	Intersección	1		1	
	T	1		1	
Efectos aleatorios	Intersección + T[a]	2	Sin estructura	3	ID_CENTRO
	Intersección + T[a]	2	Sin estructura	3	ID_CENTRO * ID_ALUMNO
Residuos				1	
Total		6		9	

Tabla AIII.1. Dimensiones del modelo

-2 log de la verosimilitud	106188,565
Criterio de información de Akaike (AIC)	106206,565
Criterio de Hurvich y Tsai (AICC)	106206,581
Criterio de Bozdogan (CAIC)	106281,356
Criterio bayesiano de Schwarz (BIC)	106272,356

a. Los criterios de información se muestran en formatos de mejor cuanto más pequeños.

Tabla AIII.2 Criterios de información

Efectos fijos

						Intervalo de confianza 95%	
Parámetro	Estimación	Error típico	gl	t	Sig.	Límite inferior	Límite superior
Intersección	247,170249	2,259585	65,657	109,387	,000	242,658400	251,682097
T	4,191970	,060405	63,503	69,398	,000	4,071279	4,312661

Tabla AIII.3 Estimaciones de los efectos fijos

Efectos aleatorios

						Intervalo de confianza 95%	
Parámetro		Estimación	Error típico	Wald Z	Sig.	Límite inferior	Límite superior
Residuos		479,955649	9,365840	51,245	,000	461,945548	498,667918
Intersección + T [sujeto = ID_CENTRO]	NE (1,1)	291,773364	57,731795	5,054	,000	197,980529	430,000346
	NE (2,1)	-2,676443	1,181183	-2,266	,023	-4,991518	-,361367
	NE (2,2)	,163227	,041220	3,960	,000	,099503	,267762
Intersección + T [sujeto = ID_CENTRO * ID_ALUMNO]	NE (1,1)	1154,863667	40,971719	28,187	,000	1077,288897	1238,024538
	NE (2,1)	-16,443125	1,524213	-10,788	,000	-19,430528	-13,455722
	NE (2,2)	,434008	,087988	4,933	,000	,291696	,645751

Tabla AIII.4 Estimaciones de los parámetros de covarianza

	Intersección | ID_CENTRO	T | ID_CENTRO
Intersección | ID_CENTRO	291,773364	-2,676443
T | ID_CENTRO	-2,676443	,163227

Tabla AIII.5 Efectos aleatorios de las escuelas asociados a la intersección y el crecimiento

	Intersección | ID_CENTRO*ID_ALUMNO	T | ID_CENTRO*ID_ALUMNO
Intersección | ID_CENTRO*ID_ALUMNO	1154,863667	-16,443125
T | ID_CENTRO*ID_ALUMNO	-16,443125	,434008

Tabla AIII.6 Efectos aleatorios de las estudiantes asociados a la intersección y el crecimiento

	Residuos
Residuos	479,955649

Tabla AIII.7. Residual

Anexo III.4 Sintaxis para calcular residuos de las escuelas

```
**Adaptación de Leyland, A. (2004)**

AUTORECODE VARIABLES = ID_CENTRO
/INTO l2id .
SORT CASES BY l2id .

**get composite residuals**

COMPUTE comp_res =  PRED_2 - FXPRED_2.

**fix_pred son las puntuaciones predichas considerando únicamente la parte fija del modelo.
**Pred es la puntuación predicha considerando tanto la parte fija como la aleatoria.

SET MXLOOP = 100 .

*Debes asegurarte de que el número de mxloop es mayor que el número de escuelas

MATRIX .
GET l2id
/FILE = *
/VARIABLES = l2id .
GET ID_CENTRO
/FILE = *
/VARIABLES = ID_CENTRO.
GET comp_res
/FILE = *
/VARIABLES = comp_res .
GET T_1
/FILE = *
/VARIABLES = T_1 .
COMPUTE temp_mat = (l2id = 1) .
COMPUTE zmat = {temp_mat} .
LOOP i = 2 TO l2id(NROW(l2id)) .
COMPUTE temp_mat = (l2id = i) .
COMPUTE zmat = {zmat, temp_mat} .
END LOOP .
COMPUTE zTz = T(zmat)*zmat .
COMPUTE zTy = T(zmat)* ID_CENTRO.
COMPUTE schl_2 = SOLVE(zTz,zTy) .
LOOP i = 1 TO l2id(NROW(l2id)) .
COMPUTE temp_mat = (l2id = i)&*T_1 .
COMPUTE zmat = {zmat, temp_mat} .
END LOOP .
COMPUTE zTz = T(zmat)*zmat .
COMPUTE zTy = T(zmat)*comp_res .
COMPUTE res_2 = SOLVE(zTz,zTy) .
COMPUTE temp_mat =IDENT(l2id(NROW(l2id)),2*l2id(NROW(l2id))) .
COMPUTE res_2_1 = temp_mat*res_2 .
COMPUTE temp_mat = {0*IDENT(l2id(NROW(l2id))),
IDENT(l2id(NROW(l2id)))} .
COMPUTE res_2_2 = temp_mat*res_2 .
SAVE {schl_2,res_2_1,res_2_2}
/OUTFILE = *
/VARIABLES = ID_CENTRO res_2_1 res_2_2 .
END MATRIX .
EXECUTE .
```

Figura AIII.3 Sintaxis para obtener los residuos de las escuelas asociados a la pendiente y crecimiento en M1_3