UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE EDUCACIÓN
Departamento de Métodos de Investigación y Diagnóstico en Educación

TESIS DOCTORAL

Efecto de la interacción en el muestreo de sujetos e ítems sobre el error de enlace|

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

Eva Expósito Casas

Director

José Luis Gaviria Soto

Madrid, 2016

© Eva Expósito Casas, 2015

UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE EDUCACIÓN

Departamento de Métodos de Investigación y Diagnóstico en Educación

EFECTO DE LA INTERACCIÓN EN EL MUESTREO DE SUJETOS E ÍTEMS SOBRE EL ERROR DE ENLACE

Memoria para optar al grado de doctor presentada por:

Eva Expósito Casas

Director:

Dr. José Luis Gaviria Soto

Madrid, 2015

UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE EDUCACIÓN

TESIS DOCTORAL

Efecto de la interacción en el muestreo de sujetos e ítems sobre el error de enlace

Eva Expósito Casas

DEPARTAMENTO DE MÉTODOS DE INVESTIGACIÓN Y DIAGNÓSTICO EN EDUCACIÓN

Madrid, 2015

A mi GRAN familia...

Agradecimientos

Desde el día que comencé mi trabajo de tesis doctoral, he ido comprendiendo poco a poco la enorme importancia de este espacio, pues el objetivo es expresar en unas breves líneas mi más sentido agradecimiento a todos aquellos que han contribuido de manera directa o indirecta a que este sueño sea hoy una realidad. El efecto de interacción, siendo el tema central de la tesis presentada, cobra su mayor relevancia en este punto, pues la existencia de este trabajo pone de manifiesto la magnitud de tal efecto pues, en esencia, se sustenta en la interacción con las personas aquí mencionadas y con tantas otras que han contribuido en cierto modo a su realización.

De forma muy especial me gustaría dar las gracias al Dr. José Luis Gaviria Soto, director de esta tesis y principal motor de mi carrera investigadora, por deslumbrar a una joven estudiante de pedagogía en segundo de carrera y por no haber dejado de hacerlo hasta el día de hoy. Su inteligencia, sus consejos, su cercanía, sus ideas, su visión de la investigación, han sido una verdadera fuente de inspiración. Gracias por dirigir con ilusión este trabajo, por disfrutar durante todo el proceso como el primer día, por enseñarme a ver las oportunidades y no los límites, por transmitirme la pasión por investigar y por constituir todo un referente personal y profesional.

En segundo lugar, me gustaría dar las gracias a la Universidad de California en los Ángeles (UCLA) por brindarme la oportunidad de realizar una estancia de investigación en el National Center for Research in Evaluation, Standards and Student Testig (CRESST), viviendo una experiencia única a nivel personal y profesional de la mano de Denisse Huang y su grupo de trabajo, quienes me acogieron como una investigadora más desde el primer día y con quienes tanto aprendí.

Quiero hacer especial mención a Esther López Martín, porque este camino es mucho más fácil cuando en él encuentras las huellas fruto de una pisada inteligente y firme que ayudan a guiar tus pasos, por ser un apoyo incondicional en todo momento y sobre todo por ser una compañera y amiga excepcional. Gracias Esther, sin lugar a dudas deberías estar presente en muchos de los apartados que aparecen a continuación, pues te has convertido en una de las personas más importantes de mi vida académica y personal.

A María Castro Morera, quién abrió mi mirada a otros horizontes, enseñándome nuevos conceptos y brindándome excelentes oportunidades. Con ella siempre he tenido la sensación de estar en casa. Gracias por enseñarnos tanto, gracias por tus sabios consejos y, sobre todo, gracias por tu cariño.

Asimismo agradezco a mis compañeros de viaje su infinito apoyo y compañía. Gracias a Enrique Navarro por aportarme siempre una visión diferente de las cosas, a Bianca Thoilliez y Eva Jiménez, porque es un verdadero placer compartir con ellas proyectos e ilusiones, a Pablo Torrijos, por ser el

ingrediente ideal para conformar una cuadrilla perfecta. Somos el mejor equipo que cualquiera pudiera imaginar y a vuestro lado he aprendido las lecciones más importantes.

Quiero dar las gracias a todos y cada uno de los miembros del grupo Medida y Evaluación de Sistemas Educativos (M.E.S.E) y del departamento de Métodos de Investigación y Diagnóstico en Educación (M.I.D.E) de la Universidad Complutense de Madrid. Gracias a Miguel Serra, por recibirme todas las mañanas con una sonrisa y por compartir con nosotros toda una filosofía de vida.

Mi inmensa gratitud a Chantal Biencinto, Coral González y Elvira Carpintero, porque siempre he encontrado en ellas una mirada de comprensión, una palabra de aliento, un abrazo de consuelo, una risa de complicidad. Gracias por tejer con cariño una red de apoyo disponible en todo momento.

Gracias a Covadonga Ruiz, por confiar en mí desde el primer día, proponiéndome mis primeros proyectos de investigación y guiando mis pasos en esas primeras gestas. Gracias a Mercedes García, porque fue un placer ser una de sus alumnas en pedagogía diferencial y aún lo es más haber seguido aprendiendo de su ejemplo. A Inmaculada Asensio, Ángeles Blanco, Luis Lizasoain y Xavier Ordoñez, quienes tanto se han preocupado por el buen desarrollo de este trabajo, por su refuerzo incondicional y por brindarme su ayuda en todo momento.

Me gustaría dar las gracias a mis compañeras del programa de doctorado Tote, Rosa, Karla y Maritza, pues con ellas inicié este recorrido y en ellas encontré excelentes personas con las que compartir esta andadura. Vuestro soporte y compañia han engrandecido mi experiencia formativa; sois todo un ejemplo de esfuerzo y sacrificio.

Gracias a todos los profesores del departamento de Métodos de Investigación y Diagnóstico en Educación II de la UNED, por haber dibujado el mejor entorno de trabajo en el que continuar enseñando y aprendiendo. En especial gracias a Daniel Anaya, Ana Patricia Fernández, Ana María González, María Teresa Martín, Mario Pena, Juan Carlos Pérez, José Manuel Suárez y Consuelo Vélaz de Medrano. Gracias a María Luisa Dueñas, con quien he tenido la suerte de compartir mi práctica docente desde que llegué a la UNED, brindándome su ayuda en todo momento y permitiéndome aprender de su experiencia.

A pesar de que los inicios parecen ya lejanos, he de reconocer que si alguien ha influido en mi formación pedagógica de forma determinante, han sido precisamente mis amigas y compañeras Marta Aguado y Lara Guerra, pues construimos una amistad infranqueable, que ha sido un pilar fundamental en todos mis proyectos académicos y personales. Hemos disfrutado grandes momentos compartiendo muchas horas de pupitre, estudio y reflexión; sin duda, todo un ejemplo de calidad humana y profesional.

No puedo dejar de agradecer su enorme apoyo a aquellos amigos que, a pesar de mis innumerables ausencias, siempre han estado ahí, preocupándose por mi trabajo y esforzándose por comprender mi alto nivel de ocupación durante todas las temporadas. Gracias a Bea, Cifu, David, Ellen, Jaime y Rodrigo por todos los momentos vividos, por vuestra comprensión, ayuda y apoyo, por demostrarme que la amistad es infinita y no entiende de edad, situación o distancia, y muy especialmente quiero dar las gracias a Tamara, porque daría lo que fuera por saber que la tendré a mi lado siempre.

Un agradecimiento muy especial merece Carlos, por caminar junto a mí y no soltar mi mano por más escarpado que sea el terreno, por infundirme infinitas dosis de confianza, alegría, amor y comprensión, por el increible destello en su mirada, por los sueños cumplidos y por los que nos quedan por alcanzar.

Y si he llegado hasta aquí, es sin duda gracias a mi familia. Gracias a mis padres, María José y Francisco Javier por su eterna entrega, por su esfuerzo y dedicación, por su bondad, por su inmenso amor y por haber construido la gran familia de la que tanto he aprendido y tan orgullosa me siento. A mis hermanos Javi, Calo, Peru, Ana y Luis gracias por ser los mejores hermanos del mundo, por vuestros cuidados, por vuestro ejemplo, por vuestra valentía, por vuestra protección, por vuestra generosidad, por estar a mi lado en todo momento acompañándome desde los primeros juegos infantiles, hasta los proyectos más difíciles a los que me he enfrentado y me enfrentaré.

A todos... ¡MUCHAS GRACIAS!

Eva

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

LISTADO DE ABREVIATURAS UTILIZADAS

AERA	American Educational Research Association
AHELO	Assessment of Higher Education Learning Outcomes
APA	American Psychological Association
BRR	Balanced Repeated Replications
CCI	Curva Característica del Ítem
CCT	Curva Característica del Test
DeSeCo	Definition and Selection of Key Competencies
DIF	Differential Item Functioning
ECM	Error Cudrático Medio
ESCS	Economic, Social and Cultural Status Index
ETS	Educational Testing Service
FDE	Función de Distribución Empírica
ICCS	International Civic and Citizenship Education Study
ICILS	International Computer and Information Literacy Study
IEA	International Association for the Evaluation of Educational Achievement
ISCED	International Standard Classification of Education
LLECE	Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación
MCMA	Multiple Choice Multiple Answer
MCSA	Multiple Choice Single Answer
NCME	National Council on Measurement in Education
OECD	Organization for Economic Cooperation and Development
OSE	Observed Score Equating
PERCE	Primera Edición Estudio Internacional Comparativo sobre Lenguaje, Matemáticas y Factores asociados
PIAAC	Programme for the International Assessment of Adult Competencies
PIRLS	Progress in International Reading Literacy Study
PISA	Programme for International Student Assessment
RMSE	Root Mean Squared Error
RECM	Raíz del Error Cuadrático Medio
SACMEQ	Southern and Eastern Africa Consortium for Monitoring Educational Quality

SEE	Standard Error of Equating
SERCE	Segunda Edición Estudio Internacional Comparativo sobre Lenguaje, Matemáticas y Factores asociados
SHP	Scaling on a Hipothetical Population
STA	Scaling to The Anchor
TALIS	Teaching and Learning International Survey
TCT	Teoría Clásica de los Tests
TERCE	Tercera Edición Estudio Internacional Comparativo sobre Lenguaje, Matemáticas y Factores asociados
TIMSS	Trends in Mathematics and Science Study
TRI	Teoría de la Respuesta al Ítem

RESUMEN

Un recorrido global por la situación actual de la evaluación educativa, permite reconocer uno de los principales desafíos a los que la investigación educativa y psicométrica se enfrenta, la comparabilidad. La necesidad de contar con medidas de progreso o cambio educativo, la posibilidad de confrontar los resultados con referentes de interés, la evaluación de conocimientos, actitudes y destrezas a distintos niveles, o los amplios dominios competenciales que se pretenden evaluar, entre otros, constituyen un claro reflejo de ello. Esta situación, contrasta con el escaso volumen de investigación al respecto, puesto de manifiesto en la ausencia de una teoría de base satisfactoria (Holland, 2013; van Der Linden, 2013).

Las fuertes asunciones metodológicas que implica la comparabilidad, así como los diseños de recogida y análisis de datos requeridos, hacen que el proceso de enlace de puntuaciones exigido para posibilitar tales comparaciones sea altamente complejo y esté sujeto a gran variedad de factores que pueden atentar contra su adecuación. De este modo, el error asociado al enlace de puntuaciones, ha permanecido oculto a la sombra de otros errores como podría ser el error muestral. Tanto es así que, el denominado error estándar de equiparación, tan solo captura la varianza fruto del muestreo de sujetos. Entre los estudios más novedosos, destacan aquellos que consideran el error asociado a

la selección de ítems. La forma convencional de cálculo del error de equiparación, basada en el error fruto del muestreo de sujetos representaría una tercera parte del error real, al que sería necesario añadir el error fruto del muestreo de ítems y el efecto combinado de tales factores (interacción). El procedimiento de Bootstrap Bidimensional, cuyo planteamiento y evaluación han constituido el marco general del presente trabajo, pretende capturar la varianza asociada al efecto de interacción entre dichos factores.

A fin de analizar la propuesta metodológica sugerida, se planteó un estudio de simulación. Las variables independientes que configuraron las cuatro condiciones experimentales analizadas fueron: Funcionamiento Diferencial del Ítem (DIF), diferencias en los grupos a evaluar, variación en la distribución del parámetro *b* de los ítems y aumento del nivel medio de dificultad de la prueba, permaneciendo constantes el tamaño muestral (2000) y el número de ítems (50). Tras aplicar el procedimiento propuesto, se analizó la existencia de efecto significativo de la interacción a través del Análisis de Varianza (ANOVA) y se juzgó la adecuación del procedimiento para su detección a través del cálculo de la Raíz del Error Cuadrático Medio (RECM).

Los resultados ponen de manifiesto un efecto significativo de la interacción en todas las condiciones experimentales analizadas, el efecto de la combinación de ambos factores difiere de la suma de los mismos considerados de forma independiente. Un hallazgo inesperado ha sido la constatación de que, a pesar de que el procedimiento implementado ha surgido en el marco del estudio de los procedimientos de enlace de puntuaciones, esta situación de interacción, y en consecuencia la necesidad de su cuantificación, estará presente en cualquier evaluación que exija una selección de sujetos y reactivos, no estando estrictamente vinculada a los procesos de enlace. Así mismo, los resultados muestran que nos encontramos ante un procedimiento que puede resultar eficiente en la detección de DIF o de diferencias en los grupos que se desean comparar, enfrentándose a dos de las principales fuentes de error sistemático reconocidas en la literatura.

ABSTRACT

An overall view of the current situation of educational evaluation enables us to identify one of the main challenges facing educational and psychometric research, that of comparability. This is clearly reflected in the need to have in place measures for educational progress or change, the possibility of confronting results with benchmarks of interest, the evaluation of knowledge, attitudes and skills at various levels, and the broad proficiency areas being evaluated, among other factors. This situation contrasts with the scarce amount of research on the subject, evidenced by the absence of a satisfactory theoretical base (Holland, 2013; van Der Linden, 2013).

The significant methodological assumptions involved in comparability, as well as the data collection and analysis designs required, make the score linking process required in order to make such comparisons a highly complex endeavour subject to a wide variety of factors that may undermine its usefulness. Errors associated with score linking tend to remain hidden behind other errors, such as sampling error. So much so, that the standard equating error only captures the variance resulting from subject sampling. Notable innovative studies include those that consider the error associated

with item selection. The conventional way of calculating equating error, based on the subject sampling error, would represent a third of the real errors, to which it would be necessary to add the item sampling error and the combined effect of these factors (interaction). The proposed Bootstrap Two-Dimensional procedure and its evaluation, which forms the general framework of this study, seeks to capture the variance associated with the effect of interaction between such factors.

In order to analyse the proposed methodology, a simulation study was set up. The independent variables making up the four experimental conditions analysed were: Differential Item Functioning (DIF), differences in the groups being assessed, variation in the *b* parameter distribution of the items and an increase in the average level of difficulty of the test, with sample size (2000) and number of items (50) remaining constant. After applying the proposed procedure, the significance of the interaction effects was analysed by using Analysis of Variance (ANOVA) and the suitability of the procedure was judged using the Root Mean Square Error (RMSE) method.

The results show a significant interaction effects under all of the experimental conditions analysed, with the effect of the combination of both factors differing from their sum taken independently. An unexpected finding was that, despite the fact that the implemented procedure arose within the framework of the study of score linking procedures, this interaction situation, and consequently the need for it to be quantified, will be present in any evaluation demanding a selection of subjects and tasks, or items, without being strictly related to linking processes. Likewise, the results show that we are looking at a procedure that could turn out to be efficient in detecting DIF or differences in the groups being compared, addressing the two of the main sources of systematic error recognised in the literature.

INTRODUCCIÓN

La metodología utilizada en investigación educativa, ha de dar respuesta a la complejidad de su objeto de estudio, la educación, considerando ésta como una realidad en movimiento que es preciso analizar atendiendo a todas y cada una de sus dimensiones. Se trata pues, de plasmar una realidad cambiante, dinámica, en permanente transformación y evolución.

El presente trabajo, pretende ahondar, desde el punto de vista teórico y metodológico, en uno de los grandes desafíos a los que la investigación educativa y psicométrica se enfrenta "la comparabilidad".

El auge de la evaluación de sistemas educativos que se viene experimentando desde hace más de tres décadas, pone de manifiesto que la evaluación educativa ha llegado para quedarse. Así, desde las concepciones básicas de la evaluación del rendimiento, entendiendo éste como resultados académico-curriculares, hasta los planteamientos más complejos, que abarcan el desarrollo de conocimientos, actitudes y destrezas a distintos niveles, se ha recorrido un largo camino. En el ámbito nacional, la legislación educativa es una clara prueba de ello, pues desde la creación del Instituto Nacional de Calidad y Evaluación, al amparado de la Ley de Ordenación General del Sistema Educativo (1990), hasta la última reforma legislativa propuesta (Ley Orgánica

para la Mejora de la Calidad Educativa, 2013), el papel otorgado a la evaluación, ha ido cobrando mayor protagonismo.

Unido a ello, numerosos organismos nacionales e internacionales vienen desarrollando una gran labor en el ámbito de la evaluación, labor ampliamente reconocida y cuya importancia creciente atañe a todas las esferas o niveles educativos. A finales de los años 70, la «*International Association for the Evaluation of Educational Achievement*» (IEA), lleva a cabo los primeros estudios periódicos, cuyo objetivo, es obtener información acerca del progreso de los estudiantes, los estudios de tendencias educativas, cobran fuerza desde ese momento y hoy por hoy, podemos apuntar a la existencia de un gran mapa de evaluación, complejo tanto por su extensión como por su diversidad. Esta realidad, nos transporta a la antigua Babilonia, a la construcción de una Torre de Babel, en la que el propósito central puede verse desmoronado a causa del uso de diferentes "lenguas". El objetivo de esta investigación, va más allá de perspectivas particulares, tratando el desafío de la comparabilidad como eje transversal que vertebra la respuesta a distintos retos en evaluación.

De este modo, la evaluación de conocimientos, actitudes y destrezas a distintos niveles, la evaluación de amplios dominios competenciales, la seguridad en los procesos de evaluación, la necesidad de contar con medidas del cambio o progreso educativo, la posibilidad de comparar con referentes de interés (países, escuelas, distritos, estados, etc.), son retos destacados en el marco evaluativo actual, el desafío de la comparabilidad o enlace de puntuaciones es un aspecto central a todos ellos.

Partiendo de la idea apuntada por Scriven (1967), la utilidad de la información proveniente de las evaluaciones depende de la posibilidad de comparar la misma con criterios de interés. La solución a los problemas de comparabilidad no es un desafío al que podamos dar respuesta de forma rápida, sin embargo, dar una solución adecuada resulta crucial, puesto que afecta directamente a la interpretación y al uso de los resultados (Gempp, 2010). A pesar de su reconocida importancia, llama la atención observar el escaso interés que esta temática ha suscitado desde el ámbito de la investigación educativa y psicométrica, poniéndose de manifiesto en el reducido volumen de trabajos realizados al respecto. Unido a ello, cabe destacar que, en sus orígenes, parecía un terreno reservado a los especialistas en psicometría, siendo los años

80 el momento en el que existe cierta apertura a otros especialistas de la medición (Kolen y Brennan, 2004). La ausencia de un marco conceptual claro, de una teoría de base satisfactoria (Holland, 2013; van der Linden, 2013) constituye un indicador clave acerca de la necesidad de trabajar con mayor profusión en este ámbito.

La posibilidad de aplicar diversos métodos a la hora de implementar distintos procedimientos para el enlace de puntuaciones, es otro de los aspectos que suele generar mayor confusión. Una primera distinción entre estos procedimientos sería la que los diferencia entre aquellos que se enmarcan bajo los supuestos de la Teoría Clásica de los Tests (TCT), y los basados en la Teoría de la Respuesta al Ítem (TRI). Cabe destacar que, las mayores diferencias entre los diferentes tipos de enlace de puntuaciones son de tipo interpretativo, no procedimental (Dorans, Moses Eignor, 2010), en este sentido, disponer de una sólida teoría de base sería el componente esencial para una mejor comprensión, uso e interpretación de los resultados obtenidos en cualquier proceso de enlace de puntuaciones.

Las fuertes asunciones metodológicas y los complejos diseños de recogida y análisis de datos, hacen del enlace de puntuaciones un proceso altamente complejo, proceso que puede verse afectado por gran diversidad de factores. En consecuencia, no podemos pasar por alto que, el enlace de puntuaciones, lleva asociado un error, error que ha sido obviado en la mayoría de los casos. A pesar de que el término «error de equiparación» es un término comúnmente aceptado, con el fin de contribuir a una delimitación conceptual más precisa, en el marco de la presente investigación se utiliza el término «error de enlace», para destacar su existencia e importancia en procesos tanto de equiparación como de escalamiento. La forma convencional de calcular el error asociado al proceso de enlace es el denominado error estándar de equiparación, estimador que tan solo captura la varianza fruto del muestreo de sujetos (error aleatorio) (Kolen & Brenan, 2004; Michaelides & Haertel, 2004). Por su parte, van der Linden (2013) señala que, los informes de equiparación típicos, ignoran el error de equiparación por completo, su error estándar es solo para calcular las fluctuaciones fruto del muestreo de sujetos. El error sistemático, consecuencia del incumplimiento de las asunciones del modelo o método seleccionado, (Wang, 2006) ha recibido escasa consideración.

En los últimos años, diversos autores han defendido la importancia asociada al error fruto del enlace, de este modo Wu (2010) asegura que, en el caso de estudios de tendencias o de crecimiento, el proceso de escalamiento puede ser el origen de distintos errores, afectando tanto a los resultados individuales como a los análisis por cohortes. Por su parte, Michaelides y Haertler (2004) ya habían indicado que, a nivel de puntuaciones grupales, el error de medida se contrae con el aumento del tamaño de la muestra, cobrando especial importancia el error de equiparación, fundamentalmente el producido por el muestreo de ítems comunes, al no verse afectado por el tamaño de la muestra. En la misma línea, Monseur y Berezner (2007) indican que el error de enlace puede superar los errores muestrales y de medida.

En este contexto, el presente trabajo de tesis doctoral, ha pretendido dar un paso más en la consideración del error de enlace, partiendo de las fuentes principales de error reconocidas en trabajos precedentes (error muestral de sujetos y de reactivos), así como del efecto de su interacción, producido por la consideración de ambos factores de manera combinada. De manera más precisa, el marco general de la presente tesis doctoral ha sido definido como «planteamiento y evaluación de una propuesta metodológica que permita la estimación del efecto de interacción entre la selección de sujetos y reactivos en procesos de enlace de puntuaciones». En concreto, se propone un método denominado «bootstrap bidimensional», caracterizado por ser una técnica intensiva con unidades de remuestreo dobles (filas y columnas), en nuestro caso (sujetos y reactivos). Para conseguir definir el marco de investigación, y estructurar el reto al que nos enfrentamos de manera conveniente, las preguntas centrales planteadas a las que se pretende dar respuesta serían las siguientes:

1. *¿Cuál es la importancia de la evaluación en el sistema educativo actual?*

2. *¿Cuáles son las principales exigencias a las que se enfrenta la evaluación educativa?*

3. *¿Qué implica la comparabilidad?*

4. *¿Qué factores pueden incidir en los procesos de enlace de puntuaciones?*

5. *¿Cuál es la perspectiva tradicional en la estimación del error de enlace?*

6. *¿Existe efecto de interacción en la selección de sujetos y reactivos?*

7. *¿Es posible mejorar los procesos de enlace de puntuaciones teniendo en cuenta el efecto de interacción en la selección de sujetos e ítems?*

8. *¿Cómo estimar la interacción entre dichas fuentes de error?*

9. *¿Resulta de utilidad el estudio de la interacción en el análisis de ítems comunes?*

10. *¿En qué condiciones muestra su efectividad el procedimiento propuesto?*

De manera más específica, la propuesta se concretaría en base a los siguientes objetivos:

1. Proponer un procedimiento para el análisis del efecto de interacción de la selección de sujetos y reactivos en procesos de enlace «bootstrap bidimensional».
    a. Justificación teórica.
    b. Diseño de una propuesta eficiente de implementación.
    c. Elaboración de sintaxis de análisis para su puesta en práctica.

2. Analizar las propiedades estadísticas globales de dicho procedimiento.

3. Comprobar el comportamiento del procedimiento ante diferentes condiciones experimentales.

    a. Funcionamiento Diferencial del Ítem.
    b. Diferencias en nivel de habilidad en los dos grupos a comparar.
    c. Variación en la distribución de los valores de *b* de los ítems que componen la prueba.

d. Desplazamiento de los valores medios del parámetro *b* de los ítems incrementando su dificultad.

La memoria que aquí presentamos se estructura en 6 capítulos. El primero de ellos, de carácter teórico y titulado "Rendimiento y evaluación", pretende realizar un análisis del marco de referencia ante el que nos encontramos, estudiando tanto el panorama actual como la evolución desde sus primeras manifestaciones hasta nuestros días.

El capítulo 2, centrado en la comparabilidad de mediciones, busca aportar una delimitación conceptual precisa que permite avanzar en un tema tan complejo como el que nos ocupa. En él, se parte del concepto general y de los orígenes de la comparabilidad, para pasar posteriormente al análisis de sus principales manifestaciones, proponiendo en última instancia una definición precisa y ajustada de todas las categorías conceptuales incluidas dentro de este ámbito. Del mismo modo, el capítulo presenta los diseños de enlace más frecuentes, aspecto central que condiciona en gran medida los procesos de comparabilidad.

En tercer lugar, el capítulo titulado "métodos de enlace", recoge una síntesis de las diferentes aproximaciones metodológicas posibles, utilizando para su estructuración la división tradicional entre aquellos métodos desarrollados en el marco de la Teoría Clásica de los Tests (TCT) y los que se encuentran bajo el paradigma de la Teoría de la Respuesta al Ítem (TRI). De manera complementaria, el capítulo cuatro aborda el problema relativo a la estimación del error de enlace, destacando los procedimientos tradicionales así como las nuevas corrientes emergentes, que comienzan a señalar la insuficiencia de los procedimientos tradicionales.

Tras ello, el apartado que recoge una descripción detallada del procedimiento llevado a cabo ha sido titulado "Propuesta de un modelo integral para el cálculo del error de enlace basado en técnicas intensivas de remuestreo de sujetos e ítems". En este quinto capítulo, se formula el problema de investigación, se definen los objetivos y se especifica el método (generación de datos, variables, hipótesis, descripción del procedimiento bootstrap bidimensional, condiciones experimentales, etc.).

El sexto capítulo ha sido dedicado a la presentación de resultados. En él, se incluye tanto la descripción detallada del procedimiento implementado, como los resultados fruto de su aplicación en las cuatro condiciones experimentales estudiadas bajo un entorno de Simulación Montecarlo.

Para finalizar, se incluye una síntesis de las principales conclusiones, así como de las limitaciones encontradas y las futuras líneas de trabajo a las que esta investigación abrirá paso.

INTRODUCTION

The methodology used in education research has to respond to the complexity of its object of study. In the case of education it is like trying to approach a moving target that must be analysed taking each and every one of its aspects into account. This entails capturing a changing, dynamic reality that is undergoing a continuous process of transformation and development.

This study sets out to take an in-depth look, from the theoretical and methodological point of view, at one of the great challenges facing education and psychometric research, "comparability".

The rise in the evaluation of education systems observed for more than three decades highlights the fact that education evaluation is here to stay. Much has been done, beginning with the basic concept of evaluating performance, understood as academic and curricular results, all the way through to the most complex approaches that address the development of knowledge, attitudes and skills at various levels. On the national scene, education legislation is clear proof of this, since from the creation of the National Institute of Quality and Evaluation, the Law of General Governance of the Education System (1990), to the most recent proposed legislative reform (Organic Law

for Improving Education Quality, 2013) the role given to evaluation has become more and more prominent.

Along with this, a number of national and international organisations have been doing some great work in the area of evaluation, a widely acknowledged task whose growing importance pertains to all spheres and levels of education. In the late 1970s, the International Association for the Evaluation of Educational Achievement (IEA) carried out the first periodical studies aimed at obtaining information about student progress; from that time, studies of educational trends began to gain strength and nowadays, we can point to the existence of a great map of education, as complex in its size as in its diversity. This reality takes us back to ancient Babylon, to the construction of a Tower of Babel, in which the central purpose can fall apart because of the use of different "languages". The aim of this research goes beyond individual outlooks, addressing the challenge of comparability as a transversal axis that runs through the response to a series of challenges in evaluation.

In this way, the evaluation of knowledge, attitudes and skills at various levels, the evaluation of wide ranging competency areas, security in the evaluation processes, the need to have measures for educational change and progress, the possibility of comparing with references of interest (countries, schools, districts, states, etc.) are major challenges in the current evaluation framework, with the challenge of comparability or score linking being a central feature of all of them.

Based on the idea put forward by Scriven (1967), the usefulness of information obtained from evaluations depends on the possibility of comparing it with criteria of interest. The solution to comparability problems is not a challenge to which we can give a quick answer, however, it is crucial that we provide an adequate solution as it directly affects the interpretation and use of results (Gempp, 2010). Despite its widely acknowledged importance, it is interesting to see the lack of interest aroused by this topic in the field of education and psychometric research, highlighted by the small number of studies carried out on the subject. Plus, it should be noted that, in its early days, it seemed to be a topic reserved for psychometric specialists, only opening up to a certain extent in the 1980s to other measurement experts (Kolen and Brennan, 2004). The absence of a clear conceptual framework, of a satisfactory theoretical base

(Holland, 2013; Van der Linden, 2013) is a clear indicator of the need for more work to be done in this field.

The possibility of applying diverse methods when implementing different procedures for score linking is another of the aspects that tends to cause greater confusion. A first distinction between these procedures would be the one that differentiates them from the ones that fall under the assumptions of Classical Test Theory (CTT) and those based on Item Response Theory (IRT). It should be highlighted that the greatest differences between the various types of score linking are interpretative, not procedural (Dorans, Moses, Eignor, 2010); in this sense, having a solid base theory would be the essential component for a greater understanding, use and interpretation of the results obtained in any score linking process.

Substantial methodological assumptions and complex data collection and analysis designs make the score linking process a highly complex one, which can be affected by a wide range of factors. Consequently, we cannot overlook the fact that score linking involves error, an error that has been ignored in most cases. Despite the term "equating error" being a commonly accepted one, in order to contribute to a more precise conceptual definition, this research study uses the term "linking error" to highlight its existence and importance in both equating and scaling processes. The conventional way of calculating the error associated with the scaling process is the so-called standard equating error, an estimator that only captures the variance arising from subject sampling (random error) (Kolen & Brenan, 2004; Michaelides & Haertel, 2004). Van der Linden (2013) pointed out that typical equating reports completely ignore equating error and that their standard error is only for calculating fluctuations resulting from subject sampling. Systematic error (resulting from the failure to comply with the assumptions of the selected model or method) (Wang, 2006) has received very little attention.

In recent years, a number of authors have defended the importance given to the error resulting from linking, so Wu (2010) argued that, in the case of studies on trends or growth, the scaling process may give rise to various errors, affecting both individual results and cohort analysis. Likewise, Michaelides and Haertler (2004) had already indicated that, at the level of group scores, the error of measurement reduces with the

increase in sample size, with particular importance acquired by equating error, basically the sort produced by sampling common items, as it is not affected by sample size. Along the same lines, Monseur and Berezner (2007) reported that linking error can exceed sampling and measurement error.

In this context, this PhD thesis study has set out to take a further step forward in tackling the issue of linking error, based on the main sources of error recognised in preceding studies (subject and item sampling error) as well as the effect of their interaction, produced by the combined discussion of both factors. More precisely, the general framework of this PhD thesis has been defined as "addressing and evaluating a methodological approach that enables an estimation of the effect of interaction between the selection of subjects and items in score linking processes". Specifically, a method known as "two-dimensional bootstrapping", characterised by being an intensive technique with double resampling units (rows and columns), in our case (subjects and items). To be able to define the research framework and structure the challenge being tackled in a suitable way, the central questions being posed to which the study aims to provide answers are the following:

1. *What is the importance of evaluation in the current education system?*
2. *What are the main demands facing Educational Evaluation?*
3. *What does comparability involve?*
4. *What factors can affect the score linking process?*
5. *What is the traditional view on estimating linking error?*
6. *Is there an interaction effect in the selection of subjects and tasks?*
7. *Is it possible to improve the score linking process taking into account the interaction effect in the selection of subjects and items?*
8. *How can the interaction between these sources of error be estimated?*

9. *Is it useful to study interaction in the analysis of common items?*

10. *Under what conditions is the proposed procedure effective?*

More specifically, the proposal is based on the following objectives:

1. To put forward a procedure for analysing the effect of interaction of the selection of subjects and tasks in "two-dimensional bootstrap" scaling processes.
    a. Theoretical argument.
    b. Design of an efficient implementation proposal.
    c. Production of an analysis syntax to be put into practice.

2. To analyse the overall statistical properties of this procedure.

3. To check the behaviour of the procedure under a range of experimental conditions.
    a. Differential item functioning (DIF).
    b. Differences in the skill level of the two groups being compared.
    c. Variation in the distribution of *b* values of the items making up the test.
    d. Displacement of the average values of the *b* parameter of the items increasing their difficulty.

The report presented here is divided into 6 chapters. The first, which is theoretical and entitled "Performance and Evaluation", aims to carry out an analysis of the framework of reference we are dealing with, studying both the current panorama and development from early instances through to the present day.

Chapter 2, focusing on the comparability of measurements, contributes a precise conceptual definition that enables progress to be made in a complex topic such as this one. The chapter deals with the general concept and the origins of comparability, before moving on to the analysis of its main instances, and eventually putting forward a precise

and accurate definition of all the conceptual categories included in this field. Likewise, the chapter presents the most frequent linking designs, a central aspect that conditions comparability processes to a great extent.

Thirdly, the chapter entitled "Linking Methods" contains a summary of the various methodological approaches possible, structured on the basis of the traditional split between methods developed as part of Classical Test Theory (CTT) and those contained in the paradigm on Item Response Theory (IRT). Complementing this, chapter 4 deals with the problem associated with linking error estimation, highlighting traditional procedures as well as new emerging trends, which are starting to underline the inadequacy of these traditional procedures.

Following on from this, the section containing a description of the procedure carried out is entitled "Proposal for an Integrated Model for Calculating Scaling Error Based on Intensive Subject and Item Resampling Techniques". In this fifth chapter, the research problem is formulated, the objectives are defined and the method is specified (data generation, variables, hypothesis, description of the two-dimensional bootstrap procedure, experimental conditions, etc.).

In the sixth chapter the results are presented. This chapter includes both the detailed description of the procedure implemented and the results from applying it under the four experimental conditions studied in a Monte Carlo Simulation environment.

Lastly, there is a summary of the main conclusions as well as the limitations found and future lines of work opened up by this research.

CAPÍTULO 1: RENDIMIENTO Y EVALUACIÓN

El objetivo central del primer capítulo, es presentar un análisis del constructo rendimiento académico, así como de la problemática general asociada a su medida. El primer apartado, se centra en la delimitación conceptual, recogiendo las diferentes perspectivas, su evolución y los factores que influyen en el mismo y que más se han identificado en la literatura al respecto, presentando una visión global de ellos. Así, habida cuenta de las implicaciones que la Administración Pública tiene en el Sistema Educativo, dentro de tal apartado, se realiza una pequeña síntesis acerca del tratamiento del concepto rendimiento desde la normativa reguladora del Sistema Educativo. Todo ello, conduce a la consideración del rendimiento dentro del paradigma de las competencias básicas, paradigma imperante en la actualidad tanto desde el punto de vista administrativo como metodológico, por este motivo, en el segundo apartado se analiza el rendimiento académico dentro del paradigma de las competencias básicas. La influencia de los aspectos tratados en estos dos primeros apartados en el ámbito de la evaluación, queda reflejada en el tercer epígrafe, en el que se trata la evaluación educativa desde la perspectiva de los resultados a la medida de las competencias. Tras ello, se presenta un análisis de las principales evaluaciones (nacionales e internacionales) llevadas a cabo en el sistema educativo, fruto del interés creciente por la medida del rendimiento académico. Por último, en el quinto apartado, se analizan las dificultades y problemática en las evaluaciones del sistema educativo asociadas a variable de respuesta.

1.1 Rendimiento: conceptualización, factores asociados y marco normativo.

El Diccionario de Ciencias de la Educación (Prellezo, 2009), considera que rendimiento es sinónimo de resultados escolares. El problema en la delimitación conceptual del término, se situará precisamente en su aplicación pedagógica. Hablar de rendimiento educativo, supone aplicar a la educación un criterio de productividad, relacionado con el resultado final o calidad del producto final (Antoni, 2002).

Si nos detenemos a analizar la literatura especializada, podemos comprobar que no existe una única definición válida de dicho concepto, éste término, puede ser interpretado de diferentes maneras en función de la dimensión en que se sitúe el énfasis. Esta es la razón que hace que el rendimiento sea considerado un término complejo, multidimensional, relativo y contextual (Salvador, Rodríguez, & Bolivar, 2004). Como consecuencia de ello, distintos autores han definido el concepto apoyándose en diferentes perspectivas, pudiendo caracterizarse en aquellas que ponen énfasis en la dimensión individual (Kaczynska, 1965; Muñoz Arroyo, 1977) o las centradas en la dimensión escolar (Pérez, 1981; Álvaro, 1990). En definitiva, y de acuerdo a lo apuntado por De la Orden, Oliveros, MafoKozi y González (2001) podemos aludir al resultado del sistema educativo (nivel macro) o al resultado específico de cada individuo (nivel micro), teniendo siempre en cuenta la naturaleza multidimensional del término (González Fernández, 1975; Tourón, 1985). En cuanto a la medida del rendimiento, Carabaña (1987), vincula la validez de las distintas medidas de rendimiento académico la definición subyacente a las mismas, de este modo, opta por una visión operacionalista, definiendo el rendimiento como "el resultado de sus mediciones social y académicamente relevantes" (Carabaña, 1987, p. 267).

Desde nuestro punto de vista, el rendimiento académico, supondrá una valoración acerca de la adquisición de conocimientos, destrezas y actitudes por parte del estudiante a lo largo de su etapa escolar siendo, por tanto, el reflejo de la actividad educativa. Los conocimientos, destrezas y actitudes que comprenderían el rendimiento, dependerán de la etapa en la que el estudiante se encuentre así como de otros aspectos derivados de los objetivos específicos de cada curso y materia. Al mismo tiempo,

permite optar por una perspectiva macro (centros, instituciones, países, etc.) o micro (a nivel de estudiante).

En definitiva, la literatura muestra diferentes enfoques desde los que se trata de definir el rendimiento académico, cuyo análisis pone de manifiesto un avance progresivo en su concepción. El análisis de los factores que han podido incidir en ello revela tres momentos clave: evolución de las funciones atribuidas al sistema educativo, valoraciones acerca del tipo de conocimiento que se ha de adquirir en la escuela y concepciones sobre el aprendizaje del alumno (Salvador, Rodríguez, & Bolivar, 2004).

Como vemos, un análisis detallado del término rendimiento, nos lleva a descubrir una realidad mucho más compleja de la intuida inicialmente, de este modo, las investigaciones al respecto deben recoger dicha complejidad, abordando el rendimiento académico con precisión teórica y metodológica. Desde el punto de vista metodológico, la tarea de evaluar el rendimiento académico se complejiza, partiendo de un simple análisis de los resultados académico-curriculares, hasta llegar a una visión más completa del término, con las implicaciones que ello supone a la hora de cuantificarlo o medirlo, "La multidimensionalidad intrínseca al rendimiento exige recurrir a una diversidad de variables, objetivos, e instrumentos, generando estrategias distintas de análisis y medición" (Matas, 2003, p. 185).

Tal y como apunta Rodríguez Espinar (1982) consideramos que, abordar de manera científica la problemática del rendimiento escolar, exige un esfuerzo por encontrar aquellos determinantes básicos de ese rendimiento, puesto que el objetivo del presente trabajo no está en el análisis de dichos determinantes, en el presente apartado nos limitaremos a recoger un breve análisis acerca de lo que dice la literatura sobre los mismos, a fin de comprender con mejor precisión la complejidad del constructo rendimiento. Así, uno de los trabajos pioneros en esta línea fue el desarrollado por Svensson (1971), en el que explora la influencia de variables como el género o el ambiente familiar en el rendimiento académico. El trabajo, muestra la dificultad que supone delimitar tales factores, de este modo, se observa cómo, el producto educativo (rendimiento académico), se ve influido por dimensiones que conciernen a las esferas personal, social, familiar, escolar, cultural, etc. de la vida del estudiante, por otro lado, no resulta sencillo determinar cuál es la importancia o aportación específica de cada una de ellas, puesto que las relaciones entre las mismas y su propia naturaleza dificultan

enormemente dicha tarea. En este mismo trabajo Svensson (1971) analiza la influencia de otras variables cuando se controlan el género y el ambiente familiar.

Una primera distinción entre los factores determinantes del rendimiento académico, podría ser aquella que contempla tanto su dimensión individual como contextual (Svensson, 1971; Instituto Nacional de Ciencias de la Educación, 1976; Walberg, 1984; Rodríguez Espinar, 1982; Álvaro, 1990; Martínez Otero, 1997; Antoni, 2002; Bruner y Elacqua, 2004) proponiendo distinciones entre variables personales, ambientales, familiares, sociales, escolares, etc.

Entre las investigaciones desarrolladas en este ámbito, destaca la realizada por Coleman y colaboradores en 1966, estudio que se convirtió en un referente tanto por su carácter novedoso como por la polémica y debate que suscitó en torno a los factores determinantes del rendimiento académico. En su estudio, Coleman y colaboradores (1966) concluyen apuntando a la mayor importancia de los factores asociados al «*background*» frente a las variables escolares, señalando que el efecto de las escuelas al progreso de los estudiantes, es similar cuando se tienen en cuenta las características del «*background*». Dicho informe, ha recibido numerosas críticas desde el punto de vista metodológico, principalmente, a causa de la utilización de la técnica de regresión paso a paso «*step by step*», que introduce como primeros predictores las variables del contexto socioeconómico y, en consecuencia, deja poca varianza por explicar a las variables escolares. De este modo, la colinealidad podría ser responsable de los resultados que otorgan mayor importancia a las características individuales en relación a las escolares.

A pesar de las críticas recibidas, posteriormente se han confirmado algunos de los hallazgos presentes en el informe Coleman, de este modo, Jencks y colaboradores (1972) afirman que, el progreso académico, se ve influido principalmente por las características de los propios estudiantes, restando importancia a los factores escolares. Para estos autores, el mayor determinante del logro educativo, es el «*Background familiar*», compuesto no solo por variables económicas. De este modo, afirman que los estudiantes se ven más influidos por lo que pasa en sus hogares, en sus barrios o por lo que ven en la televisión, que por lo que pasa en sus escuelas (Jencks et al., 1972). Al mismo tiempo, en Gran Bretaña, las investigaciones llevadas a cabo para la elaboración del informe «*Children and their Primary Schools*» más conocido como informe

«Plowden» (Plowden, 1967), mostraron cómo los factores asociados a variables familiares, tenían mayor importancia que las variables escolares (Wall & Varma, 1975).

Una respuesta a esta visión catastrofista de la escuela, aparece en Estados Unidos de la mano del movimiento de *escuelas eficaces*, que surge con el deseo por parte de investigadores, teóricos, docentes, etc. de devolver a la escuela y al docente la importancia perdida desde la publicación del informe Coleman (Piñeros & Rodríguez, 1998). De acuerdo con Murillo (2003), los dos objetivos de la investigación sobre eficacia escolar serían, en primer lugar, determinar cuánto influye la escuela en el rendimiento de los alumnos (estimación de la magnitud de los efectos escolares) y, en segundo lugar, la identificación de los factores escolares que hacen que una escuela sea eficaz.

En esta línea, diversos trabajos han analizado la influencia en el logro o resultados académicos de los estudiantes de factores referentes a variables escolares, del aula y del docente (Weber, 1971; Edmonds, 1979; Rutter, Maughan, Mortimer, & Ouston 1979; Walberg, 1984; Mortimore, Sammons, Stoll, Lewis, & Ecob, 1988; Piñeros y Rodríguez, 1998; Scheerens, 2000; Theule, 2006; Cervini, 2003).

El estudio llevado a cabo por Rutter et al. (1979) conocido con el nombre «*Fifteen thousand hours*», también tiene una importancia crucial en este ámbito. Su diseño de carácter longitudinal (estudiantes Londinenses desde el año 1970 a 1976), persigue el objetivo de analizar de los factores escolares con influencia en el comportamiento y rendimiento académico de los estudiantes (Marzano, 2000). Con el fin de detectar qué variables escolares se asocian con la efectividad de las escuelas Klitgaard y Hall (1974) desarrollan una investigación a gran escala en la que identifican aquellas escuelas que, una vez controlados los efectos de ciertas variables de «*background»* de los estudiantes, consiguen incrementar el progreso de los mismos, a pesar de que en términos promedio las características de las escuelas puedan no influir (confirmando los resultados de los trabajos previos de Jencks), consideran de interés identificar aquellas escuelas que resultan "inusualmente efectivas".

Con el fin de identificar los factores con mayor influencia en el aprendizaje escolar, Walberg (1984) utiliza modelos de productividad (propios del ámbito de la economía) aplicados al funcionamiento de las escuelas. En su trabajo, señala dentro de

su teoría de la Productividad Educativa, tres grandes grupos: factores relacionados con la aptitud de los estudiantes, con la instrucción y con el entorno. Dentro de estos tres grandes grupos se incluyen nueve sub-factores (Walberg, 1984).

El interés por los problemas metodológicos presentes en las investigaciones sobre factores determinantes del rendimiento académico, llevó a Aitkin y Longford (1986) a la publicación del artículo «*Statistical modelling issues in school effectiveness studies*», en el que reconocen la estructura anidada de los datos educativos y presentan los modelos multinivel como el procedimiento adecuado para tratar los mismos. Estos modelos multinivel permiten descomponer la varianza en niveles incluyendo variables predictoras para cada nivel, con la inclusión de efectos de interacción (Martinez Arias, Gaviria, & Morera, 2009). La metodología multinivel, ha supuesto un avance en esta línea, al permitir la diferenciación de los efectos procedentes de distintas dimensiones educativas (Ruiz De Miguel, 2009; Martínez Arias, Gaviria y Castro, 2009).

Atendiendo a la revisión realizada sobre factores determinantes del rendimiento académico (considerando éste en sus niveles micro y macro) podemos identificar cuatro grandes grupos de variables: personales, de aula, escolares y ambientales (comunidad) (Ver Tabla 1).

Tabla 1.

Factores que influyen en el rendimiento académico

Nivel			
Individual/Personal	Aula	Escolar	Ambiental/ Comunidad
o Edad o Género o Personalidad o Inteligencia o Aptitudes (verbal y numérica) o Estilos cognitivos o Intereses o Motivación o Valoración del trabajo intelectual o Rendimiento previo o Aspiraciones o Preescolarización o Habitos de estudio o Condiciones de cuidado y salud. o Familia: - Clima - Estructura o configuración familiar. - Ingresos - Ocupación de los padres - Recursos Educativos en el hogar - Actividades culturales - Estimulación de experiencias educativas - Implicación - Actitud de los padres hacia la educación - Aspiraciones educativas de los pades respecto a sus hijos	o Aprendizaje de competencias básicas o Expectativas o Tiempo efectivo de aprendizaje o Experiencia y formación docente o Comportamiento docente o Uso del refuerzo educativo o Responsabilidad o Expectativas o Estrategias o Estilo docente. o Programación de aula. o Comunicación profesor-alumno o Asingación de tareas escolares para el hogar. o Ratio o Agrupamiento	o Liderazgo/ estilo directivo o Clima o Estabilidad del profesorado o Titularidad o Calidad docente o Coordinación docente o Compromiso docente o Implicación o Evaluación de estudiantes o Evaluación docente o Tiempo de enseñanza directa o Fracaso esperado o Procedimientos de agrupación o Estructuración de la jornada escolar o Comunicación o Relación familia-escuela	o Ambiente cultural o Recursos educativos o Seguridad o Cultura del lenguaje o Curriculo o Autonomía o Financiación o Evaluación externa de escuelas. o Gasto por alumno o Investigación educativa

Fuente: elaboración propia a partir de la bibliografía citada en este apartado.

El complejo entramado de factores que inciden en el rendimiento académico, pone de manifiesto, una vez más, la problemática relacionada tanto con la delimitación conceptual del rendimiento académico, como con la investigación de aspectos relacionados con el mismo. Atender a dicha complejidad es un requisito sine qua non a todo trabajo relativo al rendimiento escolar.

De manera independiente a la perspectiva adoptada a la hora de tratar el concepto de rendimiento académico, no se puede perder de vista que, en función de los objetivos que persiga el sistema educativo, el rendimiento será entendido de una manera u otra. Tradicionalmente, los objetivos, estaban centrados en los denominados contenidos curriculares y, por tanto, el rendimiento era una medida del nivel de dominio de dichos contenidos por parte del estudiante. Si nos detenemos a analizar las diferentes Leyes reguladoras del sistema educativo, podemos comprobar cómo ha ido evolucionando la idea del producto del sistema educativo y, por tanto, la consideración de rendimiento académico.

De este modo, desde la Ley General de Educación y Financiamiento de la Reforma Educativa (1970), se hace alusión al rendimiento académico, entendido éste como resultado del proceso educativo, tanto a nivel de centros, como a nivel de estudiantes, identificando algunos de los factores que pueden influir en el rendimiento considerado a nivel de centro. No obstante, el enfoque general de la ley, se centra en la transmisión de información y conocimiento, como consecuencia de ello, el rendimiento será entendido en esta línea, y los resultados del sistema educativo valorados en función de los conocimientos adquiridos por los estudiantes.

Veinte años más tarde, la Ley de Ordenación General del Sistema Educativo (LOGSE) (1990), introduce las actitudes, los hábitos, las aptitudes y los valores como producto escolar esperado, en consecuencia, los resultados ya no conciernen sólo a la esfera de conocimientos. El rendimiento educativo se vuelve más complejo exigiendo un proceso de evaluación acorde a las características del mismo. Por otro lado la LOGSE (1990) reconoce la importancia de la evaluación del sistema educativo, y a tal efecto crea el Instituto Nacional de Calidad y Evaluación.

El interés por los resultados del sistema educativo, continúa estando muy presente en la normativa. La evaluación, por tanto, deberá tener en cuenta el rendimiento (entendido como producto del sistema educativo) en relación a los objetivos educativos planteados en dicha normativa. Posteriormente, la Ley Orgánica de Calidad de la Educación (LOCE) (2002) recoge algunas consideraciones a tener en cuenta, al presentar la primera alusión a las competencias como resultado esperado del proceso educativo. De este modo, rendimiento académico y evaluación se verán influidos por este nuevo enfoque. En la actualidad, el debate acerca de las competencias

básicas continúa abierto y ha marcado, en gran medida, el desarrollo de diversos ámbitos de la realidad educativa y, por tanto, del rendimiento y la evaluación.

Entre los aspectos de especial relevancia de la LOCE (2002), se encontraría precisamente su enfoque centrado en competencias y la importancia concedida a la evaluación educativa, lo que nos permite observar cómo el concepto de rendimiento académico continúa creciendo en su riqueza y complejidad. La posterior Ley educativa, la Ley Orgánica de Educación (LOE) (2006), tiene un carácter continuista en relación con estos dos aspectos.

Por último, en la Ley Orgánica para la Mejora de la Calidad Educativa (LOMCE) (2013), existe un marcado énfasis en la evaluación educativa, así como en las competencias clave que ha de adquirir el alumnado. La necesidad de una Ley como la articulada recientemente se ha justificado, en gran medida, atendiendo a los resultados obtenidos por España en diferentes estudios internacionales (Eurostat, 2011; PISA, 2009).

Entre las novedades más notables de dicha Ley (LOMCE, 2013), se destaca la realización de las denominadas evaluaciones externas de fin de etapa. Dichas evaluaciones, de carácter formativo y diagnóstico, tienen por objetivo garantizar tanto los niveles de aprendizaje de los estudiantes en relación con el título pretendido, como orientar a este en sus elecciones escolares en relación a sus conocimientos y competencias. Tal es el caso de la *Evaluación final de Educación Primaria* (Artículo, 21), la *Evaluación final de Educación Secundaria Obligatoria* (Artículo 29) y la *Evaluación final de Bachillerato* (Artículo, 36 bis), en estas dos últimas, la obtención de la titulación dependerá del resultado obtenido en las mismas (LOMCE, 2013). En la normativa se especifica que se trata de evaluaciones no solo de contenidos sino de competencias, en la justificación de estas evaluaciones, se señala que "Veinte países de la OCDE realizan a sus alumnos pruebas de esta naturaleza y las evidencias indican que su implantación tiene un impacto de al menos dieciséis puntos de mejora de acuerdo con los criterios de PISA" (LOMCE, 2013, p. 6). Por otro lado, en la ley se apunta que "Las pruebas serán homologables a las que se realizan en el ámbito internacional y en especial a las de la OCDE y se centran en el nivel de adquisición de las competencias" (LOMCE, 2013, p.6).

De este modo, la evaluación educativa y las competencias básicas guían, tanto el desarrollo normativo, como la práctica educativa actuales, estando estrechamente vinculados entre sí. En el siguiente apartado se presenta un análisis de la idea de competencia, visión que permitirá comprender mejor su nexo con la evaluación. La valoración de los resultados obtenidos por el sistema educativo deberá realizarse atendiendo al desarrollo de Competencias Básicas, en consecuencia, la medida del rendimiento académico vendrá marcada en esta etapa por la idea de Competencia Básica, así como por la problemática que entraña su medida, dimensión que se abordará en profundidad en el apartado 1.4. Por tanto, el producto educativo esperado (los objetivos educativos) han de contrastarse con el producto educativo en términos de resultados, este contraste entre lo esperado y lo logrado será el rendimiento educativo (De la Orden, Oliveros, MafoKozi, & Gonzalez, 2001), en este caso, las competencias básicas serán el punto de partida y de llegada.

1.2 El rendimiento en el paradigma de las competencias básicas.

Los objetivos educativos están definidos hoy en día por el desarrollo de las denominadas competencias básicas, de este modo, el rendimiento académico deberá hacer alusión al desarrollo de dichas competencias. Este enfoque, de carácter global, ha tenido fuertes implicaciones en los sistemas educativos de diferentes países, así como en sus sistemas de evaluación pero, ¿qué entendemos por competencias básicas? ¿por qué se ha optado por definir los objetivos del sistema educativo en relación a dichas competencias? ¿cuál es el alcance de dicho enfoque?

El paradigma de las competencias básicas en educación se ha instalado con fuerza en numerosos países. La globalización, caracterizada por una mayor interacción entre los ciudadanos de todo el mundo, y la influencia de la sociedad de la información, configuran una nueva realidad social, económica, política y cultural. La educación ya no busca dotar a los estudiantes de conocimientos teóricos estáticos que éstos han de memorizar, necesita que adquieran una serie de competencias que les permitan adaptarse a nuevos retos en una sociedad cambiante, seleccionar la información, buscar recursos, comunicarse de manera adecuada, trabajar en grupo, etc. Esta nueva situación

ha producido que organismos nacionales e internacionales (EURYDICE, 2002; United Nations Educational, Scientific and Cultural Organization, 1996; North Central Regional Educational Laboratory and Metiri Group, 2003; Organización para la Cooperación y Desarrollo Económicos, 2005), se preocupen por esta perspectiva emergente. En los últimos años, este enfoque se ha convertido en un eslogan global, al que se han sumado las políticas educativas de numerosos países, ello puede suponer una posibilidad instruccional para renovar el currículo escolar (Bolívar, 2010, p. 9).

La fuerte generalización del término en todos los niveles educativos, desde la educación infantil a la educación universitaria parece inevitable, por ello, debemos preocuparnos por el concepto que finalmente se asocie a dicho término. Como consecuencia de ello, se ha producido un amplio debate sobre las distintas definiciones planteadas (Millet & Sánchez, 2007). A la hora de analizar las mismas, no debemos olvidar los posibles intereses generados al respecto desde distintos ámbitos pues la separación de la discusión sobre competencias, del terreno estrictamente pedagógico, puede invadir el terreno político o social (Denyer, Furnémount, Poulain, & Vanloubbeeck, 2007). No olvidemos que, el término competencias, comenzó siendo utilizado en el ámbito de la empresa y que, posteriormente, se extendió al entorno educativo, esta situación ha causado ciertas reticencias en distintos sectores educativos que veían esta situación, en cierto modo, como una "amenaza" que podía afectar a los planteamientos pedagógicos e introducir un aire "mercantilista" en el proceso de enseñanza-aprendizaje.

En la mayoría de los trabajos sobre competencias se considera a McClelland el responsable del origen del concepto, en su artículo «*Testing for Competence Rather Than for Intelligence*» (1973), expone una serie de críticas a los tradicionales instrumentos de medida de la inteligencia así como a las consecuencias que pruebas cómo el «*Scholastic Aptitude Test* (SAT)» tienen en la vida de los sujetos así, considera las competencias como habilidades de trabajo que determinan un desempeño competente en el empleo (McClelland, 1973). Desde el punto de vista educativo, el objetivo es llegar a una definición del término que responda a las verdaderas necesidades de la educación, aprovechando los beneficios que esta visión puede traer a los distintos universos de la enseñanza, por tanto, lo importante en este punto es centrarnos en lo que podríamos considerar el «enfoque pedagógico de la educación por

competencias». En la actualidad, si revisamos la literatura especializada, podemos encontrar gran volumen de trabajos relativos a dicha delimitación conceptual.

Un punto clave en el desarrollo posterior de distintas definiciones de competencias, es la definición aportada en la «*World Declaration on Education for All: Meeting Basic Learning Needs»* (World Conference on Education for All, 1990). En dicha declaración, no se definen de manera específica ni se hace alusión a las competencias clave, sin embargo, en el Artículo 1, que versa sobre necesidades básicas de aprendizaje, en el primer párrafo, se afirma: "Cada persona – niños, jóvenes y adultos – se podrá beneficiar de las oportunidades educativas diseñadas para satisfacer sus necesidades básicas de aprendizaje. Estas necesidades se refieren a los instrumentos de aprendizaje (lectura, escritura, expresión oral, cálculo y resolución de problemas) y a los contenidos básicos de aprendizaje (conocimientos, destrezas, valores y actitudes) requeridos por los seres humanos para poder sobrevivir, desarrollar todas sus capacidades, vivir y trabajar con dignidad, participar plenamente en el desarrollo, mejorar su calidad de vida, tomar decisiones debidamente informadas y continuar aprendiendo"(p.5).. En este mismo artículo, en el Apartado cuarto, la declaración pone de manifiesto que la educación básica es más que un fin en sí mismo, es la base para el aprendizaje permanente y el desarrollo humano (World Conference on Education for All, 1990).

Posiblemente, el trabajo clave que ha influido en posteriores definiciones de competencias básicas, es la identificación de los cuatro pilares de la educación a lo largo de la vida que aparece en el informe de la UNESCO «*Learning: the treasure within*». En dicho informe, se identifican los siguientes pilares de la educación: aprender a conocer, aprender a hacer, aprender a vivir juntos y aprender a ser (Delors, 1996). En la Tabla 2 se presenta un resumen de las definiciones de competencias aportadas por distintos organismos nacionales e internacionales. En dicha tabla podemos apreciar cuáles son las notas comunes presentes en todas ellas así como las diferencias encontradas.

Tabla 2.

Definición de competencias

Organismo	Definición de competencias
OECD (2005a, p.3) Proyecto DeSeCo	Una competencia es más que conocimientos y destrezas. Involucra la habilidad de enfrentar demandas complejas, apoyándose en y movilizando recursos psicosociales (incluyendo destrezas y actitudes) en un contexto particular.
Parlamento Europeo y Consejo de la Unión Europea (2006, p.3).	Combinación de conocimientos, capacidades y actitudes adecuadas al contexto. Las competencias clave son aquéllas que todas las personas precisan para su realización y desarrollo personales, así como para la ciudadanía activa, la inclusión social y el empleo.
LOE (2006) RD 1531 (2006, p. 43053). RD 1631 (2006, p. 678).	Aprendizajes que se consideran imprescindibles desde un planteamiento integrador y orientado a la aplicación de los saberes adquiridos. Su logro deberá capacitar a los alumnos y alumnas para su realización personal, el ejercicio de la ciudadanía activa, la incorporación a la vida adulta de manera satisfactoria y el desarrollo de un aprendizaje permanente a lo largo de la vida.
LOMCE (2013) Orden ECD/65 (2015, p.6986).	«Saber hacer» que se aplica a una diversidad de contextos académicos, sociales y profesionales. Las competencias clave son aquellas que todas las personas precisan para su realización y desarrollo personal, así como para la ciudadanía activa, la inclusión social y el empleo.

Fuente: elaboración propia.

Estas definiciones muestran de manera general, alguna de las características del concepto de competencia, aportando una visión global de cuál es el objetivo educativo que se pretende lograr con la incorporación de las mismas. Como podemos observar, la definición de competencia presenta una fuerte relación con otros conceptos con los que suele venir acompañada, incluso en muchos casos puede llegar a sustituirse la parte por el todo, y considerar sinónimos algunos términos que presentan una fuerte relación con las competencias.

Por su parte, Monereo (2005) define las competencias partiendo de sus diferencias y similitudes en relación a las estrategias. Desde el análisis de la influencia de este enfoque en el ámbito universitario, Zabala y Arnau (2007) tras realizar un análisis pormenorizado de las definiciones de competencias, consideran que, una

competencia supone hacer frente a situaciones diversas de forma eficaz en un contexto determinado, para conseguir este objetivo, es necesario movilizar actitudes, habilidades y conocimientos, al mismo tiempo y de forma interrelacionada.

El sociólogo Philippe Perrenoud (1999), en su trabajo relativo desarrollo de competencias docentes, considera que éstas han de entenderse como una capacidad de movilización de diversos recursos cognitivos con el fin de hacer frente a un tipo de situaciones. Del mismo modo, el autor señala ciertos aspectos a los que se debe prestar atención en la definición de competencias, de los que destacamos lo siguiente: las competencias no son sinónimo de conocimientos, habilidades o actitudes, aunque movilizan, integran, orquestan dichos recursos; la movilización de recursos solo resulta pertinente en situación, cada situación es única, aunque se la pueda tratar por analogía con otras ya conocidas; el ejercicio de la competencia pasa por operaciones mentales complejas, sostenidas por esquemas de pensamiento, permitiendo adaptar la acción a la situación. Escamilla (2008) considera que las competencias son capacidades que se relacionan, fundamentalmente, con el saber hacer. Por su parte, Castillo y Cabrerizo (2010), apuntan que una competencia se define y se observa por: la actuación apropiada en un contexto concreto; estar asociada a un campo del saber; la adquisición diferenciada y secuenciada en niveles progresivos de dominio y las producciones o resultados.

Partiendo del análisis de las definiciones aportadas por los especialistas, consideramos que, una competencia supone una combinación de conocimientos, capacidades- habilidades, actitudes y destrezas que permiten al individuo afrontar diversidad de situaciones y problemas en distintos contextos. Las competencias tienen un carácter interdisciplinar y pueden (y deben) ser desarrolladas desde distintos ámbitos de la vida del estudiante. El desarrollo de competencias, permitirá al sujeto aprender a lo largo de toda la vida de manera autónoma y eficaz. Pero ¿cuáles son dichas competencias? ¿dónde debemos poner el énfasis? ¿es posible llegar a un acuerdo acerca de lo esencial en educación? Tras analizar el concepto de competencias, nos enfrentamos a esta nueva dificultad que, si bien es cierto, se formula por primera vez en términos de competencia, siempre ha estado presente en los distintos debates educativos, puesto que tras la selección de lo "esencial" en educación, subyacen distintas visiones del proceso de enseñanza aprendizaje.

En distintos documentos podemos encontrar las palabras "básica", "esencial" "clave", "fundamental", "imprescindible", "elemental", etc. acompañando al término competencia, con el fin de identificar qué competencias son las que tienen mayor importancia en el desarrollo integral del individuo. Tal y como recoge Escamilla (2008) en los proyectos DeSeCo y Tuning se habla de competencias esenciales, dichas competencias, tendrían un carácter medular y son los núcleos de referencia más sintéticos, pudiendo servir para delimitar los grandes principios de la dirección metodológica y organizativa en una institución. El Instituto de Evaluación (2009) indica que la normativa española ha optado por el adjetivo *básica* en su doble acepción de esencial y vinculante, sin que parezca asociable a la acepción elemental. Por tanto, en el caso de la normativa española, se hace referencia a competencias básicas sin que exista una distinción entre el término básica y esencial o vinculante.

Monereo (2005) considera que la adjetivación de las competencias como básicas, puede llevar a confusión, puesto que puede pasar a considerarse que son conductas mínimas o simples, y nada más lejos de la realidad. Además, el desarrollo de competencias no presenta límites, es decir, siempre pueden desarrollarse y mejorar. Para Millet y Sánchez (2007) hablar de competencias básicas supone hacer referencia a lo que se podría considerar la dotación cultural mínima que cualquier ciudadano o ciudadana debe adquirir, en consecuencia, el estado ha de garantizar dicha adquisición. Por su parte, Monereo y Castelló (2009), apuntan que las competencias básicas en educación obligatoria deberían permitir que los estudiantes hagan uso de diversos recursos y artefactos culturales de su entorno en tanto que ciudadanos independientes y responsables. Para Bolivar (2010) las competencias básicas serían aquellas que resultan clave y que todo el alumnado deberá dominar al término de la escolaridad obligatoria, de modo que pueda proseguir los estudios, recibir una formación profesional e integrarse socialmente sin riesgo de exclusión.

En el informe de Eurydice (2002), tras presentar un interesante análisis de la variedad en la conceptualización del término competencias, se expone que "a pesar de las diferentes concepciones e interpretaciones del término, la mayoría de los expertos parecen coincidir en que para que una competencia merezca el atributo de "clave", "fundamental", "esencial" o "básica", debe ser necesaria y beneficiosa para cualquier individuo y para la sociedad en su conjunto" (Eurydice, 2002, p. 13). En la misma línea

apuntada por Eurydice, Deakin (2008), destaca la posición de las competencias entre el carácter personal y el social.

En el ámbito de la Unión Europea, la motivación principal para introducir este nuevo lenguaje, es fruto del objetivo de lograr unos mínimos comunes en los sistemas educativos de los países miembros, con el fin de crear una potencia económica capaz de competir con la economía global. (Gimeno-Sacristán, Pérez, Torres, Angulo, & Álvarez-Méndez, 2008). En el año 2006, fueron elaboradas una serie de recomendaciones del parlamento y el consejo europeo sobre las competencias clave para el aprendizaje permanente con el fin de establecer un referente europeo, considerando que las competencias clave son "aquellas que todas las personas precisan para su realización y su desarrollo personales, así como para la ciudadanía activa, la inclusión social y el empleo" (Parlamento Europeo y Consejo de la Unión Europea, 2006, pág. 3). Este marco de referencia establece las siguientes ocho competencias básicas: comunicación en lengua materna; comunicación en lenguas extranjeras; competencia matemática y competencias básicas en lengua y tecnología; competencia digital; aprender a aprender; competencias sociales y cívicas; sentido de la iniciativa y espíritu de empresa, y conciencia y expresión culturales (Parlamento Europeo y Consejo de la Unión Europea, 2006, pág. 3). Además de estas ocho competencias, se identifican varias competencias personales que juegan un papel importante en cada una de estas ocho competencias, que serían; pensamiento crítico, creatividad, iniciativa, resolución de problemas, evaluación de riesgos, toma de decisiones y gestión constructiva de sentimientos (Parlamento Europeo y Consejo de la Unión Europea, 2006 , pág. 3).

Pero, a la hora de hablar de la selección de competencias esenciales, no podemos olvidar el trabajo realizado por el proyecto «*Definition and Selection of Key Competencies*» (DeSeCo) en el marco de la OECD. El Proyecto DeSeCo, pretende aportar un marco conceptual firme para servir como fuente de información en la identificación de competencias clave y el fortalecimiento de las evaluaciones internacionales que miden el nivel de competencia de jóvenes y adultos, tales como PISA (OECD, 2005a). El trabajo dentro de este proyecto, ha dado lugar a la definición de competencias más conocida, así como a la selección de competencias básicas más universal (Bolívar, 2010).

En el proyecto DeSeCo considera que son innumerables las competencias que necesitarían los individuos para enfrentar los complejos retos del mundo actual, por ello, intentar recoger en un listado todas y cada una de las competencias que los sujetos podrían necesitar en un contexto y en un momento determinado de sus vidas, carecería por completo de sentido práctico.

El marco conceptual del proyecto DeSeCo, clasifica las competencias clave en tres categorías interconectadas entre sí. Estas categorías, permiten estructurar el mapa de las competencias esenciales y comprender los vínculos entre ellas. Las categorías establecidas en el proyecto son: uso de herramientas de manera interactiva, interacción en grupos heterogéneos y actuar de forma autónoma (OECD, 2005a). Las competencias, estructuradas en esas tres categorías, tendrían una serie de características subyacentes que atravesarían todas ellas: motivación, compromiso, pensamiento, reflexión, creatividad, etc. Partiendo de esta idea de interacción entre las tres dimensiones, y teniendo en cuenta las diferentes combinaciones posibles en distintos contextos y situaciones, a continuación presentamos un esquema que sitúa las nueve competencias esenciales propuestas en el proyecto DeSeCo, dentro de los tres ejes considerados en dicho proyecto (Figura 1).

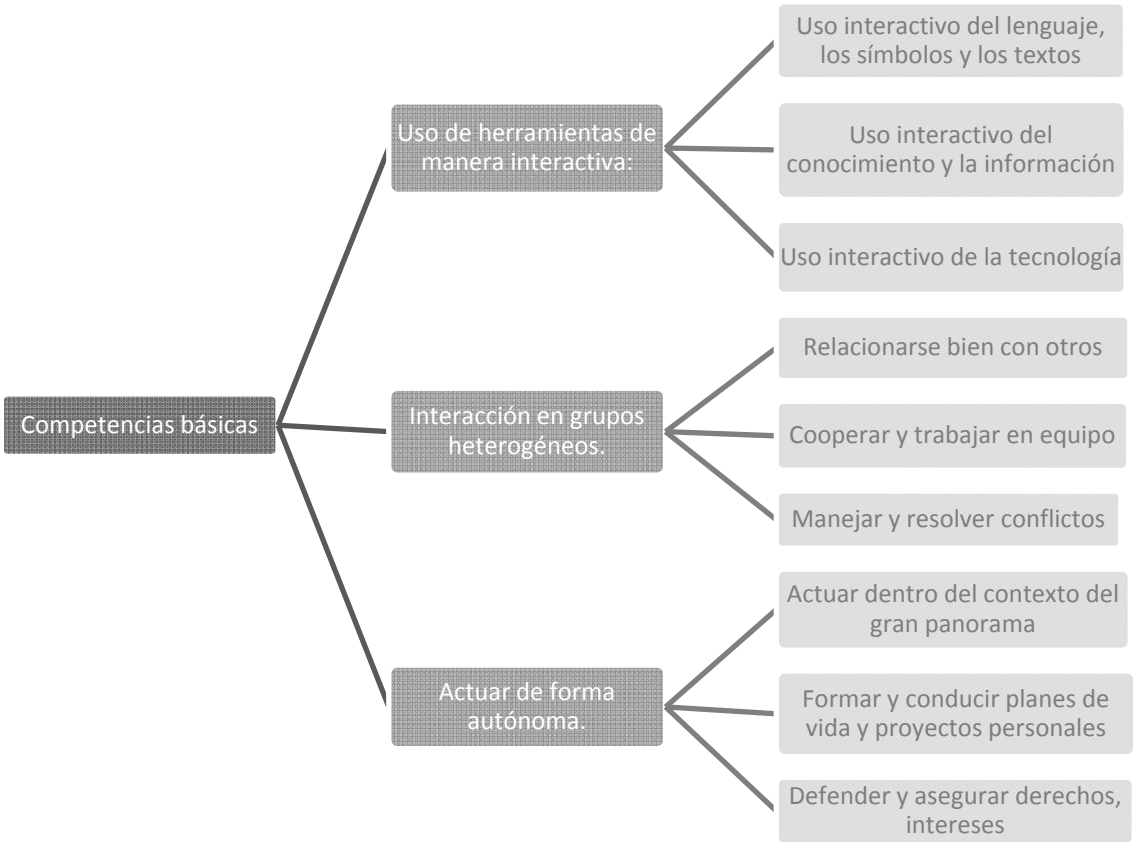


Figura 1. Estructura de competencias básicas en el proyecto DeSeCo.
Fuente: elaboración propia a partir de (OECD, 2005a).

En el caso de España, la incorporación de las competencias básicas en la enseñanza obligatoria, prevista en el artículo 6.1 de la Ley Orgánica de Educación (2006) y desarrollada en los decretos de enseñanzas mínimas de Educación Primaria y Educación Secundaria Obligatoria, ha marcado un cambio sustancial en todos los ámbitos de la Educación Obligatoria.

En los Reales Decretos de enseñanzas mínimas de Educación Primaria (Real Decreto 1513/2006, 2006) y Educación Secundaria (Real Decreto 1631/2006, 2006), es notable el protagonismo de las competencias básicas, de este modo, en ambos Reales Decretos, existe una alusión frecuente a las mismas, con el fin de contribuir a su efectiva puesta en marcha, destacando la influencia de este nuevo planteamiento en los criterios de evaluación, si el desarrollo de las competencias básicas es el objetivo del sistema educativo, debemos evaluar los resultados de los estudiantes en relación al nivel alcanzado en dichas competencias. En tales Reales Decretos, podemos comprobar cómo la definición de competencias recoge aspectos similares a los apuntados por la

recomendación del consejo europeo, de este modo, se destaca la importancia del desarrollo de las competencias básicas para la realización personal, la ciudadanía activa y la incorporación al mundo adulto (inclusión social). Pero además de definir qué se entiende por competencia en el marco normativo, también se expone en qué consiste la incorporación de dichas competencias en el currículo.

En este marco normativo se aprecia el carácter interdisciplinar y transversal de las competencias básicas, así como su potencial integrador, resultando una herramienta útil para aunar los distintos elementos del proceso de aprendizaje del estudiante. Por otra parte, se refleja el carácter práctico de las competencias así como su vinculación con el contexto en el que se aprenden y en el que se desarrollan posteriormente. Desde la LOE se aporta una visión de las competencias como eje vertebrador del proceso de enseñanza aprendizaje, implicadas tanto en el diseño curricular, en el desarrollo y en la evaluación. La LOMCE, conserva ésta filosofía, utilizando las competencias como eje vertebrador sin parearse a detallar aspectos concretos tratados en la normativa previa. Las competencias básicas que determinan estos Reales Decretos, tanto para la etapa de Educación Primaria como para la Educación Secundaria, están basadas en el marco de la propuesta realizada por la Unión Europea.

En el análisis realizado por Moya (2009, p. 22), en el que compara las competencias seleccionadas por el Ministerio de Educación y las recogidas en las recomendaciones del Consejo Europeo se puede apreciar que, a pesar de las grandes similitudes tanto en la denominación de las competencias, cómo en la definición semántica de las mismas, existen determinados aspectos en los que el Ministerio de Educación hace especial hincapié, y cómo adapta las recomendaciones globales propuestas para el conjunto de la Unión Europea a la realidad educativa de España. Llegar a una selección común de competencias básicas para el conjunto de la Unión Europea, es una tarea altamente compleja debido a la variedad de situaciones, intereses y realidades de cada uno de los países que forman la Unión. De este modo, en base a las recomendaciones generales, los organismos competentes han adaptado la propuesta a cada realidad y contexto educativo.

1.3 *Evaluación educativa: Resultados y Competencias.*

Si definimos los objetivos del sistema educativo en términos de competencias básicas, para evaluar el rendimiento de estudiantes, centros y sistemas debemos llevar a cabo una evaluación centrada en competencias. Es necesario analizar qué procedimientos podemos utilizar para evaluar el éxito del proceso educativo, y en qué grado o medida ha contribuido tanto al progreso de cada uno de los estudiantes cómo a la sociedad en su conjunto. Tras realizar un análisis del término competencia y posteriormente delimitar qué se entiende por competencia clave y cuáles son las competencias comúnmente aceptadas, nos encontramos ante nuevos interrogantes ¿qué se entiende por evaluación? ¿es posible evaluar competencias? ¿cómo se lleva a la práctica una evaluación basada en competencias? ¿qué diferencias existen frente a las evaluaciones tradicionales?

1.3.1 La evaluación educativa.

Evaluar supone estimar, apreciar o calcular el valor de algo, por tanto implica emitir un juicio para asignar o estimar dicho valor, para llegar a este juicio se analiza en qué medida se cumplen una serie de estándares o criterios, por tanto, la evaluación es esencialmente comparativa (De la Garza, 2004). La utilidad de la información que se obtiene de las evaluaciones depende de la posibilidad de comparar la misma con criterios de interés (postura adoptada ya por Scriven en 1967). En este sentido, tres serían básicamente los referentes de comparación:

a) Comparación-intra: el propio sistema comparado con él mismo en momentos previos.
b) Comparación-entre: el propio sistema comparado con sistemas de similares características.
c) Comparación-hacia: el sistema comparado con estándares externos establecidos como deseables.

La evaluación cobra cada vez más protagonismo en nuestra sociedad, de este modo, las prácticas evaluativas se han establecido con fuerza creándose sistemas de evaluación y acreditación tanto en el ámbito empresarial como en los servicios públicos

(Tiana, 2009). De la Orden (2000) señala que, "*en las sociedades más desarrolladas, la evaluación ocupa un lugar tan amplio y destacado que podríamos definir esta etapa histórica como «la era de la evaluación»*" (p. 381). La educación, no ha permanecido ajena a este boom evaluativo, y ha sido precisamente en el ámbito académico donde la evaluación ha encontrado uno de sus principales desarrollos, el incremento cuantitativo y cualitativo de las evaluaciones del sistema educativo es una prueba fehaciente de ello. Los sistemas educativos modernos, en sus orígenes, no contaban con sistemas de evaluación distintos a la evaluación realizada por parte del docente a sus estudiantes, la expansión de la educación básica hizo cada vez más necesaria la implementación de sistemas de evaluación que garantizaran la calidad de la enseñanza (Martinez Rizo, 2009). Contar con criterios objetivos para el análisis de la calidad de la educación, es un reto que han de asumir los sistemas educativos. Las evaluaciones estandarizadas pretenden afrontar dicho reto.

Tradicionalmente, se considera a Tyler el fundador de la evaluación educativa. En la definición de evaluación que elabora en el año 1950, apunta que la evaluación es el procedimiento que permite determinar hasta qué punto los objetivos educativos establecidos previamente han sido alcanzados (Tyler, 1950). En el año 1963 Cronbach aportará una nueva definición de evaluación, en la que destaca que la evaluación supone la recogida y uso de la información para tomar decisiones sobre un programa educativo (Cronbach, 1963) otorgando a la evaluación un papel esencial en el proceso de toma de decisiones en diferentes niveles. Por su parte Scriven (1967), apoya la idea de evaluación como parte del proceso de toma de decisiones, considerando la misma como ciencia de la valoración cuyo objetivo es determinar el valor de lo evaluado, apuntando además sus diferentes funciones (formativa y sumativa) y tipos (intrínseca y extrínseca).

En 1972, De la Orden señala que "evaluar en educación significa definir, determinar o valorar cualquier faceta de la estructura, el proceso o el producto educacional. Naturalmente, esta valoración se hace en función de unos criterios previamente establecidos, formulando el grado de aceptabilidad y adaptación de dicha faceta en referencia a ellos" (1972, pp. 202-203). Para Stufflebeam (1973) evaluar supone un proceso de planteamiento, recogida y obtención de información útil para la toma de decisiones. De este modo, de entre las definiciones clásicas de evaluación, podemos apuntar tres ideas fundamentales, en primer lugar el carácter procesual de la

evaluación, entendiendo ésta como una valoración (o medida) del cambio producido en el estudiante hacia unos objetivos deseados, en segundo lugar, el apoyo de la evaluación en la comparación y por último, su carácter dinámico como parte esencial de un proceso de mejora. La evaluación, no se limitará al establecimiento de un juicio comparativo o valoración, sino que irá más allá apuntando al «*feedback*» a partir de la información obtenida.

Tal y como veremos en el apartado 1.4, son numerosos los organismos nacionales e internacionales preocupados por la evaluación de sistemas educativos, las repercusiones políticas y mediáticas de dichas evaluaciones incrementan año a año, así como su magnitud y diversidad. Hablar de evaluación educativa, supone atender a la complejidad de la educación, los esfuerzos realizados en el área de evaluación están encaminados precisamente a atender dicha complejidad. El carácter intencional, objetivo y sistemático de la evaluación educativa, hace necesaria la utilización de escalas y criterios que sirvan como marco de referencia, por este motivo, evaluar implica medir, lo que supone una recogida objetiva de datos, medir es una condición necesaria para la evaluación pero no suficiente (Castillo & Cabrerizo, 2010), del mismo modo, García Ramos (1999) considera la medida como "base y elemento esencial e inseparable de los procesos de investigación y evaluación educativa" (pp. 198). La definición clásica de medida, propuesta por de Lord y Novick en 1968 apunta que medida es el procedimiento para la asignación de números (puntuaciones, mediciones) a propiedades específicas de unidades experimentales, con el fin de permitir caracterizar y preservar relaciones especificadas en un dominio de comportamiento (Lord & Novick, 2008).

Entre las múltiples funciones de la evaluación educativa, podríamos apuntar la formación, selección, certificación, ejercicio de la autoridad, mejora de la práctica docente; funciones relacionadas con la motivación y la orientación; funciones administrativas, académicas de promoción o de recuperación; de información y de retroalimentación, de control (Álvarez-Méndez, 2001). Estas múltiples funciones otorgan cierta confusión a la hora de hablar de evaluación, desdibujando sus objetivos y produciendo cierto "desconcierto" en algunos de los profesionales de la educación, situación que produce que ésta no sea bien acogida en ciertos casos.

Scriven (1967) introdujo dos funciones de la evaluación relativas a la evaluación de programas que, posteriormente, han sido generalizadas en otros ámbitos educativos y utilizadas con mucha frecuencia; la función formativa y la función sumativa. Como apunta Álvarez-Méndez (2001), la evaluación sumativa estaría próxima al movimiento de rendición de cuentas y, en consecuencia, relacionada estrechamente con los mecanismos propios del control administrativo, frente a este tipo de evaluación se situaría la formativa, encaminada a la valoración de la calidad y del valor intrínseco de los procesos de formación, sus funciones son esencialmente educativas. La función formativa, sería la función característica en el caso de la evaluación educativa, por tratarse de una evaluación que informa tanto del proceso cómo de los resultados.

De acuerdo con lo apuntado por Martín y Coll (2003) consideramos que es importante distinguir entre la función *pedagógica* y la función *social* de la evaluación "*La evaluación proporciona informaciones que son imprescindibles para reajustar el proceso de enseñanza-aprendizaje (función pedagógica); en cambio, acreditar los aprendizajes realizados por los alumnos mediante notas, certificaciones o títulos (función social) son decisiones asociadas a los resultados de la evaluación que responden a motivaciones no estrictamente pedagógicas*" (Martín & Coll, 2003, p. 50). Por tanto, a la hora de hablar de evaluación educativa debemos tener en cuenta las múltiples funciones encomendadas a la misma y ser conscientes de la importancia de conseguir los diferentes objetivos que debe cumplir, de este modo, conseguiremos procesos de evaluación completos y de calidad que permitan tanto la mejora del progreso individual de cada estudiante como de las instituciones educativas en su conjunto. Elena Martín (2009), considera que la evaluación estandarizada puede utilizarse con una finalidad diagnóstica o formativa o puede ponerse al servicio de fines asociados a procesos de acreditación de estudiantes o de instituciones.

Los criterios de clasificación de los tipos de evaluación son cuantiosos (Castillo & Cabrerizo, 2010) y en ocasiones introducen cierta confusión entre los profesionales. Desde la perspectiva de este trabajo, sin entrar a detallar los rasgos característicos y matices de dichas clasificaciones consideramos que, lo más importante, es tener en cuenta que la evaluación nos informa tanto a nivel de estudiante (nivel micro) como a niveles superiores: escuelas, áreas territoriales, comunidades, países, etc. (nivel macro).

Por otro lado, no debemos olvidar la relación existente entre el diseño de las acciones e intenciones educativas (currículo) y la evaluación educativa, puesto que en definitiva se trata de estudiar el logro de dichos objetivos educativos.

Si el diseño curricular, pone énfasis en los aspectos competenciales de desarrollo del individuo, la evaluación ha de responder a dicha necesidad. Imaginemos que nos proponemos el objetivo de conseguir no ganar peso, sería absurdo que para la comprobación de la consecución de dicho objetivo decidiésemos medir nuestra estatura, por no disponer de un instrumento adecuado (báscula) o por engañarnos a nosotros mismos acerca de la consecución de dicha meta. La evaluación educativa, por tanto, debe atender a tal necesidad considerando, en consecuencia, el rendimiento académico en términos de competencias. La función de las evaluaciones estandarizadas es comprobar si el sistema escolar está cumpliendo el encargo recogido en las intenciones educativas expresadas en el currículo, las administraciones, en su deber de garantizar el derecho de todos los estudiantes a una educación de calidad, deben supervisar el cumplimento de los objetivos educativos (Martín, 2009).

1.3.2 Evaluación y Competencias.

Volviendo a la normativa reguladora, vemos cómo en el preámbulo la Ley Orgánica de Educación (2006), se destaca la evaluación de competencias del alumnado, a través de la Evaluación General de Diagnóstico, como una de las innovaciones destacables de las incluidas en ella. Del mismo modo, la Ley Orgánica de Mejora de la Calidad Educativa (2013), pone énfasis en la evaluación de aprendizajes y competencias desde su preámbulo, apuntando a la justificación de ésta nueva normativa en base a los resultados de España en las evaluaciones internacionales, destacando las evaluaciones externas de final de etapa como una de las novedades más importantes destinadas a mejorar la calidad del sistema educativo y señalando las funciones del Instituto Nacional de Evaluación Educativa tanto en la evaluación general del sistema educativo como en las evaluaciones individualizadas (Evaluación Final de Educación Secundaria Obligatoria y la Evaluación Final de Bachillerato).

Uno de los aspectos que más cambios ha de asumir en el modelo curricular por competencias básicas es la evaluación (Castillo & Cabrerizo, 2010), como consecuencia, la forma tradicional de evaluar centrada en contenidos debe ser modificada. La evaluación tiene efectos retroactivos sobre el aprendizaje y la enseñanza, los estudiantes enfocan su aprendizaje de acuerdo a las características de la evaluación y al mismo tiempo, el docente tiende a evaluar de la misma forma que enseña (Monereo & Castelló, 2009), siguiendo esta idea, no es de extrañar que la preocupación de algunos docentes por la práctica de las denominadas "pruebas tipo PISA" haya desencadenado en ocasiones lo que podría considerarse "enseñanza tipo PISA", prueba del ineludible vínculo existente entre enseñanza, aprendizaje y evaluación. La LOMCE (2013) en el preámbulo VIII, recoge que las pruebas utilizadas en la evaluación deberán excluir la posibilidad de cualquier tipo de adiestramiento para su superación. Tal y como apunta Elena Martín (2009) las evaluaciones educativas resultan útiles al informar sobre el grado de consecución de las voluntades educativas, pero no son su origen ni el instrumento que debe utilizarse para legitimarlas.

La problemática en torno al tema de la evaluación por competencias ha sido tratada desde diferentes puntos de vista, Coll (2007) apunta que las competencias, como las capacidades, no son directamente evaluables, lo que exigiría un esfuerzo de selección de contenidos para trabajarlas y desarrollarlas, secuenciarlas, definir su progreso, sus niveles y establecer indicadores precisos de logro, con especial cuidado en la adecuación de las tareas que finalmente se le pide al alumno que realice. Evaluar por competencias supone emitir una valoración acerca de la capacidad que un alumno o alumna ha adquirido para dar respuesta a situaciones más o menos reales, problemas o cuestiones que podrían encontrar en su vida diaria. El estudiante deberá movilizar distintas combinaciones de conocimientos, habilidades, actitudes y destrezas con el fin de resolver la situación problema que se le plantean, desde la evaluación se reconocerá si el estudiante ha adquirido los niveles óptimos que le permitirán enfrentarse a una amplia variedad de situaciones (Zabala & Arnau, 2007). La evaluación de competencias debe asociarse a los conocimientos, destrezas y actitudes que se determinan en el currículo (Instituto de Evaluación, 2009).

Es preciso destacar que el desarrollo de las competencias no es una cuestión de todo o nada, es decir, las competencias admiten distintos grados de consecución y se

adquieren de manera progresiva. Por tanto, la evaluación no ha de consistir en decidir si el alumno ha adquirido o no determinada competencia, sino que es necesario establecer distintos niveles de desempeño, que permitan guiar el proceso de aprendizaje del estudiante, conduciéndole a un mayor dominio competencial de manera gradual. A tal efecto, es necesario contar con indicadores de logro pertinentes en cada competencia específica. Estos indicadores serán un referente para evaluar el grado de logro o dominio que un estudiante posee en relación a cada una de las competencias.

Evaluar por competencias, siempre implica evaluar su aplicación en situaciones y contextos próximos a los estudiantes. Los medios para evaluar competencias siempre han de ser aproximaciones a la realidad (Zabala & Arnau, 2007). Unido a ello, el término generalización haría referencia a la aplicación de conocimiento a distintos contextos, conocimiento "transcontextualizado" (Martin y Coll, 2003). Es decir, se tratará de un conocimiento aplicable a la gama más amplia posible de contextos particulares. Conviene pues planificar la evaluación de tal manera que se puedan observar comportamientos de los alumnos en tareas variadas y, en la medida de lo posible, que trasciendan los contextos meramente académicos. Otro rasgo que caracteriza a la evaluación por competencias es su carácter interdisciplinar. No es posible enseñar una competencia desde una sola materia, cómo tampoco es posible evaluarla desde una única materia.

En resumidas cuentas, al hablar de evaluación por competencias, se debe considerar que su finalidad es evaluar el desarrollo de conocimientos, actitudes y destrezas a distintos niveles, e informar a la administración educativa, a profesores, a padres y alumnos del grado de consecución de los objetivos educativos propuestos. Además, no podemos olvidar su carácter continuo, cuyo desarrollo admite diversos niveles o estadios, su contextualización y su interdisciplinariedad, que implican la adopción de nuevos procedimientos de evaluación que permitan contemplar dichas características.

Los diferentes organismos encargados de la realización de las Evaluaciones del Sistema Educativo, han utilizado distintas estrategias para concretar las evaluaciones por competencias. De este modo, el Instituto de Evaluación, en el marco de las evaluaciones diagnósticas, tiene en cuenta tres elementos clave en la evaluación por

competencias: las situaciones y contextos, los procesos y los conocimientos, destrezas y actitudes (Instituto de Evaluación, 2009).

Por otro lado, en el panorama internacional, también existe un interés global por lograr una evaluación de competencias que permita comparar los resultados obtenidos por cada uno de los países y, al mismo tiempo, poder analizar de manera comparativa la situación de cada país en relación al resto. En dichas evaluaciones, se pretende analizar el grado en que niños, jóvenes y adultos poseen los conocimientos y las destrezas que necesitan para enfrentarse a los desafíos de la vida. En apartados posteriores, analizaremos cuál es la propuesta de diferentes evaluaciones internacionales para la medida de las competencias y ante qué retos se han enfrentado a la hora de incluir éstas como resultados esperados del sistema educativo.

1.4 Evaluaciones internacionales a gran escala.

En los últimos años ha existido un incremento cuantitativo y cualitativo en la evaluación del sistema educativo, este incremento, ha estado auspiciado por diversos motivos, entre los que se encuentra, la necesidad de conocer cuál es el funcionamiento real de los centros educativos e identificar qué prácticas conducen a un mayor éxito de las escuelas. Las evaluaciones internacionales surgen, precisamente, como herramienta útil para identificar en qué países se están llevando a cabo las mejores prácticas y políticas educativas, así como para detectar posibles impedimentos o dificultades. Las evaluaciones internacionales, cuentan con instrumentos de medición estandarizados que se aplican en varios países simultáneamente, en determinados grupos o grados, y que recogen información sobre logros de aprendizaje, y sobre los factores contextuales que se considera pueden influir en el resultado académico de los estudiantes (Ferrer & Arregui, 2003).

Las primeras evaluaciones internacionales comienzan a surgir a finales de los años 50. La «*International Association for the Evaluation of Educational Achievement (IEA)*» se constituye como una entidad reconocida en el año 1967, no obstante, desde 1958, un grupo de académicos, educadores, psicólogos, sociólogos y psicómetras se reúnen en el instituto para la Educación de la UNESCO, con el fin de discutir los

problemas relacionados con la educación y la evaluación, siendo éste el germen que culmina con la constitución de la IEA. El primer estudio realizado por la IEA, conocido con el nombre «*Pilot Twelve-Country Study*» fue llevado a cabo en el año 1960 con la participación de 12 países. Poco a poco la IEA fue creciendo e incorporando mayor numero de instituciones, los éxitos de las primeras evaluaciones condujeron a la implementación de sucesivas evaluaciones con la incorporación de nuevas competencias y países. Las primeras evaluaciones desarrolladas por la IEA, junto con sus principales aportaciones y características, pueden observarse en la Tabla 3.

Tabla 3.

Primeros estudios realizados por la IEA

Año	Denominación	Características
1960	Pilot Twelve-Country Study	Participan 12 países. Se evalúa el rendimiento de estudiantes de 13 años de edad en las áreas de: matemáticas, comprensión lectora, geografía, ciencias y habilidades no verbales. La aportación más importante fue demostrar la posibilidad de llevar a cabo evaluaciones a gran escala con la participación de varios países. Se demostró la efectividad del trabajo conjunto de varios centros de investigación y la posibilidad de construir pruebas y cuestionarios comunes a través de un trabajo cross-cultural.
1964	First International Mathematics Study (FIMS)	Participan 12 países, se evalúan estudiantes de 13 años de edad y estudiantes de último año de educación secundaria. El estudio muestra el buen funcionamiento del predictor "oportunidades de aprendizaje" encontrándose diferencias en el desempeño de los estudiantes en función del mismo. El estudio también mostró los problemas de equidad entre grupos de estudiantes.
1970 1971	The Six Subject Survey	El estudio se lleva a cabo en estudiantes de 13 y 14 años. Extiende los objetivos de investigación a las áreas de ciencias, comprensión lectora, literatura, francés e inglés como segunda lengua y educación cívica. Los principales hallazgos fueron la identificación de nuevos predictores relacionados con el rendimiento de los estudiantes, tales como los intereses, la motivación, las actitudes, los métodos de enseñanza y las prácticas escolares.

Fuente: elaboración propia a partir de http://www.iea.nl/brief_history.html.

Posteriormente, la preocupación de la IEA por la medida del cambio educativo a lo largo del tiempo, lleva a finales de los años 70 a la elaboración de los primeros estudios periódicos, cuyo principal objetivo es obtener información acerca del progreso de los estudiantes. Los estudios de tendencias educativas cobran fuerza desde este momento, con las implicaciones metodológicas que ello conlleva. Este deseo, condujo a la implementación del «*Second International Mathematics Study (SIMS)*» que se llevó a cabo en 20 países en el período 1980-1982, así como el «*Second International Science Study (SISS)*» que se llevó a cabo en 24 países en 1983-1984. La aportación esencial de

estos estudios, es la demostración de que la repetición de un estudio en un periodo determinado, aporta a los países participantes información importante sobre las tendencias de sus niveles de logro, en este caso, en las áreas de matemáticas y ciencias. En permanente contacto con la realidad educativa, la IEA implementa nuevas evaluaciones relacionadas con otras preocupaciones sociales, de este modo, surge «*The longitudinal Preprimary Project (PPP*)» (1987–1989, 1992, y 1995–1997) que tiene como objetivo la evaluación del desarrollo cognitivo de los estudiantes que asisten a educación preescolar. Otros ejemplos de éste contacto con la realidad son los estudios «*Computers in Education Study (COMPED)*» (1989 y 1992), «*The Second Information Technology in Education Study Module 1 (SITES-M1)»* (1998–1999) y SITES-M2 (2001).

Del mismo modo, dentro de estas investigaciones internacionales pioneras, debemos destacar la realizada por el *Educational Testing Service* (ETS) titulada «*International Assessment of Educational Progress IAEP*» (1988), el estudió incluyó cinco países (entre los que se incluye España) y cuatro provincias canadienses, las áreas evaluadas fueron matemáticas y ciencias con estudiantes de 13 años de edad (Lapointe, Mead, & Philips, 1989). A partir de los años 90, otros organismos internacionales e instituciones comienzan a desarrollar evaluaciones con similares características. El importante papel de la IEA en el surgimiento y consolidación de los estudios internacionales de rendimiento es indudable, el diseño y desarrollo de programas de evaluación que incluyen distintos países, niveles educativos, áreas de conocimiento, competencias, etc. con un cuidadoso diseño metodológico y atendiendo a la realidad escolar, han contribuido enormemente a sentar las bases de las evaluaciones a gran escala. Posteriormente, analizaremos con mayor detalle los estudios que se mantienen en la actualidad, deteniéndonos en sus características específicas y aportaciones de los mismos.

Cada año, con la publicación de los resultados de diferentes evaluaciones internacionales del aprendizaje escolar en áreas como matemáticas, comprensión lectora o ciencias, se abre un gran debate en torno a las políticas y prácticas educativas de las entidades o países evaluados. Los medios de comunicación, se hacen eco durante un tiempo de dichos resultados, desgranando la información de los mismos en pequeñas dosis e interpretando los resultados presentes en los informes no siempre de la manera

más adecuada. Unido a ello, son numerosas las declaraciones de representantes políticos, sociólogos, politólogos, educadores, pedagogos, etc. acerca de tales resultados. Durante este periodo, el funcionamiento del sistema educativo es un tema recurrente en todos los medios de comunicación.

El informe con más repercusión mediática, política y social parece ser el informe del «*Programme for International Student Assessment* (PISA)», cuya trayectoria demuestra que, la publicación de sus resultados, produce un boom informativo en torno a los mismos, su influencia tanto en la opinión pública como en los gobiernos, crece con cada nueva edición.

De este modo, la prensa escrita y otros medios de comunicación introducen en sus contenidos una importante variedad de análisis de los resultados. A pesar de que el volumen informativo se mantiene relativamente estable durante varios meses, es en el mes posterior a la publicación de los resultados en el que se concentra el mayor interés. Si analizamos el volumen de noticias sobre el informe PISA, aparecidas en dos importantes diarios nacionales tras la publicación de los resultados de cada edición (2000/ 2003/ 2006/ 2009/ 2012) (Figuras 2 y 3) vemos como, durante el mes de diciembre (el informe se publica a principios de dicho mes), son numerosos los titulares dedicados a él. Al mismo tiempo, el interés parece ir creciendo con cada nueva edición, tal ascenso podría estar motivado tanto por el mayor conocimiento del proyecto PISA como por el aumento del valor de los datos aportados puesto que, mientras en el año 2000 solo se contaba con una primera medición, en los años 2003 y 2006 se cuenta con los primeros datos de tendencia, que permiten el análisis del progreso del sistema educativo en un sentido dinámico.

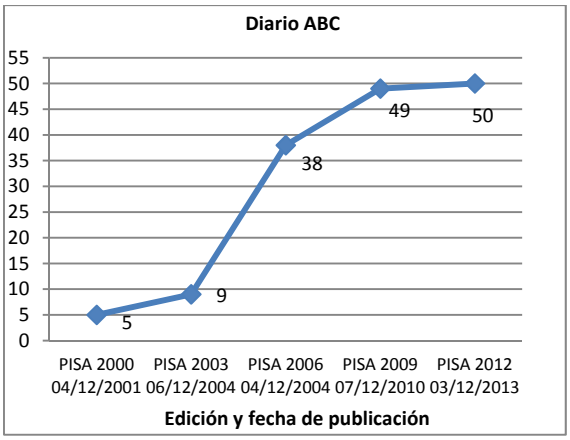


Figura 3. Número de artículos publicados en el diario ABC durante el mes posterior a la publicación del informe PISA por edición.

Fuente: elaboración propia.

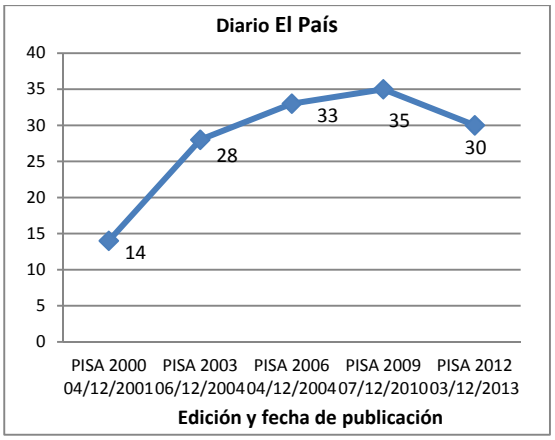


Figura 2. Número de artículos publicados en el diario El País durante el mes posterior a la publicación del informe PISA por edición.

Fuente: elaboración propia.

Unido a ello, cada año mayor número de países participa en evaluaciones internacionales del sistema educativo. Entre las razones que pueden llevar a los distintos países a tomar la decisión de participar en tales evaluaciones podrían estar:

- Conocer las experiencias educativas de otros lugares y comparar los resultados de éstas con los propios, ampliando el contexto en el que interpretar los resultados y enriqueciendo los mismos.
- Obtener información acerca de la influencia de factores familiares, contextuales y políticos en los resultados académicos.
- Interactuar con especialistas e instituciones internacionales.
- Promoción del debate educativo y fomento del interés de ciudadanos e instituciones en la enseñanza como consecuencia de la gran difusión de los resultados de este tipo de evaluaciones.
- Mejora de las propias prácticas evaluativas al contar con el asesoramiento de profesionales de ámbito internacional que permiten mejorar la calidad técnica y la eficiencia de los sistemas de evaluación.
- Incremento en la calidad de la información que reciben las administraciones educativas, al contar con una evaluación externa y comparable con otros sistemas, siendo necesario profundizar en el conocimiento de los informes transnacionales de evaluación del aprendizaje escolar (Acevedo, 2005).

- Informar a los países sobre la calidad relativa de sus sistemas educativos y, consecuentemente, sobre su competitividad en el mercado global de bienes y servicios (Ferrer & Arregui, 2003).

En definitiva, tal y como apuntan Lapointe, Mead y Philips (1989, p. 7) "*lo único que justifica la interrupción de la vida del estudiante y del profesional ocasionada por un estudio internacional es la mejora de la enseñanza. Los resultados deben proporcionar información a profesores, directores, políticos y contribuyentes que les ayude a definir las características de un rendimiento escolar de éxito y sugerir áreas que requieran una mejora o cambio*". A pesar de las bondades apuntadas anteriormente, las evaluaciones internacionales también presentan dificultades y limitaciones y han sido objeto de diversas críticas, entre las que podríamos considerar:

- Establecimiento de rankings entre los países. Efecto "liga de futbol" en el que se opta por presentar una ordenación de países sin llegar a una interpretación sustantiva de los resultados, como si de una competición deportiva se tratase.
- Interpretación aislada de los resultados sin tener en cuenta las situaciones de partida o los contextos en los que se desarrolla la actividad educativa.
- Participar en evaluaciones internacionales puede llevar a las instituciones educativas a no dar importancia a la consolidación de sistemas propios de evaluación, cuya información también resulta de gran importancia para la mejora del sistema educativo (Ferrer & Arregui, 2003).
- Limitaciones propias de los procesos de evaluación y medida. A pesar de contar con numerosos especialistas, no podemos olvidar las limitaciones intrínsecas al proceso de medición de los resultados académicos.
- La complejidad de los estudios, y la gran variedad de datos derivados de ellos, producen que la información que finalmente es difundida, resulte ambigua o contradictoria y que un mismo dato pueda ser utilizado para argumentos opuestos (en función de intereses particulares), o que sencillamente se malinterpreten los datos y se llegue a conclusiones erróneas (Husén, 1987).
- Las diferencias culturales entre países amenazan la validez de los estudios y limitan su comparabilidad. Diferencias entre naciones hacen

que las condiciones de aplicación no sean las mismas y pueden diferir los resultados como consecuencia de: familiaridad que los alumnos tienen con este tipo de pruebas estandarizadas, motivación de los estudiantes a la hora de responder la prueba y estilo de respuesta-formato de preguntas (Mislevy, 1995a).

- Los costos "explícitos" y "ocultos" que implican este tipo de evaluaciones pueden limitar la participación de los países más pobres (Goldstein, 2004).
- La metodología utilizada no es cultural o políticamente neutra, y está influida por las entidades organizadoras y financiadoras de dichos proyectos (Goldstein, 2004).

Numerosos autores han analizado las ventajas y limitaciones de la participación en evaluaciones internacionales y, como vemos, no existe un consenso claro en su determinación. Con el fin de maximizar los beneficios de este tipo de evaluaciones y evitar los posibles inconvenientes en las mismas, debemos tener en cuenta si se cumplen una serie de asunciones. Tal y como apunta Beaton (1999), existen ciertos requisitos o puntos clave que debemos analizar a la hora de juzgar la calidad o idoneidad de un estudio internacional: objetivos de la evaluación, diseño del estudio, definición de la población objetivo, muestreo, construcción de los instrumentos de medida, diseño de recogida de datos, codificación y depuración de datos y cálculo de los pesos muestrales tras dicha depuración, proceso de análisis de datos e informes elaborados a partir del estudio.

Las evaluaciones internacionales, con sus ventajas y limitaciones, tienen un fuerte impacto en las políticas educativas de todos los países participantes. Un ejemplo claro de dicho impacto podemos encontrarlo en el caso de los Estados Unidos, dónde los resultados obtenidos en los años 80 en las áreas de matemáticas y ciencias hicieron saltar la voz de alarma y los políticos llegaron a considerar que Estados Unidos era una "Nación en Riesgo" (National Commission on Excellence in Education, 1983), en dicha declaración se hacía referencia al riesgo que se sufría de perder liderato mundial si no se mejoraba el sistema de educación pública y las puntuaciones de los estudiantes en dichas áreas. Estos resultados llevaron a implementar una serie de reformas para conseguir ascender en el ranking de los sistemas educativos.

Estas evaluaciones internacionales son llevadas a cabo por diversas instituciones con características variadas (privadas, públicas, intergubernamentales, de cobertura mundial o regional). En la Tabla 4 se presenta un resumen de los principales organismos y las últimas evaluaciones implementadas por los mismos, con el fin de mostrar una panorámica general de las evaluaciones internacionales desarrolladas en los últimos años. Posteriormente analizaremos con más detalle cada una de ellas.

Tabla 4.

Evaluaciones educativas internacionales de mayor importancia

		ÚLTMAS EVALUACIONES
AGENCIA	International Association for the Evaluation of Educational Achievement (IEA)	Trends in Mathematics and Science Study (TIMSS)
		Progress in International Reading Literacy Study (PIRLS)
		International Civic and Citizenship Education Study (ICCS)
		Teacher Education and Development Study in Mathematics (TEDS-M)
	Organization for Economic Cooperation and Development (OECD)	Program for International Students Assessment (PISA).
		Teaching and Learning International Survey (TALIS)
		The Feasibility Study for the International Assessment of Higher Education Learning Outcomes (AHELO)
		Programme for the International Assessment of Adult Competencies (PIAAC)
	Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (LLECE).	Estudio Internacional comparativo sobre lenguaje, matemáticas y factores asociados, primera edición (PERCE).
		Segundo Estudio Regional Comparativo y Explicativo (SERCE)
		Tercer Estudio Regional Comparativo y Explicativo (TERCE)
	Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ)	SACMEQ I
		SACMEQ II
		SACMEQ III

Fuente: elaboración propia.

1.4.1 International Association for the Evaluation of Educational Achievement (IEA).

Tal y como veíamos al principio del presente epígrafe, la IEA surgió en el año 1958 y fue constituida formalmente en 1967. Es una asociación independiente, sin fines de lucro, cuyos miembros son universidades, institutos o agencias ministeriales dedicadas a la investigación sobre evaluación educativa, que representan el sistema educativo de su país (Instituto de Evaluación, 2010). En la actualidad está formada por miembros de más de 60 países, tanto la secretaría como la sede oficial están situadas en en Ámsterdam (Holanda) y el centro de proceso de datos e investigación en Hamburgo (Alemania).

Su objetivo es la realización de estudios a gran escala que permitan, por un lado, obtener resultados valiosos para cada uno de los países, y por otro, la comparación de las políticas y prácticas educativas entre países, con el fin de introducir mejoras en el proceso educativo. Las evaluaciones llevadas a cabo por esta organización son muy variadas e incluyen un gran abanico de temas tales como: matemáticas, ciencias, leguaje, educación cívica o comprensión lectora (Mullis, Martin, Kennedy, Trong, & Sainsbury, 2009). Las evaluaciones llevadas a cabo por la IEA se muestran en la Tabla 5, no obstante, en este apartado nos centraremos en el análisis de las evaluaciones que más impacto tienen en la actualidad (TIMSS, PIRLS y ICCS).

Tabla 5.

Estudios realizados por IEA.

Periodo	**Evaluación**
1959 — 1960	The Pilot Study
1964	First International Mathematics Study (FIMS)
1970 — 1971	The Six-Subject Study • Science • Reading Comprehension • Literature • English as a Foreign Language • French as a Foreign Language • Civic Education
1980 — 1982	Second International Mathematics Study (SIMS)
1981 — 1983	Classroom Environment
1983 — 1984	Second International Science Study (SISS)
1985	Written Composition
1990 — 1991	Reading Literacy
1989 — 1992	Computers in Education/Information Technology (COMPED)
1995	Languages in Education Study (LES)
1994 — 1995	Third International Mathematics and Science Study (TIMSS 1995)
1998 — 1999	Second Information Technology in Education Study Module 1 (SITES-M1)
1998 — 1999	Third International Mathematics and Science Study Repeat (TIMSS-R 1999)
1996 —2000	Civic Education Study (CIVED): • Phase II 1999 — 2000 • Phase I 1996 — 1997
2000 — 2001	Second Information Technology in Education Study Module 2 (SITES-M2)
2001	Progress in International Reading Literacy Study (PIRLS 2001)
1986 — 2003	Pre-Primary Education Project (PPP): • Phase III 1993 — 2003 • Phase II 1989 — 1993 • Phase I 1986 — 1994
2002 — 2003	Trends in Mathematics and Science Study (TIMSS 2003)
2005 — 2006	Second Information Technology in Education Study (SITES 2006)
2005 — 2006	Progress in International Reading Literacy Study (PIRLS 2006)
2006 — 2007	Trends in Mathematics and Science Study (TIMSS 2007)
2007 — 2008	TIMSS Advanced 2008
2007 — 2008	Teacher Education and Development Study in Mathematics (TEDS-M)
2008 — 2009	International Civic and Citizenship Education Study (ICCS)
2010 — 2011	Trends in Mathematics and Science Study (TIMSS 2011)
2010 — 2011	Progress in International Reading Literacy Study (PIRLS 2011)
2012 — 2013	International Computer and Information Literacy Study (ICILS 2013)

Fuente: elaboración propia a partir de
http://www.iea.nl/fileadmin/user_upload/IEA_Documents/IEA_Brochure.pdf.

a) Trends in International Mathematics and Science Study (TIMSS).

A pesar de que en la actualidad todos estamos familiarizados con el acrónimo TIMSS, la idea inicial era nombrar a estos estudios cíclicos haciendo referencia a su edición, así, en un primer momento se pretendía utilizar los términos «*first*», «*second*» y «*third*» y los acrónimos FIMSS, SIMSS y TIMSS, sin embargo, la cuarta edición, implicaría nuevamente el uso del acrónimo FIMSS, lo que finalmente llevó al cambio de nombre, así en 2002-03 pasó a denominarse «*Trends in International Mathematics and Science Study*»).

Como puede apreciarse en la Tabla 5, la primera aplicación del estudio TIMSS se llevó a cabo en el año 1995 (Beaton et al., 1996). La apuesta por las evaluaciones de carácter periódico por parte de la IEA condujo a la aplicación del proyecto TIMSS cada cuatro años, de este modo, la siguientes aplicaciones tuvieron lugar en los años 1998-99 (TIMSS *Repeat*) (Martin et al., 2000), 2002-03 (Mullis, Martin, González, & Chrostowski, 2004), 2006-07 (Mullis, Martin, & Foy, 2008) y 2010-11 (Mullis, Martin, Ruddock, O´Sulivan, & Preuschoff, 2009). La evaluación de logros en ciencias y matemáticas se realiza en alumnos de 4º curso de Educación Primaria y 2º curso de Educación Secundaria Obligatoria. El ciclo regular de cuatro años del estudio TIMSS aporta a los países participantes una información muy útil para la realización de comparaciones acerca del progreso de sus estudiantes en matemáticas y ciencias (Mullis et al., 2004). Además, TIMSS recoge información acerca de cómo tiene lugar el aprendizaje de las matemáticas y las ciencias en cada país, obteniendo dicha información a partir de preguntas a estudiantes, profesores y directores, en 2015, incluirá un cuestionario dirigido a las familias de los estudiantes de 4º curso. La información aportada por TIMSS, tiene en cuenta los contextos de escolarización.

Para aquellos países que participan en TIMSS desde el año 1995, el estudio TIMSS 2011 representó la quinta evaluación de tendencia. Cada año la participación ha sido mayor, si nos fijamos en la Figura 4 vemos cómo la diferencia en participantes desde el año 1995 a 2011 es de 19 países, si sumamos las entidades de comparación (localidades que participan con independencia de su país) la diferencia asciende a 33. En el año 2015 está previsto que tenga lugar la sexta evaluación TIMSS que completará un ciclo de evaluación de 20 años.

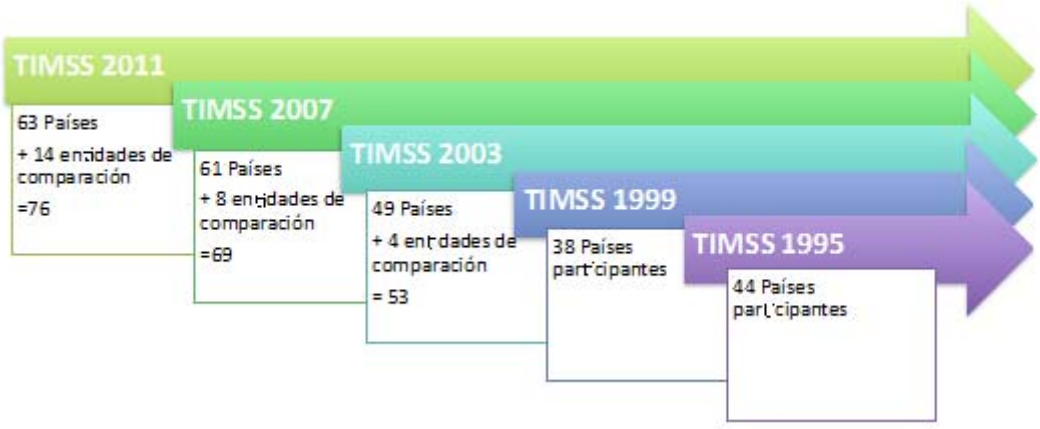


Figura 4. Evaluación TIMSS años de evaluación y número de países participantes.
Fuente: elaboración propia.

Los resultados de las evaluaciones TIMSS se realizan en dos volúmenes complementarios, uno para el área de matemáticas y otro para el área de ciencias, de este modo, para cada edición de TIMSS contamos con dos informes de resultados. TIMSS utiliza el currículo como principal elemento organizador, el modelo curricular de TIMSS tiene tres aspectos: el currículo pretendido, el aplicado y el obtenido (Mullis, Martin, Ruddock et al., 2009).

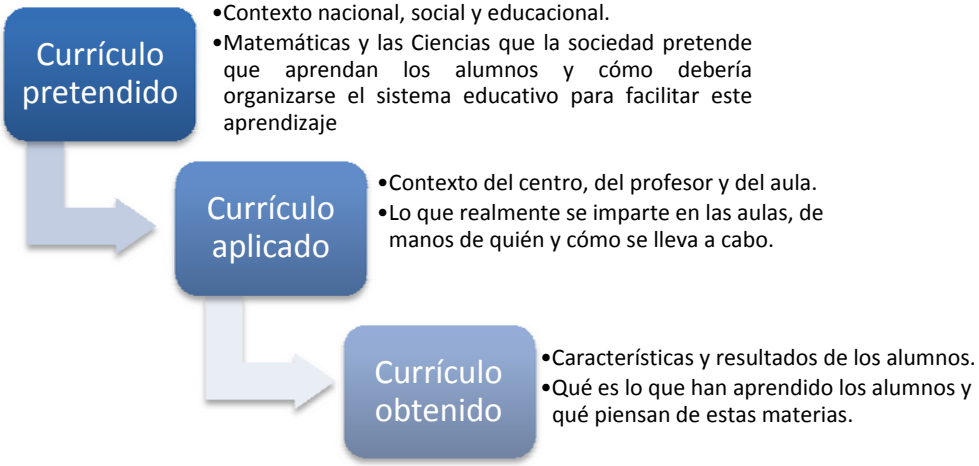


Figura 5. Modelo Curricular de TIMSS.
Fuente: adaptado de (Mullis, Martin, Ruddock et al., 2009, p. 16).

Regularmente, se revisan los marcos teóricos de la evaluación con el fin de conseguir una evaluación contextualizada que permita a los países revisar y analizar los resultados obtenidos de una forma más coherente, permitiendo su evolución gradual en función de las necesidades de futuro (Mullis, Martin, Ruddock et al., 2009).

Al ser un estudio que pretende la evaluación de tendencias en el rendimiento educativo a lo largo del tiempo, resulta crucial que no se produzcan cambios sustanciales en el contenido de las pruebas, puesto que esto, podría atentar contra la comparabilidad (tal y como veremos posteriormente). Por tanto, este trabajo conjunto de diferentes especialistas, deberá moverse en el delicado equilibrio entre la adaptación al cambio y la continuidad con las evaluaciones anteriores. Así, un análisis detallado de los dominios (de contenido y cognitivo) y las áreas temáticas evaluadas en el proyecto TIMSS a través del estudio de sus marcos de referencia. para cada curso (4ºEP y 2ºESO) y materia (Matemáticas y Ciencias), muestran las leves modificaciones implementadas en cada nueva edición. De este modo en lo relativo a los dominios de contenido de 4º y 2ª, no existen diferencias destacadas entre las ediciones de 2007, 2011 y 2015, las mayores diferencias las encontramos entre las ediciones de 1995, 1999 y 2003 y las posteriores, atendiendo como era de esperar a la natural evolución de los sistemas educativos.

Por otro lado, se debe tener en cuenta que, la evaluación TIMSS, no es una evaluación exclusiva de contenidos y en efecto, para poder responder correctamente a las cuestiones, el estudiante debe dominar el contenido matemático o científico así como poner en marcha una serie de destrezas cognitivas (Mullis, Martin, Kennedy et al., 2009). A continuación presentamos una tabla con los dominios cognitivos evaluados y el peso específico otorgado a cada uno de ellos para el caso de 2º y 4º de Educación Primaria (Tabla 6).

Tabla 6.

Destrezas cognitivas en TIMSS 4ºEP y 2º

	Matemáticas		Ciencias	
Destrezas cognitivas	4º EP	2ºESO	4º EP	2ºESO
Conocimiento	40%	35%	40%	35%
Aplicación	40%	40%	40%	35%
Razonamiento	20%	25%	20%	30%

Fuente: (Mullis & Martin, 2013).

En las distintas ediciones de TIMSS observamos diferencias en las destrezas cognitivas consideradas (Tabla 6), en este sentido, a partir de la evaluación 2003 es

cuando se comienzan a evaluar las destrezas cognitivas de forma sistemática, presentándose en este sentido un antes y un después en las evaluaciones TIMSS.

En cuanto al diseño de los cuadernillos para la evaluación del rendimiento en las áreas de matemáticas y ciencias, TIMSS diseña 14 modelos en los que se combinan diferentes bloques de preguntas, cada estudiante, responde a un solo cuadernillo. Cada bloque de preguntas aparece en dos cuadernillos con la finalidad de permitir la vinculación de las respuestas entre estudiantes, de este modo, se utilizan métodos de escalamiento basados en la teoría de la respuesta al ítem (calibración conjunta TRI), para hacer posible la elaboración de una escala integral del rendimiento para toda la población de estudiantes, a partir de las respuestas combinadas de los distintos estudiantes a los cuadernillos que se les asigna (Mullis, Martin, Kennedy et al., 2009). El objetivo con este tipo de diseño es asegurar la validez de contenido de la prueba sin que implique una excesiva sobrecarga en los estudiantes, ya que si se desea atender a las matrices de especificaciones elaboradas, el número de ítems a los que tendría que dar respuesta cada estudiante sería excesivo, incurriéndose en otros errores de medida derivados de factores como el cansancio del alumnado. Este tipo de diseño, forma parte de los denominados "muestreo matricial de ítems" que permiten la minimización del tiempo de respuesta de los estudiantes y la cobertura completa del dominio evaluado (Childs & Jaciw, 2003), basándose en la respuesta, por parte de submuestras de estudiantes a subconjuntos de ítems pertenecientes al conjunto "global" de ítems utilizados para medir el dominio. Por medio de un sistema matricial, cada bloque de ítems aparece emparejado con los restantes bloques en cada cuadernillo, permitiendo su posterior calibración conjunta.

Cada bloque está constituido por unos 10 -14 reactivos en el caso de educación primaria y entre 12 y 18 en el caso de secundaria. En las ediciones de 2011 y 2015 se cuenta con 28 bloques por curso, 14 de ciencias y 14 de matemáticas, 8 de los 14 bloques de matemáticas y 8 de los 14 bloques de ciencias se utilizan para anclar las pruebas de 2011 y 2015. Los bloques restantes fueron liberados para el conocimiento general de la comunidad educativa. Por tanto, de los 28 bloques presentes en la evaluación TIMSS 2015, se cuenta con 16 bloques de ítems de anclaje (o tendencia) (8 de Matemáticas y 8 de Ciencias) y 12 nuevos bloques (Mullis & Martin, 2013). Cada cuadernillo para el alumnado consta de cuatro bloques de ítems: dos bloques de ítems

de Matemáticas y dos bloques de ítems de Ciencias. Con el objetivo de evitar sesgos en la medida del rendimiento para alguna de las dos materias debidos a la disposición de las preguntas dentro de la prueba, en la mitad de los cuadernillos, los dos bloques de Matemáticas son los primeros, y en la otra mitad son los bloques de ciencias los que aparecen en primer lugar (ver Tabla 7). Por otro lado, la mayor parte de cuadernillos, combinan tanto ítems de anclaje con la evaluación 2011 como ítems nuevos (ver Tabla 7 en la que los bloques numerados de manera impar hacen referencia a los bloques de anclaje con la evaluación 2011). Los tipos de preguntas utilizados son opción múltiple y respuesta construida (Mullis & Martin, 2013).

Tabla 7.

TIMSS 2011. Diseño de cuadernillos

Cuadernillo	Primera Parte		Segunda Parte	
1	M01	M02	S01	S02
2	S02	S03	M02	M03
3	M03	M04	S03	S04
4	S04	S05	M04	M05
5	M05	M06	S05	S06
6	S06	S07	M06	M07
7	M07	M08	S07	S08
8	S08	S09	M08	M09
9	M09	M10	S09	S10
10	S10	S11	M10	M11
11	M11	M12	S11	S12
12	S12	S013	M12	M13
13	M13	M14	S013	S14
14	S14	S01	M14	M01

Fuente: (Mullis & Martin, 2013, p. 91).

Otro aspecto destacado de las evaluaciones TIMSS es la información recogida a través de los diferentes cuestionarios de contexto (Tabla 8). Dichos cuestionarios se aplican tanto alumnos como a profesores, directores o jefes de estudio de los centros. Del mismo modo, los países participantes también completan un cuestionario con preguntas relativas al contexto y el currículo nacional relativo a la enseñanza de las matemáticas y las ciencias. Esto permite realizar una interpretación contextualizada del resultado de los distintos países (Mullis, Martin, Ruddock et al., 2009).

Tabla 8.

Información de Contexto TIMSS. Áreas y factores asociados

Áreas contextuales y factores asociados.				
Contextos nacionales y comunitarios	Centros	Aula	Características y actitudes de los alumnos	Familia
✓Demografía y recursos. ✓Organización y estructura del sistema educativo. ✓El currículo de matemáticas y ciencias.	✓Características del centro ✓Organización escolar para la enseñanza. ✓Clima escolar. ✓Profesorado. ✓Recursos del centro. ✓Participación de los padres.	✓Formación y desarrollo del profesorado. ✓Características del profesor. ✓Características de la clase. ✓Materiales y tecnología para la instrucción. ✓Temas del currículo que se enseñan. ✓Actividades de instrucción. ✓Evaluación.	✓Demografía y antecedentes domésticos de los estudiantes. ✓Actitudes hacia el aprendizaje de las matemáticas y las ciencias.	✓Contexto Familiar. (4ºEP 2015) ✓Experiencias de aprendizaje temprano (4ºEP2015)

Fuente: elaboración propia a partir de (Mullis, Martin, Ruddock et al., 2009).

En TIMSS se abarcan cuatro grandes áreas con sus respectivos factores asociados (Tabla 8). Por primera vez, en la evaluación de 2015, se incluye un cuestionario destinado a las familias de estudiantes de 4º curso, en el que se plantean preguntas relativas al entorno familiar y experiencias de aprendizaje tempano.

Desde su primera edición, en el año 1995, TIMSS utiliza métodos de imputación de respuesta basados en la Teoría de la Respuesta al Ítem denominados "valores plausibles" para estimar los puntajes de logro. Dicha metodología, permite la asignación de puntuaciones completas en los dominios de matemáticas y ciencias a pesar de que los estudiantes sólo contestan a un sub conjunto de ítems y no a todos los reactivos que forman parte de la evaluación. Por otro lado, para aumentar la fiabilidad de las estimaciones de puntuación de los sujetos, TIMSS combina en dicha estimación tanto las respuestas de los estudiantes a los ítems de su prueba cómo las variables asociadas al «*background*» consideradas variables que condicionan el rasgo (Olson, Martin, & Mullis, 2008). Tal y como apunta Wu (2005), los valores plausibles son una imputación múltiple de la variable latente no observada para cada sujeto evaluado, podrían considerarse una representación del rango de habilidad que el estudiante razonablemente podría tener, dado un determinado patrón de respuesta condicionado por un conjunto de variables referidas a los antecedentes socio-demográficos del alumno. De este modo, para cada estudiante, se obtiene una distribución de habilidad,

un rango de posibles valores, en lugar de una habilidad concreta. Los valores plausibles se extraen de manera aleatoria de la distribución a posteriori del estudiante, no siendo apropiados para ser utilizados con el fin de analizar el progreso individual de los estudiantes (Wu, 2005).

En TIMSS se utilizan tres procedimientos diferentes de TRI en función del tipo de ítem y las características de puntuación del mismo. Cada uno es una "variable latente" dentro del modelo, que describe la probabilidad de que un estudiante responda de una manera específica a un determinado ítem en términos de la competencia de dicho estudiante (que es una variable no observada) y de varias características o parámetros de los ítems. El modelo de tres parámetros es utilizado en ítems de opción múltiple, en los que la respuesta es correcta o incorrecta. El modelo de dos parámetros, se utiliza en los ítems de respuesta construida, en los que se consideran dos opciones de respuesta, correcta o incorrecta. En ambos casos, nos encontramos ante situaciones dicotómicas (correcto/incorrecto), sin embargo, es necesaria la utilización de un modelo de crédito parcial para ítems politómicos de respuesta construida, en los que existen más de dos niveles de puntuación u opciones de respuesta (Olson, Martin, & Mullis, 2008). En lugar de realizar una estimación puntual, se realiza una distribución a posteriori de valores para cada sujeto con sus probabilidades asociadas, de esta distribución se extraen aleatoriamente 5 valores denominados, tal y como indicábamos anteriormente «valores plausibles».

Por otro lado, debemos destacar el estudio «TIMSS *advanced*» llevado a cabo por primera vez en el año 1995 en 16 países, cuyo objetivo es obtener información acerca de la preparación en matemáticas y física de los estudiantes del último año de educación secundaria (Garden, et.al, 2006). Esta extensión de TIMSS se ha realizado en los años 1995 y 2008, estando prevista su próxima aplicación en el año 2015.

Las evaluaciones TIMSS y PIRLS (de la que hablaremos en el siguiente epígrafe) resultan complementarias. Al ser realizadas ambas en 4º curso de Educación Primara, permiten a los países que participan en ambas contar con información sobre estas tres áreas tan importantes en los currículos oficiales. El año 2011 resultó especialmente importante para la evaluación internacional en 4º curso, ya que, el ciclo

de cuatro años de TIMSS está alineado con el ciclo de cinco de PIRLS, coincidiendo ambos en el año 2011 (Mullis, Martin, Kennedy et al., 2009).

b) Progress in International Reading Literacy Study (PIRLS).

Volviendo a la Tabla 5, vemos cómo en el año 1970 la IEA realizó, dentro de su proyecto «*The Six-Subject Study*» su primera evaluación sobre comprensión lectora, comprendida dentro de las 6 áreas evaluadas en el estudio. La evaluación fue realizada en 15 países. En 1991 se evalúa el área de comprensión lectora de manera independiente, reconociendo la importancia de este área en el aprendizaje escolar, dando lugar al estudio «*Reading Literacy*», en el que se evaluó la alfabetización en comprensión lectora en 32 países. Diez años después, PIRLS 2001, supone la continuación de éstos estudios, formando parte de un ciclo de evaluaciones que se realiza cada 5 años (aunque la idea inicial era su realización en ciclos de 4 años con evaluaciones en 2005 y 2009) (Campbell, Kelly, Mullis, Martin, & Sainsbury, 2001). Las ediciones de PIRLS por tanto son 2001 (Mullis, Martin, González, & Kennedy, 2003), 2006 (Mullis, Martin, Kennedy, & Foy, 2007) y 2011 (Mullis, Martin, Kennedy et al., 2009), así como la planificada para el año 2016 (Mullis & Martin, 2015).

PIRLS supone un importante estudio internacional sobre progreso en comprensión lectora, evaluando el rendimiento de los estudiantes de entre 9 y 10 años de edad (lo que correspondería con 4º curso de Educación Primaria en la mayoría de los sistemas educativos), aportando a los países una información de gran utilidad, comparable a nivel internacional, acerca del grado de dominio de la lectura tras 4 años de escolaridad (Mullis, Martin, Kennedy et al., 2009). En PIRLS la atención se centra en los años de escolaridad del estudiante y no en su edad cronológica, de este modo, la variabilidad internacional en el inicio de la escolaridad obligatoria, produce algunas diferencias en la edad de los estudiantes evaluados aunque, como bien decíamos al inicio, en la mayoría de los casos corresponde con una edad de 9-10 años. Además se recoge información relevante en relación a aquellos factores contextuales relacionados con el aprendizaje de la lectura tales como el apoyo familiar, escolar y las experiencias y políticas educativas llevadas a cabo en el aprendizaje y la enseñanza de la lectura (Mullis et al., 2003). Unido a ello, para los países que participaron en las ediciones de 2001, 2006 y 2011, PIRLS supone una riquísima fuente de información para evaluar las

tendencias de aprendizaje en el área, al ser una evaluación realizada cada cinco años la información se va enriqueciendo en cada ciclo. La evaluación PIRLS permite obtener información acerca de los siguientes aspectos:

- Rendimiento en comprensión lectora de los estudiantes tras cuatro años de escolarización (9-10 años). Seguimiento de las tendencias en el área para cada uno de los países.
- Competencias de los estudiantes en relación a los objetivos y estándares relacionados con la comprensión lectora.
- Importancia del entorno familiar y vías de fomento de la comprensión lectora.
- Organización, tiempo y materiales para aprender a leer en la escuela.
- Enfoques curriculares y métodos de enseñanza de la lectura.

Treinta y cinco países participaron en la evaluación PIRLS 2001 (ver Figura 6) (entre los que no se encuentra España), nueve de los cuales habían participado en el estudio llevado a cabo por IEA en el año 1991 (Grecia, Hungría, Islandia, Italia, Nueva Zelanda, Singapur, Eslovenia, Suecia y Estados Unidos), para estos nueve países se realizó una comparación acerca de su progreso en comprensión lectora con los datos aportados por las dos mediciones, realizando un estudio de tendencias. Para hacer posible dicha comparación, estos nueve países, contestaron a la misma versión de la prueba que en 1991, puesto que los cambios en el número de aspectos evaluados en las pruebas de 2001 no permite realizar comparaciones directas de estos resultados con los resultados de 1991 (Martin et al., 2003). En el año 2006 fueron evaluados 41 países (45 entidades si se tienen en cuenta los dos sistemas educativos de Bélgica y las 5 provincias canadienses (Mullis et al., 2007)), incluyéndose por primera vez a España. En la tercera evaluación, llevada a cabo en 2011, el número de países participantes asciende a 57 (48 países y 9 entidades de comparación) y se incorpora la evaluación pre-PIRLS, cuyo objetivo es medir los mismos aspectos que PIRLS pero en un nivel menor de desempeño, que indique las competencias básicas previas a los dominios requeridos en PIRLS (Mullis, Martin, Kennedy et al., 2009).

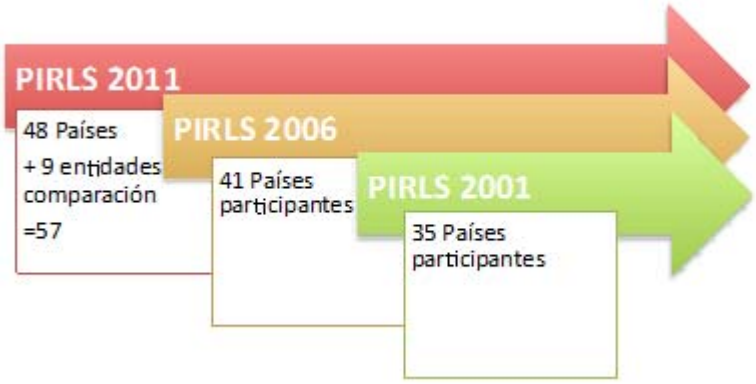


Figura 6. Evaluación PIRSLS años de evaluación y número de países participantes.
Fuente: elaboración propia.

La próxima edición de la evaluación PIRLS está programada para el año 2016 (Mullis & Martin, 2015). Las entidades que han participado en 2001, 2006 y 2011 tienen la oportunidad de medir el progreso en comprensión lectora a lo largo de cuatro puntos temporales. Para PIRLS, la habilidad lectora es "*la habilidad para comprender y utilizar el lenguaje escrito requerido por la sociedad y/o apreciado por el individuo. Los lectores son capaces de construir el significado de gran variedad de textos. Leen para aprender, para participar en comunidades de lectura en el ámbito escolar y en la vida cotidiana y leen por entretenimiento*" (Mullis & Martin, 2015, p. 12). Basándose en esta definición de partida, la evaluación se centra en tres aspectos relacionados con la competencia lectora: propósitos de la lectura, procesos de comprensión, comportamientos y actitudes ante la lectura.

Los dos primeros aspectos son evaluados a través de una prueba escrita de comprensión lectora, el último se evalúa por medio de distintos cuestionarios. Los propósitos de la lectura serían "lectura como experiencia literaria" y "lectura para la adquisición y uso de la información". En cuanto a los procesos de comprensión se tienen en cuenta los diferentes modos en que el sujeto construye el significado: selección y recuperación de información, realización de inferencias, interpretación e integración de información, examen y evaluación de contenido, lenguaje y elementos textuales. La prueba permite, a través de estas cuatro dimensiones, y combinando ambos tipos de propósitos, que los estudiantes manifiesten distintos niveles de habilidades y destrezas a partir de textos escritos. El peso específico otorgado a cada una de estas dimensiones puede observarse en la Tabla 9.

Tabla 9.

Propósitos y procesos de la Evaluación PIRLS

Propósitos de la lectura.	
Experiencia literaria	50%
Adquisición y uso de la información	50%
Procesos de comprensión	
Selección y recuperación de información	20%
Realización de inferencias directas	30%
Interpretación e integración de información	30%
Examen y evaluación de contenido, lenguaje y elementos textuales	20%

Fuente: elaboración propia.

El proceso de elaboración de pruebas es una de las fases más complejas de la evaluación, contando con un amplio grupo de especialistas encomendado a tal fin. Uno de los principales retos en la elaboración de las mismas es la eliminación de los posibles sesgos culturales, así como la inclusión de los aspectos curriculares incluidos en los diferentes países. Desde la edición de 2001, fueron elaborados un amplio conjunto de ítems, aproximadamente la mitad de los cuales eran de opción múltiple y la otra mitad de respuesta construida (Mullis et al., 2003). Para el cálculo de la puntuación final de la prueba, cada respuesta correcta corresponde a un punto (en ambos tipos de pregunta), sin embargo, aquellas preguntas que requieren una respuesta más elaborada, se califican de acuerdo a un modelo de crédito parcial, en dicho modelo, una respuesta completamente correcta podría conllevar dos o tres puntos. En el cuadernillo de preguntas de cada estudiante se consigna cuál es la puntuación de dicha pregunta y el nivel de respuesta requerido (Mullis et al., 2003).

De acuerdo a los dos propósitos de lectura considerados y atendiendo a los cuatro procesos, se elaboró un gran banco de ítems, cada conjunto de ítems hace referencia a un texto, o bien relacionado con la experiencia literaria, o con la adquisición y uso de información. Puesto que no es viable aplicar un número elevado de ítems a cada uno de los sujetos, y la complejidad del área evaluada así lo requeriría, se optó por elaborar un diseño de recogida de información utilizando «cuadernillos rotativos» con el fin de asegurar la validez de constructo de las pruebas (proceso con las mismas características técnicas que el realizado por TIMSS descrito previamente).

Cada cuadernillo está compuesto por dos textos con sus correspondientes ítems (bloques) además, la rotación y uso de algunos textos (bloques) año tras año, permite

construir la escala vertical, posibilitando el análisis del progreso en comprensión lectora entre aplicaciones, de este modo, seis de los diez bloques utilizados en 2011 fueron utilizados en anteriores aplicaciones, dos de los cuales se utilizaron en PIRLS 2001 y PIRLS 2006 y los cuatro restantes en PIRLS 2006. Por tanto, de los 10 cuadernillos utilizados en 2011, tan sólo 4 bloques son utilizados por primera. La construcción de la escala vertical de PIRLS, se basa en el anclaje por medio de los bloques de preguntas agrupados en torno a un único texto, en todas las ediciones se sigue este proceso, a excepción de la comparativa entre 1991 y 2001 ya que, la diferencia en los contenidos evaluados entre ambas aplicaciones hizo necesaria la aplicación de la misma prueba que en 1991 para aquellos países que deseaban comparar sus resultados. La escala de rendimiento de PIRLS, establecida desde el año 2001 tiene una media de 500 y una desviación típica de 100.

Tabla 10.

Textos incluidos en los cuadernillos PIRLS 2011

Propósito de la lectura	Bloques de textos				
Literario	L1	L2	L3	L4	L5
Informativo	I1	I2	I3	I4	I5

Fuente: elaboración propia.

A partir de estos bloques, son elaborados los cuadernillos a los que dará respuesta el alumno. Cada cuadernillo está compuesto por dos textos. El diseño para la elaboración de los cuadernillos en 2011 fue el que aparece en la Tabla 11.

Tabla 11.

Diseño de cuadernillos PIRLS 2011

Cuadernillo	Parte 1	Parte 2
1	L1	L2
2	L2	L3
3	L3	L4
4	L4	I1
5	I1	I2
6	I2	I3
7	I3	I4
8	I4	L1
9	L1	I1
10	I2	L2
11	L3	I3
12	I4	L4
Reading	L5	I5

Fuente: Adaptado de (Mullis, Martin, Kennedy et al., 2009).

Además, fueron aplicados una serie de cuestionarios a padres de alumnos, profesores y directores de centro, con el fin de obtener información acerca del entorno familiar y la experiencia escolar en el aprendizaje de la lectura. Al mismo tiempo, cada país contesta a una serie de preguntas relacionadas con la estructura del sistema educativo y el papel de la lectura en el desarrollo curricular (Mullis, Martin, Kennedy et al., 2009). La información que se pretende recabar con estos cuestionarios haría referencia a las dimensiones que aparecen en la Tabla 12.

Tabla 12.

Información de contexto PIRLS. Áreas y factores asociados

Áreas contextuales y factores asociados.				
Contextos nacionales y comunitarios	Hogar	Centro- Escuela	Aula	Características y actitudes de los alumnos
✓Lenguaje y alfabetización ✓Demografía y recursos ✓Organización y estructura del sistema educativo ✓Currículo relacionado con la lectura en Educación Primaria	✓Recursos económicos sociales y culturales ✓Interés de los padres en el desarrollo lector ✓Actitudes y comportamientos paternos ante la lectura	✓Características de la escuela ✓Organización de la instrucción ✓Clima de aprendizaje ✓Recursos ✓Implicación de los padres	✓Desarrollo docente y formación ✓Características y actitudes del profesorado ✓Características del grupo- clase ✓Materiales y tecnología educativa ✓Estrategias y actividades de aprendizaje ✓Orientación educativa	✓Comportamientos ante la lectura ✓Actitudes positivas hacia la lectura ✓Actitudes hacia el aprendizaje de la lectura

Fuente: elaboración propia a partir de (Mullis, Martin, Kennedy et al., 2009).

Las evaluaciones TIMSS y PIRLS tienen un importantísimo carácter complementario, al ser realizadas ambas en 4º curso de Educación Primara, aportan una panorámica internacional muy valiosa a los diferentes países acerca del progreso de sus estudiantes en matemáticas, ciencias y comprensión lectora, completando dichos datos con una exhaustiva información contextual. La coincidencia en 2011 de ambos ciclos de aplicación permite su comparabilidad (Mullis, Martin, Kennedy et al., 2009).

c) International Civic and Citizenship Education Study (ICCS).

El estudio «*The six subjet study*» llevado a cabo en el año 1971 (ver Tabla 5), puede considerarse el origen de los posteriores estudios sobre educación cívica llevados a cabo por la IEA. Dicho estudio incluía, entre las áreas evaluadas, la educación cívica. Posteriormente en el periodo 1996- 2000 se desarrollan las dos fases del estudio CIVED «*Civic Education Study*». Su objetivo era analizar de forma comparativa la preparación de los alumnos para ser ciudadanos (Torney-Purta, Lehmann, Oswald, & Schulz, 2001). El problema añadido a la hora de evaluar este área, es que la educación cívica no tiene un lugar específico bien definido en los diseños curriculares internacionales, cada país puede considerar la educación cívica como una materia específica del currículo (bajo distintas denominaciones), puede ser considerada como un apartado específico de alguna otra materia (por ejemplo historia), o considerada como área transversal que ha de estar presente en la base de todas las actividades llevadas a cabo en el centro (Schulz & Sibberns, 2004). En la parte empírica del estudio (1999) participaron unos 90.000 estudiantes de 14 años de edad procedentes de 28 países, además de unos 9.000 profesores y 4.000 directores de centros (Torney-Purta et al., 2001). Un año después alrededor de 50.000 estudiantes de 17 años de otros 16 países y 2.000 directores contestaron similares cuestionarios (Amadeo, Torney-Purta, Lehmann, Husfeldt, & Nikolova, 2002).

El estudio ICCS «*International Civic and Citizenship Education Study*», llevado a cabo en el año 2009, supone una continuación a estos dos estudios previos realizados por la IEA en torno a la educación cívica. Para aquellos países que participaron en la edición previa de CIVED, ICCS les proporcionará información de tendencia de sus estudiantes en este área, puesto que los resultados de ambas pruebas son complementarios, los diez años de diferencia entre la realización de ambos estudios

implican modificaciones necesarias en el segundo, puesto que los cambios experimentados a nivel mundial en diversos aspectos que afectan a la esfera cívica y ciudadana, implican una adaptación de las evaluaciones. De este modo, ICCS conserva aspectos que permiten la continuidad con CIVED pero, al mismo tiempo, incorpora nuevas dimensiones con importancia en la actualidad. Entre los instrumentos de evaluación del ICCS se incluye una serie de ítems cognitivos sobre tendencias fijas del CIVED, y también ítems de algunas de las escalas de conceptos y actitudes (Schulz, Fraillon, Ainley, Losito, & Kerr, 2008).

El objetivo perseguido por este estudio, es investigar el grado en que los jóvenes están preparados y dispuestos a asumir su rol como ciudadanos. Con el fin de lograr éste objetivo, se evalúa el rendimiento de los alumnos mediante una prueba de comprensión de conceptos y de competencia en lo que respecta a la educación cívica y ciudadana. Por otro lado, también recoge y analiza variables adicionales sobre las actividades de los alumnos, su disposición y su actitud ante la educación cívica y ciudadana. La recogida de datos contextuales puede ayudar a la explicación de las diferencias en las variables de resultados (Schulz et al., 2008). El estudio consta de de dos partes marco cívico y ciudadano y marco contextual.

En este estudio, participaron 38 países evaluando a unos 140.000 estudiantes de 14 años de edad y a más de 62.000 docentes procedentes de un total de 5.300 centros educativos (Schulz, Ainley, Fraillon, Kerr, & Losito, 2010). Debido a la importancia contextual en el desarrollo de la educación cívica, ICCS implementa módulos específicos regionales para Asia, Europa y América Latina. De los 38 países participantes, 35 decidieron participar en los módulos regionales, de este modo, 24 países participaron en el módulo europeo, seis en el latino americano y cinco en el asiático (Schulz, Ainley, & Fraillon, 2011). El marco de la evaluación del ICCS gira en torno a tres dimensiones: una dimensión de contenido, en la que se especifica la cuestión a evaluar dentro del civismo y la ciudadanía; una dimensión de comportamiento afectivo, que describe los tipos de percepciones del alumno y las actividades que se miden; y una dimensión cognitiva, que describe los procesos de pensamiento que se van a evaluar. Los instrumentos utilizados para obtener dicha información serán un test del alumno, que mide sus procesos cognitivos y un

cuestionario del alumno, que mide sus percepciones y comportamientos (Schulz et al., 2008). Los dominios se representan en la Tabla 13.

Tabla 13.

Dominios y subdominios de la evaluación ICCS 2009

Dominios y subdominios ICCS 2009		
Contenido	Comportamiento afectivo	Cognitivo
1. Sociedad y sistemas cívicos a) Ciudadanos. b) Instituciones estatales c) Instituciones civiles 2. Principios cívicos a) Equidad. b) Libertad. c) Cohesión social. 3. Participación cívica a) Toma de decisiones. b) Influencias. c) Cohesión social. 4. Identidades cívicas a) Autoimagen cívica. b) Conectividad cívica.	1. Creencias sobre valores: a) Ciudadanos b) Instituciones estatales c) Instituciones civiles 2. Actitudes: a) Libertad. b) Equidad. c) Cohesión social. 3. Intenciones de comportamiento a) Participación en protestas cívicas b) Futura participación en política como adultos c) Futura participación en actividades de ciudadanía 4. Comportamientos a) Actividades de carácter cívico en la comunidad b) Actividades de carácter cívico en el centro	1. Conocimiento 2. Razonamiento y análisis

Fuente: elaboración propia a partir de (Schulz et al., 2008).

Por otro lado, siguiendo con el sistema de otras evaluaciones llevadas a cabo por IEA, el estudio ICCS utiliza un diseño rotativo de bloques de cuestiones que componen los cuadernillos de la prueba con el fin de obtener una medida más precisa del dominio evaluado y así, asegurar una cobertura más amplia del marco de la evaluación sin aumentar el tiempo de realización del test de cada alumno (Tabla 14). Por otro lado, esto constituye el elemento de anclaje esencial para establecer la relación entre los distintos tests aplicando procedimientos basados en la Teoría de la Respuesta al Ítem (TRI). Uno de los bloques de preguntas proviene de la prueba CIVED, permitiendo de este modo establecer la comparabilidad entre ambas aplicaciones (Schulz et al., 2008). Se pilotaron un total de 98 ítems, 19 de los cuales procedían de la prueba CIVED. Los bloques se agruparon en 6 cuadernillos, el bloque 6 corresponde a los 19 ítems procedentes de CIVED utilizados (de un total de 98) (Schulz, Ainley, & Fraillon, 2011).

Tabla 14.

Diseño de cuadernillos prueba piloto ICCS 2009

	Posición en la prueba		
Cuadernillo	1	2	3
1	C1	C2	C4
2	C2	C3	C5
3	C3	C4	C6
4	C4	C5	C1
5	C5	C6	C2
6	C6	C1	C3

Fuente: (Schulz, Ainley, & Fraillon, 2011, p. 27).

Tras el estudio piloto, fueron 80 los ítems que pasaron a formar parte del estudio, junto con 17 ítems procedentes de la prueba CIVED. Estos 97 ítems agrupados en bloques fueron distribuidos en 7 cuadernillos, tal como muestra la Tabla 15, el bloque C7 corresponde a los 17 ítems procedentes de CIVED.

Tabla 15.

Diseño de cuadernillos pruebas definitivas ICCS 2009

	Posición en la prueba		
Cuadernillo	1	2	3
1	C1	C2	C4
2	C2	C3	C5
3	C3	C4	C6
4	C4	C5	C7
5	C5	C6	C1
6	C6	C7	C2
7	C7	C1	C3

Fuente: (Schulz, Ainley, & Fraillon, 2011, p. 29).

El estudio ICCS proporciona importante información a nivel internacional acerca de la educación cívica y ciudadana, al mismo tiempo, su riguroso diseño metodológico y la variedad de información recabada en una amplísima muestra, hacen de ICCS un referente a nivel internacional en estudios de este tipo. El enfoque psicométrico, y las técnicas utilizadas, dotan a dicho estudio de un interés adicional.

En 2016, se llevará a cabo una nueva aplicación del estudio ICCS, dicha aplicación está enlazada directamente con la evaluación de 2009, en consecuencia los países participantes en ambas evaluaciones tendrán a su disposición una valiosa información de tendencias. La aplicación se llevará a cabo entre los meses de octubre y diciembre de 2015 (hemisferio sur) y febrero y abril de 2016 (hemisferio norte), el primer reporte de resultados estará disponible en 2017.

d) Teacher Education and Development Study in Mathematics (TEDS-M).

TEDS-M representa el primer estudio internacional a gran escala y de carácter comparativo en Educación Superior, centrado en la preparación de los profesores de Educación Primaria y primer ciclo de Educación Secundaria en el área de matemáticas (Tatto et al., 2008). Este estudio representa una novedad por las siguientes razones:

- Primer estudio de educación superior realizado por la IEA.
- Primer estudio de la IEA sobre formación docente.
- Primer estudio internacional de formación docente basado en muestras nacionales representativas y probabilísticas.
- Estudio internacional de formación del profesorado que obtiene datos acerca de los conocimientos del profesorado así como de los posibles determinantes de dichos resultados.
- Estudio internacional que integra aspectos específicos de la enseñanza de las matemáticas con aspectos genéricos relacionados con las políticas y prácticas en formación del profesorado.
- Estudio internacional que analiza el currículo en el área de matemáticas y la formación del profesorado.
- Estudio comparativo a gran escala para abordar el coste de la formación del profesorado (Tatto et al., 2008).

El estudio fue llevado a cabo en 17 países, pero la aplicación no se realizó del mismo modo en todos ellos, en algunos casos (2) hubo participación de algunas provincias, 7 países participaron tan solo con profesores de primer ciclo de secundaria, 5 países, entre los que se encuentra España, evaluaron exclusivamente a profesores de Educación Primaria, en Suiza la aplicación se realizó los cantones de lengua alemana, y Tailandia y Estados Unidos incluyeron exclusivamente profesores de centros públicos.

La recogida de datos se llevó a cabo en 2008 y se contó con la participación de 700 instituciones, 7398 formadores de profesores, 15163 futuros profesores de Educación Primaria y 9389 de Educación Secundaria. Los objetivos de este estudio serían, en primer lugar, identificar cómo preparan los diferentes países a sus profesores para la enseñanza de las matemáticas en Educación Primaria y en el primer ciclo de

Educación Secundaria, y en segundo lugar, estudiar las diferencias en la naturaleza y el impacto de la formación docente en la enseñanza y el aprendizaje de las matemáticas (Tatto et al., 2012).

Con el fin de detectar aquellos factores de posible importancia para la formación y rendimiento de los futuros docentes, se realizó una exhaustiva revisión de la literatura al respecto, de este modo, se identificaron cinco fuentes de variación entre países relevantes en este aspecto: el rendimiento de los estudiantes en el área de matemáticas, diferencias en el currículo, calidad de las clases de matemáticas, naturaleza de los programas de formación del profesorado y el contenido de dichos programas de formación (Tatto, et al., 2012). A partir de estas consideraciones, se elaboraron tres preguntas de investigación acerca de la formación de los profesores de Educación Primaria y primer ciclo de Educación Secundaria del área de Matemáticas, la respuesta a tales cuestiones permitió identificar las variables a analizar así como la relación existente entre las mismas (Figura 7) (Tatto et al., 2012).

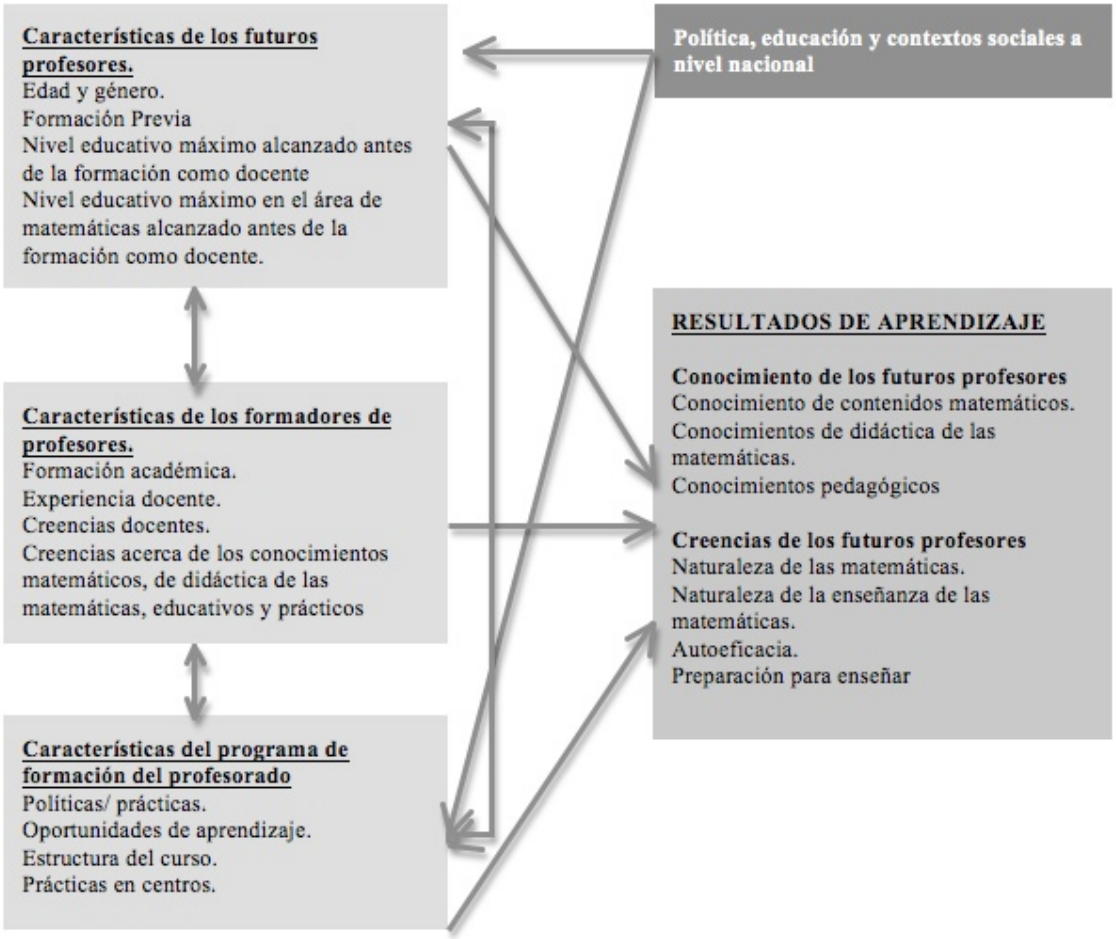


Figura 7. Interrelaciones entre las variables analizadas en TEDS-M.
Fuente: (Tatto et al., 2008, pp.14).

TEDS-M estructura la recogida de información en torno a tres componentes. El primer componente estaría relacionado con la política de formación del profesorado y el contexto, el segundo con los procesos instructivos y programas de formación y el tercero con el resultado de la formación docente. El componente I incluiría información relativa a los itinerarios de formación docente, las políticas de formación del profesorado, el currículo nacional en formación docente y el precio de la formación. El componente II recabaría información acerca del programa institucional y de los formadores de profesores. En el caso de los resultados en los futuros educadores se pretende medir tanto los resultados previstos en la formación docente como los alcanzados finalmente, tanto en el área de matemáticas como en el área de didáctica de las matemáticas. Los cuestionarios también recogen información relativa a variables de «*background*», así como acerca de las oportunidades de aprendizaje sus creencias y sus actitudes ante la enseñanza de las matemáticas (Tatto et al., 2012).

Con el fin de establecer vínculos con la prueba TIMSS, que permitan relacionar ambos resultados, a la hora de elaborar los marcos de referencia de TEDS para la medida de los conocimientos matemáticos de los profesores de Educación Primaria, se decidieron utilizar los mismos contenidos y dominios cognitivos que los utilizados en la prueba TIMSS 2007 reconociendo, al mismo tiempo, la necesidad de que los docentes hayan adquirido los conocimientos (como mínimo) de dos cursos superiores a los que deben enseñar. Siguiendo esta misma línea, la prueba de conocimientos para profesores de primer ciclo de Educación Secundaria, se desarrolló a través de los dominios de contenido de TIMSS 2007 y de TIMSS-*advanced*. De este modo, la prueba de matemáticas para ambos niveles docentes, incluye contenidos relacionados con números, geometría, álgebra y datos. Los dominios cognitivos evaluados son conocimiento, aplicación y razonamiento (Tatto et al., 2008). TEDS-M elabora las escalas de conocimientos matemáticos y de conocimientos de didáctica de las matemáticas por medio de la Teoría de la Respuesta al Ítem (TRI).

Mediante el uso de un diseño de bloques rotativos, TEDS-M consigue aumentar el alcance de medida del instrumento sin aumentar el tiempo de administración de la prueba, siguiendo de este modo con la dinámica de diseño del resto de investigaciones realizadas por la IEA analizadas en este apartado. Los bloques rotativos son utilizados como anclaje entre cuadernillos para poder establecer la escala de comparación entre cuestionarios y para estimar los cambios en el tiempo. Para la realización de proceso de calibración se utilizan modelos derivados del modelo de un parámetro (Rasch). En el caso de ítems de crédito parcial se utiliza el modelo de Masters (1982), en ambos casos se utiliza el programa ACER Conquest (Wu, Adams, Wilson, & Haldane, 2009).

En definitiva, en más de medio siglo de trayectoria en el ámbito de la Evaluación Educativa, la IEA ha enfrentado el reto de la comparabilidad en sus importantes evaluaciones, constituyendo un referente ineludible a la hora de estudiar este importante desafío.

1.4.2 Organization for Economic Cooperation and Development (OECD)

Los antecedentes de la «*Organization for Economic Cooperation and Development*» (OECD), los encontramos en la «*Organisation for European Economic Cooperation*» (OEEC), creada en el año 1947, con el objetivo de llevar a cabo el plan financiero Marshall para la reconstrucción de un continente devastado por la guerra (OECD, 2012a). Los gobernantes europeos reconocieron la interdependencia de sus economías, abriendo paso a una nueva esfera de cooperación que cambiaría el rumbo de Europa. Animados por el éxito y por las perspectivas de un escenario global, Canadá y Estados Unidos se hicieron miembros de la OEEC en el año 1960, firmando la nueva convención de la OECD el 14 de Diciembre de 1960. La Organización para la Cooperación y el Desarrollo económico, nació oficialmente el 30 de septiembre de 1961 (OECD, 2012a). Poco a poco se han ido incorporando mayor número de países, hasta llegar a los 34 integrantes que la componen en la actualidad.

Las áreas en las que trabaja la OECD para la consecución de sus objetivos son muchas, en el presente trabajo nos centraremos en las relativas a la educación, y más concretamente en las implementadas con el fin de llevar a cabo una evaluación de los sistemas educativos. En primer lugar, analizaremos la perspectiva general de la OECD en lo relativo al ámbito educativo para, posteriormente, centrarnos en sus sistemas de evaluación. La OECD aporta datos comparativos y análisis de políticas y prácticas educativas con el fin de ayudar a la construcción de sistemas educativos eficaces y eficientes así como para mejorar los resultados de aprendizaje (OECD, 2012a). Del mismo modo, proporcionan foros de discusión donde gobernantes, empresarios, ciudadanos y académicos pueden compartir y contrastar las mejores prácticas de aprendizaje. Los indicadores estadísticos aportados por la OECD aportan una fuerte evidencia para el establecimiento de las comparaciones internacionales de gran variedad de aspectos relacionados con los sistemas educativos (OECD, 2012a). Las grandes áreas en las que trabaja la OECD así como los objetivos y principales proyectos llevados a cabo en dichas áreas, pueden observarse en la Tabla 16.

Tabla 16.

Ámbitos de interés en el área de Educación de la OECD y principales proyectos desarrollados

Ámbitos de interés en el área de Educación de la OECD	
Educación infantil y cuidado de la infancia	Proyecto"*Encouraging Quality in Early Childhood Education and Care*" (inicio 1996), investiga qué define la calidad en este nivel educativo, qué políticas pueden promover y mejorar la calidad así como el modo de ponerlas en marcha.
Enseñanza	Promoción de métodos de enseñanza y recursos educativos con el objetivo de preparar a los estudiantes para el futuro. En este sentido los programas PISA (*Programme for International Student Assessment*) y TALIS (*Teaching and Learning International Survey*) ayudan a identificar buenas prácticas.
Transición educativa	Definición y desarrollo de competencias atendiendo a las nuevas exigencias del mercado laboral. El proyecto sobre la evaluación del sistema de Formación Profesional "*Vocational Education and Training* (VET)" y la publicación del estudio "*Learning for Jobs*" a finales de 2010 pretenden una revisión de la Formación Profesional comparando las distintas prácticas internacionales. El proyecto de competencias de la OECD (OECD-*Skills Strategy*) Persigue el fomento de la educación basada en competencias que permita formar ciudadanos preparados para las exigencias del mundo actual así como del mercado de trabajo.
Educación Superior	Mejora del acceso, la calidad y la relevancia de la educación superior. El estudio "*The Feasibility Study for the International Assessment of Higher Education Learning Outcomes* (AHELO)" evalúa los conocimientos de los estudiantes de Educación Superior así como aquellas tareas que son capaces de desempeñar tras la graduación. Proporciona información acerca de la relevancia y calidad del proceso de enseñanza aprendizaje en Educación Superior.
Educación de adultos	Enriquecimiento y mejora del capital humano. Trabajos pioneros en el reconocimiento de la educación formal y no formal. En 1996 se puso en marcha la estrategia "*lifelong learning for all"* en la que se contemplan el aprendizaje formal, no formal e informal, centrándose en el reconocimiento de los dos últimos tipos. El "*Programme for the International Assessment of Adult Competencies* (PIAAC), ofrece datos comparativos acerca de las competencias de los adultos de 26 países (datos 2013).
Resultados, beneficios y ganancias	Comprensión del impacto de la educación más allá de las aulas. Destacan: Programa PISA (*Programme for International Student Assessment*) que evalúa la aplicación de los conocimientos de las materias básicas a situaciones de la vida real por parte de los estudiantes. Programa AHELO (*Assessment of Higher Education Learning Outcomes)* mide los resultados de la educación superior en una escala internacional.
Equidad e igualdad de oportunidades	La OECD puso en marcha en el año 2008 un proyecto para la revisión de la educación de los estudiantes inmigrantes, centrándose en los resultados del aprendizaje de los estudiantes de primera y segunda generación así como en la integración de estos alumnos en las escuelas. El objetivo es mejorar el acceso, la participación y los resultados de los alumnos inmigrantes.
Innovación y gestión del conocimiento	El centro para la innovación educativa y la investigación de la OECD (CERI) Realiza investigaciones centradas en el aprendizaje en todas las edades y no exclusivamente en el ámbito formal, centrando su interés en la innovación y las tendencias emergentes. Dentro de los proyectos destacan: "*Innovative Learning Environments Project*" que trata de dar respuesta a la pregunta ¿Cómo pueden las escuelas de hoy en día fomentar el aprendizaje a lo largo de la vida y preparar a los estudiantes para las exigencias del siglo 21? "The Innovative Teaching for Effective Learning" Proyecto centrado en la actividad docente, tanto en los conocimientos como en las competencias pedagógicas y didácticas del profesorado. "Innovation Strategy for Education and Training" Pretende ayudar a los gobernantes a la comprensión de hacia dónde se dirige la innovación y qué implicaciones tiene en el ámbito educativo.

Fuente: elaboración propia a partir de OECD (2012).

Continuando con la selección de evaluaciones internacionales de mayor importancia en la actualidad, propuesta en la Tabla 4, dentro de las organizadas por la OECD pasaremos a analizar las evaluaciones «*Program for International Students Assessment*» (PISA), «*Teaching and Learning International Survey*» (TALIS), «*Assessment of Higher Education Learning Outcomes*» (AHELO) y «*Programme for the International Assessment of Adult Competencies*» (PIAAC).

***Programme for International Students Assessment* (PISA).**

El programa PISA se presenta como el más influyente en el ámbito de la evaluación educativa, tal y como veíamos al inicio del presente apartado, la repercusión mediática, política y educativa de los resultados de los informes de dicho proyecto se extienden año a año. El origen de este proyecto está muy ligado a otro proyecto de la OECD: el proyecto INES, dedicado a la producción de indicadores internacionales de la educación que se desarrolla desde el año 1990, publicando desde el año 1992 los indicadores producidos en el informe conocido con el título «*Education at a Glance*» (Pajares, Sanz, & Rico, 2004). El proyecto INES tomaba como indicadores de rendimiento los resultados de los proyectos llevados a cabo por la IEA (analizados en el epígrafe anterior), sin embargo, estos estudios, no se elaboraban con la frecuencia necesaria para dotar de resultados de rendimiento al sistema de indicadores de la OECD con la suficiente frecuencia (Pajares, Sanz, & Rico, 2004). De este modo, surge PISA en el año 1997 como un estudio sistemático cuyo objetivo es la evaluación internacional de los sistemas educativos por medio de la medida de competencias y conocimientos de estudiantes de 15 años de edad (entre 15 años y tres meses y 16 años y dos meses al principio de la evaluación y que hayan completado como mínimo 6 cursos de enseñanza obligatoria). Se debe destacar que la selección no se hace en función del grado o curso escolar en el que se encuentran sino en función de la edad cronológica, con el fin de evitar problemas en las comparaciones internacionales producidos por diferencias en el inicio de la escolaridad o en la estructura del sistema educativo. Hasta la fecha, más de 70 países han participado en el proyecto PISA.

Las evaluaciones PISA siguen un ciclo de tres años, en cada edición se pone el énfasis en una de las tres competencias objeto de interés (Lectura, Matemáticas y resolución de problemas y ciencias) (aproximadamente 2/3 son dedicados al área de

interés y 1/3 a las dos áreas restantes). Además, tanto los estudiantes como los directores de los centros, contestan a un cuestionario de contexto, algunos países pueden optar porque dicho cuestionario también sea contestado por los padres, de este modo PISA cuenta con información acerca del estudiante, su familia y los factores institucionales que pueden contribuir a explicar las diferencias en desempeño (OECD, 2002). Tal y como puede observarse en la Figura 8, la primera evaluación llevada a cabo por PISA se centró en la competencia lectora, área evaluada de nuevo en el año 2009. Por otro lado, debemos destacar que la evaluación de 2012, además de centrar el interés en el área de matemáticas y resolución de problemas, incorpora una prueba opcional de matemáticas y comprensión lectora por ordenador, así como una dimensión de educación financiera, también opcional, que no había sido evaluada hasta el momento, de este modo conserva su carácter continuo e innovador, adaptando las evaluaciones a las exigencias y retos del futuro.

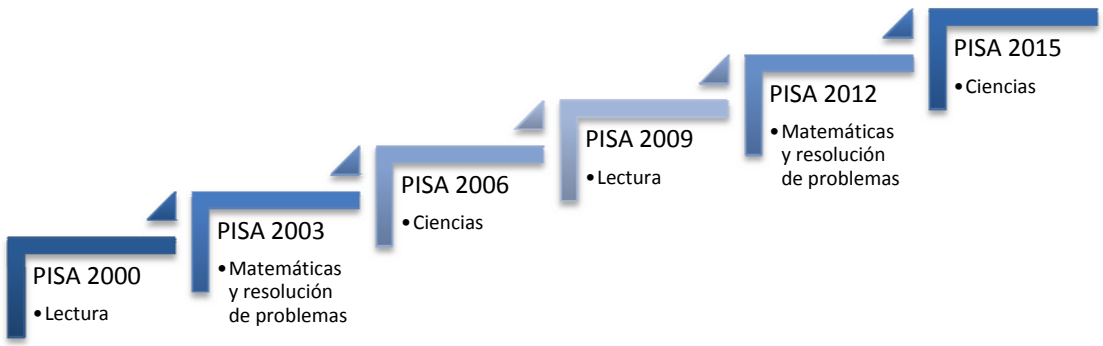


Figura 8. Evaluaciones del programa PISA
Fuente: elaboración propia.

La peculiaridad de PISA, reside en el hecho de no estar basada en los contenidos curriculares, sino en evaluar qué son capaces de hacer los estudiantes al finalizar la educación obligatoria, en qué medida son capaces de aplicar su conocimiento a la vida real y en qué medida están preparados para la participación social. PISA es una evaluación centrada en competencias para la vida. Un importante mérito de PISA es evitar el reduccionismo conceptual frecuente en las actividades de evaluación, ofreciendo como resultado la evaluación de competencias, en términos de rendimiento, en las áreas evaluadas (Monereo & Castelló, 2009). A pesar de considerar que este es el rasgo más característico, otros muchos han guiado la realización de las distintas

evaluaciones PISA, siendo recogidos en los informes de resultados de las distintas ediciones (OECD, 2010a).

De este modo, los criterios que guían la elaboración de cada una de las ediciones de PISA, hacen de esta una evaluación singular, con un importante aporte de información para los países participantes que pueden valorar el desempeño de sus estudiantes conforme a distintos criterios, mejorando las políticas y prácticas educativas a fin de lograr un mayor desarrollo competencial de sus estudiantes.

El carácter regular de PISA permite superar las perspectivas estáticas de la evaluación, reconociendo la dimensión continua del aprendizaje. Se presenta, por tanto, como un instrumento de seguimiento, al medir cada tres años las competencias indicadas y poniendo el énfasis en una de ellas en cada edición, cerrando por tanto un ciclo de evaluación cada 9 años. El mantenimiento de la estructura básica de la evaluación es el elemento que permite su comparación a lo largo del tiempo, posibilitando, a largo plazo, que los países observen sus tendencias así como las consecuencias de cambios políticos y planes de mejora de la enseñanza, complementando esta información con la aportada por los referentes internacionales (OECD, 2010a).

Posiblemente, entre estas razones se encuentre la explicación del incremento en el número de países y economías que participan en la evaluación PISA en cada edición (ver Figura 9). En la primera edición de PISA llevada a cabo en el año 2000 participaron 43 países o economías (12 de las cuales realizaron la recogida de datos un año más tarde (2001)), tres ediciones después (2009) participaron 75 (10 de los cuales realizaron la recogida de datos en 2010), en la última edición, la participación en la primera recogida de datos fue de 65 países/economías. El incremento más sustancial en el número de participantes, se ha dado precisamente en los países no miembros de la OECD que han ido incrementando su participación paulatinamente (12 países en 2000, 11 en 2003, 27 en 2006, 41 en 2009 y 31 en 2012).

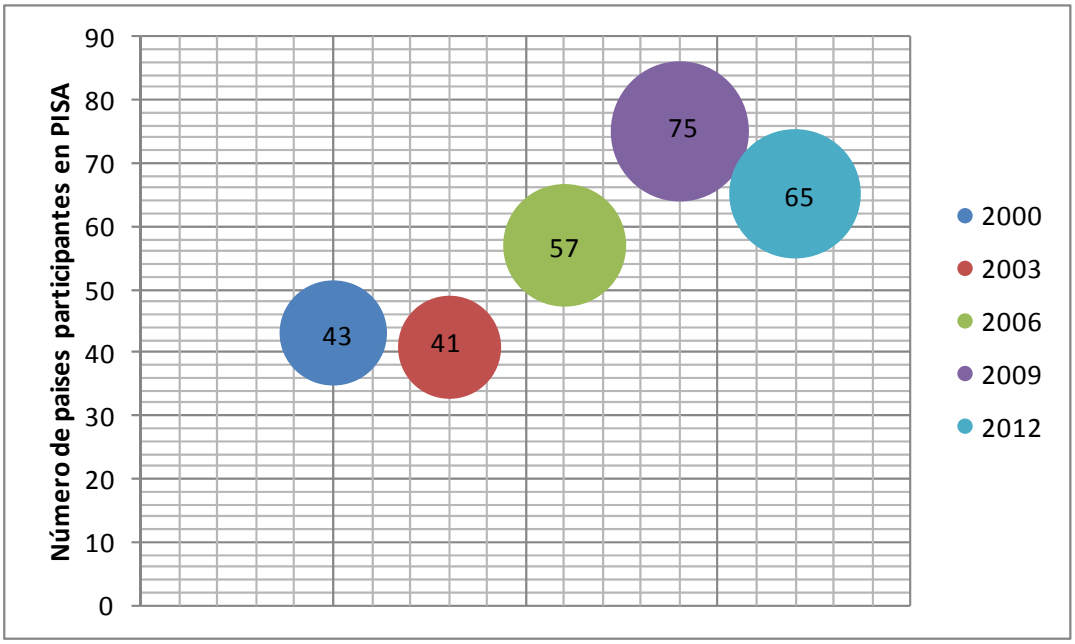


Figura 9. Número de países/economías participantes en las diferentes ediciones de PISA
Fuente: elaboración propia.

Uno de los principales retos de un estudio internacional es asegurar la comparabilidad de las poblaciones objeto de cada país/ economía, en este sentido, destaca la cuidadosa definición de población objeto realizada por PISA con el fin de eliminar sesgos que pudieran afectar a las muestras de los diferentes países/ economías, tales como el inicio de la edad de escolarización o la estructura del sistema educativo. La calidad de un estudio de este tipo, depende tanto de la información en la que están basadas las muestras nacionales, como de los procedimientos de muestreo utilizados. La mayoría de las muestras de PISA son diseñadas como muestras estratificadas de dos fases (OECD, 2010a).

El reto de la comparabilidad está muy presente a la hora de elaborar los marcos de referencia que guían la evaluación y que son actualizados en cada edición. En dicha actualización los expertos han de realizar un cuidadoso trabajo que permita conservar los elementos esenciales entre ediciones y por otro lado, incorporar nuevos elementos que hagan posible que se trate de una evaluación contextualizada, adaptada a las características del momento y a la constante evolución de los sistemas educativos. Las evaluaciones centradas en aspectos curriculares, presentan la dificultad añadida de contemplar en ellas los elementos incluidos en los currículos oficiales de los diferentes países, PISA, al ser una evaluación centrada en competencias, no tiene esa dificultad ya

que pretende evaluar las competencias deseables en todos los estudiantes al finalizar la Educación Secundaria, de manera independiente a lo contenido en los currículos de sus países, a pesar de superar esta dificultad, el trabajo de elaboración de los marcos de referencia es altamente complejo.

El marco de referencia para la evaluación de la competencia lectora proviene, en parte, del estudio realizado por IEA en el año 1992 «*Reading Literacy Study*» y del estudio realizado por el «*Educational Testing Service*» (ETS) con la colaboración de la OECD, entre otros organismos, denominado «*International Reading Literacy Survey*» (Kirsch, 2001) tomando de este último su énfasis en la importancia de la lectura en la participación social (OECD, 2009). En este estudio, se define la alfabetización en lectura como "uso de la información impresa y escrita para el funcionamiento en sociedad, para alcanzar objetivos y para desarrollar el conocimiento y el potencial" (Kirsch, 2001, p. 6.) La definición de competencia lectora utilizada en la evaluación PISA permanece estable desde la primera edición de la evaluación en el año 2000 (Tabla 17), no obstante, en el 2009 se incorporó la lectura de textos impresos y los dominios de compromiso con la lectura y metacognición.

Tabla 17.

Definición de Competencia Lectora en las evaluaciones PISA 2000 y 2009. Evolución del concepto

Marco de referencia en la evaluación de la Competencia lectora. Definición	
2000	2009
Competencia lectora es comprender, utilizar y analizar textos escritos, con el fin de alcanzar objetivos, desarrollar el conocimiento y potencial y participar en la sociedad.	La competencia lectora supone comprender, utilizar, reflexionar y comprometerse con textos escritos, con el fin de alcanzar objetivos, desarrollar el conocimiento y potencial y participar en la sociedad.

Fuente: OECD (2009).

Teniendo en cuenta el carácter multidimensional del constructo, y reconociendo la imposibilidad de atender a todas sus dimensiones desde la perspectiva de la evaluación, desde PISA se pone especial énfasis en contemplar qué leen los estudiantes y con qué propósitos y, por otro lado, conseguir evaluar un amplio rango de dificultad en el dominio (OECD, 2009). Para conseguir estos objetivos, PISA estructura la evaluación en lectura atendiendo a tres dimensiones: tipo de texto, aspecto (proceso cognitivo que determina la interacción de los lectores con el texto) y situación/ contexto, (OECD, 2009) (Tabla 18).

Tabla 18.

Tipo de texto, aspectos y contextos en la prueba de Lectura en PISA 2012

Peso relativo de tipos de texto, aspectos y contextos en la prueba de Lectura. PISA 2009				
			Formato tradicional	Formato electrónico
Textos	Ambiente (solo en formato electrónico).	De autor		70%
		Basados en Mensajes		25%
		Mixtos		5%
		Total		100%
	Formato de texto	Continuo	60%	10%
		No continuo	30%	10%
		Mixto	5%	10%
		Múltiple	5%	70%
		Total	100%	100%
Aspecto/ Proceso	Acceso y reproducción		25%	25%
	Integración e interpretación		50%	25%
	Reflexión y evaluación		25%	30%
	Complejo		0%	20%
	Total		100%	100%
Contexto	Personal		30%	30%
	Educacional		25%	15%
	Ocupacional		15%	15%
	Público		30%	40%
	Total		100%	100%

Fuente: elaboración propia a partir de OECD(2009).

En el área de matemáticas, los rasgos esenciales de la definición se conservan edición tras edición (ver Tabla 19).

Tabla 19.

Definición de Competencia Matemática en las evaluaciones PISA 2003 y 2009.Evolución del concepto

Marco de referencia en la evaluación de la Competencia Matemática. Definición	
2003	2012
Aptitud de un individuo para identificar y comprender el papel que desempeñan las matemáticas en el mundo, alcanzar razonamientos bien fundados y utilizar y participar en las matemáticas en función de las necesidades de su vida como ciudadano constructivo, comprometido y reflexivo.	Capacidad individual de formular, emplear e interpretar las matemáticas en variedad de contextos. Esto incluye el razonamiento matemático y el uso de conceptos, procedimientos, hechos y herramientas para describir, explicar y predecir fenómenos. Esto ayuda a los individuos a reconocer el papel que las matemáticas juegan en el mundo así como realizar juicios y tomar decisiones bien formadas, necesarias para ser ciudadanos constructivos, comprometidos y reflexivos

Fuente: elaboración propia a partir de los marcos de referencia para la evaluación PISA de 2003 y 2012.

No obstante, dicha definición, se ha ido completando incorporando y especificando algunos aspectos. Con el propósito de conseguir medir la competencia matemática conforme a la definición expuesta, PISA tiene en cuenta el contenido, los

procesos y los contextos, clasificando los reactivos conforme a tales dimensiones (Tabla 20).

Tabla 20.

Contenidos, procesos y contextos en la prueba de Matemáticas Pisa 2012

Peso relativo de contenidos, procesos y contextos en la prueba de matemáticas. PISA 2012		
Contenido	Cambios y relaciones	25%
	Espacio y forma	25%
	Cantidad	25%
	Incertidumbre	25%
	Total	100%
Proceso	Formulación de situaciones matemáticamente	25%
	Empleo de conceptos matemáticos, hechos procedimientos y razonamientos.	50%
	Interpretar, aplicar y evaluar resultados matemáticos.	25%
	Total	100%
Contexto	Personal	25%
	Ocupacional	25%
	Social	25%
	Científico	25%
	Total	100%

Fuente: Elaboración propia a partir de OECD (2012).

La competencia de resolución de problemas en PISA 2012 es definida como "la capacidad individual de realizar procesos cognitivos con el fin de comprender y resolver problemas y situaciones cuándo el método de solución no es explícito. Esto incluye la voluntad de comprometerse con este tipo de situaciones, con el fin de alcanzar su máximo potencial como ciudadano constructivo y reflexivo" (OECD, 2012b, p. 22).

Tabla 21.

Procesos, naturaleza de la situación y contextos en resolución de problemas PISA 2012

Peso relativo de contextos, naturaleza de la situación y procesos en Resolución de problemas. PISA 2012		
Proceso	Exploración y comprensión.	20- 25%
	Representación y formulación.	20- 25%
	Planificación y ejecución.	35- 45%
	Seguimiento y reflexión	10-20%
	Total	100%
Naturaleza de la situación	Estática	25-35%
	Interactiva	65-75%
	Total	100%
Contexto	Tecnológico	45- 55%
	No tecnológico	45- 55%
	Total	100%

Fuente: elaboración propia a partir de OECD (2012).

Lo que distingue esta competencia en las evaluaciones 2003 y 2012 no es su definición sino la incorporación en 2012 de la evaluación informática y la introducción de problemas que no pueden ser resueltos si no es por medio del uso de asistentes para su resolución (OECD, 2012b).

Tal y como veíamos en la introducción de este apartado, en la evaluación de 2012 se incluye una dimensión de competencia financiera de carácter voluntario, en la que pueden participar los países que así lo deseen. La evaluación PISA 2012 incluye este área de interés convirtiéndose en el primer estudio internacional que evalúa dicho aspecto. La competencia financiera, es entendida como el conocimiento y comprensión de conceptos y riesgos financieros así como las habilidades, motivación y confianza para aplicar ese conocimiento y comprensión con el fin de tomar decisiones efectivas en el ámbito de los contextos financieros, mejorar el bienestar económico de los individuos y la sociedad y permitir la participación en la vida económica (OECD, 2012b). El dominio se estructura en contenidos, procesos y contextos. La importancia relativa otorgada a cada aspecto, reflejada en el porcentaje de la puntuación total en la competencia que supone cada dimensión puede apreciarse en la Tabla 22.

Tabla 22.

Contenido, procesos y contextos en la prueba de Educación Financiera PISA 2012

Peso relativo de contextos, naturaleza de la situación y procesos en Resolución de problemas. PISA 2012		
Contenido	Dinero y transacciones.	30- 40%
	Planificación y gestión de finanzas.	25- 35%
	Riesgo y recompensa.	15- 25%
	Paisaje financiero.	10-20%
	Total.	100%
Proceso	Identificación de información financiera.	15- 25%
	Análisis de información en un contexto financiero.	15- 25%
	Evaluación de cuestiones financieras.	25- 35%
	Aplicar el conocimiento y la comprensión financiera.	25- 35%
	Total.	100%
Contexto	Educativo y laboral.	10-20%
	Familiar y de hogar.	30- 40%
	Individual.	35- 45%
	Social.	5- 15%
	Total.	100%

Fuente: elaboración propia a partir de OECD (2012b).

La definición de competencia científica de la edición de 2006 es conceptualmente acorde a las definiciones de las ediciones anteriores 2000 y 2003, la

única variación es que la definición de 2006 se amplió incluyendo aspectos actitudinales de las respuestas de los estudiantes, no obstante esta inclusión no afecta a la comparabilidad puesto que, estas dimensiones actitudinales, son tratadas de manera diferenciada (OECD, 2006). La organización del dominio se estructura atendiendo al tipo de conocimiento, a los procesos cognitivos implicados, y a los contextos o situaciones (Tabla 23).

Tabla 23.

Conocimientos y procesos en la prueba de Ciencias PISA 2006

Peso relativo de tipos de conocimiento y procesos cognitivos en la competencia Científica. PISA 2006			
Conocimiento	De la ciencia	Sistemas físicos.	15-20%
		Sistemas vivos.	20-25%
		Sistemas terrestres y espaciales.	10-25%
		Sistemas tecnológicos.	5-10%
		Subtotal	60-65%
	Sobre la ciencia	Investigación científica	15-20%
		Explicaciones científicas	15-20%
		Subtotal	35-40%
		Total	100%
Procesos	Identificar cuestiones científicas.		25-30%
	Explicar fenómenos científicos.		35-40%
	Utilizar pruebas científicas		35-40%
	Total		100%

Fuente: elaboración propia a partir de OECD(2006).

En la evaluación de cada una de las áreas contempladas en PISA, se tienen en cuenta gran variedad de contextos, tipos de conocimiento y procesos cognitivos, siendo el interés evaluar la competencia del estudiante en un amplio rango de dominio. Una importante característica de PISA es la utilización de diversas tipologías de ítems que, aunque pueden complicar en cierta medida el diseño de la evaluación, la aplicación, la codificación de resultados o el análisis de datos, posibilitan la evaluación de niveles de desempeño difíciles de evaluar de forma exclusiva con ítems de opción múltiple o de respuesta breve. De este modo, PISA incorpora 5 tipos de ítems en sus instrumentos de evaluación, cuyas características y rasgos principales pueden observarse en la Tabla 24.

Tabla 24.

Tipos de ítems en las pruebas PISA

TIPO	El alumno debe…	Para su corrección…	Se puntúa…
Respuesta construida abierta	…elaborar una respuesta larga, lo que hace que aparezca un amplio intervalo de respuestas individuales divergentes y de distintos puntos de vista.	…es necesario un corrector y un libro de códigos.	… normalmente siguiendo un modelo de crédito parcial en el que las respuestas parcialmente correctas o menos elaboradas son tenidas en cuenta.
Respuesta construida cerrada	…elaborar sus propias respuestas, existiendo un número limitado de respuestas aceptables.	…en la mayoría de los casos no es necesario utilizar un corrector y se realiza automáticamente.	…frecuentemente aplicando un código dicotómico de corrección (correcto/ incorrecto).
Respuesta breve	…proporcionar una respuesta breve, existiendo un número de respuestas posibles muy amplio.	…es necesario un corrector y un libro de códigos.	… siguiendo un modelo de crédito parcial o uno dicotómico.
Elección múltiple compleja	… hacer una serie de elecciones, normalmente binarias. Pueden indicar sus respuestas marcando con un círculo una palabra o una frase breve (por ejemplo sí o no) para cada punto.	…no es necesaria la intervención de un corrector.	..cada elección de manera dicotómica, ofreciendo la posibilidad de obtener el crédito máximo o créditos parciales a todo el ejercicio.
Elección múltiple	.. deben marcar una letra con un círculo para indicar su elección de entre cuatro o cinco alternativas.	…no es necesaria la intervención de un corrector.	… de manera dicotómica (correcto/ incorrecto).

Fuente: elaboración propia a partir de los informes de resultados de las ediciones 2000, 2003, 2006, 2009 y 2012.

En PISA, la evaluación en cada edición se organiza en torno a una serie de cuadernillos que han de poderse responder en un tiempo máximo de 120 minutos por estudiante (Tabla 25), eso limitaría mucho la posibilidad de evaluar de manera amplia las competencias en las tres áreas de interés, de este modo, con el fin de evitar tal situación, asegurando la validez de contenido, PISA utiliza un diseño de "bloques rotativos" entre cuadernillos, posibilitando la evaluación completa en las dimensiones analizadas y permitiendo la comparabilidad entre cuadernillos por medio de los "bloques de ítems comunes". En la descripción de los estudios llevados a cabo por la IEA, se puede apreciar como en ellos (TIMSS, PIRLS, ICCS, TEDS-M) se utiliza éste mismo tipo de diseño, formando parte del denominado muestreo matricial de ítems «*Matrix Sampling of Items*» (Childs & Jaciw, 2003). Una diferencia de PISA respecto a las evaluaciones de la IEA, es el uso para la estimación del modelo de coeficientes mixtos multinomial (forma generalizada del modelo de Rasch) (Adams, Wilson, &

Wang, 1997). Utilizando modelos diferentes en función de tipo de ítem, en el caso de ítems de respuesta binaria se utiliza el modelo logístico de un parámetro y para los ítems de escala el Modelo de Crédito Parcial (Masters, 1982).

Tabla 25.

Diseño de pruebas PISA, distribución del tiempo por evaluación y contenido

		PISA 2000	PISA 2003	PISA 2006	PISA 2009	PISA 2012
	Nº Cuader	9	13	13	13	13
Bloques	Lect.	9 bloques/30min	2 bloques/30min	2 bloques/30min	7 bloques/30min	3 bloques/30min
	Mat.	4 bloques /15min	7 bloques/30min	4 bloques/30min	3 bloques/30min	7 bloques/30min
	Cien	4 bloques /15min	2 bloques/30min	7 bloques/30min	3 bloques/30min	3 bloques/30min
MINUTOS TOTALES						
Área de interés		270	210	210	210	210
Otras áreas		120	120	180	180	180
Total		390	330	390	390	390

Fuente: elaboración propia.

Para informar sobre el desempeño de los estudiantes, PISA utiliza los valores plausibles como métodos de imputación de respuesta (llevados a cabo en evaluaciones anteriores como TIMSS descrita previamente). En PISA son extraídos de manera aleatoria cinco valores de la distribución (considerada normal) que representan el rango de habilidades de cada estudiante. El procedimiento utilizado por PISA para la estimación de los valores plausibles es similar al descrito en el caso de TIMSS, teniéndose en cuenta, del mismo modo, tanto las respuestas de los estudiantes a las preguntas de su modelo de cuadernillo como variables de «*background*» consideradas condicionantes de la habilidad o competencia. La media de la escala es de 500 y la desviación típica 100.

La información obtenida por medio de los cuestionarios de contexto, proporciona marcos de interpretación y datos para analizar los resultados. El objetivo de la información presentada en la Tabla 26, es mostrar una panorámica global tanto de las dimensiones y subdimensiones contextuales evaluadas en PISA, como de algunos ejemplos de ítems o cuestiones utilizadas para su evaluación. Nótese que, en el caso de las evaluaciones TIMSS y PIRLS, se ha incluido información acerca de las dimensiones que en dichos estudios se pretenden medir y que así son detalladas en los marcos de referencia, la diferencia en la presentación de la información relativa al contexto en los

estudios de la OECD y de la IEA responde a la información a la que se ha tenido acceso acerca de las mismas.

Tabla 26.

Información de contexto utilizada en los estudios PISA, ejemplo de cuestiones para su medida.

Información contextual utilizada en PISA y ejemplos de cuestiones para su medición.			
			Medida
Sistema educativo	Medidas de riqueza e ingresos del país y de la región		Producto interior bruto per cápita
	Situación global de los docentes		Salarios de los maestros y los beneficios relativos a otras ocupaciones con educación similar
	Implicación de la comunidad en la escuela		Extensión de la influencia de los padres y de los órganos de participación en las decisiones del centro
	Niveles de toma de decisiones		Niveles de Gobierno o participación que influyen directamente en las decisiones relativas al personal, el presupuesto, el contenido educativo y las prácticas de evaluación
	Escuelas públicas y privadas		¿La educación es pública o privada?
	Medidas de desigualdad social		Distribución de riqueza
	Rendición de cuentas		¿Se utilizan los resultados para sistemas de rendición de cuentas?
Escuela	Liderazgo		Actividades y comportamientos del director.
	Composición del alumnado		Número de estudiantes cuya primera lengua no es la lengua en la que se realiza la prueba
	Énfasis curricular		Tiempo empleado en cada materia.
	Actividades extra curriculares		Relación de actividades ofrecidas a lo estudiantes
	Tamaño de la escuela		Número total de alumnos matriculados
	Apoyo al aprendizaje y la enseñanza		Actividades y comportamientos de la dirección en relación a la enseñanza y el aprendizaje
Instrucción	Escolar	Tamaño del aula	Número de estudiantes en el aula
		Composición del aula	Antecedentes familiares (cuestionario alumno)
		Calidad docente	Actividades de enseñanza del profesor descritas por los estudiantes
		Oportunidades de aprendizaje	Frecuencia de distintas actividades de lectura
		Orden del entorno del aula	Frecuencia de alteraciones o desorden en clase
		Apoyos a la enseñanza y condiciones de aprendizaje	Percepciones de los alumnos acerca del interés del profesorado y de su apoyo
Estudiante	Estatus socioeconómico		Nivel educativo superior completado por los padres
	Condición de inmigrante		País de nacimiento del estudiante y sus padres
	Estilos de aprendizaje del estudiante		Información de cómo estudian los alumnos.
	Actitudes del estudiante (materia)		Preferencias de los estudiantes y comportamientos en actividades especificas de…

Fuente: elaboración propia a partir de los marcos de referencia de la evaluación de 2009 (OECD, 2009).

Las diferencias entre países en lo relativo a la organización de los sistemas educativos complejiza la identificación de los agentes y niveles implicados. No obstante, el sistema educativo en última instancia, depende de las interacciones y soportes al aprendizaje que se dan en el nivel fundamental: la escuela. Es a nivel escolar donde el sistema educativo interactúa con la comunidad, con los padres y con los estudiantes partiendo de especificaciones establecidas en otros niveles (OECD, 2009). En la Tabla 26 puede observarse la estructura con los niveles y las variables consideradas en cada nivel en la edición de PISA 2009.

Tal y como veíamos anteriormente, PISA recoge la información de contexto por medio del uso de dos cuestionarios, uno destinado a las escuelas (que es contestado por el director del centro o por un responsable designado) y el cuestionario del alumno (completado por los estudiantes los días de evaluación). Como complemento a esta información, los países pueden decidir participar en otros tres cuestionarios que aportan información complementaria que enriquece la información obtenida (OECD, 2009). De este modo, los cuestionarios de contexto serían: cuestionario de escuelas, cuestionario de estudiantes, cuestionario para padres (opcional), cuestionario sobre la trayectoria educativa (opcional), cuestionario de familiaridad con el uso de nuevas tecnologías de la información y la comunicación (opcional).

Algunas de las variables del cuestionario de contexto son utilizadas directamente a partir de los datos provenientes de los cuestionarios (como sucede con la variable género), sin embargo, gran número de variables son utilizadas para la construcción de índices o variables latentes que no pueden ser directamente observadas (OECD, 2012c). En estas transformaciones, se debe tener en cuenta una importante distinción entre aquellos índices que son calculados directamente por medio de una transformación aritmética o recodificando uno o más ítems, denominados índices simples, y aquellos en los que es preciso utilizar procedimientos de escalamiento. Normalmente las escalas de puntuación de estos ítems son estimadas a partir de modelos de variable latente derivados de la TRI para ítems dicotómicos o politómicos, a estos índices se los denomina de escala (OECD, 2012c) (Tabla 27).

Tabla 27.

Índices construidos a partir de la información de los cuestionarios de contexto

INDICES CONSTRUIDOS A PARTIR DE LA INFORMACIÓN DE LOS CUESTIONARIOS DE CONTEXTO	
ÍNDICES SIMPLES	ÍNDICES DE ESCALA
Cuestionario del alumno: ✓ Edad del estudiante. ✓ Programa de estudios ✓ Máximo nivel ocupacional de los padres (HISEI) a partir de códigos ISCO (4 dígitos). ✓ Nivel educativo de los padres (ISCED) ✓ Condición de inmigrante (IMMIG). ✓ Lengua utilizada en el hogar (LANGN) ✓ Estructura familiar (FAMSTRUC) ✓ Grado relativo (GRADE) ✓ Tiempo de aprendizaje (en lectura, matemáticas y ciencias). ✓ Meta - cognición. ✓ Ocupación de los padres (Blue-collar/white-collar) **Cuestionario del centro:** ✓ Tamaño del aula (SCHLSIZE). ✓ Proporción de chicas matriculadas en la escuela (PCGIRLS). ✓ Tipo de centro (SCHLTYPE). ✓ Disponibilidad de ordenadores (COMPWEB) ✓ Selección de estudiantes por parte del centro (SELSCH). ✓ Agrupación de estudiantes (ABGROUP) ✓ Responsabilidad de la escuela en la asignación de recursos (RESPRES).	**Cuestionario del alumno:** ✓ Posesiones en el hogar (HOMEPOS). - Riqueza familiar (WEALTH) - Posesiones culturales (CULTPOSS) - Recursos educativos en el hogar (HEDRES) - Recursos tecnológicos (ICTRES) ✓ Disfrute y frecuencia de la lectura (área prioritaria en 2009 lectura). ✓ Actitud ante la escuela y entornos de aprendizaje. ✓ Estrategias docentes. ✓ Disponibilidad de recursos informáticos. ✓ Uso de nuevas tecnologías. ✓ Autoeficacia en el uso de nuevas tecnologías. ✓ Actitud ante los ordenadores. **Cuestionario del centro.** ✓ Escasez de maestros (TCSHORT). ✓ Recursos educativos en la escuela (SCMATEDU). ✓ Actividades extracurriculares (EXCURACT). ✓ Liderazgo educativo (LDRSHP). ✓ Participación docente (TCHPARTI). ✓ Profesorado y clima (TEACBEHA). ✓ Estudiantes y clima (STUDBEHA). **Cuestionario de padres.** ✓ Percepción de los padres sobre la calidad de la escuela (PQSCHOOL). ✓ Implicación de los padres (PARINVOL) ✓ Recursos de lectura en casa (READRES). ✓ Apoyo al aprendizaje de la lectura (CURSUPP). ✓ Apoyo al aprendizaje de la lectura en los primeros años (PRESUPP). ✓ Motivación a la lectura de los padres (MOTREAD).

Fuente: elaboración propia a partir de OECD (2012c).

Como hemos podido apreciar a lo largo del presente apartado, la novedad fundamental de PISA se sitúa en la evaluación de competencias y en la adopción de técnicas y procedimientos psicométricos de calidad, puestos a prueba en evaluaciones anteriores de similares características. El interés creciente en los resultados de PISA, así como su evolución y mejora, convierten ésta evaluación en un referente mundial en el ámbito de la evaluación de sistemas educativos.

Teaching and Learning International Survey (TALIS)

El estudio *«Teaching and Learning International Survey»* (TALIS), se desarrolló como parte del proyecto de indicadores del sistema educativo de la OECD (proyecto INES), puesto que para dicho proyecto era esencial contar con información

sobre los profesores, la enseñanza y el aprendizaje. TALIS surge con el objetivo de proporcionar información sobre las condiciones necesarias para una enseñanza eficaz en las escuelas (OECD, 2010b). La primera edición se llevó a cabo en el curso escolar 2007-2008, siendo su diseño de carácter periódico, su última edición se llevó a cabo en 2013.

El estudio se centra en docentes de Educación Secundaria Obligatoria («*International Standard Classification of Education*» ISCED nivel 2), aunque en la edición de 2013 los países pueden decidir si ampliar el estudio a docentes de primaria (ISCED nivel 1) o de educación secundaria superior (ISCED nivel 3). Del mismo modo, los países participantes pueden ampliar el estudio aplicando el denominado "enlace" TALIS-PISA, en aquellos centros que han participado en PISA (OECD, 2014). El estudio está dirigido a centros públicos y privados. Combinando los aspectos tratados en las ediciones de 2008 y 2013 podemos decir que TALIS examina aspectos estructurados en las siguientes áreas: liderazgo, desarrollo profesional (formación permanente e inicial), evaluación de profesores y «*feedback*», prácticas pensamientos y actitudes (incluyendo prácticas de evaluación) y sentimiento de autoeficacia, satisfacción con el trabajo y clima en la escuela y en el aula.

En su primera edición, el estudio fue realizado en 24 países, participando más de 70.000 docentes y 4.000 directores (OECD, 2010b). En la edición de 2013 el número de países asciende, pasando a formar parte del estudio 108.000 profesores de 34 países (OECD, 2014). España participó en ambas ediciones. Los principios que guiaron el desarrollo del proyecto TALIS fueron relevancia política, valor añadido, orientación a indicadores, validez, fiabilidad, comparabilidad y rigor, interpretabilidad y relación costo- efectividad (OECD, 2010b; 2014).

En la Tabla 28 se muestran las principales preocupaciones políticas y dominios relacionados con los indicadores propuestos para la primera evaluación TALIS (OECD, 2010b).

Tabla 28.

Preocupaciones políticas, dominios e indicadores propuestos para la evaluados en TALIS.

Propuesta de preocupaciones políticas, dominios e indicadores para la evaluados TALIS	
Dimensiones	**Indicadores**
Atraer hacia la profesión docente	Adecuación de la oferta docente y el déficit de maestros (1). Perfil del nuevo profesorado (2). Motivaciones y experiencia de los nuevos profesores (3). Eficacia de los procedimientos de selección, contratación e incentivos (4).
Desarrollo profesional docente	Perfil de formación del docente (5). Frecuencia y distribución de educación y formación (6). Satisfacción y efectividad de la formación docente (7).
Mantenimiento de docentes en la profesión	Desgaste docente y rotación (8). Satisfacción con el trabajo y las medidas de recursos humanos (9). Reconocimiento, retroalimentación, recompensa y evaluación docente (10).
Políticas educativas y eficacia.	Liderazgo educativo (11). Clima escolar (12).
Calidad de los profesores y la enseñanza	Prácticas de los profesores, creencias y actitudes (13). Calidad del profesorado (experiencia, satisfacción, responsabilidad) (14). División del tiempo de trabajo (15).

Fuente: adaptado de (OECD, 2010b) (p.26).

Una vez establecidos estos 15 indicadores, los países participantes valoraron la importancia asignada a cada uno de ellos, con el fin de establecer cuáles deberían ser las prioridades de la evaluación, seleccionando como principales áreas de interés para la evaluación TALIS 2008: reconocimiento, retroalimentación, recompensa y evaluación docente (10); Liderazgo educativo (11) y prácticas de los profesores, creencias y actitudes (13). Además, se incluyeron los tres indicadores de la dimensión de desarrollo profesional: perfil de formación del docente (5), frecuencia y distribución de educación y formación (6), satisfacción y efectividad de la formación docente (7) y el indicador de eficacia de los procedimientos de selección, contratación e incentivos (4), por su complementariedad con los tres indicadores principales así como por su importancia para los países con el fin de obtener información identificada como clave por la Comisión Europea (OECD, 2010b).

Algunos indicadores poco valorados (clima escolar (12), división del tiempo de trabajo (15) y satisfacción con el trabajo y las medidas de recursos humanos (9)), también fueron incluidos por considerarse que podrían aportan importante información complementaria en el análisis de los temas principales (OECD, 2010b). En la edición de

2013 se mantienen globalmente estos indicadores de referencia, especificando algunos aspectos tales como la formación inicial y continua, la autoeficacia docente, satisfacción con el trabajo y clima en la escuela y en el aula.

Algunos de los ítems del cuestionario del profesor y del cuestionario del director, han sido aplicados con el fin de utilizar sus respuestas de forma directa, sin embargo, gran número de ítems fueron aplicados con el propósito de combinarlos de algún modo para obtener información acerca de variables latentes que no pueden ser observadas de manera directa (OECD, 2010b). Tal y como sucedía en PISA, podemos distinguir índices simples (aquellos que solo implican una transformación aritmética o recodificación de uno o más ítems) o índices complejos (en los que se precisa un proceso de escalamiento complejo para la medida de las variables latentes). En la Tabla 29 se pueden observar los índices simples y complejos utilizados en la evaluación TALIS 2008.

Tabla 29.

Índices simples y complejos de la Evaluación TALIS

ÍNDICES ORIGINALES CONSTRUIDOS A PARTIR DE LA INFORMACIÓN DE LOS CUESTIONARIOS A PROFESORES Y DIRECTORES EN LA PRIMERA EDICIÓN	
ÍNDICES SIMPLES	ÍNDICES COMPLEJOS
✓ Ratio profesor alumno. ✓ Número de alumnos por personal de apoyo pedagógico. ✓ Número de estudiantes por personal administrativo. ✓ Número de profesores por personal de apoyo pedagógico ✓ Número de profesores por personal administrativo ✓ Porcentaje de horas de formación de profesores obligatorias. ✓ Tamaño medio de clase. ✓ Diferentes lenguas. ✓ Nivel educativo de los padres. ✓ Ausencia de evaluaciones de la escuela. ✓ No apreciación o retroalimentación recibida por el profesorado.	**Índices derivados del análisis de componentes principales.** ✓ Índices de autonomía del centro. o Contratación y salarios docentes. o Presupuesto. o Currículum. Política estudiantil /libros. ✓ Índices de recursos. o Falta de recursos personales. o Falta de recursos materiales. **Índices derivados del análisis factorial confirmatorio.** ✓ **Liderazgo educativo.** o Enmarcar y comunicar los objetivos educativos y el desarrollo curricular. o Promocionar la mejora de la enseñanza y el desarrollo profesional. o Supervisión de la enseñanza. o Papel del director en la rendición de cuentas. o Seguimiento burocrático. ✓ **Dimensiones del liderazgo.** o Instructivo. o Administrativo ✓ **Clima escolar.** o Delincuencia estudiantil (directores). o Ánimo- moral de los docentes (directores). o Relaciones profesor alumno (profesores). ✓ **Clima de disciplina en el aula.** ✓ **Autoeficacia docente.** ✓ **Creencias sobre la enseñanza tradicional y constructivista.** ✓ **Prácticas docentes.** o Estructura. o Orientación al estudiante. o Actividades de mejora. ✓ **Cooperación del personal.** o Intercambio y coordinación para le enseñanza. o Colaboración profesional.

Fuente: elaboración propia a partir de OECD (2010).

En definitiva, el proyecto TALIS, examina con precisión y rigor metodológico aspectos importantes relacionados con el proceso de enseñanza aprendizaje a través de la opinión de docentes y directores, dicha información resulta de utilidad para todos los agentes implicados en el proceso educativo, profesores, directores, padres, gestores, políticos, etc. No obstante, no debemos olvidar que la interpretación de los resultados ha de realizarse de manera contextualizada, teniendo en cuenta que los mismos provienen de cuestionarios de opinión, con las limitaciones que ello conlleva.

***Assessment of Higher Education Learning Outcomes* (AHELO).**

El proyecto AHELO «*Assessment of Higher Education Learning Outcomes*», es el primer proyecto de estas características desarrollado a nivel internacional y uno de los últimos grandes proyectos del ámbito educativo liderados por la OECD. El objetivo principal del mismo, es obtener información acerca de lo que los estudiantes de educación superior saben y son capaces de hacer tras su graduación.

La educación superior es un ámbito de creciente importancia, tanto para el progreso de los individuos como para el crecimiento de los países, siendo un factor esencial en la innovación y el crecimiento del capital humano. Tal y como se reconoce en el preámbulo de la Declaración Mundial de la Educación Superior de la UNESCO (1998), en los últimos años se ha observado un aumento en la demanda de educación superior, unida a una diversificación de la misma y a la toma de conciencia de su importancia en el desarrollo sociocultural y económico. En dicha declaración, se expone que la segunda mitad del siglo XX ha sido la mayor época de expansión universitaria a escala mundial, el número de estudiantes matriculados se multiplicó por más de seis entre 1960 (13 millones) y 1995 (82 millones) (UNESCO, 1998). A pesar del gran número de estudiantes matriculados en estudios universitarios, y del reconocimiento de la importancia de los mismos a nivel social e individual, no existen datos comparativos de carácter internacional acerca de la calidad del proceso de enseñanza-aprendizaje en este nivel, los pocos estudios existentes son realizados a nivel nacional o autonómico y los conocidos «*rankings*» de universidades, a pesar de realizarse en un marco internacional, se basan en aspectos que no reflejan la calidad del proceso de enseñanza aprendizaje, en esencia referidos a la producción científica en revistas y al impacto de dicha producción.

Desde el año 2008 hasta el año 2012, la OECD realizó el estudio de viabilidad del proyecto AHELO (inicialmente denominado PISA para educación superior) con el fin de analizar sus posibilidades científicas y prácticas. Este estudio de viabilidad se centra en las tres áreas que pretenden ser evaluadas en el proyecto: las competencias genéricas, las competencias en el área de economía y las competencias en el área de ingeniería, así como en la posibilidad de llevar a cabo un análisis basado en los modelos de valor añadido (Tremblay, Lalancette, & Roseveare, 2012).

Dentro de la dimensión de competencias genéricas se encontrarían el pensamiento crítico, el razonamiento analítico, la resolución de problemas y la comunicación escrita (Tremblay, Lalancette, & Roseveare, 2012). Los países participantes en el análisis de las competencias genéricas fueron Colombia, Egipto, Finlandia, Corea, Kuwait, México, Noruega, Eslovaquia y Estados Unidos (Connecticut, Missouri y Pensilvania) (Tremblay, Lalancette, & Roseveare, 2012). En el área de economía, se pretenden examinar las competencias específicas en los graduados en economía y los países participantes fueron Bélgica (Flandes), Egipto, Italia, Méjico, Holanda, Rusia y Eslovaquia (Tremblay, Lalancette, & Roseveare, 2012). El marco de referencia provisional y las pruebas, fueron elaboradas por docentes especialistas en economía, planteando como objetivos la evaluación de los conocimientos que los estudiantes deberían alcanzar al finalizar el titulo de grado, demostrando su conocimiento y comprensión. Del mismo modo, en el área de ingeniería, el objetivo es evaluar las competencias específicas en ingeniería, habiendo sido diseñado el marco de referencia y el instrumento provisional por ingenieros especialistas en educación, planteando el objetivo de evaluar los conocimientos y el grado de comprensión que los estudiantes deberían alcanzar al finalizar sus estudios de ingeniería. Los países participantes en este caso fueron Abu Dhabi, Australia, Canadá (Ontario), Colombia, Egipto, Japón, Méjico, Rusia y Eslovaquia (Tremblay, Lalancette, & Roseveare, 2012).

La primera fase del estudio consistió en el análisis del marco teórico de referencia y la viabilidad de la elaboración de marcos e instrumentos de evaluación con la validez suficiente para el análisis de diversas realidades lingüísticas, culturales e institucionales (Tremblay, Lalancette, & Roseveare, 2012, p. 88).

Además de esta evaluación acerca de los resultados de aprendizaje tanto genéricos como específicos, AHELO incorpora variables contextuales que ayudarán a una mejor comprensión y análisis de los resultados. La información contextual proviene de varios niveles de desagregación como pueden ser la institución, el programa o los estudiantes. Dicha información, es obtenida a partir de los datos y documentación existentes en cada país y a partir de tres cuestionarios de contexto dirigidos a: estudiantes, institución y facultad.

Por último, en el análisis de viabilidad del proyecto AHELO, se estudió la posibilidad de medición de las mejoras y progresos en los resultados de los estudiantes, teniendo en cuenta la contribución de la institución (medidas de valor añadido). Por tanto, este eje complementario, investigará acerca de la viabilidad y utilidad de las medidas de valor añadido en educación superior, que pretenden analizar cuál es el efecto de las instituciones de educación superior en el aprendizaje de los estudiantes. El estudio de viabilidad del proyecto AHELO se centró precisamente en el análisis pormenorizado de aspectos metodológicos, técnicos y psicométricos que puedan dar lugar a un estudio de calidad en el futuro.

Entre las principales conclusiones del estudio destacan el buen funcionamiento de gran número de ítems y la detección de funcionamiento diferencial en alguno de ellos, siendo especialmente sensibles los ítems de respuesta construida. Todos los instrumentos alcanzaron niveles adecuados de validez de constructo y los instrumentos de economía e ingeniería también obtuvieron una adecuada validez de contenido, los niveles de validez aparente en las tres dimensiones evaluadas también han resultado óptimos, sin embargo, los resultados acerca de la validez concurrente son menos concluyentes (Tremblay, Lalancette, & Roseveare, 2013). En relación a la fiabilidad, los tres instrumentos presentan buena fiabilidad global, fiabilidad que desciende en los niveles de institución o país, apuntando a la necesidad de ciertas modificaciones para su mejora. La fiabilidad inter-jueces puede ser considerada buena en las tres áreas evaluadas (Tremblay, Lalancette, & Roseveare, 2013). En definitiva, el estudio de viabilidad AHELO permitió demostrar que es posible desarrollar instrumentos con suficiente fiabilidad y validez para distintos países, lenguajes, culturas e institucionales (Tremblay, Lalancette, & Roseveare, 2013).

Los sistemas de evaluación implementados por la OECD mantienen, desde sus orígenes, su interés por la medida del progreso o cambio educativo desde una perspectiva dinámica, atendiendo a la realidad educativa desde un enfoque ajustado a sus características y posibilitando la realización de comparaciones internacionales de diversa índole. Así, el rigor metodológico requerido para asegurar la comparabilidad es una preocupación transversal que caracteriza a todas sus propuestas evaluativas.

Programme for the International Assessment of Adult Competencies (PIAAC).

El Programa Internacional para la Evaluación de las Competencias de la Población Adulta (PIAAC), iniciado en 2008, está diseñado para proporcionar información acerca de la disponibilidad de las denominadas competencias clave en la sociedad y la forma en que estas competencias son utilizadas en el trabajo y en la vida cotidiana (OECD, 2013). Las competencias evaluadas son lectura, matemáticas y resolución de problemas en entornos informatizados (esta última no en todos los países), siendo consideradas "clave" en el procesamiento de información.

El estudio recoge información de cómo son utilizadas dichas competencias en casa, en el trabajo y en la comunidad; cómo son desarrolladas, mantenidas o perdidas a lo largo de la vida; como están relacionadas con la actividad laboral, los ingresos, la salud y la participación social y política (OECD, 2013). En consecuencia, vemos como el estudio PIAAC, persigue un objetivo complejo, pues en última instancia, deberá proveer información de interés a los países y entidades participantes, tal y como figura en el informe de resultados, el objetivo último es que la información proveniente del estudio PIAAC, ayude a los legisladores a:

- Examinar la importancia de la competencia lectora, matemática y de resolución de problemas en una serie de resultados económicos y sociales.
- Evaluar la adecuación de los sistemas de educación y formación y de las prácticas laborales y sociales en el desarrollo de las habilidades requeridas por el mercado laboral y por la sociedad en general.
- Identificar los instrumentos políticos que pueden permitir reducir las deficiencias en competencias clave (OECD, 2013).

En el estudio, cuya recogida de datos se llevó a cabo en 2011/ 2012, participaron unos 166.000 adultos, de edades comprendidas entre los 16 y los 65 años, pertenecientes a 24 países o regiones entre los que se encuentra España (OECD, 2013).

La valiosa información proporcionada en los primeros resultados, publicados en 2013, confirma lo ya apunado anteriormente, la importancia de esta evaluación en la estrategia global de la OECD así como sus implicaciones políticas, sociales e

institucionales. Entre los resultados destacados en el primer informe, se pone de manifiesto que: en la mayor parte de los países existe una elevada proporción de población adulta situada en los niveles más bajos de las competencias evaluadas; gran parte de la población no tiene experiencia o no posee las competencias requeridas para el uso de nuevas tecnologías: tan solo un porcentaje situado entre el 2.9% y el 8.8% de los adultos demostraron el más alto nivel en la competencia de resolución de problemas en entornos informatizados (OECD, 2013).

Además de el análisis primario de los resultados, el estudio PIAAC, ha dado lugar a interesantes análisis secundarios por parte de la comunidad científica. Así, a pesar de la reciente publicación de resultados, debemos destacar el trabajo de Hanushek, Schwerdt, Wiederhold y Woessmann (2015). En dicho estudio, los autores señalan que mayores niveles de competencias cognitivas (medidas a través de la competencia matemática, lectora y de resolución de problemas), están sistemáticamente relacionadas con salarios más altos en los 23 países analizados (Hanusheck et al., 2015). Del mismo modo, ponen de manifiesto la importancia de subrayar que casi toda la discusión internacional en torno a la política educativa se centra en la calidad de la escuela y en el rendimiento del estudiante, sin embargo, para entender las implicaciones económicas completas de estas discusiones, es necesario considerar los resultados del mercado de trabajo en vinculación con los resultados escolares (Hanusheck et al., 2015).

Otro trabajo interesante publicado recientemente, es el llevado a cabo por Meroni, Vera-Toscano y Costa (2015), estudio en el que analizan la influencia de las competencias docentes en los resultados de los estudiantes, a través de la vinculación de los resultados de PISA y PIAAC. Otra aproximación que analiza la vinculación de los resultados de PISA y PIAAC, utilizando exclusivamente la muestra Alemana, es la realizada Cortina (2015). Así, en los meses posteriores a la publicación de los resultados, la evaluación PIAAC está dando lugar a interesantes análisis secundarios que utilizan diferentes aproximaciones a los mismos e incluso vinculan sus datos con los de otras evaluaciones como PISA.

La riqueza de la propuesta, así como la novedad de sus resultados, harán de PIAAC todo un referente en evaluación de competencias en población adulta.

1.4.3 Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (LLECE).

El Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (LLECE) es una red de sistemas de evaluación de la calidad de la educación en América Latina y el Caribe. Es coordinado por la Oficina Regional de Educación de la UNESCO para América Latina y el Caribe y tiene su sede en Santiago de Chile. Fue creado en el año 1994, suponiendo una gran oportunidad para el desarrollo de nuevas formas de cooperación internacional (Valdés et al., 2008). Dentro de sus objetivos se encuentran la producción e información sobre logros de aprendizaje de los alumnos y el análisis de los factores asociados a los mismos, generar conocimiento sobre evaluación de sistemas educativos, aportar nuevas ideas y enfoques sobre evaluación de la calidad de la educación y contribuir al fortalecimiento de las unidades locales de evaluación. Los países que forman parte del LLECE actualmente son: Argentina, Brasil, Chile, Colombia, Costa Rica, Ecuador, Guatemala, Honduras, México, Nicaragua, Panamá, Paraguay, Perú, República Dominicana y Uruguay, junto al estado mexicano de Nuevo León (Pedraza et al., 2013).

Estudio Internacional Comparativo sobre Lenguaje, Matemáticas y Factores asociados, primera edición (PERCE, 1997), segunda edición (SERCE, 2006) y tercera edición (TERCE, 2013).

El primer estudio internacional comparativo sobre lenguaje, matemáticas y factores asociados fué realizado en el año 1997, partiendo del acuerdo de 13 países de la región y suponiendo un gran hito al permitir la obtención, por primera vez, de información comparativa sobre los logros de aprendizaje de los alumnos de América Latina y el Caribe. Los cursos evaluados fueron 3º y 4º grado y además de la evaluación de las áreas de matemáticas y lenguaje, se incluyeron cuestionarios de factores contextuales dirigidos a alumnos, directores, profesores y tutores (Casassus, Froemel, Palafox, & Cusato, 1998). Una de sus ventajas más destacadas es la posibilidad de contar con datos comparativos entre los países con un carácter contextualizado y con mayor referencia a los currículos oficiales, apoyándose en las similitudes y diferencias existentes entre ellos. El objetivo de la comparabilidad se presenta como la nota definitoria del estudio, con los compromisos metodológicas que ello conlleva. La

necesidad de realizar la medida de logro de las áreas de matemáticas y lenguaje a través de una única prueba (en cada área) que contemple los aspectos curriculares de todos los países participantes, constituye una aportación sin precedentes (Casassus et al., 1998).

La población objetivo del estudio está compuesta por todos los niños y niñas que están cursando tercer y cuarto grado de Educación Básica (Casassus, Cusato, Froemel, & Palafox, 2001). El procedimiento de muestreo utilizado en este primer estudio fue un diseño estratificado, no proporcional a la población total de cada país, de selección bi-etápica aleatoria, con probabilidades iguales para todos los alumnos y no auto-ponderada (LLECE, 2001).. La prueba de lenguaje fue aplicada a un total de 54.589 estudiantes y la prueba de matemáticas a 54.417 alumnos, obteniéndose información contextual complementaria de 41.088 tutores (personas a cargo del niño), 3.675 maestros, 1.387 directores y 1.509 escuelas (LLECE, 2001).

Tabla 30.

Dimensiones y niveles de desempeño PERCE

Dimensiones y niveles de desempeño			
Lenguaje		Matemática	
Competencia comunicativa Comprensión e interpretación textual		Competencia matemática. Resolución de problemas	
Niveles de desempeño	I. lectura literal-primaria. II. lectura de carácter literal en modo de paráfrasis. III. lectura de carácter inferencial.	Niveles de desempeño	I. Reconocimiento y utilización de hechos y relaciones matemáticas básicas. II. Reconocimiento y utilización de estructuras matemáticas simples. III. Reconocimiento y utilización de estructuras matemáticas complejas

Fuente: elaboración propia a partir de (LLECE, 2001).

Para la construcción de las escalas tanto de matemáticas como de lenguaje (Tabla 30), así como para la equiparación de las distintas formas de las pruebas, se utilizó el modelo de Rasch. La escala fue establecida con una media internacional de 250 puntos y una desviación estándar de 50 (Casassus et al., 1998). Para los análisis de resultados se utilizó metodología multinivel (modelos jerárquicos lineales). Se consideraron tres niveles: alumno, escuela y país (Casassus et al., 1998).

Continuando con la misma filosofía y objetivos, el Segundo Estudio Regional. Comparativo y Explicativo (SERCE), fue llevado a cabo en el año 2006, con el propósito de generar conocimiento acerca de los aprendizajes en el área de Matemáticas

y Lenguaje (Lectura y Escritura) en 3º y 6º grado y en Ciencias de la Naturaleza para estudiantes de 6º (Valdés et al., 2008). En esta segunda edición participaron 196,040 estudiantes pertenecientes a 3.065 escuelas (Valdés et al., 2008). A pesar de sus similitudes con PERCE, es necesario destacar la imposibilidad de realizar comparaciones para evaluar el progreso de los países analizando los resultados de ambos estudios por diversas razones: diferencias en los cursos evaluados 3º/4º y 3º/6º, incorporación de las habilidades para la vida en la segunda edición, diferencias en los países participantes (tan solo 8 tienen puntuaciones en ambas ediciones), diferencias en la muestra, diferencias en los instrumentos utilizados, en la estandarización de los procedimientos y en el proceso de supervisión y aplicación (Valdés et al., 2008). Como vemos, el problema de comparabilidad se pone de manifiesto a la hora de trabajar con dos de las ediciones de una misma evaluación, por problemas tanto en el diseño como en los objetivos y contenidos.

Tabla 31.

Dominios y niveles de desempeño SERCE

Dominios y niveles de desempeño			
	Lenguaje y Comunicación	Matemáticas	Ciencias
Dominio	Lectura Metalinguistico Escritura	Numérico Geométrico Medición Estadístico Variacional	Seres vivos Tierra y medio ambiente Materia y energía Ciencia, Tecnología y Sociedad
Niveles de desempeño	• Literal. • Inferencial • Crítico intertextual	• Reconocimiento de objetos y elementos • Solución de problemas simples • Solución de problemas complejos	• Reconocimiento de conceptos y elementos • Solución de problemas simples • Solución de problemas complejos

Fuente: elaboración propia a partir de (Bogoya Maldonado (Coord), 2005).

En la Tabla 31, podemos observar los dominios y niveles de desempeño evaluados en Lenguaje y Comunicación, Matemáticas y Ciencias, las diferencias en relación a la primera edición son notables, constituyendo una de las dificultades esenciales a la comparabilidad entre ambos.

La tercera evaluación del aprendizaje en Educación Primaria, llevada a cabo por el LLECE, lleva por título Tercer Estudio Regional Comparativo y Explicativo

(TERCE). En dicho estudio, se evalúan las áreas de matemáticas y lenguaje en tercer y sexto grado, y el área de ciencias en sexto grado. Las pruebas, de acuerdo con el trabajo llevado a cabo en ediciones previas, hacen referencia a elementos comunes en los currículos escolares. La muestra efectiva es de 3.065 escuelas pertenecientes a 15 países, participando un total de 195.752 estudiantes y 9.965 docentes. Una de las aportaciones específicas del TERCE es su posibilidad de comparación con el estudio precedente (SERCE), situación que permitirá a los países participantes en ambas ediciones comparar su puntaje y analizar las tendencias en los sistemas educativos. Como puede observarse en la comparativa de las Tablas 31 y 32, en esta última edición, se han conservado ampliamente los dominios y niveles de desempeño en las tres áreas evaluadas, con el objetivo de afrontar el reto de la comparabilidad y conseguir una utilidad mayor en los resultados.

Tabla 32.

Dominios y niveles de desempeño TERCE

Dominios y niveles de desempeño			
	Lenguaje y Comunicación	Matemáticas	Ciencias
Dominio	Comprensión textual (Lectura) (a) Metalinguistico (b) Escritura (c)	Numérico Geométrico Medición Estadístico Variacional	Salud Seres Vivos Ambiente La tierra y el sistema solar Materia y energía Ciencia, Tecnología y Sociedad
Niveles de desempeño	• Literal (a) • Inferencial (a) • Crítico intertextual(a) • Pertinencia de las categorías gramaticales en los textos (b) (c) • Conceptualización sobre los tipos de texto (b) (c) y la organización formal de los mismos (c)	• Reconocimiento de objetos y elementos • Solución de problemas simples • Solución de problemas complejos	• Reconocimiento de información y conceptos • Comprensión y aplicación • Pensamiento científico y resolución de problemas

Fuente: elaboración propia a partir de (Pedraza Daza et al., 2013).

Uno de los aspectos esenciales en las tres ediciones de los estudios del laboratorio latinoamericano de evaluación de la calidad educativa, radica en el análisis de las variables contextuales y su relación con el desempeño. A modo de síntesis, la Tabla 33, agrupa los principales índices utilizados en las dos primeras ediciones del estudio.

El LLECE, tal y como sucedía con la IEA o la OECD, presentan en sus evaluaciones una nota característica, la comparabilidad de puntuaciones como

componente esencial de sus propuestas. Así, desde la realización del Primer Estudio Internacional Comparativo sobre Lenguaje, Matemáticas y Factores asociados (PERCE, 1997) dieciocho años atrás, el LLECE se ha enfrentado al reto de mantener sus escalas, permitir las comparaciones internacionales, elaborar distintas formas de una misma prueba por cuestiones de seguridad en el proceso, evaluar un amplio dominio competencial, etc.

Tabla 33.

Factores contextuales simples y compuestos por áreas de observación considerados por el LLECE

ÍNDICES CONSTRUIDOS A PARTIR DE LA INFORMACIÓN DE LOS CUESTIONARIOS DE CONTEXTO	
ÍNDICES SIMPLES. ÁREAS DE OBSERVACIÓN	ÍNDICES COMPUESTOS.ÁREAS DE OBSERVACIÓN
Del Alumno ✓ Género. ✓ Grado. ✓ Indígena. ✓ Asistencia a preescolar. ✓ Trabajo. ✓ Repetidor. **Contexto familiar** ✓ Nivel educativo de los padres. ✓ Tiempo en casa de la madre los días de trabajo. ✓ Recursos de lectura en el hogar, bi-parentalidad en el hogar. ✓ Frecuencia de participación en las actividades relacionadas con la escuela. ✓ Conocimiento del profesor. ✓ Frecuencia de asistencia a reuniones escolares. **Del maestro y el ámbito educativo** ***Escuela*** ✓ Rural. ✓ Urbana. ✓ Titularidad. ***Currículo:*** ✓ Tiempo instruccional por materia. ***Práctica Pedagógica:*** ✓ Años de experiencia enseñando. ✓ Años de actividades de capacitación. ✓ Números de cursos de pedagogía en los últimos tres años. ✓ Trabajo paralelo. ✓ Tipo de evaluación. ✓ Criterios de agrupación de los alumnos. ✓ Satisfacción con el salario. ✓ Liderazgo del director. ✓ Condiciones de trabajo, satisfacción laboral, adecuación del horario de trabajo, autonomía para la práctica pedagógica. ✓ Causas percibidas de los resultados. ✓ Expectativas de rendimiento. **Del Director y el Microcosmos Escolar.** ✓ Tipo de grupo (simple o multi-grado). ✓ tasa de alumnos maestro. ✓ Infraestructura. ✓ materiales de instrucción. ✓ Número de libros en la biblioteca. ✓ Autonomía de la escuela. ✓ Autonomía para tomar decisiones administrativas. ✓ Amistad entre los compañeros. ✓ Grado en que los compañeros se molestan mutuamente. **De las Autoridades Públicas y el Macrocosmos** ✓ Tipo de gestión por parte del estado (dependencia administrativa). ✓ Educación inicial (asistencia a institución preescolar).	**Del alumno y su contexto familiar** ✓ Nivel socioeconómico y cultural (ISEC). - Nivel educativo de los padres. - Tiempo en casa de la madre los días de trabajo - Recursos de lectura en el hogar. - Bi-parentalidad en el hogar. ✓ Educación preescolar. ✓ Involucramiento de los padres (1ªedicion)/ Participación de la familia en la escuela. - Participación en actividades escolares. - Conocimiento del profesor. - Frecuencia de asistencia a reuniones escolares. ✓ Prácticas educativas del hogar. - Frecuencia con que la familia lee al niño. - Profesor Privado. - Hacer la tarea y hablar de la escuela. - Lectura independiente. ✓ Clima positivo/ Clima negativo. ✓ Disponibilidad de libros y materiales. **Del maestro y el ámbito educativo** ✓ Actitudes de los profesores/ Satisfacción docente. - Satisfacción con el salario. - Liderazgo del director. - Condiciones de trabajo, satisfacción laboral, adecuación horario, autonomía pedagógica. ✓ Atribuciones de rendimiento. - Causas percibidas de los resultados. - Expectativas de rendimiento. ✓ Organización del aula (sexto grado). **Del Director y el Microcosmos Escolar** ✓ Autonomía del director. - Autonomía de la escuela. - Autonomía para tomar decisiones administrativas. ✓ Clima del aula. - Amistad entre los compañeros. - Grado de molestias mutuas entre compañeros. **Índices compuestos creados a partir de la información de diversas áreas** ✓ Cultura Escolar. - Trabajo paralelo. - Actitudes de los profesores. - Autonomía del director. ✓ Práctica del Aula. - Tipo de grupo. - Tipo de evaluación y criterios de agrupación. - Tiempo instruccional (Lenguaje y Matemáticas). - Involucramiento de los padres. - Clima en el aula. ✓ Recursos de la Escuela. - Años de experiencia del profesorado enseñando. - Años enseñando en la escuela. - Tasa de alumnos maestro. - Infraestructura, materiales de instrucción. - Número de libros en la biblioteca. **Índices a nivel de escuela** ✓ Instalaciones, servicios básicos, infraestructura. ✓ Programas compensatorios (programas asistenciales gratuitos).

Fuente: elaboración propia a partir de (Casassus et al., 2001; Treviño, Katherine, Gempp, & Donoso Rivas, 2013).

1.4.4 Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ)

El «*Southern and Eastern Africa Consortium for Monitoring Educational Quality»* (SACMEQ) es una organización internacional sin ánimo de lucro, compuesta inicialmente por 7 Ministerios de Educación de países de la zona sur y este de África y que se ha ido ampliando progresivamente hasta contar con 15 en su último proyecto. El objetivo principal es compartir experiencias y conocimientos con el fin de aplicar métodos científicos para evaluar las condiciones de escolarización y la calidad de la educación, contando con la asistencia del «*International Institute for Educational Planning* (IIEP)» de la UNESCO (Southern and Eastern Africa Consortium for Monitoring Educational Quality, 2010). Los 15 Ministerios de Educación que forman parte de SACMEQ pertenecen a los siguientes paises: Botswana, Kenya, Lesotho, Malawi, Mauritius, Mozambique, Namibia, Seychelles, South Africa, Swaziland, Tanzania (Mainland), Tanzania (Zanzibar), Uganda, Zambia y Zimbawe. Los proyectos llevados a cabo hasta el momento (SACMEQ I, SACMEQ II Y SACMEQ II) se encuentran estrechamente enlazados ya que SACMEQ I proporciona la línea base para SACMEQ II y SACMEQ III.

SACMEQ I (1995- 1998)

En este proyecto participaron siete Ministerios de Educación (Kenya, Malawi, Mauritius, Namibia, Tanzania (Zanzibar), Zambia y Zimbawe), cada uno de ellos encargado de la realización del informe nacional. Las pruebas aplicadas fueron utilizadas para explorar aspectos como: estudio inicial de inputs educativos, condiciones generales de escolarización o nivel educativo de los estudiantes de sexto grado en las áreas de comprensión lectora y matemáticas. Aproximadamente 20.000 estudiantes distribuidos en 1000 escuelas de primaria participaron en este primer proyecto (Hungi et al., 2010).

SACMEQ II (1998- 2004)

El doble de Ministerios de Educación (14) participó en esta segunda edición del proyecto SACMEQ (Botswana, Kenya, Lesotho, Malawi, Mauritius, Mozambique,

Namibia, Seychelles, South Africa, Swaziland, Tanzania (Mainland), Tanzania (Zanzibar), Uganda y Zambia). Este proyecto aporta una medida del cambio en las condiciones de escolarización y la calidad de la educación para seis países entre los años 1995 y 2000. En este segundo proyecto se contó con la participación de unos 40.000 estudiantes, 5.300 docentes, 2.000 directores y 2.000 escuelas (Hungi et al., 2010).

SACMEQ III (2005- 2010)

En la tercera edición del proyecto participaron 15 países (Botswana, Kenya, Lesotho, Malawi, Mauritius, Mozambique, Namibia, Seychelles, South Africa, Swaziland, Tanzania (Mainland), Tanzania (Zanzibar), Uganda, Zambia y Zimbawe). Este tercer proyecto aportó información sobre el cambio en las condiciones de escolarización y el rendimiento escolar entre 1995 y 2000 para seis Ministerios de Educación y entre 2000 y 2007 para 14. En este tercer proyecto participaron unos 61.396 estudiantes, 8.026 profesores y 2.779 escuelas (Hungi et al., 2010).

En los proyectos SACMEQ la población objeto de estudio la constituyen todos los estudiantes matriculados en sexto grado. La población es descrita de este modo debido a las dificultades de realizar la selección por edad, puesto que se dan altas tasas de escolarización tardía y repetición. Seleccionar a los sujetos participantes por edad implicaría la participación estudiantes de gran variedad de cursos escolares, situación que complicaría enormemente la aplicación (Hungi et al., 2010). Los proyectos SACMEQ utilizan el modelo de Rash para el cálculo de la puntuación en matemáticas y comprensión lectora de los estudiantes, posteriormente, a través de una transformación lineal, dichas puntuaciones se convierten a una escala con media 500 y desviación típica 100 con el fin de mejorar la interpretación de los resultados (Hungi et al., 2010).

Los ítems que componen las pruebas de matemáticas y comprensión lectora son organizados en un primer momento en función de su índice de dificultad. En segundo lugar, cada ítem es analizado para identificar qué competencia específica se requiere para su correcta resolución (Hungi et al., 2010). Posteriormente, se agrupan los ítems de similar dificultad y que comparten un "tema" común, en relación a la competencia requerida para dar respuesta correcta al reactivo (Hungi et al., 2010). Se identificaron

ocho competencias principales para cada una de las pruebas, dichas competencias se presentan en la Tabla 34.

Tabla 34.

Niveles competenciales para Comprensión Lectora y Matemáticas en las evaluaciones SACMEQ

	COMPRENSIÓN LECTORA	MATEMÁTICAS
1	Pre Lectura	Pre matemáticas
2	Nivel emergente	Nivel emergente
3	Nivel básico	Nivel Básico
4	Lectura de significado	Nivel Principiante
5	Lectura interpretativa	Nivel Competente
6	Lectura inferencial	Nivel Experto
7	Lectura analítica	Resolución de problemas concretos
8	Lectura crítica	Resolución de problemas abstractos

Fuente: Elaboración propia a partir de Hungi, et al. (2010).

Para la variable de localización de la escuela, los directores informaron acerca de su percepción sobre la misma, la razón que llevo a los organizadores a utilizar dicha fuente de información es la enorme variabilidad entre directores y entre países acerca de lo considerado "rural" o "urbano" (Hungi et al., 2010). Por otro lado, el cálculo del índice de estatus socio económico, ha resultado especialmente complejo en las evaluaciones SACMEQ. Tal y como apunta Dolata (2005), la mayor parte de la literatura existente acerca del índice de estatus socio económico, analiza el mismo en el contexto de países industrializados, medidas transfronterizas estandarizadas como el ESCS («*Economic, Social and Cultural Status Index*») de PISA, funcionan bien, pero fueron ideadas para países ricos y miembros de la OECD, estos no incluyen ningún país africano o asiático como India o China y sin embargo estos países cuentan con una alta población infantil. Muy pocos trabajos han analizado el «S*ocioeconomic Status*» SES en países en vías de desarrollo.

A pesar de la escasa literatura se pueden identificar dos principales elementos relacionados con los padres: nivel educativo e ingresos. El estatus económico de los padres es calculado utilizando variables sobre las que los niños pueden informar de forma apropiada (Dolata, 2005). Las posesiones en el hogar suelen variar sustancialmente entre países, sin embargo, contar con electricidad y agua corriente aparecen con regularidad, en África estas variables son muy importantes ya que no todo el mundo tiene acceso a ellas y son cuestiones de gran importancia en la vida del

individuo (Dolata, 2005). El índice ESCS que incluye PISA tiene en cuenta factores como el número de ordenadores o la conexión a internet que no tienen relevancia en países en vías de desarrollo (Dolata, 2005).

Las variables utilizadas para el cálculo del SES en las evaluaciones SACMEQ son: nivel educativo de los padres, número de libros en casa, recursos básicos (periódico, revistas, radio, televisión, VCR, radio casete, teléfono, refrigerador, coche, motocicleta, bicicleta, agua corriente, electricidad y una mesa para escribir), calidad estructural de la casa/calidad de los materiales utilizados para su construcción y cantidad de ganado en propiedad (pregunta abierta que no se pudo utilizar por las respuestas inusuales emitidas por los estudiantes) (Dolata, 2005).

El análisis del cambio y la posibilidad de comparación entre los países, son los aspectos sustanciales en estas evaluaciones. La metodología empleada en las mimas está condicionada por esta necesidad de comparación, contar con referentes de comparación en países en vías de desarrollo es el aspecto más destacado de dicha evaluación por ajustarse a las características particulares de estos países.

En definitiva, la descripción de las evaluaciones realizada en el apartado 1.4 nos ha permitido apreciar la importancia de la comparabilidad como característica común a todas ellas, independientemente del organismo o institución encargado de su diseño y puesta en marcha. Así, la importancia de la evaluación a lo largo del tiempo queda reconocida en todas ellas.

1.5 Problemas de comparabilidad en la Evaluación Educativa.

Tal y como hemos podido apreciar en la revisión de las principales evaluaciones internacionales, llevada a cabo en el capítulo precedente, existen numerosas condiciones que han de cumplirse si se desea asegurar la comparabilidad de los resultados. Como veíamos, por cuestiones de seguridad en la aplicación, es frecuente que sea preciso elaborar varias formas de una misma prueba dentro de una evaluación. De este modo, la seguridad en el proceso implica la posibilidad de comparación de distintas formas de un

mismo test, construidas con el propósito de evaluar el mismo constructo con el mismo nivel de dificultad.

En esta misma línea, el amplio dominio competencial que las diferentes pruebas pretenden abarcar, también supone un importante reto de cara a la comparabilidad. En las evaluaciones educativas, la intención es evaluar un amplio rango del dominio de interés, sin embargo, para conseguir una prueba completa, deberían incluirse gran cantidad de reactivos que permitiesen abarcar dicho dominio de forma exhaustiva. La aplicación de tales pruebas sería compleja, puesto que los estudiantes deberían contestar a pruebas excesivamente largas, viéndose afectadas las propiedades psicométricas de las mismas. En consecuencia, evaluaciones educativas como TIMSS, PIRLS o PISA (tal y como se recoge en el apartado anterior) implementan un sistema de "bloques" que permiten abarcar un amplio rango del dominio evaluado sin que sea necesario construir formas excesivamente largas, de este modo, bloques de anclaje rotativos permiten la comparación entre los cuadernillos de un mismo curso o nivel.

Unido a ello, cambios curriculares o cambios en el dominio evaluado dificultarían la posibilidad de comparación entre distintas ediciones de una misma prueba, exigiendo la puesta en marcha de diferentes técnicas de enlace de puntuaciones. Por otro lado, la construcción y mantenimiento de bancos de ítems exigirá poner en funcionamiento técnicas para la comparabilidad.

La unión de evaluación y comparación, ha sido puesta de manifiesto en los apartados precedentes, sin embargo, parte de las evaluaciones educativas tradicionales, han utilizado un solo momento para la medida del logro académico, con el objetivo de elaborar modelos que expliquen las diferencias en el rendimiento de los alumnos (diferencias interindividuales) sin preocuparse por las diferencias que se producen en un mismo alumno (diferencias intraindividuales), es decir, el cambio en el aprendizaje. Desde los primeros trabajos relacionados con la comparabilidad, se destaca precisamente esta necesidad (Flanagan, 1951).

Un análisis de los estudios internacionales expuestos en el apartado anterior, muestra cómo, en la mayoría de ellos, se utiliza como variable de respuesta el rendimiento (en términos de competencias) y en sus primeras ediciones hacen uso de una única medida del rendimiento de los alumnos en la materia evaluada, por ejemplo

TIMMS (1999 y 2003) en el ámbito de las Ciencias y las Matemáticas, y PISA (2000, 2003 y 2006) en las áreas de Matemáticas y resolución de problemas, Lectura y Ciencias. Esta primera medición, es el inicio del proceso, ya que la aplicación de las pruebas en dichas materias tiene un carácter cíclico, con el fin de conseguir la comparabilidad. Una única toma de datos no nos informa del progreso en el aprendizaje, es necesario, por tanto, partir de medidas repetidas del rendimiento que nos permitan un análisis del cambio. La utilización de una única toma de datos, supone el no reconocimiento de una de las principales características de la educación, ser una realidad en continuo cambio y evolución. Los estudios en los que se lleva a cabo una única medición en un momento puntual del tiempo, sin tener en cuenta la situación previa de los alumnos (estudios transversales), permiten conocer cuál es el estatus de un alumno en un momento determinado del tiempo, no aportando información de otros aspectos que revisten mayor importancia en el ámbito educativo, por ejemplo, no podemos conocer cuál ha sido la evolución de un estudiante dentro de un curso académico puesto que no contamos con otras mediciones de su desempeño.

Por tanto, para la evaluación del cambio en el aprendizaje, se precisa la realización de estudios longitudinales. De acuerdo con la definición que propone Goldstein, los estudios longitudinales son aquellos que están compuestos por más de dos mediciones a lo largo de un seguimiento, todo estudio de cohortes tiene como mínimo dos mediciones, una al principio y otra al final del seguimiento (Goldstein, 1997). En definitiva, los estudios longitudinales implican mediciones repetidas de los sujetos a lo largo del tiempo. Siempre que sea posible, las evaluaciones educativas deberán seguir este diseño, pues permite obtener información acerca de la evolución o cambio producido, aspecto importantísimo en el ámbito educativo. Igualmente, resulta recomendable utilizar éste tipo de diseño en el caso de que el objetivo del investigador sea el estudio de tendencias, cambios o desarrollos a través del tiempo, o bien, en los casos en que se busque la secuencia temporal de fenómenos.

No obstante, este diseño no está exento de dificultad, también existen ciertas limitaciones inherentes a los estudios longitudinales, tales como la mortalidad experimental o el elevado coste de realización (entre otras). Un aspecto muy concreto de la problemática de los diseños longitudinales es precisamente el referido al enlace de las puntuaciones de los sujetos en distintas pruebas o en pruebas aplicadas en distintos

momentos (equiparación/escalamiento), siendo este el núcleo del trabajo presentado. Si el complejo proceso para la obtención de resultados comparables no es llevado a cabo correctamente, corremos el riesgo de realizar medidas del rendimiento inadecuadas, en las que las puntuaciones no sean realmente comparables, por ello, el proceso técnico de enlace de puntuaciones es una cuestión de interés en el ámbito de la evaluación educativa en general y de forma particular en estudios longitudinales, en análisis de medidas repetidas, en estudios de tendencias, etc.

Por otro lado, no debemos olvidar la importancia que en las evaluaciones a gran escala tiene la posibilidad de comparación con referentes tales como escuelas, ediciones, distritos, estados, países, etc. Poder estudiar tales diferencias permite detectar y analizar las mejores prácticas educativas, de este modo, los sistemas educativos pueden identificar líneas de cambio y mejora.

En definitiva, tal y como apunta De la Orden (2000, p. 382) "La evaluación constituye esencialmente un juicio de valor sobre una realidad y, como todo juicio, se apoya en una comparación. Comparación y juicio son, pues, los componentes esenciales de la evaluación, su núcleo conceptual". En el diseño e implementación de evaluaciones educativas, es esencial tener en cuenta la posibilidad de comparación, pudiendo encontrar los siguientes problemas:

- Necesidad de comparación entre las distintas formas de una misma prueba.
- Adaptación a necesidades y cambios curriculares.
- Medir un amplio rango de dominio de la competencia evaluada.
- Elaboración y mantenimiento de bancos de ítems.
- Medida del cambio.
- Comparación con criterios de interés: ediciones, escuelas, distritos, estados, países, etc.

En el siguiente apartado se tratarán en profundidad los problemas relativos a la comparabilidad de mediciones, analizando en detalle las distintas técnicas de enlace y su utilización en el ámbito educativo.

CAPÍTULO 2: COMPARABILIDAD DE MEDICIONES

Tal y como se ha apuntado en el capítulo 1, en la actualidad existe una preocupación creciente por la evaluación educativa. Destacando la idea expuesta por Scriven en 1967, consideramos que la utilidad de la información que se obtiene de las evaluaciones depende de la posibilidad de comparar la misma con criterios de interés. En consecuencia, la comparabilidad de mediciones, es un área cuya importancia para investigadores, gestores, docentes y, en definitiva, para todos aquellos con implicaciones en el ámbito educativo, y más concretamente en la evaluación, va en aumento.

En el presente capítulo se expone un primer acercamiento al concepto y orígenes de la comparabilidad, en segundo lugar se incluye un apartado en el que se tratan las diferencias entre los distintos tipos de enlace de puntuaciones: predicción, escalamiento y equiparación. En tercer lugar se profundiza en los aspectos relativos a la equiparación horizontal y el escalamiento vertical, a continuación se dedica un apartado a los requisitos de la equiparación y al cumplimiento de alguno de estos supuestos por otras técnicas de enlace, finalmente, se presenta una secuencia orientativa de pasos a seguir en los procesos de equiparación y escalamiento, destacando aquellas decisiones que se han de tomar durante el proceso y que pueden incidir en los resultados finales.

2.1 Concepto y orígenes.

Desde la década de los 80 la evaluación de sistemas educativos ha experimentado un auge, mejorándose tanto su frecuencia como la calidad de los procesos evaluativos. No obstante, dicha evaluación ha de estar basada en una serie de asunciones que aseguren que el proceso ha sido el adecuado. Una de esas asunciones, exige la introducción de un sistema que permita la comparación de puntuaciones procedentes de diversos instrumentos o de instrumentos aplicados en distintos momentos.

A la hora de desarrollar una prueba estandarizada, es necesario tener en cuenta multitud de elementos, los problemas técnicos a los que se ha de responder suelen presentar altos grados de complejidad. Entre estos problemas, se encuentran los relativos a la comparabilidad de puntuaciones, resolverlos de forma adecuada resulta crucial puesto que afectan directamente a la interpretación y al uso de las puntuaciones (Gempp, 2010). Analizar la evolución de realidades como la educativa, implica la posibilidad de comparar los resultados procedentes de ediciones distintas de la misma evaluación. De este modo, evaluaciones a gran escala como las señaladas en el apartado 1.4 perderían un valor sustancial si no se asegurase la posibilidad de comparar los resultados procedentes de las mismas en distintas ediciones, no dando respuesta a la demanda de conocer las tendencias o evolución de una realidad en continuo cambio. Tal y como señalan Dorans, Moses y Eignor (2011), la equiparación es esencial para cualquier programa de evaluación que continuamente produce nuevas ediciones y cuya expectativa es mantener el mismo significado a lo largo del tiempo.

Estas comparaciones entre los resultados de pruebas o evaluaciones distintas, suele venir garantizada por el proceso técnico conocido comúnmente como «*equiparación*» [del inglés *equating*][1], no obstante, debemos tener en cuenta que no siempre es posible realizar dicha equiparación y que en numerosas situaciones han de utilizarse otros procedimientos que, no pudiendo ser considerados equiparación en sentido estricto, por las razones que se expondrán más adelante, si permitirán realizar

[1] La traducción al castellano del término *equating* como equiparación está comúnmente aceptada por los especialistas del área, no obstante, debemos destacar que es posible encontrar, especialmente en el contexto latinoamericano, otras expresiones como “igualación” o “equivalencia”.

ciertas comparaciones entre puntuaciones de distintos instrumentos o realizadas en distintos momentos.

Por tanto, la palabra anglosajona «*equating*», ha sido designada para hacer referencia a las técnicas necesarias para contar con instrumentos de evaluación cuyos resultados sean estrictamente intercambiales, siendo en este punto donde encontraremos la principal distinción entre esta y otras técnicas que hacen posible la comparabilidad, solo hablaremos de equiparacion cuando los resultados de las pruebas equiparadas sean estrictamente comparables e intercambiables. El uso generalizado del término en los últimos años ha contribuido a evidenciar la necesidad de introducir ciertas aclaraciones en su definición, con el propóstio de evitar confusiones derivadas del mal uso del mismo. El objetivo del proceso de equiparación es el establecimiento de una equivalencia entre las puntuaciones de los tests a equiparar, de tal modo que, las puntuaciones de una prueba serán expresadas en términos de la otra, siendo indiferente a cuál de ellas haya dado respuesta el sujeto, siempre y cuando la equiparación haya sido realizada adecuadamente, de este modo, la equiparación tendría la función de asegurar la adecuada medida de la habilidad subyacente del sujeto evaluado independientemente de la forma a la que ha contestado dicho sujeto (von Davier, 2011). En definitiva, tras la realización del proceso de equiparación las puntuaciones procedentes de los tests equiparados serán totalmente intercambiables, teniendo en cuenta las restricciones al respecto que apuntaremos más adelante.

Imaginemos que se aplica una prueba de rendimiento en Comprensión Lectora (CL) a una muestra representativa de alumnos de 1º de ESO, obteniéndose una media de 250 puntos. Un año más tarde, se diseña una nueva prueba de CL y se aplica a una nueva muestra representativa de alumnos de 1º de la ESO obteniéndose un resultado de 260 puntos. La diferencia en dichos resultados puede deberse a dos motivos: en primer lugar, a que realmente los alumnos del segundo año tengan un rendimiento mayor en comprensión lectora o, en segundo lugar, a que la prueba realizada para el segundo año sea ligeramente más fácil. En un primer momento, se podría considerar que una posible solución pasaría por volver a aplicar la prueba realizada el primer año en la siguiente aplicación. En este caso, el problema relativo a la diferencia en la dificultad entre ambos instrumentos quedaría resuelto y contaríamos con dos pruebas de idéntica dificultad, esta solución, sin embargo, incorporaría un nuevo problema, relativo al conocimiento,

por parte de los alumnos de la nueva muestra o de sus profesores, de los ítems que forman parte del instrumento. Una situación similar nos encontraríamos en el caso de dos estudiantes que compiten por la obtención de una beca de estudios, el principal criterio para otorgar dicha beca a uno u otro es la puntuación obtenida en una prueba de rendimiento académico, supongamos que dichos estudiantes contestan la prueba en días diferentes, el estudiante 1 obtiene una puntuación superior a la del estudiante 2. En primer lugar podríamos explicar dicha diferencia en puntuación como consecuencia de que el estudiante 1 presenta mayor nivel que el estudiante 2, sin embargo ¿no podrían deberse tales diferencias a diferencias en la dificultad de las pruebas? ¿qué habría sucedido si los estudiantes hubiesen contestado a la otra forma? ¿qué estudiante es justo merecedor de la beca? En esta situación, la diferencia entre los dos estudiantes podría explicarse por una diferencia en la dificultad de la prueba en lugar de a una diferencia en la competencia de los estudiantes, si asignamos la beca de estudios siguiendo este criterio se estaría cometiendo un importante error.

La realidad, muestra cómo en diversas situaciones prácticas se han encontrado problemas relativos a la equiparación de puntuaciones con fuertes repercusiones. Un ejemplo de ello lo encontramos en Estados Unidos, donde al equiparar las versiones antigua y nueva del denominado «*Armed Services Vocational Aptitude Battery* (ASVAB)», se encontraron con que gran número de participantes fueron seleccionados, tras nombrar a una comisión para el análisis de la situación, y tras cuatro años de estudios, la comisión concluyó que, por problemas relativos al proceso de comparabilidad, desde 1976 hasta 1980, habían sido seleccionados 350.000 individuos no cualificados para formar parte del servicio militar (Kolen, 2001).

Para evitar situaciones como las descritas anteriormente, que producirán un error en las interpretaciones de los datos, los sistemas de evaluación que aplican diferentes pruebas tienen el difícil objetivo de construir instrumentos que sean, diferentes pero cien por cien comparables. Las diferentes formas de un test suelen construirse intentando conservar al máximo la igualdad tanto en el contenido de las mismas como en las propiedades psicométricas, los constructores de pruebas se esfuerzan notablemente en conseguirlo (Kolen & Brennan, 2014). A pesar de dichos esfuerzos, la diferencia fundamental entre formas suele residir en factores como la dificultad, la fiabilidad o el contenido. El proceso de equiparación es utilizado en situaciones en las

que existen formas alternativas de un mismo test y las puntuaciones obtenidas en las diferentes formas han de ser comparadas entre sí, la equiparación ajusta diferencias en dificultad, no diferencias en contenido (Kolen & Brennan, 2014).

Con el fin de conseguir pruebas estrictamente comparables, es condición sine qua non que se cumplan determinados requisitos técnicos, a los que han hecho referencia multitud de autores en distintos trabajos (ver Apartado 2.4). El problema relativo al cumplimiento de dichos supuestos aumenta cuando se trata de comparar niveles diferentes de las variables medidas, es decir, cuando se trata, por ejemplo, de comparar las puntuaciones de los sujetos entre distintos niveles educativos, en estos casos no se utilizará el término equiparación, siendo más adecuado hablar de «*scaling*» o «escalamiento», distinción en la que entraremos en detalle más adelante.

A pesar de su importancia y sus fuertes repercusiones en la calidad de los procesos de medición educativa y psicológica, el escalamiento y la equiparación de puntuaciones, no han sido temas tratados en profundidad desde la psicometría clásica. En el año 1951 Flanagan expone la necesidad de contar con procedimientos que permitan la comparabilidad entre puntuaciones de diferentes tests en el capítulo titulado «*Units, scores, and norms*» como parte de la primera edición del libro «*Educational Measurement*» editado por Lindquist. En él, destaca la especial importancia de la comparabilidad en el ámbito de la evaluación educativa, apuntando procedimientos básicos para su consecución y diferenciando con claridad entre la comparabilidad de puntuaciones de pruebas construidas con el mismo propósito e igual fiabilidad (equiparación) y aquellas que no poseen estos requisitos (Flanagan, 1951), sentando de algún modo las bases de la terminología utilizada con posterioridad. En la segunda edición de esta misma obra editada por Thorndike (1971), se publica, de la mano de Angoff, uno de los trabajos pioneros con mayor impacto en el ámbito de la equiparación, dicho capítulo titulado «*Scales, Norms and Equivalent Scores*» será editado en el año 1984 en forma de libro por el «*Educational Testing Service*» (ETS) (ver Tabla 35). En 1980, Lord dedica un capítulo a la equiparación de puntuaciones dentro de su libro "*Aplications of item response theory to practical testing problems*", en el que detalla los requistos de la equiparación, siendo un referente hasta nuestros días.

De este modo, la mayoría de las revisiones (Martínez Arias, 1995; Navas, 1996; 2000; Muñiz, 1997; 2003; Kolen, 2004; Pacheco Villamil, 2007), citan los trabajos de Angoff (1971; 1984) y Lord (1980) como los primeros referentes en este área, ya que en ellos se introduce un análisis detallado con cierto impacto en la comunidad científica. Hasta entonces, la equiparación era un asunto que preocupaba exclusivamente a los especialistas en psicometría, a partir de los años 80 comienza a ser un aspecto de interés para diversos especialistas en el ámbito de la medición (Kolen & Brennan, 2004). De manera global, en dichos trabajos pioneros, la equiparación es considerada un proceso cuyo objetivo es la consecución de un sistema de conversión de las unidades de pruebas diferentes, consiguiendo que los resultados entre ambas sean comparables o equivalentes.

Tabla 35.

Principales publicaciones en los inicios del estudio de la comparabilidad de puntuaciones

AUTOR	AÑO	TÍTULO	TIPO DE CONTRIBUCIÓN
Angoff	1971	Scales, Norms and Equivalent Scores	Capítulo en el libro Educational Measurement
Lord	1977	Practical Applications of item characteristic curve theory	Artículo: Journal of Educational Measurement
Lord	1980	Practical applications of item response theory.	Libro
Kolen	1980	Comparison of traditional an item response theory methods for equating tests.	Artículo: Journal of Eduational Research
Holland y Rubin	1982	Test Equating	Libro
Skaggs y Lissitz	1982	Test Equating: Relevant Issues and Review of Recent Research	Comunicación presentada en el Congreso anual de *American Educational Research Association*
Angoff	1984	Scales, Norms and Equivalent Scores	Publicado como libro el capítulo del año 1971
Standards for Educational and Psychological Testing	1985	Standards for Educational and Psychological Testing	Libro
Applied Psychological Measurement	1987	Test equating	Revista: Monográfico

Fuente: elaboración propia.

En el mes de abril de 1980 tiene lugar la conferencia titulada «*Test Equating*» organizada por el «*Educational Testing Service»* (ETS), esta conferencia da lugar a la publicación del primer libro monográfico sobre el tema en el año 1982, marcando otro hito histórico a considerar en el estudio de la comparabilidad de puntuaciones. El libro,

titulado «*Test Equating*», es editado igualmente por el ETS a partir de las aportaciones y discusiones del congreso de 1980 (Holland & Rubin, 1982). La obra, centrada en los aspectos estadísticos de la equiparación, supone una completa revisión de los trabajos realizados hasta la fecha, contando con la colaboración de autores como Angoff, Braun, Flanagan, Lord y Thorndike, entre otros. Asimismo, el libro presenta, un apartado de discusión sobre los trabajos presentados (Holland & Rubin, 1982). Durante este periodo se evidencia el enorme interés del ETS por el estudio de los problemas relativos a la equiparación.

Como se puede observar, la década de los 80 se perfila como un momento clave en el estudio de la equiparación. Unido a los trabajos anteriores, los «*Standards for Educational and Psychological Testing*» (1985) dedican varios «*standards*», que irán aumentando en sucesivas ediciones, de este modo, vemos como en el año 1990 se presentan 4 «*standards»*, en 1999 pasan a ser 8 y en su última edición, publicada en 2014 son 9 (American Educational Research Association, 1990; 1999; 2014). Del mismo modo, como se aprecia en la Tabla 35, en 1987 la revista «*Applied Psychological Measurement»* edita un número especial que incluye importantes trabajos que abordan aspectos muy diversos de la equiparación de puntuaciones (Angoff, 1987; Brennan y Kolen, 1987a; Brennan y Kolen, 1987b; Cook, & Petersen, 1987; Fairbank, 1987), en los que se analizan cuestiones teórico prácticas desde la perspectiva clásica y desde la Teoría de la Respuesta al Ítem. Un año más tarde, es publicado por «*Educational Measurement»*, el trabajo de Kolen titulado «*An NCME Instructional Module on Traditional Equating Methodology*» cuyo objetivo es promover la comprensión conceptual de los métodos clásicos de equiparación (Kolen, 1988).

Estos trabajos son prueba del creciente interés suscitado por la comparabilidad de puntuaciones a partir de la década de los 80, en este punto, cabe preguntarse ¿cuál fue el desencadenante de dicho interés por parte de la comunidad científica? ¿cuáles fueron los hechos que contribuyeron a este auge por el estudio de la comparabilidad? La mayoría de los autores coinciden en apuntar, principalmente, al uso masivo de los tests en Estados Unidos, con fines tales como el acceso y elección universitarias, el servicio militar, puestos de trabajo, promociones, certificaciones, etc. Situación que obligó a contar con formas alternativas de un mismo test, introduciendo de este modo el

problema de comparar las puntuaciones, con el riesgo de cometer injusticias comparativas con importantes repercusiones en la vida de los individuos (Muñiz, 2003).

Profundizando en esta idea, tal y como recoge Martínez Rizo (2004), el surgimiento en Estados Unidos de la necesidad de regular los procedimientos de admisión a las universidades, llevó a la creación del denominado «*College Board*» en noviembre del año 1900. Hasta 1925 las pruebas utilizadas consistían en preguntas de tipo ensayo; en 1926 se aplicó por primera vez la prueba denominada «*Scholastic Aptitude Test*» *(SAT)* (Martínez Rizo, 2004). Al inicio de la década de 1980, los Estadounidenses, se enfrentan a los primeros problemas asociados con la comparabilidad de dos de sus pruebas nacionales de mayor importancia el «*Scholastic Aptitude Test*» *(*SAT) y el «*American College Testting*» (ATC). En relación al SAT, desde mediados de los años sesenta, se observó un descenso en los resultados obtenidos por los estudiantes, dicho descenso se repetía año tras año, consiguiéndose unos resultados promedios ligeramente inferiores cada año respecto al anterior, esta tendencia continuó hasta mediados de la década siguiente (Martínez Rizo, 2004). Del mismo modo, la preocupación por el descenso en los resultados del ATC, era un tema recurrente (Kolen, 2001). Esta situación generó cierta preocupación en la comunidad educativa en su conjunto ¿existía un empeoramiento real del rendimiento educativo? ¿eran dichos resultados realmente comparables? Estas dudas sobre la comparabilidad de resultados hicieron que se formara una comisión para el estudio sobre el descenso de los resultados del SAT («*Advisory Panel on the Scholastic Aptitude Test Score Decline*») y al mismo tiempo, el ETS, formó comisiones para estudiar si la equiparación podría ser una de las causas subyacentes al descenso en las puntuaciones del ACT (Kolen, 2001). En ambos casos, se descartó la posibilidad de que existieran problemas de equiparación entre dichas puntuaciones, concluyendo con el convencimiento de que las pruebas sí eran comparables. El descenso en los resultados debía atribuirse, a un cambio real en el rendimiento del alumnado.

Por tanto, y de acuerdo con lo expuesto por Muñiz (2003), podemos destacar dos razones principales del aumento del interés en este campo de investigación, en primer lugar, la continua crítica y discusión social del sistema generalizado de tests, obligando a los constructores a justificar y explicar públicamente sus métodos de equiparación. En segundo lugar, los nuevos modelos basados en la teoría que domina la psicometría

actual, Teoría de la Respuesta al Ítem (TRI) que permitían un tratamiento más adecuado superando algunas de las limitaciones de la Teoría Clásica (Muñiz, 2003). A estos aspectos se unirían los apuntados por Lawrence en el prólogo de Dorans, Pommerich y Holland (2007) en el que indica que, el aumento de la atención en los procesos utilizados para la comparabilidad de mediciones se debe, entre otros motivos, al incremento en el número y variedad de programas de evaluación y al reconocimiento por parte de los expertos de la necesidad de establecer un escalamiento entre ellos, así como el auje de los movimientos de «*accountability»* relacionados con hacer la evaluación mucho más visible. La equiparación de puntuaciones, es esesencial para cualquier programa de evaluación que continuamente produce nuevas ediciones de sus pruebas, pretendiendo que el significado de las puntuaciones sea el mismo a lo largo del tiempo, dichas pruebas pueden ser construidas utilizando la misma matriz de especificaciones, sin embargo, las propiedades psicométricas de las mismas, pueden presentar variaciones en distintos aspectos (Dorans, Moses, & Eignor, 2010).

El enlace de puntuaciones es, por tanto, un tema de actualidad, en el que es necesario arrojar claridad si queremos aumentar la calidad de las evaluaciones del sistema educativo, realizando interpretaciones de datos correctas a partir de la comparación adecuada de resultados de distintas evaluciones o realizados con distintos instrumentos.

2.2 Enlace de puntuaciones: prediccion, escalamiento y equiparacion.

Existe cierta confusión práctica y conceptual en el ámbito de la comparabilidad o enlace de puntuaciones, diversos especialistas en el área utilizan aproximaciones inadecuadas. Algunos autores, pueden llegar a considerar que, haciendo uso de la técnica estadística adecuada, cualquier par de mediciones puede equipararse a todo evento, olvidando que la equiparación es un objetivo que en ocasiones no puede lograrse empíricamente (Gempp, 2010). Tal y como apunta Kolen (1988), desafortunadamente no siempre es posible llevar a cabo una adecuada equiparación. Del mismo modo, la consideración errónea de las técnicas de equiparación como medios para igualar dos pruebas diferentes, es una creencia bastante extendida, sin tener en

cuenta el verdadero propósito de la equiparación que, según lo señalado por Kolen y Brennan (2014), está en el ajuste del nivel de dificultad entre distintas pruebas.

Existen procesos similares a la equiparación que no suponen una equiparación en sentido estricto, de este modo, se suele utilizar el término «*linking*» "enlace" (Mislevy, 1992; Linn, 1993; Feuer, Holland, Green, Bertenthal & Hemphill, 1999; Dorans & Holland, 2000; Kolen, 2004; Pommerich & Dorans, 2004; Holland & Dorans, 2006; Holland; 2007; von Davier, 2011; Dorans, Moses, & Eignor, 2011; American Educational Research Association, 2014) como término genérico para hacer referencia a la variedad de enfoques que permiten comparar los resultados procedentes de diferentes pruebas. Los «*Standards for Educational and Psychologial Testing»* (1990; 1999) ya reconocen esta especificidad al hablar de «*scaling to achieve comparability»* (escalamiento para la comparativa) como término para englobar aproximaciones a la comparabilidad que no suponen una equiparación en sentido estricto. En la edición de 2014, los «*standards»* utilizan el término «*Score Linking*» como término general que engloba la equiparación y otros métodos de transformación de puntuaciones cuyo objetivo es la comparabilidad, incluyendo los métodos de escalamiento vertical «*vertical scaling methods*» (American Educational Research Association, 2014).

Por tanto, se ha de tener en cuenta que, las condiciones para establecer una equiparación entre puntuaciones, son muy exigentes, existiendo otros procedimientos que permiten establecer un enlace entre resultados (y no una equiparación de los mismos). En este apartado, intentaremos arrojar luz sobre el tema introduciendo un pequeño análisis de las diferentes técnicas de enlace de puntuaciones, entre las que se encuentra la equiparación.

El problema de la comparabilidad entre mediciones se evidencia en aquellas situaciones en las que se precisa la construcción de algún tipo de regla de correspondencia que permita expresar el resultado de una evaluación o prueba en la métrica de otra, introduciendo por tanto una metodología de enlace de puntuaciones (Gempp, 2010). Se trata de realizar una transformación estadística que permita encontrar la equivalencia entre las puntuaciones de distintas pruebas, la idea general, por tanto, es la transformación entre las puntuaciones de dos tests (Holland & Dorans, 2006).

En diferentes trabajos (Flnagan, 1951; Angoff , 1971; Linn, 1993; Mislevy, 1992; Feuer et al., 1999; Dorans, 2000; 2004; Dorans & Holland, 2006; Holland, 2007), se presenta un completo análisis de los distintos tipos de enlace de puntuaciones. La confusión conceptual en este sentido es notoria, haciéndose cada vez más necesario establecer una clasificación clara de las diferentes terminologías empleadas. Para van der Linden (2013), el aumento del conocimiento reside en la capacidad de explicar más con menor número de conceptos, aportando una visión crítica de los trabajos de clasificación de otros autores, considerando que se utilizan términos innecesarios imposibles de explicar a personas que se encuentren fuera del ámbito de la medición educativa y que introducen mayor confusión. Sin embargo, tanto el citado trabajo de van der Linden (2013) como el de Holland (2013), a pesar de sus posiciones enfrentadas, coinciden en señalar la ausencia de una teoría de base satisfactoria en el ámbito de la equiparación.

Existen gran variedad de términos utilizados en relación a la comparabilidad o enlace de puntuaciones como «*anchoring», «benchmarking», «calibration», «equating», «prediction», «projection», «scaling», «statistical moderation», «social moderation», «verification»,* y *«auditing»*. Algunos de estos conceptos tienen asociados adecuados significados y requisitos técnicos, sin embargos otros no los tienen, un mismo término puede ser utilizado con diferentes significados en distintos contextos (Linn, 1993). Van der Linden (2013) apunta que, a pesar de las diferencias terminológicas, cada uno de los diferentes tipos de enlace de puntuaciones utilizan casi exactamente los mismos diseños de equiparación utilizados en el procedimiento equipercentil, considerando que en realidad dicha terminología no responde a una verdadera clasificación de métodos diferenciados. En nuestra opinión, tal y como apuntan Dorans, Moses y Eignor (2010), las mayores diferencias entre los distintos tipos de enlace de puntuaciones son de tipo interpretativo, no procedimental.

A la hora de trabajar en el ámbito de la comparabilidad, cobra especial importancia tener un esquema general de las metodologías de enlace de puntuaciones y saber reconocer en qué condiciones son válidas. Para conocer cuáles son estos procedimientos nos basaremos en la taxonomía descrita exhaustivamente por Holland y Dorans, (2006) y por Holland (2007) incorporando las aportaciones de otros autores. Las metodologías de enlace que vamos a describir basándonos en estos trabajos, se

caracterizan por tres aspectos fundamentales, la validez, la precisión y la reversibilidad con que se pueden transformar las puntuaciones. Basándose en estas características los autores considera tres grandes grupos de enlaces: predicción («*prediction*» o «*expected*»), escalamiento («*scaling*») y equiparación de puntuaciones («*equating*»).

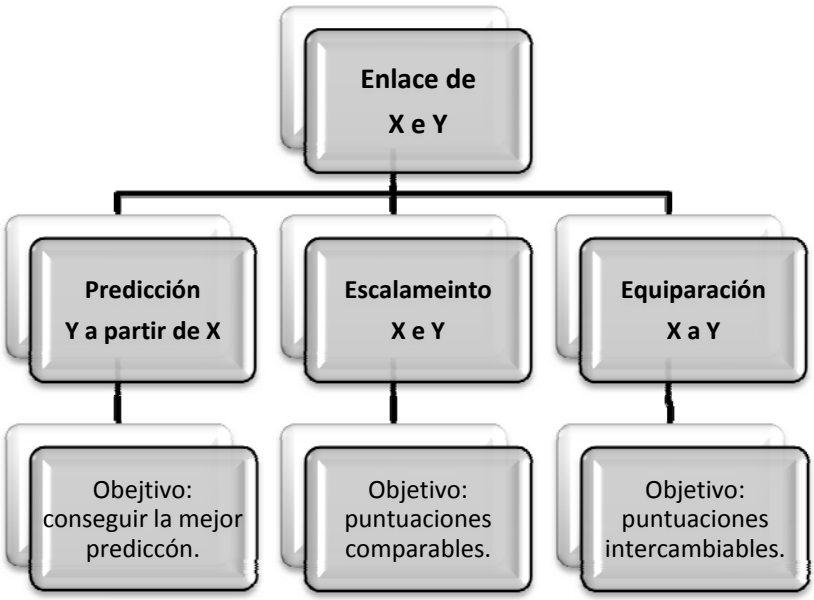


Figura 10. Categorías generales básicas de los métodos de enlace y sus objetivos
Fuente: Adaptado de (Holland & Dorans, 2006, p. 188) (Holland, 2007, p. 6).

Estas tres categorías básicas (Figura 10), al mismo tiempo, se dividen en sub categorías de acuerdo a factores como las especificaciones de las pruebas, el diseño de recogida de datos, las características de la población, etc. A pesar de tratarse de la taxonomía más aceptada por la comunidad investigadora, el último trabajo de van der Linden (2013) propone una simplificación, considerando que solo sería necesario hablar de equiparación y predicción. En cualquier caso, es importante distinguir de manera unívoca entre los diferentes procedimientos reconocidos ya que, en muchas ocasiones, pueden parecer similares sin serlo. El investigador ha de reconocer las diferencias entre estas categorías y seleccionar en qué circunstancias es más adecuado utilizar una u otra.

A) Predicción.

De acuerdo con Holland, Dorans y Petersen (2007) esta es la forma más antigua de enlace entre puntuaciones, y su confusión con la equiparación se ha dado desde los principios de la psicometría.

Su objetivo, es predecir la puntuación esperada en una evaluación a partir de otra información relevante. Dicha información relevante puede proceder de una o varias pruebas o, por ejemplo, de variables socio-demográficas relacionadas con el constructo a predecir (Holland & Dorans, 2006). Un ejemplo clásico sería el intento de predecir el resultado en una prueba de selección universitaria (en el caso de España la Prueba de Acceso a la Universidad (PAU)) basándonos en los resultados de evaluaciones previas del aprendizaje durante el bachillerato. Los enlaces por predicción, generalmente se usan en los estudios de validez predictiva y en la construcción de modelos de crecimiento escolar. De acuerdo con Dorans (2000), de los tres tipos de enlaces que presenta esta taxonomía este es el menos restrictivo y exigente. En resumidas cuentas, la predicción trata de minimizar la imprecisión en la predicción de una puntuación a partir de otra (u otras) puntuaciones. Es preciso resaltar una de sus características más destacadas: la asimetría entre la puntuación predicha y las variables predictoras, es decir, la predicción opera en una única dirección (Holland & Dorans, 2006). La utilización de la regresión lineal es la causante de ello, donde la recta de regresión para predecir Y en función de X no es la inversa de la recta para predecir X a en función de Y (Galton, 1988). Siendo precisamente la simetría entre predicciones uno de los principales requisitos en el caso de la equiparación.

Pueden considerarse varios tipos de predicción, la predicción de puntuaciones observadas individuales, la proyección de distribuciones de puntuaciones observadas o la predicción de puntuaciones verdaderas. En el primer caso, el cálculo de la función de enlace nos permite predecir el rendimiento de otros sujetos evaluados a posteriori cuando provienen de una población equivalente de aquella con la que fue realizado el estudio. El objetivo es predecir la puntuación de un examinado en un test, a partir de alguna información diferente de dicho examinando, tal información podría ser la puntuación en otro test, variables de tipo sociodemográfico, personal, etc. (Holland, 2007). La eficacia de la predicción está sujeta al cumplimiento de los supuestos del modelo de regresión.

La proyección de distribuciones de puntuaciones observadas, consiste en la estimación de la distribución de puntuaciones de una prueba a partir de los resultados de otra. Para ello se parte de una población de evaluados que contesta a ambas pruebas, a partir de aquí, se intenta predecir la distribución condicional de la prueba B para cada

puntuación observada en la prueba A, mediante ciertas consideraciones de los métodos de regresión lineal (Holland, 2007). Por tanto no se predicen valores individuales sino una distribución de puntuaciones. Posteriormente, dichas puntuaciones pueden ser agregadas con el fin de obtener un resultado total, como sucede en los valores plausibles de las pruebas PISA (Gempp, 2010). Por último, es necesario destacar la predicción de puntuaciones verdaderas, en este sentido, el trabajo pionero es el desarrollado por Kelley (1927) en el que se propone una fórmula para estimar la puntuación verdadera en Y a partir de la puntuación observada en Y bajo la perspectiva de la TCT. Por otro lado Holland y Hoskens (2003) analizan la predicción de la puntuación verdadera en un test a partir de la puntuación observada en ese mismo test así como, la predicción a partir de la puntuación observada en otro test que no ha sido construido con el fin de ser escalado con el primero, distinguiendo por tanto entre lo que denominan "predicción directa de puntuación verdadera" y "predicción indirecta de puntuación verdadera".

La predicción no es un método óptimo para realizar comparaciones entre puntuaciones, este hecho ha sido evidenciado por numerosos autores. No obstante tiene importancia destacar su presencia como uno de los métodos primitivos de enlace entre puntuaciones, utilizado con mucha frecuencia en la investigación educativa. En numerosos trabajos se analiza la distinción entre predicción y equiparación (Flanagan, 1951; Angoff , 1971; Mislevy, 1992; Linn, 1993; Holland & Dorans, 2006).

B) Escalamiento.

Originalmente, los métodos de escalamiento fueron denominados "métodos para la creación de puntuaciones comparables" (Holland, 2007). Los actualmente denominados métodos de escalamiento «*scale aligning*» tienen el objetivo de obtener puntuaciones comparables a partir de la transformación de los resultados de tests diferentes a una escala común. Dichos métodos ocupan la segunda posición en antigüedad en relación a los diferentes métodos de enlace de puntuaciones (Holland, 2007), siendo considerados igual de antiguos que los propios métodos de equiparación (Holland & Dorans, 2006). En el caso de la predicción el objetivo era conseguir predicciones lo mas exactas posibles, sin embargo, en este caso el objetivo es transformar las puntuaciones de dos tests diferentes a una escala común (Gempp, 2010). El proceso para obtener puntuaciones comparables se realiza de manera indirecta, es

decir, utilizando una conexión entre los dos tests externa a ambos, como puede ser un tercer test o un test de anclaje, estas conexiones externas entre los tests a comparar pueden haber sido creadas con otros fines, y no específicamente con el objetivo del escalamiento entre pruebas (Holland, 2007).

El estudio de estos métodos se inició a principios del siglo XX, pero su desarrollo ha sido desordenado y no ha seguido una secuencia lógica demasiado clara (Gempp, 2010). Los métodos de escalamiento y equiparación son confundidos con frecuencia debido a que los procedimientos estadísticos utilizados en escalamiento también pueden ser utilizados en equiparación (Dorans, Moses, & Eignor, 2011). Diferentes trabajos (Dorans, 2000; 2004; Holland & Dorans, 2006; Holland, 2007), presentan un acercamiento a una posible taxonomía o clasificación de este tipo de métodos, donde la confusión conceptual sigue presente al utilizarse en diversos trabajos distintos términos. Es en los trabajos de Holland y Dorans (2006) y Holland (2007) en los que puede encontrarse la clasificación más clara, atendiendo a tres criterios principales: propósito, diseño y características técnicas de las pruebas. Tomando como punto de partida tales criterios, realizan un esquema, en el que se incluyen 6 categorías dentro de este gran grupo como: escalamiento de baterías «*battery scaling*» propuesto por Kolen (2004), escalamiento de anclaje «*Anchor scaling*» que incluye el escalamiento a una población hipotética y el escalamiento vertical «*vertical scaling*» propuesto por Kolen y Brennan (2004), calibración «*calibration*» y concordancia «*concordance*» propuesto por Dorans (2004). En la Figura 11 podemos observar una adaptación del esquema propuesto por los autores en el que se incorporan todos los tipos de escalamiento mencionados anteriormente.

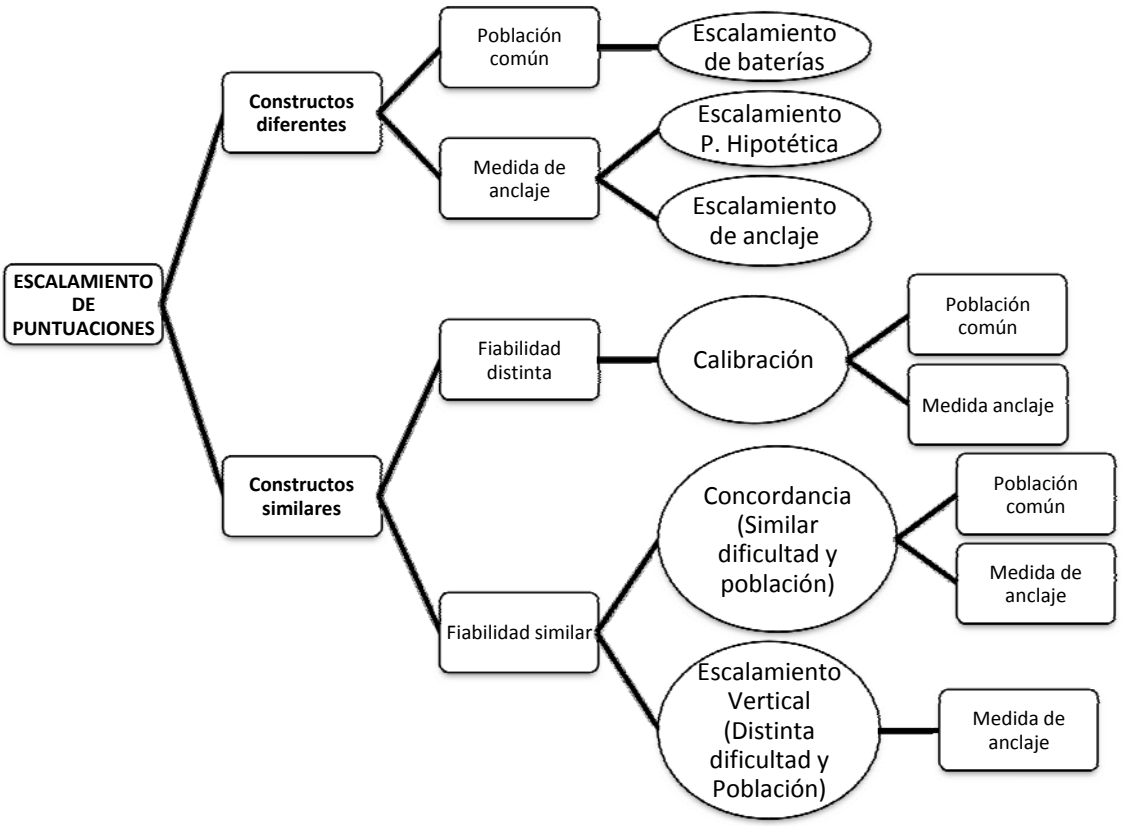


Figura 11. Tipos y subtipos de escalamiento.
Fuente: Holland & Dorans (2006, p. 190); Holland (2007, p. 13).

Como podemos observar en el esquema representado en la Figura 11, en primer lugar se distinguen dos grandes grupos de procedimientos, en función de si las pruebas que se desean comparar hacen referencia a constructos diferentes o similares. A continuación pasamos a describir brevemente cada uno de ellos.

- **Escalamiento de baterías: diferentes constructos en una misma población.**

Cuando dos o más tests, que miden constructos diferentes, se administran a una población común, las escalas de cada test pueden transformarse con el objetivo de que posean una distribución comparable, este procedimiento fue denominado «*test battery scaling*» por Kolen (2004).

El escalamiento de baterías permite conseguir escalas con una distribución comparable (no puntuaciones equivalentes) en pruebas que evalúen constructos distintos

en una población común de examinandos. El proceso de recogida de datos se suele realizar de dos formas a) una muestra de examinados toman las dos formas de manera completa, o b) muestras equivalentes de sujetos tomados de una población común toman uno de los test, en ambos casos ambos tests son contestados por grupos equivalentes en relación a la población de referencia (Holland, 2007).

Por último, es preciso destacar dos aspectos, en primer lugar, tal y como apunta Gempp (2010), el error de este tipo de escalamiento es conceptualmente mayor que en los métodos que alinean pruebas dirigidas a evaluar constructos similares, en segundo lugar, el error de alineamiento es inversamente proporcional a la magnitud de la correlación entre las pruebas utilizadas (Gempp, 2010).

- **Escalamiento mediante medidas de anclaje: diferentes constructos y diferente población.**

En segundo lugar, si queremos hacer comparables las mediciones de constructos diferentes que, además, han sido evaluados en poblaciones distintas, podemos recurrir al denominado "Escalamiento mediante medidas de anclaje″ («*anchor measure*»). Consiste en utilizar una tercera medición que si ha sido aplicada a ambas poblaciones y que correlaciona con los constructos evaluados (Holland, 2007). Por tanto, la medida de anclaje es un "puente" que nos permite situar las puntuaciones en una escala con distribución común. Algunos autores como Linn (1993) y Mislevy (1995b) utilizan el término "moderación estadística″ («*statistical moderation*») para hacer referencia a estos casos, considerando que la moderación estadística se da en aquellas situaciones en las que se desea comparar resultados procedentes de distintas fuentes. Un ejemplo de ello sería el uso de una prueba externa con el fin de ajustar las calificaciones asignadas por un profesor (Linn, 1993).

Existen dos aproximaciones estadísticas que implican un uso diferenciado de las medias de anclaje. La primera de ellas es el "escalamiento a una población hipotética" («*Scaling on a Hipothetical Population*» (SHP)) (ver Holland y Dorans,2006) y la segunda el "escalamiento hacia el anclaje″ («*Scaling to the Anchor*» (STA)) (ver Holland, 2007).

Hasta aquí, hemos hablado de hacer comparables pruebas que miden constructos diferentes en poblaciones comunes y no comunes. A continuación abordaremos las posibilidades que encontramos cuando el objetivo radica en comparar pruebas que miden los mismos o similares constructos.

- **Calibración: mismo constructo, diferente fiabilidad en la misma población de examinados.**

En el esquema de la Figura 11, vemos como la clasificación que proponen Holland y Dorans (2006) distinguen entre aquellas pruebas que, intentando medir constructos similares presentan distinta fiabilidad y las que presentan una fiabilidad similar. En el caso de que el objetivo sea comparar el mismo constructo, en la misma población, con pruebas de diferente fiabilidad o dificultad, se puede emplear la técnica de calibración.

A lo largo de los años, así como en la actualidad, el término calibración ha sido utilizado con diferentes significados en el ámbito psicométrico, dicha situación dificulta en cierta medida su uso. Angoff (1971) lo utiliza para hacer referencia al proceso de escalamiento entre las puntuaciones de tests que miden el mismo constructo pero con diferente dificultad o fiabilidad, es decir, lo utilizan para hacer referencia a lo que comúnmente denominamos "escalamiento vertical". Al igual que Angoff, Linn (1993) entiende que la calibración se utiliza en aquellas situaciones en las que dos tests han sido diseñados para medir el mismo constructo pero en diferentes niveles o con distinta fiabilidad, considerando que es más apropiado hablar de calibración que de "equiparación vertical" por no cumplirse los criterios para una equiparación. Por otro lado, gran número de especialistas utilizan el término "calibrar" para denominar el proceso de obtención de los parámetros de los ítems analizados bajo el enfoque psicométrico de la Teoría de la Respuesta al ítem (TRI) (Lord, 1980). En el caso de la taxonomía que nos ocupa (Dorans & Holland, 2006; Holland, 2007) la calibración es entendida como la técnica utilizada en aquellas situaciones en las que las pruebas miden el mismo constructo, con similares niveles de dificultad pero con diferencias en la fiabilidad (y en ocasiones en la longitud de la prueba), el término calibración, es utilizado en el mismo sentido por Kolen y Brennan (2014).

Normalmente las diferencias en fiabilidades en este tipo de pruebas se suele dar por diferencias en la longitud del instrumento. Poner las puntuaciones del test de menor longitud, en la escala del más fiable y de mayor longitud, no incrementará la fiabilidad de la forma breve, esta es una apreciación que no siempre es tenida en cuenta y por tanto puede crear alguna confusión (Holland, 2007).

- **Concordancia: similar constructo, dificultad, población y fiabilidad.**

Cuando contamos con constructos similares, evaluados con pruebas de igual fiabilidad y dificultad en la misma población, la técnica utilizada para enlazar los resultados es conocida como "concordancia". Si los tests que se pretenden vincular no han sido elaborados conforme a las mismas especificaciones técnicas, el enlace entre ambos tests debe considerarse "concordancia″ en lugar de equiparación (Pommerich, Hanson, Harris, & Sconing, 2004). En esta situación los tests objeto de análisis intentan medir el mismo constructo pero han sido construidos de acuerdo a diferentes especificaciones técnicas, en la mayoría de los casos suelen presentar similar dificultad, longitud y fiabilidad (Holland, 2007). El uso del proceso de escalamiento permite aportar un valor adicional a las puntuaciones de ambos tests, al hacer posible la expresión de los resultados de cada uno de ellos como si fueran puntuaciones del otro (Dorans, Pommerich, & Holland, 2007).

En el escalamiento por concordancia, el resultado final suele ser una tabla de concordancia que permite demostrar la correspondencia entre las puntuaciones de ambas pruebas. Estas tablas facilitan a los estudiantes que han tomado dos pruebas la posibilidad de comparar sus resultados en ellas, en los casos en los que el estudiante haya tomado solo una prueba, podrá hacerse una idea de cuál hubiera sido su resultado en la otra (Gempp, 2010). Tal y como apuntábamos en anteriores tipos de escalamiento, en este caso también existe el riesgo de creer que las pruebas son equivalentes y sus resultados intercambiables, pero como veremos más adelante solo los procedimientos de equiparación pueden garantizar que los resultados de una prueba sustituyan los resultados de otra (Pommerich, 2007).

Debido a que los tests que están siendo enlazados, pueden medir algún dominio diferente o haber sido construidos utilizando distintas vías, la concordancia entre las

puntuaciones es especialmente sensible a la población que se utilizó para establecer la función de concordancia (Dorans, Pommerich, & Holland, 2007).

- **Escalamiento vertical: constructos similares, similar fiabilidad, distinta dificultad en diferente población.**

Los tests aplicados a los estudiantes año tras año a lo largo de diferentes cursos escolares, pretenden medir el mismo constructo (comprensión lectora, matemáticas, ciencias, etc.) pero en un amplio rango de dificultad, desde los cursos más bajos, en los que los niveles de dificultad serán menores, a los cursos superiores con mayores niveles de dificultad, puesto que van dirigidos a poblaciones de diferentes edades y con diferente dominio de las áreas evaluadas. Año tras año cobra mayor importancia la idea de poder comparar los resultados de pruebas de rendimiento realizadas en distintos grados (ver Apartado 1.4). En posteriores apartados se presenta un análisis en profundidad de sus características e implicaciones, en este punto el objetivo es mostrar un panorama general dentro de la taxonomía de los diferentes tipos de enlace de puntuaciones.

El escalamiento vertical, trata de establecer un enlace de puntuaciones entre pruebas que evalúan constructos similares (no exactamente los mismos dados los cambios curriculares que se dan en los distintos niveles) con la misma fiabilidad y requerimientos técnicos pero en poblaciones distintas (estudiantes de distinto grado) y con diferentes rangos de dificultad (las pruebas de los grados superiores son inevitablemente más difíciles que las de los niveles inferiores). Este tipo de enlace es comúnmente conocido como escalamiento vertical («*vertical scaling*») (Kolen & Brennan, 2004), no obstante, también se utilizan otras denominaciones como "calibración de test en distintos niveles de habilidad″ (Angoff, 1971), ″escalamiento para la comparativa" (AERA, APA, NCME, 1985), "calibración" (Misslevy, 1992; Linn, 1993), "equiparación vertical" (Skaggs & Lissitz, 1982, 1986; Crocker & Algina, 1986), "enlace vertical" (Carlson, 2011) o "escalamiento transversal" (Carlson, 2011).

El escalamiento vertical, fue durante años un tema tratado por un reducido grupo de especialistas en psicometría, responsable del desarrollo de baterías de tests estandarizados de rendimiento a nivel nacional, aplicadas en educación primaria y

secundaria, sin embargo, el aumento del interés en este ámbito, surge con la creación, por parte de los estados, de sus propias evaluaciones (Dorans, Pommerich, & Holland, 2007). La capacidad de medir a los estudiantes a lo largo de un continuo es cada día más importante, especialmente a partir de la ley «*No Child left Behind Act*» (2001) (Harris, 2007). En las evaluaciones educativas, los resultados de cada curso no son directamente comparables entre sí por estar anclados a su año especifico de origen, situación que limita el valor de la evaluación en diferentes ámbitos, como la medida del progreso individual de los estudiantes durante los años de escolaridad (Gempp, 2010). En consecuencia, el objetivo central del escalamiento vertical estriba en conseguir una escala común en la que situar los resultados de cada uno de los grados, de este modo, es posible evaluar el cambio en la variable medida a lo largo del tiempo. El escalamiento vertical, comprende una variedad de técnicas utilizadas para desarrollar y mantener escalas verticales cuya naturaleza es evolutiva (Carlson, 2011).

La dimensionalidad de las pruebas, es uno de los grandes retos en escalamiento vertical, cuando el objetivo es evaluar un determinado constructo, como por ejemplo matemáticas o comprensión lectora, a lo largo de un amplio rango de cursos, existe cierta vinculación de los mismos a dimensiones curriculares específicas, de este modo, podemos encontrar diversas sub-dimensiones en cada uno de los constructos evaluados, dimensiones que fluctúan entre cursos. Dicha dimensionalidad puede presentar variaciones importantes en función del área objeto de estudio, existiendo áreas con fuertes diferencias entre cursos. No olvidemos que el objetivo de los tests que se desean enlazar es medir un mismo constructo (por ejemplo matemáticas o comprensión lectora) con similar fiabilidad (Dorans, Pommerich, & Holland, 2007). No obstante, Carlson (2011) destaca que la mayoría de los diseños llevados a cabo para la realización de escalamiento vertical utilizan contenidos apropiados para los grados adyacentes.

C) Equiparación.

Comúnmente es considerado el más importante y estadísticamente robusto de todos los tipos de enlace de puntuaciones pero, al mismo tiempo, es el más exigente y el que requiere la asunción de mayor numero de requisitos estadísticos. De acuerdo con la clasificación que hemos seguido hasta el momento (Dorans & Holland, 2006; Holland, 2007) las formas más "primitivas" de enlace de puntuaciones, expuestas en primer lugar

(predicción, escalamiento de baterías) no exigían ningún tipo de requisito en aspectos relativos a la relación entre las pruebas que se desean enlazar. Sin embargo, a medida que avanzamos en dicha clasificación, encontramos procesos de escalamiento más exigentes (concordancia, calibración o equiparación). En el caso de la equiparación se cuenta con numerosas especificaciones técnicas que han de ser tenidas en cuenta. Su objetivo es poner en la misma escala los resultados de una o más pruebas que evalúan el mismo constructo, con la misma fiabilidad y bajo especificaciones técnicas similares, con el objetivo de que puedan ser utilizadas de manera intercambiable.

En el presente trabajo, se dedicará mayor atención al estudio en profundidad de los procesos de equiparación y escalamiento vertical, por las fuertes implicaciones de los mismos en las evaluaciones a gran escala. Por otro lado, es preciso destacar, tal y como apuntan Kolen y Brennan (2004, 2014) que, a pesar de las distinciones mencionadas acerca de la equiparación y el escalamiento, los procedimientos estadísticos utilizados con frecuencia son similares en ambos casos, a pesar de sus diferentes propósito.

2.3 Equiparacion Horizontal y Escalamiento Vertical.

A la luz de la síntesis conceptual presentada en los apartados previos, puede observarse la complejidad que subyace al establecimiento de un marco conceptual claro al hablar de procedimientos para el enlace de puntuaciones. Tradicionalmente, los investigadores han señalado dos tipos de equiparación, horizontal y vertical, denominando equiparación horizontal, tal y como veíamos en el apartado anterior, al proceso llevado a cabo cuando los tests a equiparar se intentan construir con igual dificultad, éste es el caso de la equiparación entre cuadernillos de una misma prueba, la equiparación de resultados entre años en una medición estandarizada, el desarrollo de cuadernillos de pruebas equivalentes para evaluar intervenciones educativas o la equiparación de ítems dentro de un banco. En otros casos la dificultad de los tests a equiparar es claramente distinta ya que se pretende comparar, por ejemplo, la misma habilidad en distintos cursos, en éste caso tradicionalmente se ha utilizado el término equiparación vertical (Skaggs & Lissitz, 1982). Como se puede comprobar, nos encontramos ante objetivos diferentes, en la equiparación horizontal se pretende

comparar tests que miden el mismo constructo, con el mismo nivel de dificultad, mientras que en el caso de la denominada inicialmente equiparación vertical los instrumentos difieren en dificultad. No obstante, en el marco de una misma investigación, es posible que estén presentes ambos objetivos.

Una de las primeras definiciones de estos tipos de enlace, que destaca por su gran claridad, la podemos encontrar en la comunicación presentada en 1982 por Skaggs y Lissitz titulada «*Test Equating: relevant issues and a review of recent research*», dichos autores consideran que en la equiparación horizontal se cuenta con formas de una prueba que están interconectadas, construidas para medir el mismo constructo con similares propiedades psicométricas, para la misma población de examinados. En la equiparación vertical las pruebas son diseñadas para evaluar un mismo constructo pero con diferentes niveles de dificultad, siendo aquí el propósito la equiparación de pruebas que miden la misma habilidad pero en un amplio rango de dificultad. Asimismo, estos autores consideran que en el caso de la equiparación vertical, el problema es más complicado, puesto que el objetivo es desarrollar puntuaciones que unan la misma dimensión en diferentes niveles, cuando los test han sido diseñados con distinta dificultad con la intención de medir a grupos con diferentes habilidades.

Es preciso recordar, tal y como se apuntaba anteriormente que, de acuerdo con las normas incluidas en los «*Standards for Educational and Psychological Testing»* (AERA, APA y NCME, 1999) es más correcto referirse a este tipo de proceso como "escalamiento para la comparativa" o hablar de "métodos de escalamiento vertical" tal y como figura en la última revisión de los «*standards»* (AERA, APA y NCME, 2014). A medida que se ha ido avanzando en este campo de estudio se ha hecho más necesario distinguir entre los términos equiparación y escalamiento para lograr la comparabilidad, reservando el término equiparación para los casos en los que se derivan puntuaciones estrictamente intercambiables, en tests elaborados con las mismas especificaciones para medir el mismo constructo y donde se cumplen los principios de simetría e invarianza que veremos más adelante.

Tanto la equiparación horizontal como el escalamiento vertical (denominación más utilizada) han de ser llevados a cabo en numerosas situaciones prácticas que van más allá del rendimiento académico. A continuación, en base a lo expuesto por distintos

autores, y atendiendo a la idea central expuesta por Prieto y Dias (2003), presentamos algunas situaciones prácticas generales en las que se manifiesta la necesidad de dicho proceso:

a) Pruebas para el ingreso en la universidad, selección de personal, evaluación académica o certificación profesional (Prieto & Dias, 2003). El uso de distintas formas de un mismo test puede ayudar a evitar el fraude y aumentar la seguridad en el proceso.

b) Medición del cambio producido en un atributo (aptitudes, actitudes, disfunciones psicológicas o progreso educativo) por la maduración o la intervención psicológica y educativa (Prieto & Dias, 2003). En el ámbito educativo, tal y como apuntan Patz y Yao (2007) el desarrollo apropiado de una escala vertical, añade valor sustancial a las evaluaciones de rendimiento, posibilitando la estimación y seguimiento del crecimiento a lo largo del tiempo.

c) Construcción de bancos de ítems (elemento clave en los nuevos modelos psicométricos) (Prieto & Dias, 2003). Su uso puede requerir la incorporación periódica de nuevos reactivos, con el consiguiente proceso de equiparación vinculado a dicha incorporación. Del mismo modo, las escalas verticales permiten realizar comparaciones entre los diferentes ítems que componen los tests, permitiendo la selección más adecuada de los ítems que formarán parte de las pruebas de cada grado (Patz & Yao, 2007) (Figura 15).

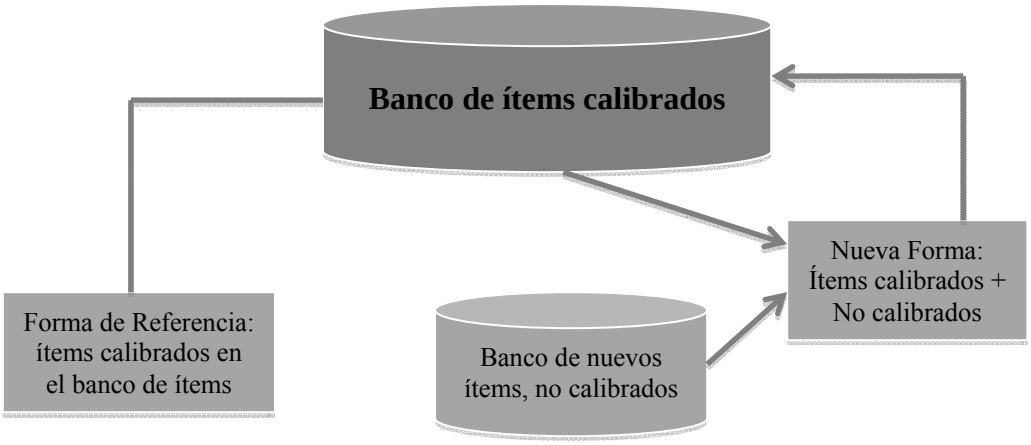


Figura 12. Banco de ítems.

Fuente: Petersen (2007) (p.65).

d) Validación intercultural de tests**.** La construcción y adaptación de instrumentos de evaluación psicológica, ha de dar respuesta al creciente interés por la investigación intercultural y el aumento de pruebas globales de dimensión internacional (Prieto & Dias, 2003).

e) Análisis del funcionamiento diferencial de los ítems (FDI) (Prieto & Dias, 2003). La detección del FDI en el marco de la TRI, se realiza a partir de la comparación de las curvas características de los ítems obtenidas en los distintos grupos, pero antes de realizar dicha comparación, es necesario que las estimaciones de los parámetros estén en la misma escala (Navas, 2000).

f) Construcción de pruebas que abarquen un amplio dominio del constructo a evaluar, sin la necesidad de incluir un número excesivo de reactivos que dificulte la aplicación y funcionamiento adecuado de las mismas. La introducción de bloques comunes para la equiparación entre distintos cuadernillos es frecuente en diversas evaluaciones a gran escala, exigiendo la aplicación de la técnica de enlace adecuada.

g) Mantenimiento de las escalas de puntuación («*score scales»*) de diferentes evaluaciones cuando se desarrollan nuevas formas (Kolen, Tong, & Brennan, 2011). Para que la equiparación sea posible, las formas han de ser desarrolladas bajo el mismo contenido y especificaciones estadísticas que la forma original utilizada para la construcción de la escala (Kolen, Tong, & Brennan, 2011).

La importancia del establecimiento de técnicas adecuadas para la equiparación y el escalamiento de puntuaciones es enorme, lo que ha producido una creciente implicación de áreas diversas de la psicometría y la educación, con el consiguiente incremento (cuantitativo y cualitativo) de trabajos realizados en esta línea. Navas (1996) señala que, la realidad con la que se cuenta en las actuales situaciones de evaluación, en las que usualmente se utilizan diferentes instrumentos de medida, implica que ambos procesos (equiparación/ escalamiento de puntuaciones) sean indispensables, ya que representan el instrumento básico de que se dispone para asegurar la adecuada comparación de las puntuaciones obtenidas en distintas pruebas. En esta misma línea, von Davier (2007), apunta que, una buena equiparación se asemeja a la buena cocina:

comienza con buenos ingredientes, utiliza las herramientas adecuadas y requiere un poco de talento.

Centrados en el ámbito educativo, no debemos olvidar la importancia que cobra año tras año el desarrollo de modelos que permitan evaluar tanto el progreso individual de los estudiantes, como el establecimientos de modelos de crecimiento («*growth models»*) y algunas medidas de valor añadido. La principal razón para la creación de una escala vertical, es la necesidad de medir el aprendizaje a lo largo del tiempo (Harris, 2007). Estos aspectos, afectan a todos los implicados en el proceso educativo: estudiantes, padres, instituciones responsables, investigadores, etc. Tal y como veíamos en el Capítulo I, dicha importancia se está viendo plasmada en la creciente extensión de diferentes programas de evaluación llevados a cabo por diversos organismos, nacionales e internacionales, dichos programas, necesitan evolucionar a lo largo del tiempo para fortalecer o consolidar su relación con las reformas educativas, con los cambios curriculares, con las poblaciones de referencia y, en definitiva, con la sociedad en su conjunto (Liu & Walker, 2007). En los entornos educativos, que se centran fundamentalmente en el logro educativo, los sistemas de evaluación se encuentran continuamente en un momento de transición, o al menos deberían hacerlo (Brennan, 2007). Una de la ironías de la medición educativa es que estos cambios en los programas de evaluación (incluso cuando dichos cambios suponen una mejora) ponen hasta cierto punto en peligro la comparabilidad de las puntuaciones (Brennan, 2007).

2.4 Requisitos para la Equiparación.

Existe una importante paradoja teórica dentro del proceso de equiparación, pues tal y como han reconocido diversos autores, para hablar propiamente de establecer una equiparación entre las puntuaciones de dos tests, ambos han de medir la misma variable con similares propiedades psicométricas. Si se trata de variables distintas no es posible hablar de equivalencia y, del mismo modo, si las formas a equiparar no tienen la misma fiabilidad, se comentará el error de equiparar un test con otro menos fiable asumiendo que las puntuaciones en ambos son intercambiables (Muñiz, 2003, p. 276). Si una equiparación rigurosa exige medir lo mismo con la misma fiabilidad, la paradoja está clara, tal y como Lord (1980) afirma, una equiparación estricta sólo es posible cuando

es innecesaria. Por su parte, Harris y Crouse (1993) consideran que, cuando dos pruebas son muy similares, no equiparar puede ser la mejor opción ya que, la equiparación, podría introducir más error del que puede eliminar. Cuando los métodos de equiparación son aplicados a pruebas que difieren en contenido, dificultad o confiabilidad, las puntuaciones resultantes no pueden ser utilizadas como intercambiables (Yin, Brennan, & Kolen, 2004).

Gran variedad de autores han tratado los criterios de equiparación (Angoff, 1971; Lord, 1980; Morris, 1982; Dorans & Holland, 2000; Kolen & Brennan, 1995, 2004, 2014; Holland & Dorans; 2006; van der Linden, 2013). El trabajo de Lord (1980) sentó las bases de multitud de trabajos posteriores en relación a las propiedades y condiciones que han de ser tenidas en cuenta en la equiparación, señalando los que podríamos denominar criterios clásicos, dichos criterios, subyacen en los trabajos desarrollados con posterioridad.

Los cuatro criterios clásicos propuestos por Lord (1980) son los que aparecen a continuación, el cuarto criterio "misma característica", no es especificado en el listado de criterios principales, siendo incorporado de forma complementaria en dicho apartado:

1. Equidad: el principio de equidad apunta que si la equiparación de los tests *x* e *y* es equitativa con cada individuo, debería ser indiferente, en todos los niveles de habilidad, a qué test haya dado respuesta el sujeto (*x* o *y*) (Lord, 1980). La distribución de frecuencias condicional sobre el test *y* después de la transformación es la misma que la distribución de frecuencias condicional sobre el test *x*. Una vez que se ha procedido a la equiparación, las puntuaciones pueden utilizarse de manera intercambiable, siendo indiferente a cuál de ellas ha contestado el sujeto evaluado.

2. Invarianza poblacional: es otra de las propiedades apuntadas por Lord y recogidas por diferentes autores en distintos trabajos. En ella se considera que la transformación, producto del proceso de equiparación, será la misma independientemente de la muestra con que se obtenga. La equiparación deberá ser la misma sin importar la distribución de habilidad en las dos

poblaciones equivalentes (Lord, 1980). Por otro lado también nos permite asegurar que la transformación debe ser invariante para cualquier sub-población de evaluados, es decir, la equiparación no está sesgada para diferentes grupos de interés que puedan tomar la prueba (genero, nacionalidad, etnia, etc.).

3. Simetría: Lord (1980) señala que, un requisito básico de la equiparación, es que el resultado sea el mismo sin importar a qué test llamemos *x* y a que test llamemos *y*, requisito que no puede satisfacerse cuando predecimos la puntuación de un test a partir de la puntuación en otro. La propiedad de simetría, por tanto, requiere que las puntuaciones equiparadas sean simétricas, es decir, que la función que se utilice para la transformación de la forma *x* a la escala de la forma *y*, sea la inversa de la función usada para transformar las puntuaciones en la forma *y* a la escala de la forma *x*. Estas reglas de la propiedad de simetría en equiparación no se cumplen en regresión, puesto que, por lo general, la regresión de *y* sobre *x* es diferente a la regresión de *x* sobre *y* (Kolen & Brennan, 1995). Es decir, los resultados serán los mismos utilizando $Y^* = f(X)$ ó $X^* = f(Y)$.

4. Misma característica: Lord destaca que las puntuaciones de los tests *x* e *y* no podrán ser equiparadas a menos que se satisfagan dos condiciones: que las dos medidas sean perfectamente fiables o que los dos tests sean estrictamente paralelos (Lord, 1980). Diferentes formas de una misma prueba deben ser construidas para medir el mismo contenido y con las mismas propiedades estadísticas si se pretende realizar una equiparación, de lo contrario, independientemente de los métodos utilizados, las puntuaciones no serán realmente intercambiables. Esta propiedad es esencial para poder considerar las puntuaciones de formas alternativas de un mismo test intercambiables (Kolen & Brennan, 1995). Tal y como señala Lord (1980), los dos test deben medir las mismas características (aptitud, rasgo latente, destreza, etc.).

En numerosos trabajos podemos encontrar análisis detallados acerca del cumplimiento de dichos requisitos. De este modo trabajos como los de Morris (1982),

Pommerich, Hanson, Harris, y Sconing, (2004) o Dorans y Holland (2000), Holland y Dorans (2006), Gemmp (2010), Dorans (2012, 2013) van der Linden (2013), Holland (2013), entre otros, analizan la importancia de los requisitos de equiparación clásicos e incorporan nuevas aportaciones.

La pregunta en torno a ello es clara "*¿Es posible cumplir en la práctica con tales exigencias?"* El difícil cumplimiento de las condiciones anteriormente expuestas ha propiciado el surgimiento de numerosos debates y desacuerdos acerca de qué es la equiparación y sobre qué métodos deben usarse (Martínez Arias, 2005), del mismo modo, el grado de importancia concedido a cada uno de estos requisitos, suele ser el tema más debatido entre los distintos autores. Las condiciones en torno a las que se da mayor consenso son en las de *simetría e invarianza poblacional*, condiciones que la mayoría de los autores destacan por su importancia, sin embargo, el mayor desacuerdo surge en torno a la condición de *igualdad* y en menor medida en la de *la misma característica* (Martínez Arias, 2005). En torno a ésta última dimensión la polémica está en si ha de ser unidimensional o por el contrario podría ser de naturaleza multidimensional. En la práctica, la condición de *igualdad* es improbable que se satisfaga (es prácticamente imposible construir formas de igual fiabilidad en cada nivel de aptitud) y como consecuencia, se ha propuesto sustituirlo por una condición de igualdad débil, que se concretaría en que la media condicional en cada nivel de aptitud del test Y sea igual que la media condicional del test X (Martínez Arias, 2005). Este criterio debilitado, permitiría equiparar formas que difieren en dificultad y/o fiabilidad (de especial importancia en escalamiento vertical).

En función del cumplimiento de dichos requisitos, se podrá hablar de equiparación de puntuaciones o será preciso referirse a otras técnicas de enlace de puntuaciones que no pueden considerarse equiparación en sentido estricto por no cumplir uno o varios de los requisitos de la equiparación. En algunos casos, estas condiciones son dicotómicas y fácilmente comprobables (por ejemplo la invarianza o la fiabilidad), sin embargo, el problema está en que en otros casos el investigador ha de hacer un juicio cualitativo respecto a cuánto nos podemos acercar al cumplimento del requisito (Gempp, 2010) (Tabla 36).

Tabla 36.

Cumplimiento de los requisitos de la equiparación por las diferentes técnicas de enlace de puntuaciones

REQUSITO	**Equiparación**	**Escalamiento**					Predicción
		Baterías	Anclaje	Calibración	Concordancia	Vertical	
Constructos	=	≠	≠	=	≈	≈	≈ o ≠
Fiabilidad	=	≠	≠	≠	=	≈	≈ o ≠
Simetría	Necesaria	Necesaria	Necesaria	Necesaria	Necesaria	Necesaria	No existe
Equidad	Necesaria	No existe	No existe	No existe	No existe	No existe	No existe
Invarianza	Necesaria	Necesaria	Necesaria	Necesaria	Necesaria	Necesaria	Posible

Fuente: Adaptado de Gempp (2010).

2.5 Consideraciones prácticas.

2.5.1 Secuencia en el proceso de enlace.

Tal y como se ha podido apreciar a lo largo del presente capítulo, la complejidad práctica y conceptual en relación a los diferentes procedimientos de enlace de puntuaciones es notable, de este modo, el buen desarrollo de los mismos dependerá en gran medida de las decisiones tomadas durante su evolución. La falta de consenso en la literatura al respecto, dificulta la tarea de establecer criterios, pautas, procedimientos, etc. que ayuden a una correcta toma de decisiones para asegurar la idoneidad del proceso. Con el fin de reflejar una idea global acerca de qué supone el proceso de equiparación, Kolen y Brennan (2004, pp. 7-8) proponen una serie de pasos:

1. Decidir el propósito de la equiparación.
2. Construir formas alternativas.
3. Elegir un diseño para la recogida de datos (los tipos de diseño se presentarán en el Apartado 2.5.2).
4. Aplicar el diseño de recogida de datos.
5. Seleccionar una o más definiciones operativas de equiparación.
6. Seleccionar uno o más métodos estadísticos de estimación.
7. Evaluar los resultados de la equiparación.

Son muchas las decisiones a las que se enfrenta el investigador a la hora de llevar a cabo un proceso de equiparación. Del mismo modo, cuando se trata de un proceso de escalamiento vertical, es preciso considerar las decisiones que el investigador ha de tomar, decisiones que exigen gran cuidado en la planificación del

proceso. Autores como Kolen y Brennan (2004), Harris (2007) o Gempp (2010), han analizado las distintas decisiones a considerar en el desarrollo de una escala vertical. A continuación, en base a las propuestas de dichos autores, se presenta un recorrido a lo largo de las decisiones más destacadas que pueden condicionar los resultados finales.

1. Nivel de análisis de los datos: decidir sobre si se desea información a nivel individual, agregado (profesores, escuelas, áreas territoriales, etc.) o ambas opciones (Gempp, 2010).

2. Especificaciones de los tests a equiparar:

    a) Distancia entre los grados que se pretenden escalar: la validez, fiabilidad y precisión de una escala vertical es inversamente proporcional a la distancia entre los cursos que se desean enlazar (Gempp, 2010).
    b) Contenido a evaluar y cobertura del instrumento: algunas áreas evaluadas en las pruebas de desempeño están muy ligadas al currículum escolar, en estos casos, los estudiantes tienden a obtener mejores resultados al finalizar el curso en el que se introducen esos nuevos contenidos que al final del curso anterior (Kolen & Brennan, 2004). Sin embargo, en pruebas que no están ligadas tan estrechamente al currículo, se espera que la media de crecimiento de los estudiantes en un curso escolar, sea similar a lo largo de las sub-áreas de contenido (Kolen & Brennan, 2004).
    c) Dimensionalidad: este aspecto está estrechamente relacionado con el descrito en el apartado b de esta sección, la variable medida, no deberá sufrir grandes modificaciones a lo largo de la escolaridad o de los niveles que se pretenden medir (Gempp, 2010). Del mismo modo, será preciso decidir cómo se define el contenido a lo largo de los grados especificando el contenido compartido por grados adyacentes (Harris, 2007). La dimensionalidad está en parte relacionada con la idea de crecimiento. La construcción de las distintas formas (en particular su contenido y especificaciones estadísticas) tienen un importante impacto en los resultados de la escala vertical, incluyendo los efectos de suelo y techo y la superposición entre pruebas (Harris, 2007).

d) Tipo de ítems a incluir: opción múltiple, respuesta abierta, crédito parcial, etc.

3. Definición de crecimiento: este es considerado un aspecto crucial en la construcción de una escala vertical, siendo imprescindible el desarrollo de una definición conceptual de crecimiento, especialmente en pruebas relacionadas con el currículo escolar (Kolen & Brennan, 2004). De acuerdo con estos autores existen dos enfoques principales: definición de dominio de crecimiento y definición de crecimiento grado a grado (Kolen & Brennan, 2004). Los estudiantes contestan a pruebas específicas de su nivel.

4. Cuestiones técnicas: las diferentes elecciones técnicas durante el proceso pueden dar lugar a distintos resultados en las escalas verticales pudiendo derivar en diversas interpretaciones de crecimiento (Tong & Kolen, 2007). Aspectos tales como el tipo de diseño de recogida de datos, la composición de las pruebas y la calidad de los reactivos, el modelo de estimación de la habilidad, o el método de escalamiento, pueden repercutir en los resultados y sus interpretaciones (Tong & Kolen, 2007). Harris (2007) destaca que son numerosos los aspectos que podrían considerarse bajo la categoría de "cuestiones técnicas" tales como el método de calibración, la elección de la técnica de enlace entre los parámetros de los ítems para situarlos en la misma escala, la elección del modelo, etc. Sin embargo, no existe un estudio comparativo definitivo y los investigadores no cuentan con unas directrices claras e inequívocas a seguir.

5. Articulación de la escala vertical y las escalas horizontales: el investigador puede optar por presentar escalas diferentes para la evaluación horizontal y la vertical o unir ambas en una escala común. La decisión acerca de la escala debe evitar resultados paradójicos (Gempp, 2010).

Finalmente, debemos tener en cuenta la idea apuntada por Harris (2007) que señala que, en lugar de centrar la atención en qué método particular de escalamiento es mejor, deberíamos analizar qué conjunto de opciones funcionan mejor en relación al propósito que se pretende conseguir y bajo qué condiciones particulares.

2.5.2 Diseños.

Es inevitable unir los diseños de recogida de datos con los métodos de equiparación y escalamiento, pues ambos están estrechamente relacionados, la base de un adecuado procedimiento para la comparabilidad está en el correcto diseño de recogida de información que nos permitirá la utilización de alguno de los métodos que se propondrán más adelante. Asimismo, algunas investigaciones exigen un tipo u otro de diseño de recogida de datos que será determinante a la hora de la elección del procedimiento para la comparabilidad de mediciones a llevar a cabo. Los diseños de escalamiento y equiparación de puntuaciones constan de dos componentes principales, el primero de ellos es el diseño de recogida de la información y el segundo es el modelo estadístico utilizado para conseguir la comparabilidad entre los tests de interés (Cook, 2007).

Los diseños de recogida de la información, de acuerdo con Holland (2007) se diferencian en dos grandes grupos, aquellos que utilizan poblaciones comunes y aquellos basados en ítems comunes. Dentro de los diseños que utilizan poblaciones comunes, tal y como exponen Holland y Dorans (2006) se encontrarían el diseño de un solo grupo («*single group desing*»), el diseño con grupos equivalentes *(«equivalent group desing»*) y el diseño con contrabalanceo («*counterbalanced desing*»), todos ellos los veremos con más detalle en la presente sección. La utilización de diseños con ítems comunes surge de la necesidad de controlar las diferencias en la habilidad de los sujetos cuando existen dos poblaciones diferenciadas, en lugar de solo una (Holland, 2007). Si tenemos dos poblaciones (P y Q), en el diseño con ítems comunes, la forma X, y un conjunto de ítem comunes (o test de anclaje) A, sería aplicado a la población P, y el test Y, junto con los ítems de anclaje A, sería aplicado en la población Q (Holland, 2007). Este procedimiento es denominado diseño de grupos no equivalentes con test de anclaje «*nonequivalent groups with anchor test*» (NEAT) (von Davier, Holland, & Thayer, 2004; Holland & Dorans, 2006) o diseño con ítems comunes para grupos no equivalentes (Kolen, 2007).

Los diseños que aparecen con más frecuencia en las investigaciones empíricas son fundamentalmente 3: *el diseño con grupos equivalentes, el diseño de un solo grupo con contrabalanceo y el diseño con test de anclaje*. La elección del diseño condicionará

tanto cuestiones prácticas como estadísticas (Kolen & Brennan, 2004), por ese motivo es importante prestar especial atención a dicho aspecto. Las diferencias en el diseño, pueden ser más importantes que el procedimiento estadístico utilizado para el escalamiento (Kolen & Brennan, 2004).

En primer lugar, el *diseño con grupos aleatorios («Random Groups Desing»*) (Kolen & Brennan, 1995, 2004, 2014) o *diseño con grupos equivalentes* (Holland & Dorans, 2006) (Figura 13), consiste en la asignación aleatoria de los sujetos a cada una de las formas que serán administradas utilizando con frecuencia un diseño en espiral (Kolen & Brennan, 2004). En este tipo de diseño, diferencias encontradas entre ambos grupos serán consideradas un indicador de diferencias en la dificultad entre ambas formas, el diseño se basa en la generación de muestras equivalentes gracias al azar (Kolen & Brennan, 2004).

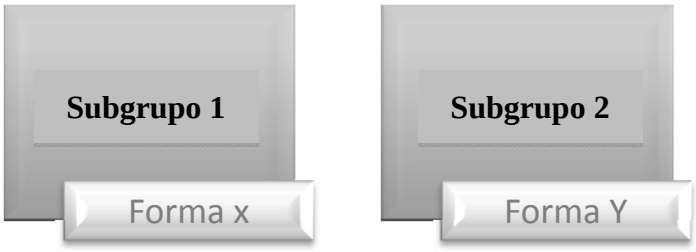


Figura 13. Diseño de dos grupos al azar.
Fuente: Kolen y Brennan (2004, p. 14).

Kolen (2007) apunta que, en el caso del escalamiento, en el que las test han sido construidos con la intención de medir el mismo constructo, posibles diferencias en las condiciones de medida dificultarían la asignación aleatoria, por ejemplo, el tiempo de respuesta al test X podría diferir respecto al tiempo de respuesta del test Y, lo que impediría realizar ambas pruebas en el mismo aula, con la consiguiente modificación en las condiciones de aplicación de ambas formas, en este sentido, una posible solución sería la asignación aleatoria de escuelas, con la dificultad añadida del gran número de sujetos que sería preciso evaluar, Kolen denomina este procedimiento («*random groups desing- randomization by school»*) (Kolen, 2007).

En el *diseño de un solo grupo* se cuenta con una muestra aleatoria de sujetos a la que se aplicarán los tests a equiparar, es decir, todos los sujetos responderán a ambos

tests. Uno de los problemas que puede surgir en este tipo de diseños, es el derivado del orden de aplicación de las pruebas, pues sería lógico pensar que, factores como el cansancio o la familiaridad con la prueba, podrían afectar al instrumento aplicado en segundo lugar, de manera positiva o negativa (Kolen & Brennan, 2004), para poder resolver este inconveniente, una medida comúnmente utilizada es el contrabalanceo, basado en la división de la muestra en dos partes aleatorias aplicando en ambas los tests en distinto orden (Figura 14) (Kolen & Brennan, 2004).

Figura 14. Diseño de un solo grupo con contrabalanceo.
Fuente: Kolen & Brennan (2004, p. 14).

Del mismo modo que sucedía en el diseño con grupos equivalentes, en el caso de que el objetivo sea el escalamiento (y no la equiparación), se han de tener en cuenta ciertas especificaciones del diseño. En el escalamiento, los tests son construidos con la intención de medir constructos similares, asumiendo por tanto, la existencia de posibles variaciones en el contenido de ambos test y, en consecuencia, variaciones en las condiciones de medida que podrían dificultar la aplicación de ambas pruebas en la misma sesión o aula (Kolen, 2007). Como consecuencia de dichas dificultades, en el caso del escalamiento se utilizan variaciones de este tipo de diseño. Entre las posibles variaciones se encontraría la asignación aleatoria a la condición (forma que se toma primero) realizada por escuela, denominado por Kolen «*single group design with counterbalancing for linking- randomization by school*» (Kolen, 2007) y la búsqueda de grupos naturales en que se haya administrado la prueba en ambos órdenes «*single group design with counterbalancing for linking- naturally occurring groups*» (Kolen, 2007).

En la práctica, la utilización del diseño de un solo grupo resultaría muy difícil de justificar, siendo el diseño de un solo grupo con contrabalanceo la opción más frecuente y que resulta más ajustada (Kolen, 2007) pero, ¿en qué aspectos basaríamos la decisión entre la elección del método de un solo grupo con contrabalanceo y el de dos grupos equivalentes? En el ámbito aplicado, el diseño de un solo grupo con contrabalanceo, de acuerdo con lo expuesto por Kolen y Brennan, (1995, 2004) podría ser aplicado en lugar del diseño de dos grupos al azar, cuando se cumpliesen las siguientes condiciones: la administración de las dos formas del test a los mismos sujetos sea operacionalmente posible, no esperar que existan verdaderos efectos devidos al orden de aplicación, dificultad de contar con la participación de suficiente número de examinados para realizar el estudio de equiparación utilizando el diseño de dos grupos al azar.

Por último debemos hablar del *diseño con tests (o ítems) de anclaje para grupos no equivalentes*. Este es el diseño más utilizado, consiste en aplicar los dos tests a equiparar a dos muestras (uno a cada una) pero además ambos instrumentos cuentan con un número determinado de ítems comunes, a los que se denomina ítems de anclaje (o test de anclaje), y que son exactamente iguales en ambos instrumentos. En este caso las muestras no tienen por qué ser equivalentes (ver Figura 15).

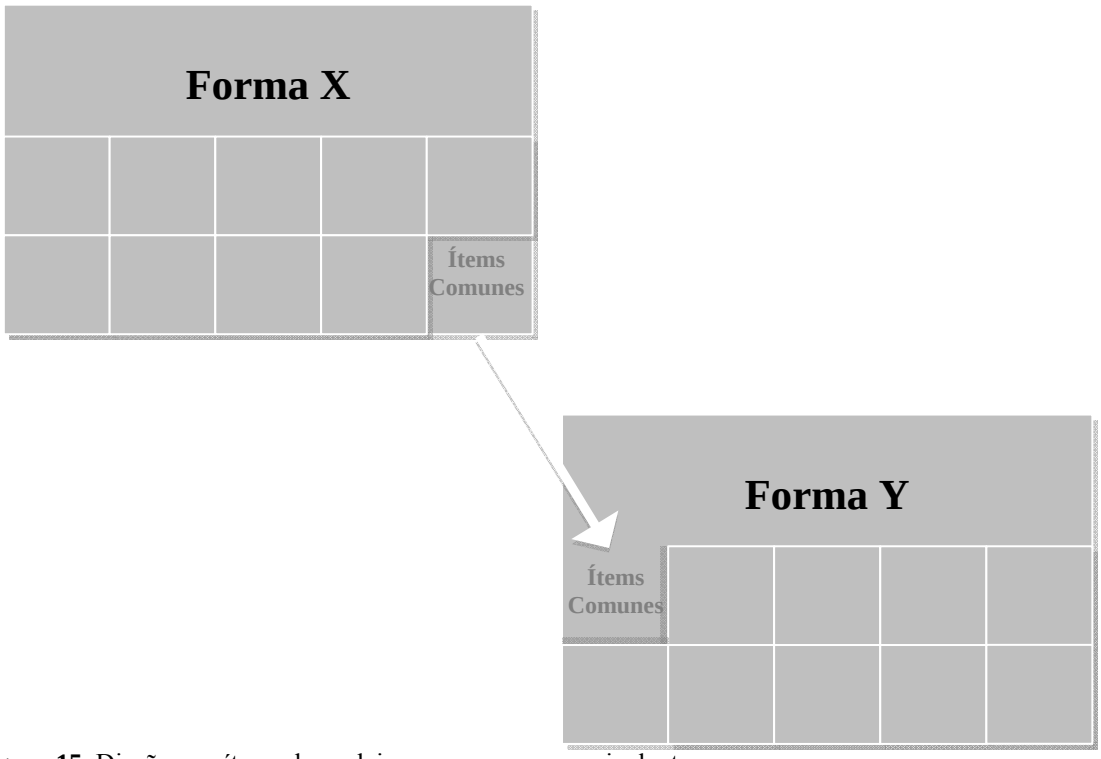


Figura 15. Diseño con ítems de anclaje para grupos no equivalentes

Fuente: Kolen & Brennan (2004, p. 14).

Se pueden diferenciar dos tipos de test de anclaje, *interno o externo*. Un test de anclaje *interno* es un subconjunto de elementos contenidos en ambos tests cuyas puntuaciones se utilizan al calcular las puntuaciones de las dos pruebas en las que está presente. El test de anclaje interno no precisa de más tiempo de aplicación que el diseño de grupos equivalentes. Un test de anclaje *externo* es una prueba separada que se administra a cada examinado junto con la forma que le corresponda. Sus puntuaciones no se usan en el cálculo de las puntuaciones globales de los tests. Requiere más tiempo de aplicación que el diseño de grupos equivalentes, aunque si el test es corto esta diferencia de tiempo no es relevante (Kolen & Brennan, 2004). Tal y como apunta Martínez Arias (2005), las investigaciones que usan ítems de anclaje suelen optar por la modalidad de anclajes externos, pues ello posibilita contar con referentes fuera del propio instrumento que presentan la ventaja de poder ser utilizados como independientes pero al mismo tiempo con una fuerte correlación con la variable medida.

Las puntuaciones en el test de anclaje pueden usarse para estimar el rendimiento del grupo de examinados en las dos formas, simulando, por medio de procedimientos estadísticos, la situación en que el mismo grupo de examinados cumplimentase las dos formas del test, los ítems comunes nos aportan elementos para la transformación. Idealmente, el test de anclaje estará formado por elementos similares a los de las dos formas a equiparar. El significado de la puntuación de anclaje no cambia de una forma a otra. Cuanto mayor sea la correlación entre este test de anclaje y las dos formas, más útil resultará este procedimiento. No obstante, cuando los grupos se han formado por aleatorización, es factible el uso de tests de anclaje que no miden la misma aptitud que las formas a equiparar, ya que estos pueden proporcionar información útil y reducir la varianza de error de la equiparación (Lord, 1980). Tal y como apunta Cook (2007), el diseño con test de anclaje, es seguramente el que tiene una mayor prevalencia en la investigación práctica pero, al mismo tiempo, es uno de los diseños más complejos de implementar. Una de las razones de esta difícil implementación es que, este tipo de diseños, funcionan mejor bajo ciertas condiciones: similitud entre las muestras de sujetos a los que se aplican las formas a comparar, similitud entre las dos formas a equiparar y fuerte relación entre las puntuaciones a equiparar en ambas formas y el test de anclaje (Cook, 2007).

En este punto, es preciso destacar la principal diferencia que nos encontramos entre el diseño con test de anclaje para grupos no equivalentes en equiparación y en escalamiento. Tal y como hemos observado en los diseños anteriores en el caso de la equiparación, los tests a equiparar han sido construidos para medir el mismo constructo, siendo uno de los principales requisitos del diseño que el contenido de los ítems comunes represente adecuadamente el contenido de las formas a equiparar, sin embargo, cuando el contenido de los tests a equiparar difiere (escalamiento) resulta difícil suponer que el contenido del test de anclaje representará el contenido de ambas formas, por tanto, en el caso del escalamiento, el diseño con tests de anclaje para grupos no equivalentes se utiliza sin considerar estrictamente dicho criterio (Kolen, 2007). En esta situación, cuando se usa el diseño con ítems de anclaje para grupos no equivalentes en escalamiento, los resultados pueden estar condicionados por el uso de diferentes ítems de anclaje, siendo un diseño sensible a la elección de los mismos. Para dar respuesta a esta situación los métodos estándar pueden ser modificados para el uso de múltiples test de anclaje en un mismo escalamiento (Kolen, 2007).

En este sentido, la elección de los ítems que formarán parte de los test de anclaje, tiene una importancia crucial en este tipo de diseño, tanto en la equiparación como en el escalamiento. De este modo, Weeks y Domingue (2013) señalan que las características de los ítems comunes tienen un importante impacto en la calidad del enlace de puntuaciones, apuntando que éstos deben ser representativos del constructo o constructos evaluados, deben ser seleccionados intencionalmente y tienen que estar representados en un número suficiente. De acuerdo con lo expuesto por Martínez Arias, Hernández y Hernández (2006, pp. 427-428) debemos tener en cuenta una serie de cuestiones a la hora de escoger los ítems de anclaje.

a) *Incluir un número elevado de la forma de referencia. En ocasiones se habla de 20 pero esto depende de la calidad y el número de ítems.*
b) *Los ítems deben ser similares en formato y contenido a los de las formas.*
c) *Incluir cuestiones representativas de todos los rangos de dificultad.*
d) *Las cuestiones incluidas deben ser idénticas en las dos formas, no se permite ninguna variación.*
e) *Cuando hay varios ítems ligados a un estímulo común (por ejemplo un texto, el enunciado de un problem, etc.) debe incluirse el bloque completo.*

f) Los ítems no deben mostrar efectos de funcionamiento diferencial.

Hay otras normas que no son tan esenciales pero si recomendables:

g) No incluir cuestiones del final del test, a menos que el tiempo del test sea muy amplio.
h) Los ítems deben estar en posiciones similares en las dos formas, no necesarieamente en la misma, pero si parecidas, ya que los sujetos pueden responder de forma diferente según la posición del ítem.
i) En igualdad de otras circunstancias, seleccionar ítems que correlacionen alto con la puntuación total.

En esta línea, Kolen y Brennan (2004) señala que, aumentar el número de ítems comunes, reduce el error aleatorio de equiparación, de este modo, indican que la experiencia sugiere que el conjunto de ítems de anclaje debería suponer, al menos, un 20% del total de ítems, sin olvidar aspectos tales como la heterogeneidad del contenido de las pruebas. Es recomendable comparar las propiedades psicométricas de los ítems en ambos grupos a fin de identificar si algunos ítems presentan funcionamiento diferencial, en cuyo caso no podrían formar parte del anclaje (Kolen & Brennan, 2004). Para dicho análisis se podría utilizar una aproximación clásica o basada en TRI, las posibles modificaciones a posteriori, evidencian una vez más, la necesidad de contar con un número suficiente de ítems en el anclaje ya que, si contamos con números demasiado ajustados, no podríamos suprimir del anclaje reactivos que no han funcionado correctamente.

Otros autores (Chong & Osborn Poopp, 2005), hacen una revisión exhaustiva de ciertas características que pueden afectar a los ítems de anclaje, analizando que tipo de ítems son los más adecuados para dicha función, de este modo juzgan la adecuación de distintos tipos de ítems. En su análisis, destacan que los ítems de opción y respuesta múltiple «*Multiple Choice Multiple Answer»* (MCMA) son preferibles a los ítems de opción múltiple y única respuesta «*Multiple Choice Single Answer»* (MCSA). Consideran que el acierto por adivinación en el segundo tipo de ítems es demasiado alto, teniendo que ampliar excesivamente el número de opciones para conseguir una reducción. Sin embargo, en el primer tipo, el acierto por adivinación se reduce por el

número de combinaciones posibles en la respuesta. En consecuencia, el parámetro de dificultad de los ítems de tipo MCMA, es una mejor estimación de la verdadera dificultad del ítem. No obstante, los autores señalan que, disponer de estos ítems de anclaje, no siempre es la mejor opción (por ejemplo, en niveles elementales su uso es problemático), en ocasiones estos ítems son más apropiados para medir conocimientos concretos que habilidades generales relacionadas con una materia o asignatura. Utilizar ítems de tipo MCSA construidos con buenos distractores reduce el acierto por adivinación y, en la mayoría de los casos, se suelen obtener buenos resultados cuando se utilizan como ítems de anclaje (Chong & Osborn Poopp, 2005).

Del mismo modo, Chong y Osborn Poopp recuerdan la vulnerabilidad de éstos tipos de ítems (MCMA y MCSA) a la memorización por parte de los estudiantes así como al posible intercambio de información entre aquellos que toman la prueba antes y los que la toman después, situación que podría poner en peligro su valía como ítems de anclaje. Ítems que no son de opción múltiple podrían reducir este inconveniente (Chong & Osborn Poopp, 2005).

Por otro lado, en este mismo trabajo, los autores consideran que, los ítems que están agrupados en torno a un estimulo común, podrían resultar problemáticos cuando se utilizan como ítems de anclaje, ya que estos ítems podrían presentar altas correlaciones, lo que violaría el supuesto de independencia local, esencial en los modelos de teoría de la respuesta al ítem. Unido a esto, cuando se utiliza un conjunto de ítems que hacen referencia a un mismo estimulo, cada uno de los reactivos presentará diferentes características técnicas, si el investigador decide utilizar este conjunto de ítems, se verá obligado a incorporar como ítems de anclaje algunos ítems que no presenten las adecuadas características técnicas (inestabilidad o imprecisión en los parámetros de los mismos) (Chong & Osborn Poopp, 2005). En esta misma línea, Zu y Liu (2010), analizan el efecto del uso de ítems independientes («*discrete anchor items*») e ítems basados en un estímulo común («*passage-based ítems*»), observando un peor funcionamiento cuando se utilizan éstos últimos.

En último lugar, Chong y Osborn Poopp (2005) señalan que, los ítems de crédito parcial, no deben ser utilizados como ítems de anclaje. Este tipo de ítems implican una sucesión de pasos, en la que cada paso se asocia a un nivel de dificultad a la hora de

obtener mejor o peor puntuación. Cuando se utilizan ítems de crédito parcial como ítems de anclaje, debemos ser conscientes de la substancial perdida de información que se produce, ya que deberíamos considerar un índice de dificultad para cada uno de los niveles de desempeño del ítem (Chong & Osborn Poopp, 2005).

En definitiva, la importancia de la selección de los ítems de anclaje en este tipo de diseño es un aspecto muy importante a considerar. Las implicaciones de los mismos en los procesos de equiparación/escalamiento exigen prestar especial atención a la hora de su selección y utilización en las formas a comparar.

Por último, otro aspecto de importancia en cualquier diseño de recogida de datos empleado, es el relativo al tamaño de la muestra. El tamaño muestral influye directamente en el error de estimación. Las fórmulas utilizadas en el ámbito de la equiparación para el cálculo del error estándar de equiparación («*standard error of equaitng»* (SEE)), o error estándar de escalamiento («*standard errror of linking»* (SEL)) pueden ser utilizadas para estimar el tamaño de la muestra necesario para conseguir un nivel óptimo de precisión en el proceso de equiparación o escalamiento (Liu & Walker, 2007).

Tal y como se ha observado en el apartado anterior, existe una estrecha relación entre los diseños de recogida de datos y los procedimientos de equiparación que es posible utilizar. Por tanto, antes de adentrarnos en la metodología de equiparación, es preciso clarificar como ha sido el procedimiento de recogida de datos y diseñar un proceso de equiparación adecuado, o viceversa, diseñar una investigación que cuente con un diseño de recogida de datos, que nos permita una adecuada comparabilidad de los resultados, mediante el uso de procedimientos de equiparación pertinentes.

CAPÍTULO 3: MÉTODOS DE ENLACE

Existe un amplio abanico de procedimientos estadísticos que permiten la obtención de puntuaciones comparables, en este sentido, es preciso destacar, tal y como hacíamos en el Apartado (2.2) lo apuntado por Kolen y Brennan en 2004, a pesar de las diferencias teóricas e interpretativas mencionadas acerca de la equiparación y el escalamiento, los procedimientos estadísticos utilizados con frecuencia son similares en ambos casos, así mismo Dorans, Moses y Eignor (2010) apuntan que, las mayores diferencias entre los diferentes tipos de enlace de puntuaciones, son de tipo interpretativo, no procedimental.

A la hora de clasificar los distintos métodos, es posible diferenciar dos grandes grupos, por un lado estarían aquellos métodos considerados clásicos, desarrollados principalmente bajo los supuestos de la Teoría Clásica de los Test (transformación en la media, método lineal y equipercentil) y los basados en la Teoría de la Respuesta al Ítem (TRI). A pesar de que gran parte de la investigación especializada en el tema se ha centrado en el estudio de las diferencias entre los distintos métodos, no resulta sencillo decidir sobre el uso de uno u otro, al no existir trabajos concluyentes al respecto. Como se ha podido apreciar a lo largo de los distintos apartados del capítulo 2, son muchos los condicionantes que pueden influir en el enlace de puntuaciones, de este modo, factores como el tamaño de la muestra, la naturaleza del constructo evaluado, el diseño de las

pruebas, el diseño de recogida de información, la distancia entre pruebas, la definición de crecimiento, la calidad de los reactivos, etc., son variables que condicionarán el proceso, de este modo, se destaca, una vez más, la idea apuntada por Harris en 2007, al señalar que, en lugar de centrar la atención en qué método particular de enlace es mejor, se debería analizar qué conjunto de opciones funcionan mejor en relación al propósito que se pretende conseguir y bajo qué condiciones particulares. Tal y como apunta von Davier (2011), la selección del método utilizado para conseguir la comparabilidad importa principalmente cuando la necesidad es mayor (cuando las formas difieren en dificultad), todos los métodos producen similares resultados en situaciones en las que las formas y las poblaciones son idénticas.

En el presente capítulo se expone, en primer lugar, una breve aproximación inicial a los diferentes métodos enmarcados bajo estas perspectivas, posteriormente, se incorporan dos apartados específicos en los que se detallan de forma más precisa dichos procedimientos, el objetivo central de esta sección está en mostrar una panorámica general de tales métodos. El último apartado, está dedicado a los posibles criterios utilizados a la hora de valorar la calidad del proceso de enlace, entre los que se encuentra, como procedimiento fundamental, el cálculo del error de equiparación o escalamiento, cuyo análisis exhaustivo se realizará en el capítulo 4.

3.1 Aproximación conceptual.

La investigación en el ámbito del enlace de puntuaciones ha sido desarrollada desde diferentes perspectivas y marcos de referencia. De este modo, von Davier (2011) señala que, la equiparación de puntuaciones, puede ser llevada a cabo utilizando procedimientos basados en las puntuaciones observadas («*observed-score equating*») (OSE) y en la Teoría de la Respuesta al Ítem (TRI) (von Davier, 2011). Siguiendo la clasificación más común adoptada por distintos autores (Kolen, 1988; Hambleton, Swaminathan, & Rogers, 1991; Yan & Houang, 1996; Navas, 1996, 2000) entre otros, en este apartado presentaremos los procedimientos estadísticos clásicos y los basados en la Teoría de la Respuesta al Ítem, teniendo en cuenta que, la elección entre los distintos métodos, estará condicionada, principalmente, por los propósitos del enlace y las condiciones particulares de la investigación. Crocker y Aligna (1986), en relación a la elección entre los métodos lineal y equipercentil, señalan que, para decidir el método adecuado, será preciso tener en cuenta tres criterios:

1. Analizar si las asunciones subyacentes al modelo son loables.
2. Valorar si el procedimiento resulta práctico.
3. Juzgar los resultados del proceso.

Los métodos clásicos han sido muy populares en la literatura especializada durante años, debido a la sencillez conceptual de sus pasos y a sus adecuados resultados, sin embargo, los métodos clásicos no siempre cubrían las necesidades (Yang & Houang, 1996). En cuanto a las asunciones subyacentes al modelo, es preciso destacar que, los modelos basados en TRI, requieren robustas asunciones difíciles de conseguir en la vida real (Hambleton & Swaminathan, 1985; Hambelton , Swaminathan & Rogers, 1991).

De forma global, los actuales métodos difieren en la necesidad de selección y asignación aleatoria de los sujetos a los grupos, la administración de las pruebas y el uso de anclajes. En el caso de haber sido utilizada una asignación aleatoria de los sujetos a los grupos o si los sujetos toman ambas formas en un orden aleatorio, la aproximación clásica sería adecuada, en el resto de situaciones, los métodos basados en la TRI resultarían más apropiados (Yang & Houang, 1996). Del mismo modo, Navas (2000)

sugiere utilizar el método lineal cuando contamos con pequeños tamaños muestrales, con test similares y cuando solo es necesario obtener gran precisión en un área de la escala próxima a la media. Por otro lado, indica que es recomendable hacer uso del método equipercentil o de procedimientos basados en laTRI en el caso de contar con tamaños muestrales grandes y cuando existe necesidad de precisión y rigor en toda la escala (Navas, 2000). De otro modo, von Davier (2011) apunta que, la elección del método, dependerá fundamentalmente del modelo de medida utilizado. Así, sugiere que los modelos de medida y equiparación que utilizan las puntuaciones totales y los modelos que utilizan la interacción ítem-sujeto por medio de los parámetros, en ocasiones, se superponen o se apoyan mutuamente (von Davier, 2011).

3.2 La equiparación en el marco de la Teoría Clásica de los Test.

Los tres métodos utilizados por la psicometría clásica para equiparar pruebas son fundamentalmente: *Equiparación en la Media, Lineal y Equipercentil* (Kolen, 1988). Los métodos de equiparación basados en Teoría Clásica fueron descritos en detalle principalmente por Angoff (1971) y Kolen (1988). En los métodos tradicionales de equiparación, la correspondencia entre puntuaciones es establecida mediante características de la distribución de puntuaciones igualadas para ambos grupos (Kolen, 1988).

En evaluación, cuando se utilizan métodos basados en las puntuaciones observadas (OSE), la información sobre el ítem es agregada a través de los sujetos evaluados y la distribución de puntuaciones se utiliza como base para equiparar las distintas formas, considerando las formas variables aleatorias (von Davier, 2011). Los procedimientos basados en puntuaciones observadas utilizan como unidad de medida la puntuación total en el test (independientemente de la forma en que se obtuvo) y la equiparación es realizada mediante el emparejamiento de las dos distribuciones de puntuación (ya sea en términos percentiles o mediante la media y la desviación típica) (von Davier, 2011).

3.2.1 Equiparación en la media.

En el método denominado comúnmente e*quiparación en la media*, las medias de las dos formas a equiparar se igualan para un grupo particular de examinados (Kolen, 1988). Se trata del método que podríamos considerar más básico o elemental, habiendo sido utilizado con escasa frecuencia en la literatura.

En este tipo de equiparación, se considera que la forma X difiere de la forma Y en una cantidad constante a lo largo de toda la escala de puntuación (Kolen & Brennan, 2014). Supongamos que la media de un test (X) es de 20 puntos y la de otro (Y) es de 25 ¿cómo se equipararían ambas puntuaciones haciendo uso de este procedimiento? Sencillamente bastará con sumar 5 puntos a las puntuaciones de los sujetos en el test X, o bien restar 5 puntos a las puntuaciones de los sujetos en el test Y. Esta diferencia se considera constante para todo el rango de puntuación, es decir, en todas las puntuaciones, ya sean altas o bajas, se considerará dicha diferencia entre ambas formas.

Las diferencias entre X e Y se justifican en base a diferencias en la dificultad de las pruebas. Se trata de un método simple e intuitivo pero que apenas se utiliza en la actualidad, posiblemente a consecuencia de las fuertes asunciones acerca de la distribución de puntuaciones si se pretende que la equiparación posea algún sentido (Muñiz, 1997). A pesar de que la consideración de una diferencia constante puede ser muy restrictiva, en numerosas situaciones de evaluación el uso de éste procedimiento es útil para ilustrar algunos conceptos importantes relativos a la equiparación (Kolen & Brennan, 2014).

3.2.2 Transformación lineal.

En el caso de la *transformación lineal*, las medias y desviaciones típicas de las dos formas a equiparar para un grupo determinado de examinados se igualan (Kolen, 1988). En resumidas cuentas, se trata de equiparar las puntuaciones con típicas iguales. Tal y como veíamos en el apartado anterior, la transformación en la media consideraba que, las diferencias en dificultad entre ambas formas, son constantes a lo largo de todo el rango de puntuación, sin embargo, en la transformación lineal, se tiene en cuenta la

variabilidad de dicha diferencia, considerando que, en función del nivel evaluado, las diferencias entre ambas formas pueden variar. De este modo, el elemento diferencial entre ambos procedimientos estriba en este punto. La equiparación lineal permite que la diferencia entre las dos formas no sea la misma para sujetos de un nivel bajo en el rasgo evaluado que en el caso de aquellos que tienen mayor nivel. Para los tests X e Y (Kolen y Brennan, 2004):

$$Z_x = Z_y \quad (1)$$

Explícitamente:

$$\frac{x - \mu(X)}{\sigma(X)} = \frac{y - \mu(Y)}{\sigma(Y)} \quad (2)$$

Despejando Y:

$$l_y(x) = y = \sigma(Y) + \left[\frac{x - \mu(X)}{\sigma(X)}\right] + \mu(Y) \quad (3)$$

Donde $l_y(x)$ es la ecuación para la transformación lineal de x a la escala de la forma y. Si reordenamos los términos:

$$l_y(x) = y = \frac{\sigma(Y)}{\sigma(X)}x + \left[\mu(Y) - \frac{\sigma(Y)}{\sigma(X)}\right]\mu(X) \quad (4)$$

De este modo llegamos a la transformación lineal a partir de la pendiente y el intercepto:

$$\text{Pendiente} = A = \frac{\sigma(Y)}{\sigma(X)} \quad (5)$$

$$\text{Intercepto} = B = \left[\mu(Y) - \frac{\sigma(Y)}{\sigma(X)}\right]\mu(X) \quad (6)$$

$$l_y(x) = A(x) + B \quad (7)$$

Aplicando la ecuación de transformación vemos como las puntuaciones de la forma X expresadas en la escala de la forma Y vienen dadas por una transformación lineal de éstas, permitiendo la variación en función del rango o nivel de puntuación. Esta conversión cumple el principio de simetría (a diferencia del método de la regresión) (Angoff, 1971).

En la equiparación lineal, la asunción principal es que la distribución de puntuaciones de los dos test difiere tan solo respecto a las medias y las desviaciones típicas. En tal caso, las puntuaciones estándar serán iguales en ambas situaciones. Cuando se cumple este supuesto, el método de transformación lineal puede considerarse un caso especial de equiparación equipercentil o una aproximación a dicho procedimiento (Hambleton, Swaminathan, & Rogers, 1991).

Igual que sucede en todos los procedimientos estadísticos, la equiparación está sujeta al error aleatorio, como consecuencia de las variaciones en la estimación de las medias y desviaciones típicas de ambas formas fruto del muestreo de sujetos (Angoff, 1971). Del mismo modo Angoff (1971) indica que el error típico de medida para las puntuaciones equiparadas con este procedimiento, en los *diseños de dos grupos equivalentes*, viene dado por:

$$S_e = \sqrt{\frac{2S_Y^2}{N_t}(Z_X^2 + 2)} \qquad (8)$$

Donde N_t es el número total de sujetos considerando ambas muestras y

$$Z_X = {(X - \bar{X})}\Big/{S_X} \qquad (9)$$

El error típico de equiparación aumenta cuando las puntuaciones equiparadas se alejan de la media (Muñiz, 2003).

Cuando el diseño utilizado es el diseño de *grupos equivalentes* a cada grupo de sujetos se le aplica una forma del test, siendo posible la utilización directa de la ecuación de equivalencia a partir las medias y desviaciones típicas de ambas formas

(Navas, 1996). Sin embargo, en el *diseño de un solo grupo con contrabalanceo*, como ya se ha señalado en el apartado de diseños (2.5.2), los tests suelen aplicarse en distinto orden a cada una de las sub-muestras en las que se divide la muestra, con el fin de contrabalancear los posibles efectos derivados del orden de aplicación (Muñiz, 2003). Esta situación, implica la necesaria consideración de cálculos accesorios que permitan la obtención de los datos requeridos para poder utilizar la ecuación de transformación en una segunda fase. De forma específica, los valores globales de la muestra: $\bar{X}, \bar{Y}, S_x \; y \; S_Y$, deben ser obtenidos a partir de los valores de las sub-muestras, teniendo en cuenta que cada sub-muestra contestó al test en distinto orden, tras el cálculo de los estadísticos, el procedimiento es el mismo que para el caso de diseño de dos grupos equivalentes (Muñiz, 2003).

El error típico de medida de las puntuaciones equiparadas (Angoff, 1971) en el diseño de dos grupos vendría dado por:

$$S_e = \sqrt{\frac{\left(S_y^2\right)\left(1 - r_{xy}\right)\left[Z_X^2\left(1 + r_{xy}\right) + 2\right]}{N_t}} \qquad (10)$$

El error típico es menor que el asociado para el diseño de dos grupos equivalentes, en consecuencia, para conseguir la misma precisión con ambos aquél requiere más sujetos que éste (Muñiz, 2003).

En el caso del diseño con *test de anclaje* (ver apartado 2.5.2), se dispone de dos muestras, 1 y 2, y se aplica a cada una de ellas uno de los tests a equiparar, sea el test X al 1 y el test Y a 2. Además un test Z se aplica en ambas muestras. Pueden considerarse dos aproximaciones básicas para la transformación lineal bajo el diseño con ítems comunes los denominados "métodos basados en el grupo sintético" (Braun & Holland, 1982) y los denominados "métodos basados en los datos" (Navas, 1996).

El grupo sintético estaría formado por dos estratos ponderados exhaustivos y mutuamente excluyentes (grupos 1 y 2/ estratos 1 y 2). A partir de los datos en éste grupo sintético, se estimarían las medias y desviaciones típicas de los tests a equiparar (X e Y) (Braun & Holland, 1982).

$$Grupo\ Sintético = f1 + (1 - f)2 \quad (11)$$

La ecuación de equivalencia se construiría en base a los valores de media y desviación típica de X e Y en el grupo sintético (Kolen, 1988).

$$\mu_{Sin}(X) = f_1\mu_1(X) + f_2\mu_2(X)$$
$$\mu_{Sin}(Y) = f_1\mu_1(Y) + f_2\mu_2(Y) \quad (12)$$

$$\sigma^2{}_{Sin}(X) = f_1\sigma^2{}_1(X) + f_2\sigma^2{}_2(X) + f_1f_2\big(\mu_1(X) - \mu_2(X)\big)$$
$$\sigma^2{}_{Sin}(Y) = f_1\sigma^2{}_1(Y) + f_2\sigma^2{}_2(Y) + f_1f_2\big(\mu_1(Y) - \mu_2(Y)\big) \quad (13)$$

Pero en las ecuaciones 12 y 13 no conocemos todos los términos, pues ignoramos los valores de $\mu_2(X)$, $\mu_1(Y)$, $\sigma^2{}_2(X)$ y $\sigma^2{}_1(Y)$, para calcular estos cuatro parámetros se puede utilizar gran variedad de métodos, entre los más frecuentes se encuentra el método de Tucker, el método de Levine para test con la misma y distinta fiabilidad y el método de estimación de frecuencias (Navas, 1996).

A partir de los procedimientos señalados, una vez obtenidos los valores $\mu_2(X)$, $\mu_1(Y)$, $\sigma^2{}_2(X)$ y $\sigma^2{}_1(Y)$ es posible calcular el error típico de medida por medio de la siguiente ecuación (Muñiz, 2003):

$$S_e = \sqrt{\frac{2S_Y^2(1 - r^2)[(1 + r^2)Z_x^2 + 2]}{N_t}} \quad (14)$$

Donde se asume que:

$$r = \frac{b_{XZ(A)}}{S_x} = \frac{b_{YZ(B)}}{S_Y} \quad (15)$$

En segundo lugar, los métodos basados en los datos, equiparan utilizando únicamente la información empírica, es decir, no se estima la puntuación que obtendrían los sujetos en pruebas a las que no han dado respuesta realmente (Navas, 1996). La

ecuación de equivalencia, se construye a partir de éstos datos empíricos. Entre los procedimientos más habituales destaca el método propuesto originalmente por Lord, basado en máxima verosimilitud para la estimación de $\mu(X)$, $\mu(Y)$, $\sigma(X)$ y $\sigma(Y)$, para su exposición nos basamos en la didáctica propuesta de Kolen (1988).

$$\mu(X) = \mu_1(X) + \alpha_1(X/Z)\big(\mu(Z) - \mu_1(Z)\big)$$
$$\mu(Y) = \mu_2(Y) + \alpha_2(Y/Z)\big(\mu(Z) - \mu_2(Z)\big) \quad (16)$$

$$\sigma^2(X) = \sigma_1^2(X) + \alpha_1^2(X/Z)\big(\sigma^2(Z) - \sigma_1^2(Z)\big)$$
$$\sigma^2(Y) = \sigma_2^2(Y) + \alpha_2^2(Y/Z)\big(\sigma^2(Z) - \sigma_2^2(Z)\big) \quad (17)$$

El error típico de equiparación de este método viene dado por la siguiente expresión:

$$\sigma(Y^*/X) = \sqrt{\frac{2\sigma^2(Y)}{N_1 + N_2}(1 - \rho^2)\big((1 + \rho^2)Z_X^2 + 2\big)} \quad (18)$$

Donde;

$$\rho = \frac{\alpha_1(X/Z)\sigma(Z)}{\sigma(X)} = \frac{\alpha_2(Y/Z)\sigma(Z)}{\sigma(Y)} \quad (19)$$

La equiparación lineal puede presentar el problema de proporcionar puntuaciones fuera de rango (mayores que la puntuación máxima de la escala), los resultados dependen en gran parte de los niveles de habilidad de los grupos en los que se realiza la transformación (Martínez Arias, Hernández, & Hernández, 2006).

3.2.3 Procedimiento Equipercentil.

La equiparación con el método de percentiles o *equiparación equipercentil* es el método más usual, esta situación ha producido que, en ocasiones, se hayan llagado a definir como puntuaciones equivalentes aquellas con iguales percentiles (Muñiz, 2003).

A pesar de su sencillez, la equiparación equipercentil suele producir mejores resultados que las anteriores. La función de equiparación equipercentil, es considerada por algunos autores como la más importante de entre los diferentes procedimientos basados en las puntuaciones observadas (von Davier, Holland, & Thayer, 2004).

El método consiste en equiparar aquellas puntuaciones de ambos tests cuyos percentiles son iguales. Supongamos que tenemos dos tests que miden la misma habilidad X e Y, el procedimiento equipercentil consistiría en equiparar las puntuaciones en X e Y cuyos percentiles sean iguales. De este modo, si una puntuación directa de 20 puntos en el test X corresponde al percentil 70, y en el test Y al percentil 70 le corresponde a la puntuación 22, la puntuación de 20 puntos en el test X se equipara a la puntuación de 22 en el test Y, ya que a ambas puntuaciones directas les corresponde el mismo rango percentil. La equiparación equipercentil puede considerarse incluso más general que la equiparación lineal (Kolen & Brennan, 2014). El desarrollo de la equiparación equipercentil propuesto por Braun y Holland (1982), contando con que X e Y son variables aleatorias continuas, y la distribución de frecuencias relativas acumuladas para cada forma es $F_{(X)}$ y $G_{(y)}$, se resumiría en la siguiente formulación general:

$$Equi_y(x) = F^{-1}\left(G_{(y)}\right) \quad (20)$$

A grandes rasgos, el procedimiento para llevar a cabo un proceso de transformación equipercentil, podría estructurarse en dos pasos generales a los que se alude con frecuencia en la literatura:

1. Cálculo de la distribución de frecuencias relativas acumuladas, es decir, porcentaje de casos por debajo de un intervalo para los dos tests a equiparar. Cálculo de los rangos percentiles de las formas X e Y.

2. A raíz de estas distribuciones se obtienen las puntuaciones equiparadas, tomando aquellas que tienen idénticas frecuencias acumuladas.

Sin embargo, lo más frecuente, es contar con distribuciones de puntuación discretas, no continuas, complejizándose la definición de equiparación equipercentil

(Kolen & Brennan, 2014). En este sentido, en el ámbito de la medición, existe la tradición de considerar puntuaciones discretas como continuas utilizando los rangos percentiles (Kolen & Brennan, 2014). "*El rango percentil de una puntuación entera es el rango percentil en el punto medio del intervalo que contiene dicha puntuación, en la práctica, suelen redondearse las puntuaciones al entero más próximo*" (Martínez Arias, Hernández, & Hernández, 2006, p.431). En esta situación, por ejemplo una puntuación de 22 puntos se considera dentro del rango percentil de 21,5 y 22,5, los sujetos con una puntuación de 22 puntos se distribuyen uniformemente dentro de dicho rango. En la terminología utilizada con mayor frecuencia, el rango percentil de una puntuación entera es el rango percentil situado en el punto medio del intervalo que contiene tal puntuación (Kolen & Brennan, 2014). El método utilizado más ampliamente, con el fin de salvar esta dificultad del carácter discreto de las puntuaciones, es el *método de los rangos percentiles* descrito anteriormente, como alternativa el *método Kernel*, remplaza F_x y G_y con aproximaciones continuas suavizadas respecto al procedimiento clásico de rangos percentiles (von Davier, Holland, & Thayer, 2004).

Tanto en los *diseños de dos grupos* como en el *diseño de un solo grupo* (ver apartado de diseños 2.5.2) el procedimiento de equiparación equipercentil es el mismo, ajustándose de manera global a los pasos generales descritos con anterioridad. Sin embargo, en el *diseño con test de anclaje,* dicho procedimiento se complejiza. A continuación presentamos las características del procedimiento equipercentil para el caso particular de diseño con ítems comunes, por ser éste el que presenta una diferencia substancial respecto al procedimiento general anteriormente descrito.

Tal y como sucedía en el caso de la trasformación lineal, cuando contamos con ítems de anclaje en el procedimiento equipercentil, realizaremos la misma distinción; métodos basados en el *grupo sintético* y métodos basados *en los datos* (Navas, 1996). En el caso de los métodos basados en el grupo sintético (población compuesta por dos estratos ponderados), el procedimiento sería el mismo que el descrito en la equiparación lineal (ver ecuaciones 11, 12 y 13). La principal diferencia es que, tras la estimación de las distribuciones de X e Y, no es necesario calcular sus medias y varianzas, bastará con llevar a cabo la equiparación equipercentil como si nos encontráramos ante el diseño de un solo grupo (Navas, 1996).

Dentro de los métodos basados en los datos, encontraríamos el método directo o encadenado, el método del anclaje predictor y el método de anclaje predicho (Navas, 1996). En el primero de los casos se realiza un procedimiento equipercentil como el descrito para un solo grupo entre X y V e Y y V, considerando equivalentes aquellas puntuaciones en X e Y con igual puntuación en V (Navas, 1996). El riesgo en esta aproximación reside en la doble equiparación que implica, ya que puede producir un incremento de la varianza de los errores de equiparación (Navas, 1996). El método de anclaje predictor utilizaría los siguientes pasos: a) construcción del diagrama de dispersión del grupo 1(X y V) y del grupo 2 (Y y V); b) cálculo para cada puntuación V de las correspondientes puntuaciones en X e Y y c) generación de la tabla de conversión (Navas,1996).

Por último, el método del anclaje predicho, se llevaría a cabo siguiendo los siguientes pasos: a) construcción del diagrama de dispersión para el grupo 1 (X y V) y para el grupo 2 (Y y V); b) cálculo de la media de las correspondientes puntuaciones en V para cada posible valor en X; c) cálculo de la media de las correspondientes puntuaciones en V para cada posible valor en Y y d) generación de la tabla de conversión considerando equivalentes las puntuaciones en las formas a equiparar que corresponden al mismo valor de V (Navas, 1996).

Una de las ventajas de éste procedimiento frente a otros, reside precisamente en su carácter práctico, es decir, en el destino u objetivo final al que se destinan las puntuaciones de ambos test; En aquellas situaciones en las que el objetivo es establecer puntos de corte (situación que se da con gran frecuencia), este procedimiento permitirá que la proporción de examinados seleccionados que hayan contestado al test X sea igual a la proporción de examinados seleccionados que dieron respuesta al test Y (Martínez Arias, 2005). En pruebas cuyo fin es dar lugar a una selección de sujetos, utilizar las puntuaciones de pruebas sin equiparar, supone un gran riesgo, puesto que, tal y como se señalaba anteriormente, resulta extremadamente difícil que pruebas diferentes posean exactamente el mismo nivel de dificultad, ante esta situación, el número de sujetos seleccionados será mayor entre aquellos que hayan contestado la prueba más sencilla. Este procedimiento, no establece supuestos en relación a los tests que se equiparan; su aplicación, simplemente permite hacer coincidir la distribución de puntuaciones de dos tests (Martínez Arias, 2005). En esta línea van der Linden (2013), apunta que sería más

apropiado hablar de «*Quantil transformation*», considerando que el denominado procedimiento equipercentil permite equiparar la forma de cualquier distribución de puntuaciones con cualquier otra, lo que no implica necesariamente que se equiparen las puntuaciones observadas de diferentes formas de tests.

3.3 La Equiparación en el marco de la Teoría de la Respuesta al Ítem.

En el ámbito de la psicometría en general, así como en la equiparación en particular, ha existido un uso generalizado de la TCT, poniendo de manifiesto su utilidad práctica. Sin embargo, la debilidad en los supuestos generales en los que se basa, cuya plausibidad real está en entredicho, constituye tanto su fuerza (generalidad de aplicación) como su flaqueza (Martínez Arias, 2005). Siguiendo a Muñiz y Hambleton (1992), entre los problemas planteados por la TCT se podrían considerar fundamentalmente los siguientes:

1. *Las fuentes de error y su operativización*.
2. *Invarianza de las mediciones y de las propiedades de los instrumentos de medida*.

Tal y como revela el nombre de este enfoque (Teoría de la Respuesta al Ítem), la unidad cardinal o elemental de análisis del test pasa a ser el ítem, en lugar de las puntuaciones totales, tal y como ocurría en la TCT (ver Apartado 3.2). La TRI aporta un nuevo enfoque de carácter probabilístico al dilema que plantea la medida de rasgos y constructos latentes (no observables), agitando los pilares sobre los que se sustentaba el enfoque clásico. Las repercusiones de éste nuevo enfoque supusieron toda una revolución en el área de la psicometría, de este modo, la introducción de los modelos de TRI condujo a la mejora de las técnicas de escalamiento (McArdle & Grimm, 2011). Dejar atrás algunas de las limitaciones de la TCT es uno de los principales avances que nos aporta la TRI en el ámbito de la equiparación, de este modo, la propiedad de *simetría* conlleva que la equiparación no se vea afectada por el test usado como referencia, el requisito de *invarianza,* implica que el procedimiento de equiparación sea independiente de la muestra, ambas condiciones, son tenidas en cuenta por la TRI, al

menos desde un punto de vista teórico (Martínez Arias, 2005). Así, si nos centramos en la propiedad de invarianza de los parámetros de la TRI, de los cuatro requisitos a satisfacer para derivar puntuaciones equivalentes, nos quedaríamos con solo uno: todo los test deben de medir el mismo constructo (Navas, 1996). Tal y como apunta Haebara (1980), la equiparación es un proceso de extrema importancia en estudios que implican un modelo logístico y un modelo de rasgo latente en general.

La medida de la habilidad latente del individuo, es obtenida a través de la interacción entre las características de los ítems y de los sujetos que contestan la prueba (von Davier, 2011). La dificultad inherente a estos procesos de cálculo basados en la probabilidad, fue una de las razones por las cuales este procedimiento se retrasó hasta la llegada del uso generalizado de computadores con mayores prestaciones y capacidades de cálculo (Martínez Arias, 2005).

La obtención de medidas independientes de los instrumentos utilizados es una de las aportaciones principales de la TRI, esto nos lleva a preguntarnos *¿por qué es necesario realizar un proceso de equiparación si podemos obtener medidas independientes del instrumento?* (Martínez Arias, 2005). Sin duda para dar respuesta a esta cuestión es imprescindible tener en cuenta que, puesto que los parámetros de los ítems y la aptitud son desconocidos, es necesario realizar una estimación de los mismos, en consecuencia, la indeterminación de la escala demanda que se establezca una métrica común en la estimación, y ésta estará ligada al grupo concreto en el que se realiza la calibración (Martínez Arias, 2005). De este modo, la métrica suele instaurarse por medio de la estandarización de los parámetros de dificultad o los de aptitud. La solución al problema de indeterminación ha de preceder a la comparación de puntuaciones, que implica la transformación lineal para expresar puntuaciones de habilidad (θ) y parámetros de los ítems en la misma métrica para los dos tests (Martínez Arias, 2005). El problema de la indeterminación lineal, consustancial a estos modelos, hace necesario que se sitúen en la misma métrica las estimaciones obtenidas en distintas ocasiones (Navas, 1996).

Tal y como apunta von Davier (2011), asumiendo que el modelo se adecúa a los datos, el ajuste en las diferencias entre dos formas es llevado a cabo, en primera instancia, por medio del enlace de los parámetros de los ítems, trasladándose

posteriormente a las puntuaciones de los sujetos. De este modo, afirma que el atractivo de los modelos de TRI se encuentra dentro de la teoría psicométrica: se trata de modelos matemáticos de los test que permiten inferir la habilidad del sujeto y clasificar a éste conforme a la misma, la vinculación de los parámetros de los ítems para ajustar las diferencias entre formas, es un aspecto inherente al propio modelo teórico (von Davier, 2011). Los modelos de TRI consiguen incrementar su flexibilidad gracias a las fuertes asunciones estadísticas del modelo pero, probablemente, dichas asunciones no se sostengan con precisión en situaciones reales de evaluación (Kolen & Brennan, 2014). En consecuencia, el estudio de la robustez del modelo, de acuerdo a la violación de los supuestos de partida, así como el estudio del ajuste, es un aspecto crucial en el ámbito de la TRI (Kolen & Brennan, 2014).

Los procedimientos de equiparación basados en TRI, son los que presentan mayor uso en estos momentos en situaciones en las que se mantienen los supuestos del modelo y los tamaños muestrales son suficientemente grandes (especialmente adecuados con grupos no equivalentes e ítems de anclaje) (Martínez Arias, Hernández, & Hernández, 2006).

Por su parte Kolen y Brennan (2014), señalan que la equiparación dentro de la TRI suele caracterizarse por seguir tres pasos: a) estimación de los parámetros de los ítems; b) utilizando una transformación lineal los parámetros se transforman a una escala base TRI y c) en el caso de utilizar el número de respuestas correctas, el número de puntuaciones correctas en la nueva forma es transformada al número de respuestas correctas en la vieja forma y después a la escala de puntuación. La escala de habilidad θ, con media 0 y desviación típica 1, puede resultar poco intuitiva, especialmente cuando el destinatario de la información no está familiarizado con la psicometría, situación que puede darse con frecuencia en el ámbito educativo, en consecuencia, transformar la escala θ a otra escala más intuitiva o frecuente, suele ser recomendado cuando se trabaja en el ámbito de la TRI (Navas, 1996).

Dentro de los modelos de TRI, el investigador deberá decidir cuál de ellos se ajusta de forma más adecuada a las características de sus datos y al objetivo de su trabajo. El modelo logístico de tres parámetros utiliza la información de los parámetros *a* (discriminación) *b* (dificultad) y *c* (adivinación) para definir la curva característica del

ítem. El modelo logístico de dos parámetros considera el parámetro *c* igual a 0, suponiendo una simplificación del modelo. El conocido como modelo de Rasch (Rasch, 1960), considera 0 el parámetro *c* y 1 el parámetro *a*, siendo el modelo de trabajo más sencillo, destacando por la simplicidad en los procesos de estimación. A pesar de que los modelos logísticos de tres parámetros (3PLM) y un parámetro (1PLM) son los más utilizados en el ámbito del enlace de puntuaciones, otros modelos como el Modelo de Respuesta Graduada (Samejima, 1969), el Modelo de Respuesta Nominal (Bock, 1972) o el Modelo de Crédito Parcial (Masters, 1982) también pueden ser utilizados, ejemplo de ello son los trabajos de Baker (1992; 1993), Kim y Cohen (2002) y Kim (2010).

Existen dos aproximaciones para la estimación de los parámetros de los ítems y de la habilidad, el procedimiento de *máxima verosimilitud* y el procedimiento de *máxima verosimilitud marginal* (Kolen & Brennan, 2014). El procedimiento de máxima verosimilitud marginal se diferencia por el establecimiento a priori de la distribución de probabilidad en la población (por lo general una distribución normal), los parámetros de los ítems son estimados teniendo en cuenta dicha distribución a priori (Kolen & Brennan, 2014). En el diseño para grupos no equivalentes, los parámetros estimados a través de los procesos descritos anteriormente, se encuentran en diferente escala TRI (Kolen & Brennan, 2014). Por ejemplo, si estimamos los parámetros de la forma X, en base a la población 1, y los parámetros de la forma Y en base a la población 2 (siendo ambas poblaciones no equivalentes), obtendremos las habilidades para cada grupo en una escala con media 0 y desviación típica 1 (escala θ), incluso si los grupos difieren en nivel de habilidad, por tanto la transformación es necesaria (Kolen & Brennan, 2014).

En consecuencia, una vez estimados los parámetros de los ítems, conforme a alguno de los modelos y procedimientos de estimación descritos anteriormente, el problema al que ha de enfrentarse la equiparación de puntuaciones está en el establecimiento de las *constantes en la transformación lineal (α y β)* (solo β en el caso de modelos de un parámetro). Tal y como apunta Navas (1996, p. 328), "*la equiparación en el marco de la TRI supone precisamente la determinación de las constantes en la transformación lineal que permitirá poner en la misma métrica las estimaciones obtenidas en dos ocasiones*". En este sentido Navas (1996) destaca la existencia de un fuerte paralelismo entre los métodos de transformación lineal (descritos en el apartado 3.2.2) y los métodos basados en la TRI, utilizándose en algunos casos

métodos lineales además de los métodos propios desarrollados dentro de la TRI. Por tanto, para el cálculo de dichas constantes se han empleado distintos métodos que dan lugar a los diferentes procedimientos de equiparación.

El método de equiparación seleccionado dependerá en gran medida del diseño de recogida de datos utilizado (Kolen & Brennan, 2014). De este modo, en el *diseño con grupos aleatorios*, los parámetros de las formas X e Y son estimados de forma independiente (utilizando una escala común por ejemplo con media 0 y desviación típica 1), y pueden ser considerados en la misma escala sin necesidad de realizar transformaciones complementarias, ya que los grupos son equivalentes (Kolen & Brennan, 2014). En el *diseño de ítems comunes con contrabalanceo*, los parámetros para todos los examinados en ambas formas pueden ser estimados de forma conjunta (para las dos formas y para los mismos sujetos) lo que permite asumir que están en la misma escala. En el *diseño con ítems comunes para grupos no equivalentes*, los sujetos que toman la forma X y aquellos que toman la forma Y no son equivalentes, por tanto, la estimación de los parámetros de ambas formas no está en la misma escala, sin embargo, la existencia de ítems comunes en ambas formas permitirá crear la función de transformación (Kolen & Brennan, 2014). Como alternativa, el procedimiento denominado *calibración conjunta* («*concurrent calibration*»), permite estimar los parámetros de los ítems de las formas X e Y de manera simultánea, indicando qué ítems son comunes a las dos formas y a qué forma contesta cada sujeto. Otra alternativa sería fijar los parámetros de los ítems comunes en la forma de referencia (vieja forma), cuando se calibran los ítems de la nueva forma, este procedimiento se denomina «*fixed parameter calibration*» y puede introducir mayor error cuando existen grandes diferencias entre ambos grupos (Kolen & Brennan, 2014). De este modo, una de las principales distinciones entre los procedimientos de enlace dentro de la TRI es hablar de «*concurrent and separate calibration*» (Hanson & Béguin, 2002).

En consecuencia, en el diseño para grupos no equivalentes en el ámbito de la TRI, la equiparación consistirá en el cálculo de las constantes para la transformación lineal, constantes que serán la base que permitirá establecer una misma escala para las estimaciones de los parámetros obtenidas en dos ocasiones distintas (Navas, 1996). De este modo, la transformación de los parámetros de los ítems en la TRI se basaría en los siguientes supuestos generales de partida. Si definimos las escalas *I* y *J* dentro de un

modelo logístico de tres parámetros, relacionadas linealmente, los valores de θ de ambas escalas vendrían relacionados de acuerdo a la siguiente expresión (Kolen y Brennan, 2014):

$$\theta_{Ji} = A\theta_{Ii} + B \quad (21)$$

Donde A y B hacen referencia a las constantes en a transformación lineal y θ_{Ji} y θ_{Ii} representan los valores de θ para el sujeto *i* en las escalas *J* e *I*. Los parámetros de los ítems en ambas escalas estarían relacionados conforme a la siguiente expresión:

$$a_{Jj} = \frac{a_{Ij}}{A}, \quad (22)$$

$$b_{Jj} = Ab_{Ij} + B \quad (23)$$

y

$$c_{Jj} = c_{Ij} \quad (24)$$

Donde a_{Jj}, b_{Jj} y c_{Jj}, son los parámetros para el ítem *j* en la escala J, y a_{Ij}, b_{Ij} y c_{Ij}, son los parámetros en la forma I para el ítem *j*. El parámetro *c* es independiente de la transformación, tal y como puede observarse en la ecuación.

Para dos sujetos *i* e i^*, las constantes A y B, para los ítems *j* y j^*, vendrían dadas por la siguiente expresión (Kolen y Brennan, 2014):

$$A = \frac{\theta_{Ji} - \theta_{Ji^*}}{\theta_{Ii} - \theta_{Ii^*}} = \frac{b_{Jj} - b_{Jj^*}}{b_{Ij} - b_{Ij^*}} = \frac{a_{Ij}}{a_{Jj}} \quad (25)$$

y

$$B = b_{Jj} - Ab_{Ij} = \theta_{Ji} - A\theta_{Ii} \quad (26)$$

En ocasiones resulta de mayor utilidad expresar la relación en términos de grupos de ítems o sujetos (Kolen y Brennan, 2014).

$$A = \frac{\sigma(b_J)}{\sigma(b_I)}$$
$$A = \frac{\mu(a_J)}{\mu(a_I)} \quad (27)$$
$$A = \frac{\sigma(\theta_j)}{\sigma(\theta_I)}$$

y

$$B = \mu(b_J) - A\mu(b_I)$$
$$B = \mu(\theta_J) - A\mu(\theta_I) \quad (28)$$

Siguiendo a Hambleton, Swaminathan, y Rogers (1991) cuatro son los métodos para el cálculo de las constantes *A* y *B* incluidos en la panorámica general de la TRI: método de la regresión, método media/ sigma, método robusto media/sigma y método de la Curva Característica del Ítem, el primero de ellos no debería ser considerado por el incumplimiento de la propiedad de simetría. Kolen y Brennan (2014) clasifican los procedimientos dentro de la TRI en dos grandes grupos: los procedimientos Media/Media y Media/Sigma y los procedimientos basados en la Curva Característica. Por su parte Navas (1996) clasifica los diferentes procedimientos en tres categorías: Métodos basados en los Momentos, Métodos basados en la Curva Característica y Otros métodos. A continuación, atendiendo a ésta última clasificación, presentamos los procedimientos de estimación para el cálculo de las constantes dentro de la TRI.

3.3.1 Métodos basados en los momentos.

Tal y como apunta Navas (1996, p.329) "*los métodos basados en los momentos, derivan el valor de las constantes a partir de los primeros momentos de las distribuciones de las estimaciones de los parámetros*". Dentro de esta categoría estarían los procedimientos Media/ Media, Media/ Sigma, Método robusto Media/Sigma y método iterativo de la media y la desviación típica robustas y ponderadas.

Uno de los procedimientos más sencillos para transformar escalas, en el diseño con ítems comunes para grupos no equivalentes, es sustituir las medias y desviaciones típicas de los parámetros de los ítems de las ecuaciones por los de los parámetros de los

ítems comunes, el procedimiento Media/Sigma consistiría precisamente en esto (Kolen & Brennan, 2014).

$$A = \frac{\sigma(b_{Jc})}{\sigma(b_{Ic})} \quad (29)$$

$$B = \mu(b_{Jc}) - A\mu(b_{Ic}) \quad (30)$$

En el método denominado Media/Media, la media del parámetro *a,* estimada para los ítems comunes, es utilizada para estimar la constante de transformación lineal *A*.

$$A = \frac{\mu(a_{Jc})}{\mu(a_{Ic})} \quad (31)$$

Posteriormente, la media del parámetro *b*, estimada en los ítems comunes, es utilizada para estimar la constante *B* conforme a la siguiente ecuación:

$$B = \mu(b_{Jc}) - A\mu(b_{Ic}) \quad (32)$$

Una vez estimadas las constantes *A* y *B* podemos proceder a la transformación de las puntuaciones de los sujetos y de los parámetros estimados en cada caso, sustituyendo los valores de *A* y *B* en las ecuaciones 21, 22 y 23. No obstante, cuando se utilizan estimaciones en lugar de los parámetros reales, o cuando no se cumplen con precisión las especificaciones del modelo de TRI, los procedimientos media/media y media/sigma podrían producir diferentes resultados (Kolen & Brennan, 2014). Una de las razones por las cuales suele ser más utilizado el procedimiento media/sigma, frente al procedimiento media/media, es la mayor estabilidad en la estimación de los parámetros *b* frente a los parámetros *a*. Sin embargo, otros autores, han apuntado a la mayor conveniencia del uso del procedimiento media/media frente al procedimiento media/sigma, por la mayor estabilidad en la estimación de las medias frente a las desviaciones típicas, que suelen resultar más inestables, en este sentido, no existe investigación concluyente acerca de la mayor adecuación de uno u otro procedimiento (Kolen & Brennan, 2014).

Linn, Levine, Hastings y Wardrop (1981), propusieron *el método robusto media sigma* como solución a los problemas planteados por el procedimiento media/sigma tradicional. La idea central del procedimiento es la consideración de los distintos errores de estimación de los parámetros, dando mayor peso a aquellos parámetros de dificultad cuya estimación parece más robusta, ítems con parámetros *b* pobremente estimados tendrán menos peso (Linn, Levine, Hastings y Wardrop, 1980) . En este sentido, el procedimiento exige que inicialmente se realice una ponderación de cada par de ítems comunes *i*, ($b_{y_{ci}}$, $b_{x_{ci}}$) por su peso:

$$w_i = \left[máximo\left\{\sigma^2(b_{y_{ci}}), \sigma^2(b_{x_{ci}})\right\}\right]^{-1} \quad (33)$$

Utilizado el par de valores ($b_{y_{ci}}$, $b_{x_{ci}}$) de los ítems comunes *i* en los test X e Y se ponderan dichos valores con la inversa de la mayor varianza de las dos estimaciones. De este modo, los pares de valores con grandes diferencias reciben bajos pesos, mientras que los pares de valores con bajas diferencias reciben los pesos más altos, por tanto, las mejores estimaciones del parámetro *b*, tienen mayor importancia en el modelo (Linn et al., 1981). La varianza del parámetro de dificultad se obtiene por medio de la inversa de la matriz de información, tomando los valores de la diagonal principal.

La secuencia a seguir en el caso de utilizar el método media/sigma robusto se detalla con claridad en los apéndices A y B del trabajo mencionado (Linn et al., 1980) un resumen del procedimiento sería el siguiente:

i. Cálculo del peso w_i para el parámetro b ($b_{y_{ci}}$, $b_{x_{ci}}$):

$$w_i = \left[máximo\left\{\sigma^2(b_{y_{ci}}), \sigma^2(b_{x_{ci}})\right\}\right]^{-1} \quad (34)$$

Donde $\sigma^2(b_{y_{ci}})$ y $\sigma^2(b_{x_{ci}})$ son las varianzas estimadas en los ítems comunes para el parámetro b.

ii. Normalización de los pesos.

$$\acute{w}_i = {w_i}\Big/{\sum_{j=1}^{k} wj} \quad (35)$$

Donde *K* representa el número de ítems comunes en las formas X e Y.

iii. Estimación de los valores de *b*:

$$\acute{b}_{y_{ci}} = \acute{w}_\iota b_{y_{ci}}$$
$$\acute{b}_{x_{ci}} = \acute{w}_\iota b_{x_{ci}} \quad (36)$$

iv. Determinación de las medias y desviaciones típicas de los parámetros estimados (tras la ponderación previa descrita).

v. Cálculo de los valores de A y B (constantes de transformación) por medio de las medias y desviaciones típicas de los parámetros estimados y ponderados.

Por su parte, Stocking y Lord (1983), partiendo de la estimación de A y B del procedimiento de la media y la desviación típica robustas, proponen el denominado método iterativo de la media y desviación típica robustas y ponderadas («*robust iterative weighted mean and sigma method*»). En dicho procedimiento, a partir de los parámetros *A* y *B,* se opera de forma iterativa, para ello en cada iteración son utilizados pesos robustos con base en la distancia perpendicular de las estimaciones de dificultad de ambas formas (b_x *y* b_y) a la línea de conversión, procedimiento que permite, de forma simultánea, una depreciación del peso de los valores considerados atípicos y de los parámetros cuyas estimaciones han resultado más pobres (Navas, 1996), es decir, a partir de los valores iniciales estimados para *A* y *B* el proceso se repite hasta que los valores de *A* y *B* no varían. Sin embargo, los propios autores no se mostraron muy convencidos con los resultados de este nuevo procedimiento, apuntando una lógica de superioridad de los métodos basados en la curva característica (Stocking & Lord, 1983).

3.3.2 Métodos basados en la Curva Característica.

Los procedimientos descritos en el apartado anterior, pueden producir similares curvas característica del ítem para reactivos que en realidad presentan diferentes

propiedades (Kolen & Brennan, 2014), la limitación común a tales procedimientos está en la consideración del parámetro de dificultad sin considerar los parámetros *a* y *c* a la hora de estimar las constantes en la transformación A y B (Navas, 1996). Lord (1975) afirma que, en sentido estricto, los métodos basados en la curva característica del ítem, serían los únicos apropiados cuando los test a equiparar (en el diseño con ítems comunes) han sido administrados a sujetos con distinto nivel de habilidad. Estos métodos están basados en las habilidades estimadas $\hat{\theta}$, obtenidas a partir de la aplicación de la teoría de la Curva Característica del Ítem (Lord, 1975).

En respuesta a los problemas planteados por los métodos basados en los momentos, Haebara (1980) y Stocking y Lord (1983) proponen sendos métodos alternativos, éstos procedimientos permiten obtener puntuaciones más estables que los métodos basados en los momentos (Kim & Kolen, 2007). Tal y como apunta Haebara (1980), una ventaja del procedimiento propuesto frente a los anteriores es, que los ítems con mayor discriminación, afectan más al proceso de equiparación que los ítems con menor discriminación. Ambos procedimientos se basan en la consideración simultánea de todos los parámetros de los ítems (Kolen & Brennan, 2014).

La puntuación verdadera τ_{xa} de un examinado con una habilidad θ_a en *K* ítems comunes en el test X, vendría dada por la expresión (Socking y Lord, 1983).

$$\tau_{xa} = \sum_{i=1}^{k} P(\theta_a, b_{x_{ci}}, a_{x_{ci}}, c_{x_{ci}}) \quad (37)$$

Del mismo modo, la puntuación verdadera τ_{ya} de un examinado con una habilidad θ_a en *K* ítems comunes en el test Y, vendría dada por la expresión:

$$\tau_{ya} = \sum_{i=1}^{k} P(\theta_a, b_{y_{ci}}, a_{y_{ci}}, c_{y_{ci}}) \quad (38)$$

Para el conjunto de ítems comunes:

$$b_{y_c} = Ab_{x_{ci}} + B \quad (39)$$

$$a_{y_{ci}} = \frac{a_{x_{ci}}}{\propto} \quad (40)$$

$$c_{Y_{ci}} = c_{x_{ci}} \quad (41)$$

Las constantes A y B son elegidas por un procedimiento iterativo de tal modo que se minimice la función F cuando:

$$F = \frac{1}{N}\sum_{a=1}^{N}(\tau_{X_a} - \tau_{y_a})^2 \quad (42)$$

Donde N es el número de sujetos examinados. La función F también es un indicador de la discrepancia entre τ_{ya} y τ_{xa}. Para realizar el proceso iterativo en el que se consigan los valores de A y B que minimicen la función F, es decir, la discrepancia entre los valores de τ_{ya} y τ_{xa}, es necesario utilizar programas informáticos especializados.

En el diseño con test de anclaje, el número de ítems comunes es muy importante, así como la correlación de éstos con el total de la prueba. Características como estas pueden condicionar en gran medida la calidad del proceso de equiparación. Por ejemplo, si los ítems de anclaje son muy fáciles para un grupo y muy difíciles para otro, los parámetros estimados en los dos grupos podrían ser inestables y, por tanto, la equiparación será pobre. De este modo, es importante que los ítems comunes se encuentren en un rango aceptable de dificultad para los dos instrumentos. La evidencia empírica sugiere que los mejores resultados son obtenidos si los ítems comunes representan el mismo contenido en los dos test que se someterán a escalamiento/equiparación.

Por otro lado, es aconsejable asegurarse de que la distribución de habilidad es razonablemente similar en ambos grupos, como mínimo respecto a los ítems comunes. La propuesta de Hambleton, Swaminathan y Rogers (1991) se basa en que una regla de

oro para determinar el número de ítems de anclaje sería que éstos supongan aproximadamente el 20% o el 25% del número total de ítems de la prueba.

Como decíamos anteriormente, dentro de los procedimientos basados en la Curva Característica, los más conocidos son los propuestos por Haebara (1980) y Stocking y Lord (1983). Ambos tienen en cuenta todos los parámetros de los ítems (*a, b* y *c*). El que más difusión ha tenido es el de Stocking y Lord (Martínez Arias, Hernández, & Hernández, 2006). A continuación pasamos a describir brevemente las propiedades características de cada uno de estos procedimientos.

a) Método de Haebara

El conocido como método de Haebara (Haebara, 1980) que estima la diferencia entre las curvas características de los ítems (CCI), se concreta en la suma de la diferencia existente entre la CCI de cada ítem al cuadrado para los sujetos de un valor determinado de habilidad, es decir, dado un valor de θ_i, el sumatorio de la diferencia entre los ítems al cuadrado, puede ser expresada del siguiente modo (Kolen & Brennan, 2014):

$$H_{dif}(\theta_i) = \sum_{j:V} \left[p_{ij}(\theta_{Ji}; \hat{a}_{Jj}, \hat{b}_{Jj}, \hat{c}_{Jj}) - p_{ij}\left(\theta_{Ji}; \frac{\hat{a}_{Ij}}{A}, A\hat{b}_{Ij} + B\hat{c}_{Ij}\right)\right]^2 \quad (43)$$

El sumatorio es realizado para el conjunto de ítems de anclaje (j:v), posteriormente, en una segunda fase, el valor de *Hdif* , es sumado para los sujetos del nivel de habilidad θ_i, el proceso de estimación se basa en el cálculo de las constantes A y B que minimicen el siguiente criterio (Kolen & Brennan, 2014):

$$H_{crit} = \sum_i H_{dif}(\theta_i) \quad (44)$$

b) Método de Stocking y Lord.

Del mismo modo que el procedimiento anterior, este método tiene su origen en el cómputo de las diferencias al cuadrado que se dan entre las Curvas Características de los tests que se desean comparar en diferentes niveles de aptitud (Martínez Arias, 2005). La discrepancia de éste método en relación al de Haebara, es que Stocking y Lord (1983) usan el cuadrado de la diferencia entre las sumas, de acuerdo a la siguiente expresión (Kolen & Brennan, 2014):

$$SL_{dif}(\theta_i) = \left[\sum_{J:V} p_{ij}(\theta_{Ji}\,; \hat{a}_{Jj}, \hat{b}_{Jj}, \hat{c}_{Jj}) - \sum_{j:V} P_{ij}\,(\theta_{Ji}; \frac{\hat{a}_{ij}}{A}, A\hat{b}_{Ij} + B, \hat{c}_{Ij})\right]^2 \qquad (45)$$

En la estimación se procede buscando la combinación de A y B (las constantes que han de ser utilizadas en la transformación lineal) que minimizan el siguiente criterio (Kolen & Brennan, 2014):

$$SL_{crit} = \sum_i S\,L_{dif}(\theta_i) \qquad (46)$$

$SL_{dif}(\theta_i)$ es el cuadrado de la diferencia entre la curva característica del test dado un nivel de habilidad θ_i, en contraste, la expresión $H_{dif}(\theta_i)$, es la suma de las diferencias entre las curvas características de los ítems dado un nivel de habilidad θ_i al cuadrado (Kolen & Brennan, 2014). El procedimiento de Stocking y Lord es muy similar al de Haebara, excepto por la sustitución de las Curvas Características del Ítem (CCI) por las Curvas Características del test (CCT) (Kim & Kolen, 2007).

El procedimiento es iterativo y requiere el uso de algún programa informático. Kolen y Brennan (2014) proponen el uso de los programas ST y POLYST para la suma de las diferencias para los sujetos evaluados.

Actualmente, los procedimientos más utilizados son los basados en TRI, con el diseño con test de anclaje y especialmente con el modelo logístico de un parámetro (Martínez Arias, 2005). Varias investigaciones han contrastado la idoneidad de distintos procedimientos de TRI, sin embargo, los resultados no son concluyentes. De este modo Hanson y Béguin (2002), realizaron un estudio de simulación analizando el

funcionamiento de la calibración conjunta y por separado, dentro de la calibración por separado estudiaron tanto procedimientos basados en los momentos como procedimientos basados en la Curva Característica, los resultados apuntan a un menor error en la Calibración Conjunta, sin embargo dicho resultado no se mantiene en todas las condiciones. Por su parte, Kim y Kolen (2007) apuntaban a un funcionamiento similar de los procedimientos Stocking-Lord y Haebara, frente a cierta mejora en la estabilidad del procedimiento de calibración conjunta, sin embargo, Lee y Ban (2010) analizan el funcionamiento de cuatro procedimientos de enlace basados en la TRI (calibración conjunta, calibración por separado Stocking-Lord, calibración por separado Haebara y transformación de aptitud), los principales resultados muestran un mejor funcionamiento de los procedimientos de calibración por separado y en concreto ligeramente a favor del procedimiento de Haebara.

3.3.3 Otros métodos.

Como alternativa, existen métodos en los que no es necesario realizar una transformación lineal posterior, este es el caso de los procedimientos denominados calibración conjunta/concurrente «*concurrent calibration*» y calibración fija o método de las b´s fijas «*fixed parameter calibration (FPC)*».

El primero de los casos, el procedimiento de calibración concurrente (Lord, 1975), se basa en la estimación simultánea de los parámetros de los ítems de los test de ambas formas, siendo el procedimiento que podríamos considerar más sencillo (Lord, 1975). En este caso, teniendo en cuenta que partimos de un diseño con test de anclaje, se considera que tenemos una única muestra de sujetos, en esta muestra, los sujetos del subgrupo 1 han dado respuesta a la forma X y a los ítems de anclaje V, en la submuestra 2, los sujetos han contestado a la forma Y más los ítems de anclaje pero, a la hora de la estimación, consideraremos que tenemos una única muestra (1+2) y una única prueba (X+Y+V). En el momento de estimar los parámetros se consideran valores perdidos la respuesta de los sujetos de la submuestra 2 a la forma X y, del mismo modo, se considera valor perdido la respuesta de los sujetos de la submuestra 1 a los ítems específicos de la forma Y (pues se trata de ítems a los que no han respondido los sujetos), una vez configurada la matriz de datos, compuesta por una única muestra y una

única prueba, se procede al cómputo simultaneo de los parámetros de los ítems, consiguiendo parámetros en la misma escala para los ítems de la forma X, Y e V, lo que evita la necesidad de una transformación lineal posterior.

En segundo lugar, el procedimiento denominado calibración fija «*fixed parameter calibration (FPC)*», se apoya en el establecimiento de los parámetros de los ítems de anclaje como fijos, para el posterior cálculo de los parámetros de los ítems en la nueva forma, no implicando pasos de enlace intermedios, fijando los parámetros de la forma operativa previamente estimados y calibrados (Kim, 2006). En este procedimiento contamos con una forma considerada operativa o de referencia (X), en la que se estiman los parámetros de la forma X y los ítems de anclaje de manera simultánea, posteriormente, se procede a la estimación de los parámetros de los ítems de la nueva forma Y, estimando de forma libre los ítems específicos de dicha forma y fijando los parámetros de los ítems de anclaje (V), con los valores obtenidos de la calibración anterior (forma X+V), por tanto, en este procedimiento contamos con dos muestras y dos conjuntos de ítems, siendo los parámetros de los ítems de anclaje (V), los que son estimados en un primer paso con la forma de referencia u operativa (X) y fijados en un paso posterior para la estimación de los parámetros de la forma Y. Tal y como apuntan Kolen y Brennan (2014), este procedimiento puede introducir mayor error cuando existen grandes diferencias entre ambos grupos, sin embargo, algunos autores afirman que, en general, los métodos basados en la Curva Característica, así como la calibración conjunta, se han considerado más apropiados que los métodos basados en los momentos, y en la mayoría de los casos, la calibración conjunta ha mostrado un menor error de equiparación (Li, Jiang, & von Davier, 2012).

CAPÍTULO 4: EL ERROR DE ENLACE

En evaluación educativa, el análisis de la magnitud y fuentes del error resulta un factor esencial. Como veíamos en el Capítulo 1 el creciente interés por las evaluaciones a gran escala y los estudios de tendencia o de crecimiento, exige un tratamiento en profundidad de los orígenes específicos del error asociado a este tipo de mediciones. El enlace de puntuaciones (equiparación y escalamiento), es un elemento inherente a tales evaluaciones y cuyo error asociado ha sido obviado en la mayoría de los casos.

En el presente capítulo, tras una breve aproximación inicial en la que se expone de forma global la naturaleza, características e importancia del error de enlace, pasaremos a analizar, de manera más precisa, los procedimientos de cálculo frecuentemente utilizados en su estudio. Cabe destacar que, el término «error de equiparación», es un término comúnmente aceptado, sin embargo, de acuerdo con lo expuesto en el Capítulo 2 del presente trabajo, así como con los objetivos de esta investigación, consideramos más adecuado hablar aquí de «error de enlace» con el fin de poner de manifiesto su importancia tanto en los procesos de escalamiento como de equiparación.

4.1 Naturaleza, características e importancia.

A pesar del creciente interés por las evaluaciones a gran escala, los estudios de tendencia y de crecimiento, aún queda mucho terreno por explorar en lo relativo a la idoneidad de los procesos de medida implementados en los mismos. La posibilidad de comparar puntuaciones, componente esencial en estos trabajos, exige fuertes asunciones metodológicas así como complejos diseños de recogida y análisis de datos (Capítulos 2 y 3). El error asociado a este proceso, ha recibido escasa consideración en la literatura especializada, haciéndose necesario un análisis en profundidad que ayude a determinar la magnitud y origen del error de enlace y, en consecuencia, a un diseño e interpretación de resultados más preciso.

En función de los objetivos perseguidos, cobrarán mayor importancia fuentes particulares de error. Los numerosos pasos a seguir durante el proceso de medida y el reporte de resultados en estas evaluaciones, llevan consigo determinadas fuentes de imprecisión asociadas a cada paso, identificar el origen y la magnitud de los errores en el proceso de medida resulta necesario, pues estos, pueden atentar contra la validez de los resultados finales (Wu, 2010). De este modo, Wu (2010) destaca que, si se está interesado en informar de los resultados individuales de los estudiantes, el error de medida será el más importante, por otro lado, si el objetivo es el cálculo del nivel medio de desempeño de una cohorte, el error muestral será la principal fuente de imprecisión. En el caso de estudios de tendencias o de crecimiento, el proceso de escalamiento puede ser el origen de diversas causas de error, afectando tanto a los resultados individuales como a los análisis por cohortes. En este sentido, los resultados de Michaelides y Haertel (2004) indican que, a nivel de puntuaciones individuales, el porcentaje de varianza de error que supone el error de equiparación es muy pequeño, sin embargo, a nivel de grupo el efecto del error de medida se contrae con el aumento del tamaño de la muestra, cobrando mayor importancia el error de equiparación, especialmente el relativo al muestreo de ítems comunes, ya que éste no se ve afectado por el tamaño de la muestra. En esta misma línea, el trabajo de Monseur y Berezner (2007) señala que, el error de enlace, puede superar a los errores muestrales y de medida.

En este punto, es preciso aclarar que, el componente de error por el que se ve afectado cualquier proceso de enlace, puede descomponerse en dos elementos: *error*

sistemático y *error aleatorio* (Kolen, 1988; Kolen & Brennan, 2014). Para una fácil distinción entre estos dos tipos de error, podemos decir que, el error sistemático, normalmente, se produce cuando existen fallos en las asunciones del modelo o método seleccionado y, por otro lado, el error aleatorio, está causado por las características muestrales (Wang, 2006). Tal y como apuntan Qian, Jiang, y von Davier (2013), numerosos factores pueden causar variabilidad en los procesos de escalamiento y equiparación basados en TRI, tales como la variabilidad entre grupos de examinados y/o ítems, la estacionalidad, diferencias regionales, diversidad en la lengua nativa, género, así como otras variables demográficas.

Los estudios internacionales con mayor impacto en la actualidad (TIMSS, PIRLS y PISA) presentados en el Apartado 1.4, difieren en el modo de tratar dicho error. Tal y como apuntan Monseur y Berezner (2007), los estudios llevados a cabo por la IEA (TIMS, PIRLS) y la OECD (PISA), difieren en el tratamiento y comunicación del error asociado al estudio de tendencias. En los estudios de la IEA el error estándar no incluye la incertidumbre asociada al proceso de escalamiento, sin embargo, PISA sí reconoce esta fuente de error, considerando que las estimaciones realizadas en los estudios de tendencia pueden verse afectadas por la selección de ítems comunes (Monseur & Berezner, 2007). Sheehan y Mislevy (1988) advertían del problema que supone la práctica extendida de ignorar la incertidumbre asociada al proceso de escalamiento cuando se realizan inferencias que utilizan diferentes conjuntos de ítems, situación que se pone de manifiesto de forma contundente en los estudios de cambio o tendencias. En esta misma línea, Michaelides y Haertel (2004) por medio de la comparación de estas dos fuentes de error, observaron cómo sus magnitudes eran similares, apuntando que la forma convencional de cálculo del error de equiparación refleja aproximadamente solo la mitad del error de equiparación real.

En la práctica, evaluar los resultados de un proceso de equiparación y escalamiento resulta difícil, al no estar disponible ningún criterio que permita investigar los sesgos potenciales de los resultados de dicho enlace (Jiang, von Davier, & Chen, 2012). Tal y como ya apuntaban Harris y Crouse (1993), en el ámbito de la equiparación, no se han desarrollado criterios que permitan juzgar de forma adecuada la idoneidad de las diferentes opciones o alternativas posibles durante el proceso. En la última década, varios autores (Michaelides y Haertel, 2004; Wang, 2006; Monseur &

Berezner; 2007; Wu, 2010; Jiang, von Davier & Chen, 2012; Qian, Jiang & von Davier, 2013) han tratado aspectos relacionados con el cálculo del error de enlace y sus implicaciones en el proceso de evaluación, sin embargo, el volumen de investigación es todavía escaso, especialmente si consideramos el tratamiento del error sistemático. Esta falta de tratamiento en profundidad que venimos observando, puede estar producida por la inexistencia de una solución analítica sencilla y las dificultades de implementación de los métodos intensivos de remuestreo como *bootstrap* y *jackknife* (Harris & Crouse, 1993). Por otro lado, la forma convencional de cálculo del error estándar de equiparación tan solo captura la varianza fruto del muestreo de sujetos (error aleatorio) (Kolen & Brennan, 2004; Michaelides & Haertel, 2004), de este modo, van der Linden (2013) señala que los informes de equiparación típicos ignoran el error de equiparación por completo, su error estándar es solo para calcular las fluctuaciones fruto del muestreo.

El estudio del impacto que la violación de los supuestos de la TRI tiene en la estimación de puntuaciones tampoco ha recibido la atención que debería, no incluyéndose el error de equiparación o escalamiento en las estimaciones propias de los estudios de tendencias (Wu, 2010). Si bien es cierto, bajo los supuestos de la TRI, las propiedades específicas de los ítems serán plenamente consideradas a través de la estimación de sus parámetros, sin embargo durante el proceso de escalamiento se producen erróneas especificaciones que alteran dicha situación tales como: pequeños cambios en los ítems, efectos de posición, efectos del currículo, etc., en consecuencia, la utilización de diferentes sets de ítems comunes podría generar diferencias en las transformaciones realizadas, incluso cuando la muestra de sujetos es grande (Monseur & Berezner, 2007). Tal y como apunta Dorans (2012), los modelos de TRI, desafortunadamente, ignoran la influencia potencial de las condiciones de medida.

El cálculo del error en el enlace de puntuaciones es, por tanto, un aspecto relevante en la investigación educativa y psicométrica actual. Los avances en el cálculo del error son importantísimos ya que pueden facilitar la difícil tarea de decidir qué método utilizar o incluso si resulta conveniente llevar a cabo un proceso de enlace o no. A través de un análisis del error estándar, podemos predecir que tamaño muestral necesitaríamos con el fin de alcanzar determinados niveles de precisión en la equiparación y, en consecuencia, realizar comparaciones entre los distintos

procedimientos (Liou & Cheng, 1995). Por otra parte, el análisis del error sistemático, nos ayudará a realizar diseños más precisos en los que se cuantifique el error asociado al incumplimiento de los supuestos del modelo, ayudando a la toma de decisiones en aspectos esenciales como el número de ítems de anclaje a incluir o la dimensionalidad de las pruebas. Debido a la magnitud que puede alcanzar el error de enlace, sus consecuencias en la interpretación de resultados de distintas evaluaciones es un aspecto a destacar, tal y como apuntan diversos autores (Monseur & Berezner, 2007; Wu, 2010) las evaluaciones a gran escala ponen de manifiesto de forma clara este problema, al presentar substanciales cambios en las unidades de análisis al considerar el error asociado al proceso de escalamiento. En efecto, hay situaciones en las que el enlace de puntuaciones introduce más error del que eliminaría, no obstante, una vez tomada la decisión sobre la conveniencia de utilizar un procedimiento de enlace, habrá que decidir que método utilizar, cuál será la vía más recomendable en relación a las características particulares del estudio (Navas, 2000). El procedimiento más adecuado para tomar este tipo de decisiones es sin duda el basado en el cálculo del error.

A continuación se presenta una breve descripción de los dos componentes del error: aleatorio y sistemático; componentes que analizaremos posteriormente con mayor detalle (apartados 4.2 y 4.3).

El error aleatorio está presente cuando las puntuaciones de los examinados son muestras de una población y, a partir de tales puntuaciones, se calcula la equivalencia o equiparación. Es decir, el error aleatorio está influido por la muestra y sus características. En el caso de disponer de datos poblacionales, y no de una muestra de dicha población, este tipo de error no estaría presente. Por tanto, la cantidad de error aleatorio en la estimación disminuye a medida que aumenta el tamaño de la muestra (Kolen & Brennan, 2014).

Por otro lado, el error sistemático puede presentar diferentes orígenes, el método de estimación de las relaciones de equiparación puede ser uno de ellos, introduciendo sesgos en la estimación de las relaciones de equiparación, del mismo modo, el error sistemático puede ser fruto del incumplimiento de las asunciones estadísticas propias del modelo de escalamiento utilizado (simetría, unidimensionalidad, diseño de recogida de datos mal implementado, funcionamiento diferencial) (Kolen & Brennan, 2014). Al

aumentar el tamaño de la muestra el error aleatorio disminuye, situación que no afecta al error sistemático, siendo ésta la principal distinción entre ambos tipos de error (Kolen & Brennan, 2014).

A continuación, se analiza con más detalle cada uno de estos dos componentes del error de escalamiento, deteniéndonos en estudiar los métodos comúnmente utilizados para el cálculo de cada uno de ellos.

4.2 Cálculo del error aleatorio.

El denominado "Error Estándar de Equiparación", es el índice usualmente utilizado para medir el error aleatorio de equiparación, no considerando el error sistemático. Tal y como apuntan Michaelides y Haertel (2004), la forma convencional de cálculo del error estándar de equiparación, tan solo captura la varianza fruto del muestreo de sujetos. Este término procede de la traducción al castellano del término anglosajón «*Standard Error of Equating*» (SEE)**.** Se concibe como la desviación estándar de las puntuaciones equiparadas en diferentes replicaciones hipotéticas del procedimiento de equiparación con distintos tamaños muestrales y poblacionales (Kolen & Brennan, 2004). Así, Kolen y Brennan (2004) indican que, en cada replica hipotética se especifica tanto el tamaño de la muestra como el de la población. Las puntuaciones equivalentes de la forma Y en la escala de la forma X, son estimadas en varios niveles de puntuación con el método particular de equiparación que ha sido utilizado. El error estándar de equiparación, para cada nivel de puntuación, es la desviación estándar, calculada por medio de varias replicaciones, de la puntuación de Y equivalente en la forma X a ese mismo nivel. El error estándar normalmente difiere a lo largo de los niveles de puntuación.

La utilidad del cálculo del SEE es enorme, ya que puede ser empleado tanto para estimar el tamaño necesario de la muestra como para comparar la precisión de diferentes métodos de equiparación. Tal y como recogen Kolen y Brennan (2004) para el cálculo del SEE deben ser especificados cada una de los siguientes apartados:

- El diseño para la recogida de datos.

- La definición de las equivalencias.
- El método utilizado para la estimación de las equivalencias.
- La población de examinados.
- El tamaño de la muestra (tanto para la vieja forma nueva como para la nueva).
- Los niveles de puntuación de interés.

Dado un caso particular, se define $\widehat{eq_y}(x_i)$ como una estimación de la puntuación en X equivalente en la forma Y, y $E\left[\widehat{eq_y}(x_i)\right]$ como la equivalencia esperada, donde E representa la esperanza bajo diferentes muestras aleatorias de la población (Kolen & Brennan, 2004). En definitiva, el error estándar de equiparación, es calculado por medio de la diferencia entre la puntuación equivalente en la forma Y y la puntuación equivalente esperada:

$$\widehat{eq_y}(x_i) - E\left[\widehat{eq_y}(x_i)\right] \quad (47)$$

Supongamos que la equiparación es replicada un gran número de veces, de tal modo que, cada réplica del proceso de equiparación, se basa en diferentes muestras aleatorias de los examinados pertenecientes a esa población (Kolen & Brennan, 2004). La varianza de error de equiparación para el nivel de puntuación (x_i) sería:

$$var\left[\widehat{eq_y}(x_i)\right] = E\left\{\left[\widehat{eq_y}(x_i)\right] - E\left[\widehat{eq_y}(x_i)\right]\right\}^2 \quad (48)$$

El error estándar de equiparación (SEE) es definido como la raíz cuadrada de la varianza de error.

$$se\left[\widehat{eq_y}(x_i)\right] = \sqrt{var\left[\widehat{eq_y}(x_i)\right]} = \sqrt{E\left\{\left[\widehat{eq_y}(x_i)\right] - E\left[\widehat{eq_y}(x_i)\right]\right\}^2} \quad (49)$$

Tal y como se ha apuntado en los apartados precedentes, éste índice se utiliza para el cálculo del error aleatorio, cuyo origen es la utilización de submuestras de sujetos en el establecimiento de las relaciones de equiparación, este tipo de error no estaría presente si se utilizasen datos de la población en su conjunto, como puede observarse en la siguiente ecuación:

$$\widehat{eq_y}(x_i) = E\left[\widehat{eq_y}(x_i)\right] \quad (50)$$

El error estándar podría ser tenido en cuenta para la especificación del diseño de recogida de datos. Conceptualmente, este método de cálculo del error de equiparación se basa en que se seleccionan muestras hipotéticas y el proceso de equiparación es repetido un amplio número de veces, la variabilidad de cada puntuación es contabilizada para obtener el error estándar en cada diseño, incorporando las variaciones oportunas en función de la situación específica.

En el análisis presentado por Kolen y Brennan (2004, 2014), que estudiaremos en detalle en el presente apartado, se asume que la población de examinados es infinita o, al menos, de un tamaño muy amplio. Dichos autores apuntan que, en ocasiones, concebir la población como infinita tiene sentido, como en aquellos casos en los que se considera que ésta está formada por todos los estudiantes pasados, presentes y futuros que van a tomar las pruebas que se pretenden escalar. Por otro lado señalan que, algunos trabajos, estiman que el grupo de examinados es la población, en este caso, no podría hablarse de error aleatorio de equiparación porque no estaría implicada ninguna muestra de sujetos (Kolen & Brennan, 2004). El error aleatorio de equiparación está presente en la mayoría de los casos, pues la dificultad de acceso a la población en su conjunto suele ser un factor común en las distintas evaluaciones.

Las investigaciones llevadas a cabo han utilizado tanto procedimientos de remuestreo (Kolen & Jarjoura, 1987; Zeng, 1991) como procedimientos analíticos (Lord,1982a; Lord, 1982b; Osgwood, 2001). En el primero de los casos, se trata de procedimiento computacionales intensivos (Noreen, 1989), que calculan los estadísticos en miles de replicas, estos procedimientos han sido desarrollados gracias a la evolución en economía y rapidez de los ordenadores (Efron, 1990). En el caso que nos ocupa, estos procedimientos se basan en la extracción de sucesivas muestras, a partir de las cuales, es posible calcular el error estándar. Existen diversos procedimientos intensivos de remuestreo, tales como jackknife (Tukey, 1958), las pruebas de aleatorización (tests de permutación o permutación estocástica), la validación cruzada (Moser, 1951), o el bootstrap (Efron, 1979). El método de remuestreo más utilizado en el análisis del error de equiparación ha sido bootstrap (Kolen & Jarjoura, 1987; Zeng, 1991; Tsai, Hanson,

Kolen, & Forsyth, 2001; Cui & Kolen, 2008), sin embargo, algunos trabajos también han optado por la utilización de jackknife (Fairbank, 1987; Sheehan & Mislevy, 1988; Haberman, Lee & Qian, 2009; Xu & von Davier, 2010). En el ámbito de la Evaluación Educativa, también podemos encontrar el procedimiento «*Balanced Repeated Replications»* (BRR) (replicación repetida balanceada) (OECD, 2005b). La replicación repetida balanceada, utilizada en PISA, es un método empleado para la estimación del error estandar, especialmente indicado cuando se utilizan datos procedentes de muestras complejas donde la agrupación imposibilita la independencia de las observaciones (Kish y Frankel, 1970), consiste en la selección al azar de una unidad (centro educativo) dentro de cada pseudoestrato al que se le atribuye un peso de 0 y se doblan los pesos de los restantes centros (OECD, 2005b). PISA utiliza la adaptación de Robert Fay, consistente en el suavizado a partir de los pesos (Rao y Sao, 1999), este procedimiento, estará indicado para aquellas situaciones en las que se trabaje con muestras estratificadas complejas.

Posiblemente el mayor uso del procedimiento bootstrap en el análisis del error de equiparación, se deba a la claridad con la que recoge los procedimientos que subyacen a los métodos de remuestreo, siendo el procedimiento conceptualmente más sencillo (Efron, 1990), del mismo modo, la frecuencia de su uso, puede deberse al mayor desarrollo teórico y aplicado de éste procedimiento frente a otros (López & Elosua, 2004) o la existencia de su versión no paramétrica, que se adapta de manera adecuada a los ámbitos en los que se desconoce la distribución de probabilidad de las variables aleatorias (tal y como sucede en las ciencias del comportamiento) (Solanas & Sierra, 1992). En consecuencia, de forma global, el procedimiento bootstrap ha ido sustituyendo de manera progresiva a otros métodos anteriores como jackknife (Revuelta & Ponsoda, 2003).

El segundo tipo de procedimientos son los denominados analíticos, estos se basan en los estadísticos de la muestra, el desarrollo de las ecuaciones suele llevar mucho tiempo y sus resultados pueden llegar a ser muy complejos, habiendo sido menos utilizados en la literatura al respecto, especialmente cuando se trabaja con la TRI ya que, tal y como apuntan Tsai et al. (2001), las ecuaciones apropiadas para el cálculo del error estándar todavía tienen que ser desarrolladas para la mayoría de los procedimientos de equiparación basados en la TRI. Los trabajos de Lord (1982b) u

Ogasawara (2001) resultan el referente en dicha línea. El denominado "método delta" (Kendall & Stuart, 1977) es uno de los procedimientos analíticos clásicos.

Ambos tipos de métodos pueden ser usados en muchas situaciones, la decisión entre uno u otro dependerá de la información de la que dispongamos y del uso que vayamos a hacer de la información procedente del cálculo de los errores estándar. En el presente trabajo, nos centraremos en el procedimiento de remuestreo bootstrap por las razones expuestas anteriormente y de acuerdo con la principal línea de trabajo de los investigadores del área, en consecuencia, el análisis de dicho procedimiento será el más exhaustivo de los presentados.

El método bootstrap, es un método de cálculo por ordenador intensivo para la estimación de los errores estándar en una gran variedad de estadísticos. Implica la utilización de múltiples muestras aleatorias (pseudo-aleatorias) (Kolen & Brennan, 2004), la muestra aleatoria para este procedimiento es diseñada con reemplazamiento. El uso de dicho método es más común que el de los procedimientos analíticos, debido a que las ecuaciones adecuadas para el cálculo de los errores estándar aún no se ha desarrollado para muchos de los procedimientos de equiparación basados en TRI (Tsai et al., 2001), así como por el menor esfuerzo requerido (Kolen & Brennan, 2004). De acuerdo con Kolen y Brennan (2004), los pasos a llevar a cabo en la estimación de los errores estándar utilizando el procedimiento bootstrap en una muestra serían los siguientes:

1. Determinar el tamaño muestral.
2. Diseño de la muestra aleatoria, con reemplazamiento de los valores muestrales.
3. Calcular los estadísticos de interés para la muestra bootstrap.
4. Repetir los pasos 2 y 3 R número de veces.
5. Calcular la desviación estándar para los estadísticos de interés bajo las R muestras bootstrap. Esta desviación estándar, es la estimación del error estándar bajo dicho método.

En primer lugar, ilustraremos el uso del procedimiento bootstrap para el cálculo del error estándar en la equiparación de dos formas dentro del diseño de grupos

equivalentes utilizando el método equipercentil (Kolen & Brennan, 2004). En esta situación, contamos con las puntuaciones de una muestra de sujetos N_x con puntuaciones en la forma X, así como con una muestra de sujetos N_y con puntuaciones en la forma Y. Siguiendo los pasos apuntados anteriormente el procedimiento consistiría en:

1. Extracción mediante bootstrap de una muestra aleatoria con reemplazamiento N_x de la muestra N_x.
2. Extracción mediante bootstrap de una muestra aleatoria con reemplazamiento N_y de la muestra de N_y.
3. Estimación, mediante el procedimiento equipercentil de la puntuación equivalente x_1, a partir de las muestras obtenidas en los pasos 1 y 2.
4. Repetir los pasos de 1 a 3 R número de veces. obteniendo $\hat{e}_{y_1}(x_i)$, $\hat{e}_{y_2}(x_i)...\hat{e}_{y_R}(x_i)$.
5. Estimación del error estándar de acuerdo a la ecuación:

$$\widehat{se}_{boot}[\hat{e}_Y(x_i)] = \sqrt{\frac{\sum_r [\hat{e}_{Yr}(x_i) - \hat{e}_{Y.}(x_i)]^2}{R-1}} \quad (51)$$

donde

$$\hat{e}_{Y.}(x_i) = \frac{\sum_r \hat{e}_{Yr}(x)}{R} \quad (52)$$

El interés está en estimar el error estándar en todo el rango de puntuación. Uno de los problemas que puede surgir es el relativo al ajuste en el cálculo de aquellas puntuaciones con baja frecuencia. En esta situación sería recomendable utilizar un procedimiento bootstrap paramétrico. La diferencia fundamental está en el ajuste previo de las distribuciones empíricas tanto en la forma X como en la forma Y, es decir, en el procedimiento bootstrap paramétrico, las muestras de los pasos 1 y 2 son extraídas de distribuciones ajustadas, produciendo estimaciones más estables de los errores estándar, el problema es la introducción de sesgos cuando el modelo paramétrico no es una estimación adecuada de la distribución poblacional, siendo el procedimiento bootstrap paramétrico muy poco utilizado en la práctica (Kolen & Brennan, 2004).

Tal y como hemos analizado en apartados precedentes, en gran parte de las evaluaciones llevadas a cabo en diferentes ámbitos, la equiparación no es un proceso puntual, sino que implica lo que se podría considerar una "cadena de equiparación", esto hace que no siempre se equiparen dos formas de un test, sino que es posible equiparar un test con una cadena de equiparación anterior, numerosos ejemplos de esta situación los tenemos en evaluaciones nacionales e internacionales realizadas en diferentes países, siendo éste aspecto uno de los de mayor interés en la actualidad. El error en una cadena de equiparación se puede calcular también utilizando el método bootstrap, permitiendo el cálculo incluso en cadenas largas de equiparación, no obstante, a medida que aumenta la cadena el proceso de cálculo se vuelve más complejo (Kolen & Brennan, 2004).

Kolen y Brennan (2004) destacan la utilidad que en ocasiones puede suponer el uso de un valor agregado del SEE. A este valor se le denomina media del error estándar de equiparación (mean SEE). Este valor agregado permite la comparación del error cometido bajo diversos procedimientos de equiparación. En el caso de la equiparación equipercentil se calcularía utilizando la siguiente ecuación.

$$\sqrt{\sum_i f(x_i)\, se^2\left[\hat{e}_y(x_i)\right]} \qquad (53)$$

El método bootstrap también puede ser fácilmente utilizado en el diseño de ítems comunes con grupos no equivalentes. En este tipo de diseño, la muestra N_X está formada por los examinados que contestan a la forma X y N_y por los sujetos a los que les fue administrada la forma Y. El proceso de remuestreo será realizado un amplio número de veces y los errores estándar serán las desviaciones estándar de las estimaciones bajo las distintas muestras, tal y como sucedía en el caso del diseño de grupos equivalentes (Kolen & Brennan, 2004).

La complejidad del uso del procedimiento bootstrap reside, principalmente, en la complejidad de la realización de las réplicas requeridas en cada caso, el avance en la capacidad de procesamiento de los modernos equipos informáticos ha ayudado en gran medida a resolver este problema. Sin embargo, en el caso de los procedimientos basados en la TRI las limitaciones técnicas son cuantiosas, requiriendo equipos con grandes

capacidades de cálculo, en esta línea Kolen y Brennan (2004) apuntan las limitaciones de la aplicación de estos métodos en el caso de la TRI por dichas dificultades técnicas.

En el pesente trabajo, no se utilizarán procedimientos analíticos para el cálculo del error de enlace, de este modo, con el fin de ilustrar de forma completa la presente propuesta, analizaremos brevemente las principales características de los procedimientos análiticos, en concreto las de su mayor representante (método delta), sin detenernos en detalles concretaos acerca del mismo.

Los métodos analíticos en general, suelen ser utilizados cuando el tiempo de procesamiento informático necesita ser reducido o cuando se desea estimar el tamaño necesario de la muestra para realizar una equiparación (Kolen & Brennan, 2004) . El método delta está basado en la extensión de las series de Tylor (Kendall & Stuart, 1977).

Los pasos a seguir para la utilización del método delta son:

1. Especificar las varianzas y covarianzas de error para cada θ_j.
2. Encontrar la derivada parcial de la ecuación de equiparación para cada θ_j.
3. Sustituir las varianzas y las derivadas parciales en la ecuación:

$$var\left[\widehat{eq_y}(x_i)\right] \cong \sum_j eq^{2\prime}{}_{yj} var\left(\hat{\theta}_j\right) + \sum_{j \neq k} \sum eq'{}_{YJ} eq'{}_{YK} cov\left(\widehat{\theta_J}, \widehat{\theta_k}\right) \quad (54)$$

Los resultados de los errores estándar serán expresados en términos de los parámetros. Los parámetros estimados se utilizan en lugar de los parámetros que han sido obtenidos al estimar los errores estándar (Kolen & Brennan, 2004). El cálculo de los errores estándar de equiparación utilizando el método delta es muy complejo y las expresiones de los resultados requieren gran esfuerzo.

4.3 El error sistemático.

El error sistemático es el resultado de la violación de las condiciones de equiparación o escalamiento o de las asunciones estadísticas que requiere el método empleado. Este error es más difícil de controlar y de calcular que el error aleatorio, en consecuencia, el volumen de investigación a este respecto, ha sido inferior. La reducción del error sistemático podría lograrse con el diseño cuidadoso de los test, la adecuada implementación del diseño de equiparación, el uso apropiado de las técnicas estadísticas, etc. En realidad, si conocemos el error sistemático que está afectando a nuestra investigación, lo controlaríamos sin dudarlo. Sin embargo, la detección resulta más compleja, pues en evaluación educativa, debemos ser conscientes de las innumerables fuentes de error a las que la evaluación se expone. Como veremos en el presente apartado, la dificultad es aún mayor en estudios longitudinales. Kolen y Brennan (2014) proponen como principales fuentes de error sistemático las siguientes:

1. Técnicas de suavizado. Las técnicas de suavizado ayudan a la reducción del error aleatorio, sin embargo, son una fuente de error sistemático. En este sentido, la utilidad de las técnicas de suavizado radica en la reducción del error aleatorio por encima del error sistemático que introducen.
2. Violación de las asunciones estadísticas del modelo empleado: simetría/ unidimensionalidad.
3. Problemas en la puesta en práctica del diseño de recogida de información. Diferencias en el funcionamiento de los ítems en los grupos a equiparar (por ejemplo efectos de posición).
4. Diferencias sustanciales en la conducta de los grupos a equiparar.

Diversos trabajos desarrollados en los últimos años, se han preocupado por el análisis del error sistemático en las evaluaciones a gran escala o estudios de tendencias, como consecuencia de la violación de los supuestos de los modelos o errores en el diseño e implementación de los mismos (Sheehan & Mislevy, 1982; Michaelides & Hartler, 2004; Monseur, Sibberns & Hastedt, 2006; Monseur & Berezner, 2007; Wu, 2010). Las violaciones en los supuestos del modelo incluirían la multidimensionalidad de las pruebas, el funcionamiento diferencial de los ítems, efectos de posición en los reactivos e inconsistencias en los registros (Wu, 2010). Sin embargo, tal y como

apuntábamos anteriormente, el estudio del impacto que la violación de tales asunciones tiene en este tipo de evaluaciones, no ha recibido el suficiente interés por parte de la comunidad científica, debido probablemente a las dificultades asociadas a su medida, prueba de ello, es la falta de información acerca del error de equiparación o escalamiento en las estimaciones propias de los estudios de tendencias (Wu, 2010).

La TRI, considera que los parámetros estimados para cada ítem recogen plenamente sus propiedades específicas, sin embargo, durante el proceso de escalamiento pueden existir especificaciones erróneas que alteren dicha situación (cambios en los ítems, efectos de posición, efectos curriculares, efectos culturales, etc.) (Monseur, Sibberns, & Hastedt, 2006). En consecuencia, la utilización de diferentes sets de ítems comunes podría generar diferencias en las transformaciones realizadas, incluso cuando la muestra de sujetos es grande (Monseur & Berezner, 2007). En el mundo real, no es apropiado pensar que necesariamente la selección de ítems comunes tenga un mínimo efecto en el proceso de escalamiento (Haberman, Lee, & Qian, 2009), del mismo modo, el tipo de ítems utilizados, podría afectar notablemente al error sistemático de escalamiento (Zu & Liu, 2010). En el año 1988, Sheehan y Mislevy ponían de manifiesto el problema derivado de la práctica extendida de ignorar la incertidumbre asociada al proceso de escalamiento cuando se realizan inferencias que utilizan diferentes conjuntos de ítems.

El análisis del error sistemático se presenta como una de las principales vías de análisis en la mejora y desarrollo de las evaluaciones a gran escala, estudios de tendencia o de crecimiento. La cuantificación de dicho error, supondrá una notable ayuda tanto en el diseño como en la aplicación de tales estudios, mejorando la toma de decisiones sobre aspectos tales como el número de ítems de anclaje a incluir, el tipo de ítems o la dimensionalidad de las pruebas. Del mismo modo, el cálculo del error sistemático presenta notables implicaciones en la interpretación de los resultados, pudiendo influir significativamente en los resultados del proceso evaluador en diferentes niveles de análisis, especialmente en el tratamiento de valores agregados. Tal y como apuntan diversos autores (Monseur, Sibberns, & Hastedt, 2006; Monseur & Berezner, 2007; Wu, 2010) las evaluaciones a gran escala ponen de manifiesto de forma clara este problema, al presentar substanciales cambios en las unidades de análisis al considerar el error asociado al proceso de escalamiento.

Kolen y Brennan (2014) afirman que, la mejor vía para controlar el error sistemático, es a través de un cuidadoso diseño de las pruebas, adecuado diseño y puesta en marcha del procedimiento de recogida de datos y el uso de las técnicas estadísticas apropiadas. Por tanto, la reducción del error sistemático de enlace, ha de estar presente en distintas fases del proceso, desde el diseño de las pruebas hasta su puesta en marcha y posteriores análisis. De este modo, Kolen y Brennan (2014) aportan una serie de pautas a considerar en las distintas fases del proceso. La Tabla 37 muestra un resumen de las principales dimensiones en las que se centra dicha propuesta.

Tabla 37.

Aspectos a considerar en la reducción del error sistemático

Aspectos a considerar en la reducción del error sistemático	
Construcción de la prueba	o Especificaciones iniciales y construcción de la prueba. o Cambio en las especificaciones de la prueba. o Características de los ítems comunes seleccionados.
Recogida de datos: diseño e implementación	o Elección entre los distintos diseños descritos en el apartado (2.5.2). o Desarrollo del plan de enlace. o Características de los sujetos pertenecientes a los grupos a equiparar. o Tamaño de la muestra
Selección del método de enlace	o Media o Lineal o Equipercentil o Tres parámetros o Rasch o Calibración conjunta (Ver Apartado 3.3.3).

Fuente: elaboración propia a partir de Kolen & Brennan (2014).

La utilización de la Raíz del Error Cuadrático Medio («*Root Mean Squared Error*, RMSE») se presenta como una buena alternativa para el cálculo de la combinación de los errores sistemático y aleatorio de enlace en estudios de simulación (Zu & Liu, 2010). El RMSE, representa la desviación típica de las diferencias entre los valores predichos y los observados.

El funcionamiento diferencial de los ítems podría considerarse uno de los aspectos clave que puede incidir en el error de enlace, especialmente cuando se trabaja en el ámbito de evaluaciones a gran escala. El estudio del denominado sesgo en los ítems comenzó a realizarse de forma consistente durante los años 60 (Angoff, 1993).

Camilli y Shepard (1994, p. 8) definen el sesgo como "*invalidez o error sistemático puesto de manifiesto en cómo un test evalúa a los sujetos de un determinado grupo*". Las connotaciones de dicho término apuntan a la necesidad de un replanteamiento de su uso, ahondado en la necesidad de utilizar términos más precisos, de este modo, Holland y Thayer (1988), señalan su preferencia por el uso de términos más neutrales como funcionamiento diferencial del ítem («*Differential Item Functioning*» (DIF)). Los «*Standards for Educational and Psychological Testing*», consideran que existe funcionameinto diferencial del ítem cuando, existe disparidad entre los grupos de sujetos evaluados, con el mismo nivel de habilidad general o similar estatus en determinado criterio, difieren en términos promedio en su respuesta a un ítem particular (American Educational Research Association, 2014). De acuerdo con estas definiciones, podemos afirmar que un item posee DIF cuando sujetos pertenencientes a distintos grupos, con el mismo nivel en el rasgo o atributo evaluado, tienen diferentes probabilidades de responder correctamente al ítem, es decir, la probabilidad de responder correctamente al ítem no depende del nivel de habilidad o dominio del sujeto en el rasgo medido por el reactivo. Tal y como afirma Martinez Arias (2005), no podemos usar de forma intercambiable los términos DIF y sesgo, el sesgo estaría enmarcado en un contexto más global de validez de contenido mientras que el DIF haría referencia al procedimiento estadístico de cálculo de las diferencias. En esta linea Teresi y Jones (2013), señalan que el DIF está adherido a los hallazgos estadísticos, mientras que el sesgo haría referencia a un proceso valorativo más amplio.

Existen diversos procedimientos para la detección del DIF. Entre las características comunes a los distintos procedimientos estaría el uso de los resultados globales del test como criterio para detectar el DIF, la consideración de que los conjuntos de ítems que componen el test son homogeneos y la unidimensionalidad (Martínez Arias, 2005), de este modo, Teresi y Jones (2013) afirman que la mayoría de los procedimientos de detección de DIF asumen los supeustos de independencia local y unidimensionalidad.

De especial importancia en el ámbito del enlace de puntuaciones sería la teoría multidimensional del DIF, según la cual, el DIF se produce cuando, bajo ciertas condiciones, se incumple el supuesto de unidimensionalidad del test. Así pues, la multidimensional es la causa del DIF. Como pioneros y principales garantes de esta

teoría podemos citar a Ackerman (1992), Camilli (1992) y Kok (1988). Existe una habilidad principal (la que se desea medir) y habilidades espúreas (que no se desean medir pero se están midiendo). En evaluaciones como las descritas en el Apartado 1.4, la unidimensionalidad puede ser un factor clave en el diseño de enlace de puntuaciones.

Efectivamente, la reducción del error sistemático podría lograrse con el diseño cuidadoso de los test, si sabemos las fuentes que pueden estar afectando al error sistemático podríamos actuar directamente sobre el mismo. Sin embargo, en situaciones reales de evaluación esto no es posible y se suele trabajar sobre hipótesis posteriores de funcionamiento de los reactivos en situaciones de evaluación. Es frecuente pasar por alto numerosas situaciones en las que el error sistemático pudiera estar presente, pues su detección pasa por la formulación de dichas hipótesis de partida. Del mismo modo, el análisis pormenorizado de los datos no suele arrojar mayor claridad sobre estos aspectos, pues se suelen realizar análisis posteriores que se basan en suposiciones previas acerca de la existencia de tales fuentes de error, sin embargo ¿qué sucede en aquellos casos en los que no contamos con tales hipótesis? Una de las conclusiones más importantes a la que podemos llegar, es la necesidad de atender a la magnitud del error, independientemente de las consideraciones del investigador acerca de su existencia, pues sin lugar a dudas, tener en cuenta todas las fuentes de error resultaría del todo imposible.

CAPÍTULO 5: PROPUESTA DE UN MODELO INTEGRAL PARA EL CÁLCULO DEL ERROR DE ENLACE BASADO EN TÉCNICAS INTENSIVAS DE REMUESTREO DE SUJETOS E ÍTEMS

Tras el análisis presentado en el capítulo 4, en el que se detalla la problemática asociada al cálculo de los errores de enlace, en el presente capítulo se expone un modelo integral para el cálculo del error. Dicho procedimiento, está basado en la utilización de técnicas intensivas de remuestreo (bootstrap) de sujetos e ítems de forma combinada, este procedimiento, permite estimar el efecto del muestreo de sujetos e ítems así como el efecto de su interacción.

A pesar del reconocimiento generalizado de dichas fuentes de error, los procedimientos implementados para su estimación son todavía escasos y sus resultados tratan de forma independiente los errores asociados a la selección de sujetos y a la selección y funcionamiento de los reactivos utilizados. El modelo de estimación sugerido, se presenta como una valiosa vía en la estimación del efecto conjunto de dichas fuentes de error, mostrando una perspectiva más realista que podría evitar la frecuente infraestimación del error de enlace.

5.1 Formulación del problema de investigación.

Retomando la idea propuesta por Wu (2010), resulta imprescindible identificar el origen y cuantificar la magnitud de los errores de medida, puesto que estos pueden atentar contra la validez de los resultados finales de cualquier estudio. Una vez presentadas las alternativas metodológicas más utilizadas en el ámbito de la estimación del error de enlace, nos encontramos en el momento adecuado para ofrecer una alternativa de análisis que dé respuesta a las principales carencias encontradas, en concreto la propuesta deberá atender al principal problema detectado en la revisión precedente, es decir, a la consideración del efecto combinado de la selección de sujetos y reactivos.

Tal y como veíamos en el capítulo 4, el denominado «error aleatorio de equiparación/ escalamiento» se presenta cuando las puntuaciones de sujetos pertenecientes a una muestra se utilizan para establecer las relaciones de anclaje, si toda la población estuviese disponible, no existiría esta fuente de error.

Por otro lado, en apartados precedentes se ha expuesto que, teniendo en cuenta las asunciones teóricas de la TRI, la misma función de anclaje será obtenida independientemente de los ítems comunes que hayan sido utilizados, puesto que las propiedades específicas de los ítems son plenamente contabilizadas por los parámetros de los mismos. Sin embargo, en la práctica, son múltiples los problemas que pueden suceder en la especificación del modelo, tales como pequeños cambios en los ítems, efectos de posición, efectos asociados al currículum o a la dimensionalidad de las pruebas, las condiciones de aplicación, etc. De este modo, bloques alternativos de ítems comunes, podrían generar diferentes resultados en el enlace de puntuaciones. Por tanto, la selección de los ítems utilizados en una evaluación puede ser una de las principales fuentes de error, especialmente en estudios que cuentan con muestras de sujetos muy elevadas, el error muestral es pequeño en comparación con el error fruto del muestreo de ítems, puesto que éstos siempre representan una proporción muy pequeña del universo de ítems posibles.

En consecuencia, los estudios de tendencias o crecimiento, podrían verse especialmente afectados por tales factores, haciéndose imprescindible la estimación del

error asociado al procedimiento empleado para la comparabilidad de puntuaciones. La no inclusión del error de enlace en estos trabajos, puede suponer una considerable infraestimación del error estándar. Esta estimación, deberá considerar tanto el error asociado a la muestra de sujetos, como el error asociado a los ítems utilizados, sin olvidar la interacción entre ambas fuentes de error. En la literatura especializada, se ha venido trabajando en la estimación del sesgo asociado al muestreo de ítems y sujetos, sin embargo sería preciso profundizar en estimaciones que permitan considerar la interacción existente entre ambos factores. La finalidad de esta tesis, es comprobar si el procedimiento propuesto (bootstrap bidimensional), es adecuado para la estimación del error de enlace a partir de la consideración del error fruto del muestreo de sujetos, del muestreo de ítems y de su interacción, así como determinar bajo qué condiciones ofrece mejor funcionamiento estadístico.

En numerosas situaciones prácticas podemos apreciar la importancia del estudio de la influencia de distintos factores de forma independiente y combinada, considerando el efecto de la interacción producida entre los factores objeto de estudio. Por ejemplo, si estamos realizando una investigación sobre el denominado mal agudo de montaña (mal de altura, mal de páramo, etc.) producido por la falta de adaptación del organismo a situaciones de deficiencia de oxigeno en condiciones de altitud, podemos observar una relación directa de sus efectos con la velocidad de ascenso y la altitud alcanzada. Sin embargo, la gravedad de los síntomas sufridos (incluso la existencia o ausencia de los mismos), no depende exclusivamente de dichos factores, siendo imprescindible analizar la "interacción" entre los mismos y las características y situaciones particulares de los sujetos analizados, pues el padecimiento de dicho mal, no depende tan solo de los factores principales reconocidos, siendo fruto, en gran medida, de la interacción entre éstos. Así, imaginemos que estamos analizando los síntomas del mal de altura en dos grupos de sujetos (A y B), el grupo A, se desplaza desde la provincia de Manabí (131 metros sobre el nivel del mar) a la ciudad de Quito (2850 metros sobre el nivel del mar) utilizando como medio de transporte el avión, desplazamiento en el que emplean 1 hora de viaje. El grupo B, realiza el mismo desplazamiento entre dichas ciudades ecuatorianas pero utiliza el tren como medio de transporte, recorriendo dicha distancia en 5 horas de trayecto distribuidas a lo largo de 1 día. Al llegar a su destino, en ambos grupos se observan los síntomas del "mal de altura", no obstante, la gravedad de los mismos será más acusada en los sujetos pertenecientes al grupo A, como consecuencia

del efecto de interacción producido entre los metros ascendidos (2715 en ambos grupos), y la velocidad de ascenso, mucho mayor en el caso del grupo A. Así, si la ciudad de destino fuese La Paz (3640 metros de altitud), y el grupo A se desplazase en avión, empleando unas 11 horas de viaje y el grupo B realizase dicho trayecto en tren, empleando 4 días), seguiríamos observando mayor severidad de los síntomas en el grupo A, destacando la existencia de efecto de interacción entre las dos variables consideradas pues, a pesar de que ambos grupos han ascendido 3509 metros, la interacción entre dichos factores es un elemento crucial a considerar. El efecto producido por la consideración combinada de éstos factores, no podría detectarse si no atendemos a los mismos conjuntamente.

En el caso que nos ocupa, resulta imprescindible reconocer esta fuente de variación, pues sería arriesgado pensar que no existe este efecto de interacción entre las características de los sujetos evaluados y las pruebas o ítems utilizados.

En el marco de la presente investigación, los efectos que nos interesaría analizar serían tanto los efectos principales (sujetos/ítems) como su interacción. El efecto principal sería definido como el efecto directo de un determinado factor, sobre una variable dependiente. En el caso de la interacción, la definición resulta algo más compleja, pues podríamos considerar varias situaciones: a) ciertas combinaciones de dos factores producen efectos más allá de los efectos observados cuando dichos factores se consideran por separado, b) las diferencias medias entre los niveles de factor A no son constantes y por lo tanto, dependen de los niveles de factor B, c) hay un efecto conjunto de los factores A y B en la variable dependiente y, por último, d) Hay un efecto único que no podría predecirse a partir del conocimiento sólo de los efectos de factores principales (Lomax & Hahs-Vaughn, 2012). En consecuencia, hablar del efecto de interacción entre factores admitiría varias acepciones, una definición desde el punto de vista no formal, nos permitiría afirmar que se da interacción en aquellas situaciones en las que el resultado de la combinación de dos factores difiere de la suma de los efectos principales de tales factores sobre la variable dependiente objeto de interés (Pardo & San Martín, 2010).

Una de las herramientas más utilizadas para esclarecer el concepto de interacción, es la representación de dicho efecto a través de un gráfico de líneas. A

continuación, ilustraremos la importancia de considerar el efecto del muestreo de sujetos, de la selección de ítems y de su interacción a través de una representación gráfica simplificada. En los gráficos que se presentan a continuación, el factor A (muestra de sujetos), se representa en el eje X, contando con 3 niveles de dicha variable (tres muestras aleatorias de sujetos), el eje Y representa la habilidad estimada y las distintas líneas representan el segundo factor considerado (conjunto de ítems seleccionados A, B, C y D).

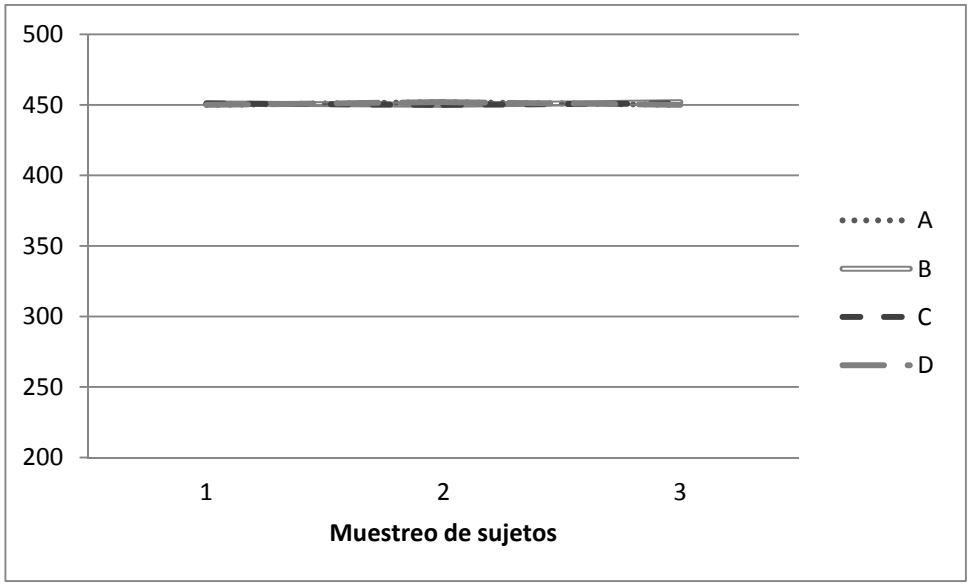


Figura 16. Ausencia de efecto de los factores considerados.
Fuente: elaboración propia.

En la Figura 16, se representa una situación en la que ninguno de los factores considerados (muestra de sujetos 1, 2 o 3) y prueba utilizada (A, B, C, o D) incide en los resultados, no observándose efecto significativo de dichos factores sobre el nivel de habilidad estimado, en consecuencia, la estimación que realice no dependerá de la prueba que seleccione ni de la muestra de sujetos que utilice, pues en cualquier caso obtendré los mismos resultados. Sin embargo, en situaciones reales de evaluación difícilmente encontraremos una situación como la descrita en la Figura 16.

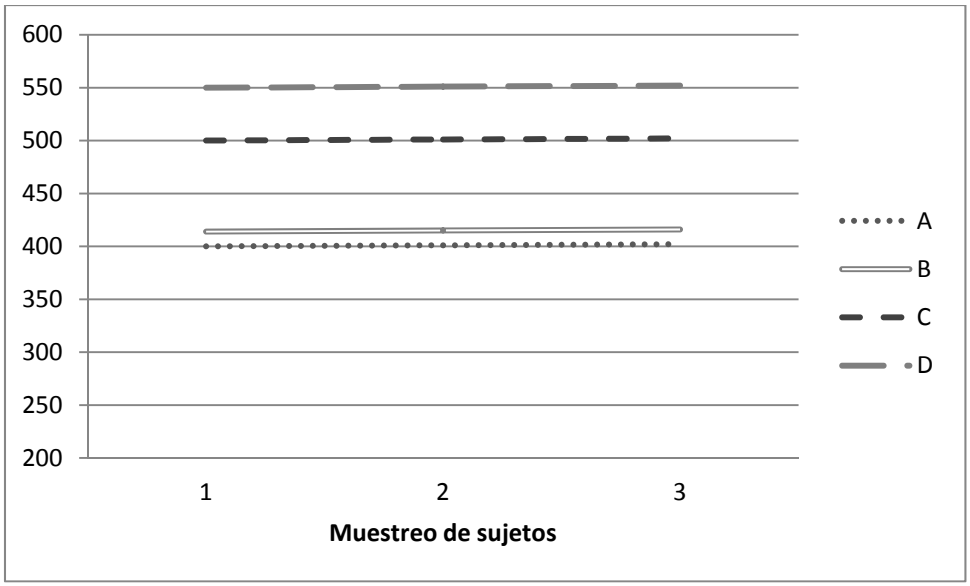


Figura 17. Efecto del factor "muestreo de ítems" en la habilidad estimada.
Fuente: elaboración propia.

La Figura 17, representa la situación en la que, el factor que parece incidir en la habilidad estimada para los estudiantes es el conjunto de ítems seleccionados, pues dada una muestra de sujetos (1), las habilidades pueden variar considerablemente en función de dicho factor (400 puntos en el caso de responder a la prueba A y 550 en el caso de responder a la prueba D). Tales diferencias, son observadas en todas las muestras analizadas (1, 2 y 3), observándose el mismo funcionamiento de manera independiente a la muestra de sujetos utilizada.

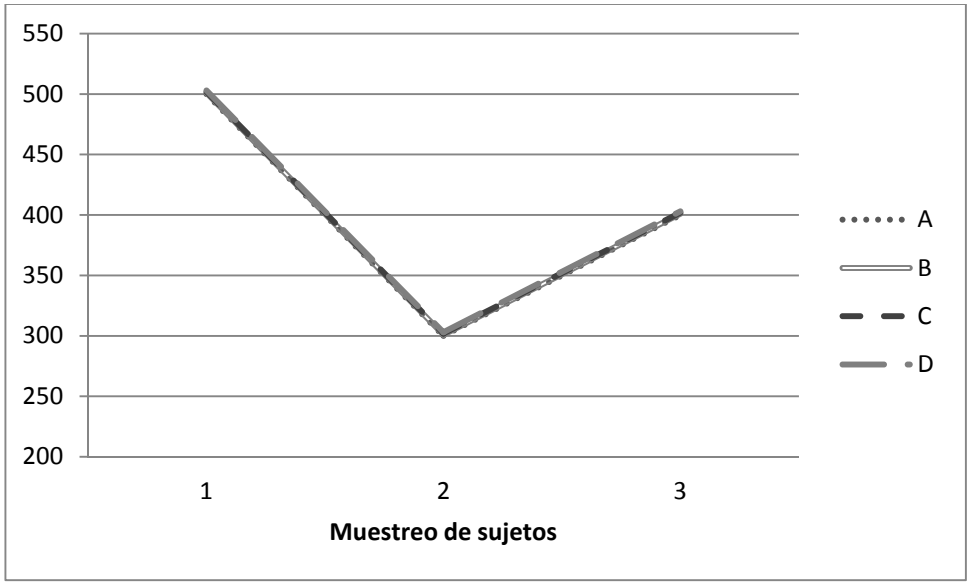


Figura 18. Efecto del factor "muestreo de sujetos" en la habilidad estimada.
Fuente: elaboración propia.

Sin embargo, en la Figura 18, el factor que produce diferencias es precisamente el relativo a la muestra de sujetos objeto de análisis, pues se observa que, independientemente de la prueba utilizada (A, B, C ó D), los resultados dentro de cada grupo son los mismos, sin embargo, las diferencias en nivel de habilidad entre los grupos son las más destacadas, siendo el factor A (selección muestral) el que está produciendo diferencias sustanciales en la estimación de habilidad.

En situaciones convencionales, la consideración por separado del error fruto del muestreo de sujetos y del muestreo de ítems, está dejando a un lado la posible influencia de un efecto de interacción, un efecto combinado producto de la acción conjunta de los dos factores principales considerados. Este efecto, puede repercutir de manera significativa en la habilidad estimada en los sujetos, alterando considerablemente los resultados del proceso de enlace de puntaciones y, en consecuencia, del proceso evaluador.

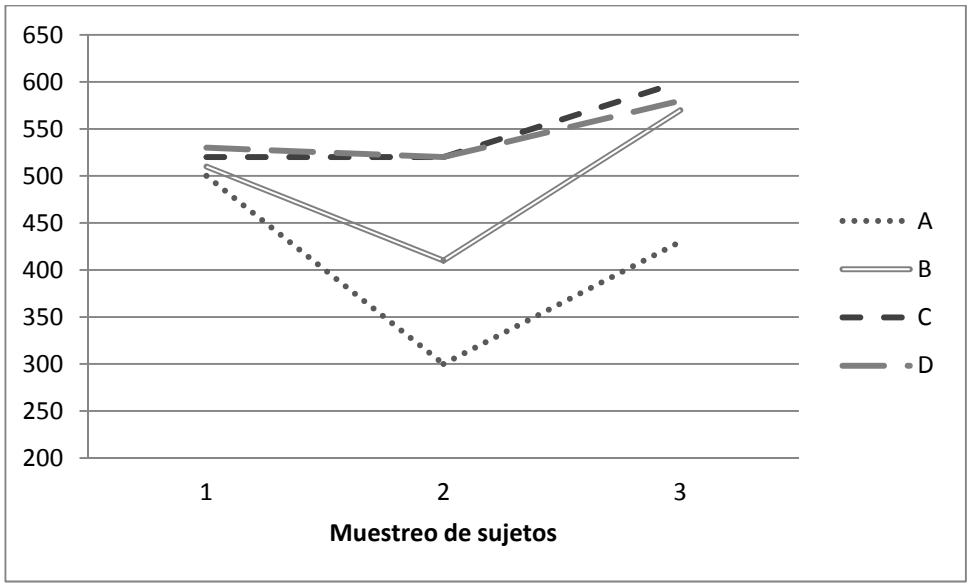


Figura 19. Efecto de interacción entre los factores A (muestreo de sujetos) y B (muestreo de ítems). Fuente: elaboración propia.

En el cuarto gráfico (Figura 19), se representa dicha interacción, en la representación observamos cómo existe un efecto producido por la selección de sujetos (diferencias observadas entre los grupos 1, 2 y 3), un efecto producido por el conjunto de ítems seleccionados (A, B, C ó D), pues dentro de un mismo grupo se aprecia una notable diferencia producida por la prueba utilizada, así como un efecto de interacción,

pues la combinación de la selección muestral (1, 2 ó 3) y el efecto de las pruebas utilizadas (A, B, C ó D), produce diferencias destacadas en los resultados. De este modo, la prueba A parece estar funcionando convenientemente en el grupo 1, sin embargo no parece estar produciendo los mismos resultados en los grupos 2 y 3.

En consecuencia, la estimación del error de enlace, deberá considerar el error asociado a la muestra de sujetos, el error asociado a los ítems utilizados, así como el efecto producido en la interacción de ambos factores. El marco general del problema sería el planteamiento y evaluación de una propuesta metodológica que permita la estimación de éste efecto de interacción, en concreto, el problema de investigación planteado podría precisarse en las siguientes cuestiones:

¿Existe efecto de interacción en la selección de sujetos y reactivos?¿Es posible mejorar los procesos de enlace de puntuaciones teniendo en cuenta el efecto de interacción en la selección de sujetos e ítems?¿Cómo estimar la interacción entre dichas fuentes de error?¿Resulta de utilidad el estudio de la interacción en el análisis de ítems comunes?¿En qué condiciones muestra su efectividad el procedimiento propuesto?

5.2 Objetivos de la investigación.

El objetivo central de este trabajo de investigación es la presentación y evaluación, mediante simulación experimental, del comportamiento y los efectos que pueden incidir en el procedimiento para la medida del efecto de interacción (sujetos/ reactivos) propuesto, al que hemos denominado *bootstrap bidimensional*, de forma que se pueda analizar tanto su funcionamiento global como bajo qué condiciones puede rendir mejor el mismo, estudiando los factores que podrían incidir en su articulación y su capacidad de estimación.

De forma más específica, el trabajo de investigación que presentamos se centra en la concreción de un procedimiento de «bootstrap bidimensional» para la estimación del efecto que la selección muestral de sujetos, la selección de ítems y su interacción tienen sobre la estimación del nivel de habilidad del sujeto. La posibilidad de realizar

una estimación sobre el efecto individual y combinado de dichos factores, es un aporte de interés para la práctica evaluativa actual. La utilización de una simulación experimental nos permitirá, tanto la presentación detallada del procedimiento planteado, como el análisis de las propiedades estadísticas que subyacen al mismo, a través del análisis de distintos factores que podrían estar interrelacionados con el mismo.

Los objetivos en los que podríamos concretar esta investigación serían:

1. Proponer un procedimiento para el análisis del efecto de interacción de la selección de sujetos y reactivos en procesos de enlace «bootstrap bidimensional».
    a. Justificación teórica.
    b. Diseño de una propuesta eficiente de implementación.
    c. Elaboración de sintaxis de análisis para su puesta en práctica.
2. Analizar las propiedades estadísticas globales de dicho procedimiento.
3. Comprobar el comportamiento del procedimiento ante diferentes condiciones experimentales.
    a. Funcionamiento Diferencial del Ítem.
    b. Diferencias en nivel de habilidad en los dos grupos a comparar.
    c. Variación en la distribución de los valores de *b* (dificultad) de los ítems que componen la prueba.
    d. Desplazamiento de los valores medios del parámetro *b* (dificultad) de los ítems incrementando su dificultad.

5.3 Método.

Tal y como se ha venido adelantando en apartados precedentes, el principal rasgo que caracteriza la metodología del presente trabajo es la aplicación de la técnica intensiva de remuestreo (bootstrap) para la extracción de muestras pseudo-aleatorias de sujetos e ítems que permiten la posterior estimación del error de enlace. Tal y como apuntan Harris y Crouse (1993), existen importantes dificultades de implementación de las técnicas intensivas de remuestreo. En la misma línea, Kolen y Brennan (2014) afirman que, la combinación de técnicas intensivas de remuestreo en el ámbito de la TRI, complica enormemente su uso. Algunos autores como Haberman, Lee y Qian (2009), con el propósito de salvar estas dificultades en los procesos de cálculo, han utilizado procedimientos como el Jackknife Agrupado «*Grouped jackknifing*» con el fin de realizar técnicas de remuestreo en ítems y sujetos.

A pesar de la complejidad que añade la utilización simultánea del remuestreo en la matriz de ítems y sujetos, dicha técnica permite la obtención de una estimación que combina ambas fuentes de error, de forma independiente (filas= efecto del muestreo de ítems/ columnas= efecto del muestreo de sujetos) y conjunta (efecto combinado de la selección de reactivos y sujetos), utilizar procedimientos de estimación independiente, limitará de forma notable la información obtenida en la aplicación, pues tal y como se ha argumentado previamente, la interacción puede constituirse como un elemento clave en el cálculo del error.

La aplicación de dichas técnicas implica el cálculo de la puntuación en rendimiento utilizando sub-muestras en las que se extrae una muestra de ítems y de sujetos en cada caso. En un segundo paso, partiendo de la variabilidad entre estas estimaciones, es calculado el error de enlace. En el caso que nos ocupa, nuestro remuestreo es doble, por tanto, deberán estimarse las puntuaciones en rendimiento utilizando sub-muestras en las que se combine un conjunto de ítems y sujetos al mismo tiempo.

A fin de evaluar la adecuación del procedimiento de cálculo propuesto, así como los posibles factores asociados, se ha diseñado un experimento de simulación Monte Carlo, en el que se han definido determinadas constantes (variables independientes,

inputs del modelo) y una variable dependiente (habilidad, θ), simulando varias condiciones experimentales con el objetivo de estudiar la precisión de las estimaciones así como los factores asociados.

En la simulación de tipo Monte Carlo, en cada condición experimental hay más de una muestra, situación que da cabida a la existencia de error muestral, constituyendo estudios que se consideran equivalentes a los estudios experimentales clásicos, siendo la unidad experimental la muestra (Castro Morera, 1997). De este modo, en la presente investigación nos encontramos ante un estudio experimental clásico, en el que observaremos la influencia de las variables independientes seleccionadas por su mayor vinculación con el efecto de interacción, sobre la variable dependiente objeto de interés (error cuadrático medio en la estimación de θ). La generación de los datos, así como los cálculos posteriores, se ha llevado a cabo por medio del programa R, en su versión 3.12 (R Core Team, 2013).

En el presente apartado, pasaremos a describir en primer lugar el proceso de generación de datos, la selección de variables y las hipótesis (5.2.1), para detallar posteriormente las características principales del procedimiento propuesto «bootstrap bidimensional» (5.2.2), así como la potencialidad del mismo, al permitir la consideración del error producido por el muestreo de sujetos y de ítems y lo que es más importante, la interacción entre ambos factores (5.2.3).

5.3.1 Generación de datos, variables e hipótesis.

Los estudios de simulación, pretenden reproducir características y comportamientos reales y, al mismo tiempo, ejercer un control sobre factores que pueden resultar objeto de interés, gracias a la determinación a priori de la situación a analizar. La libertad en el diseño que caracteriza este tipo de estudios, lleva consigo un importante riesgo, el derivado de que las condiciones del estudio de simulación sean poco realistas y, en consecuencia, se produzca una limitación en la generalización de sus resultados (validez externa) (Revuelta & Ponsoda, 2003), tales aspectos, han sido tenidos en cuenta a la hora de definir el entorno de simulación del presente estudio.

Primeramente, con el fin de evitar el efecto de la varianza muestral, se utilizará un tamaño muestral de 2000 sujetos en cada una de las condiciones experimentales, este tamaño muestral ha sido seleccionado por su ajuste con tamaños habitualmente utilizados en Evaluación Educativa. La muestra, compuesta por 2000 estudiantes, ha sido extraída de una población que se ajusta a una distribución normal con media 0 y desviación típica 1. El tamaño de la prueba, en las 4 condiciones experimentales planteadas, será de 50 ítems, estos 50 ítems se consideran una selección muestral del universo de ítems posibles.

El empleo de un estudio de simulación, dentro de la presente investigación, queda plenamente justificado puesto qué, lejos de buscar el análisis de un fenómeno educativo concreto, el trabajo aquí presentado pretende realizar un análisis pormenorizado de una propuesta metodológica determinada, analizando su posibilidad de implementación, sus ventajas, limitaciones, efectos asociados, implicaciones prácticas, etc. (Castro Morera, 1997). El algoritmo general de simulación se dividiría en los siguientes apartados principales:

A) Selección muestral de sujetos (N=2000) y de reactivos (*i*=50). Los valores de theta de la población de la que se extrae la muestra de sujetos siguen una distribución normal con media 0 y desviación típica 1. En relación a los valores de b de la población de ítems de la que se seleccionan los 50 reactivos que formarán parte de la prueba, tiene una media de 0 y una desviación típica de 0.5 para las condiciones 1, 2 y 4 (Anexos 1, 2 y 4) y de 0,10 para la condición 3 (ver Anexo 3).

B) Generación del patrón de respuestas. La estimación de las puntuaciones de los estudiantes se realiza siguiendo el modelo de Rasch (1960), en el que se considera que, la puntuación de un sujeto en determinado instrumento de medida (suma total de las respuestas del sujeto al conjunto de ítems), y la puntuación en un ítem (suma de las respuestas dadas por los sujetos a un ítem), permiten la estimación de los parámetros del modelo. En definitiva, la probabilidad de respuesta correcta por parte de un sujeto a determinado reactivo, depende del rasgo estimado y de la dificultad del ítem, en el caso de ítems dicotómicos la probabilidad de acertar un determinado ítem sería la siguiente:

$$P_i(\theta) = \frac{exp^{D(\theta - b_i)}}{1 + exp^{D(\theta - b_i)}} \quad (55)$$

Teniendo en cuenta los valores de theta y b provenientes del paso 1, se genera la matriz de respuestas (1/0).

En la simulación de datos planteada contamos con una muestra de 2000 sujetos (N_x=1000 y N_y=1000), con puntuaciones en 50 reactivos, En las dos primeras condiciones experimentales, utilizaremos el grupo X como grupo de referencia, haciendo variar las condiciones para los sujetos del grupo Y. En las dos condiciones restantes variarán las condiciones para la muestra global.

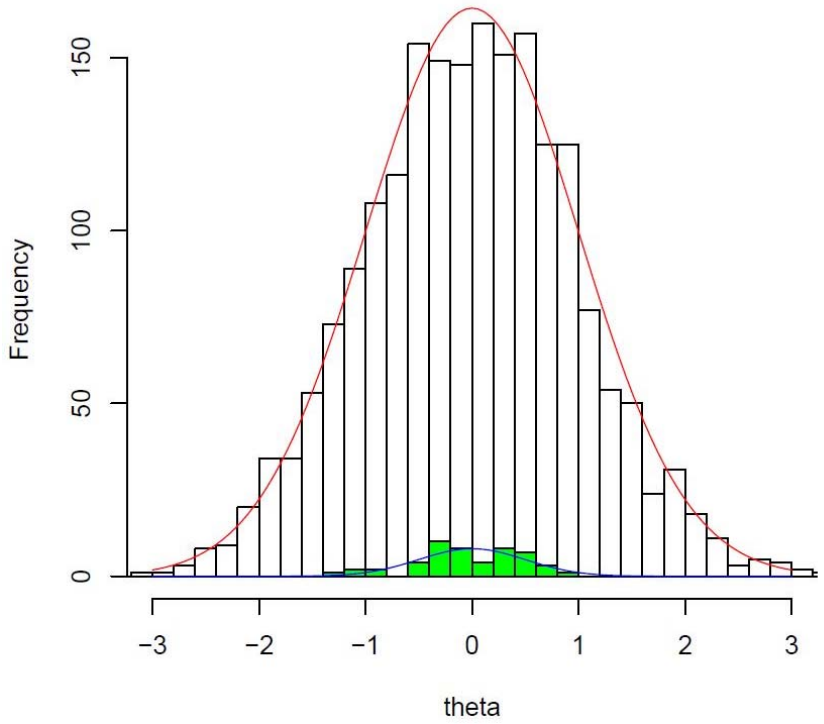


Figura 20. Distribución de Thetas y b´s en la iteración 0. Condiciones experimentales 1, 2 y 4.
Fuente: elaboración propia.

En la Figura 20 puede observarse el ajuste a la distribución normal de habilidades y parámetros b en los datos simulados para la situación de partida (iter= 0) en la simulación de datos para las condiciones experimentales 1, 2 y 4, en las que la desviación típica de la distribución del parámetro b es de 0.5, esta iteración es la base utilizada en el inicio de cada condición experimental. En la Figura 21, se muestra esta

misma distribución para la iteración 0 en el caso de la condición experimental 3, en la que el valor de la desviación típica para la distribución del parámetro *b* es de 0.1.

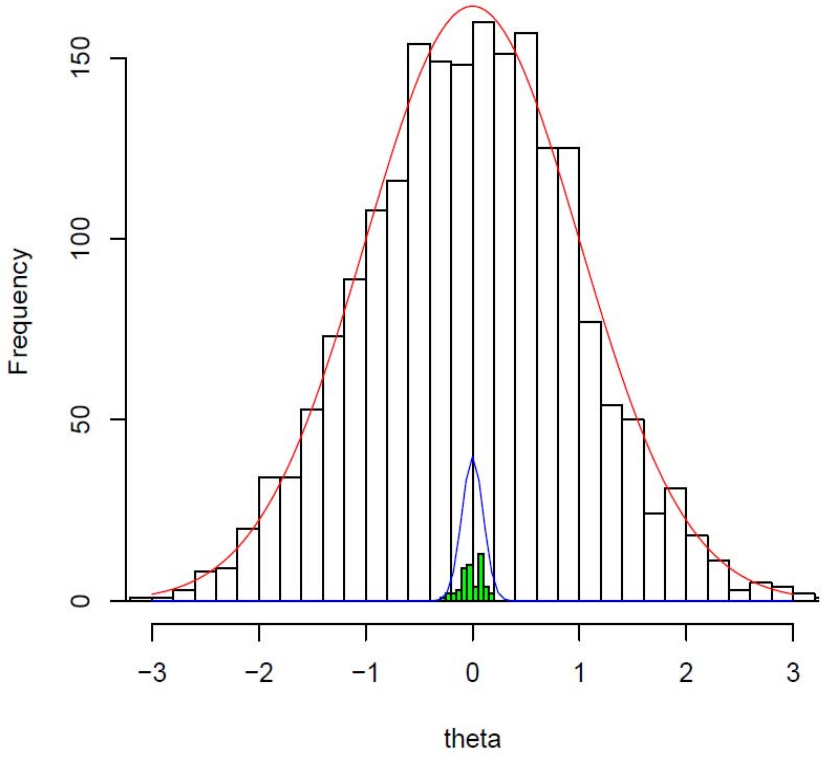


Figura 21. Distribución de Thetas y b´s en la iteración 0. Condición experimental 3.
Fuente: elaboración propia.

C) Partiendo de las 4 condiciones experimentales diseñadas en la presente investigación (cada una de ellas analiza una de las variables independientes/ experimentales objeto de estudio) se generan 10 iteraciones, a partir de la condición de partida (iter 0), situación que permite analizar el funcionamiento del procedimiento en las 4 condiciones experimentales contempladas en la presente investigación. El objetivo del apartado que nos ocupa, es caracterizar los rasgos generales de la simulación realizada, en consecuencia, es importante destacar la generación de éstas 10 réplicas en las que se manipula, de forma constante, la variable independiente objeto de interés, dando como resultado 11 archivos de datos por cada condición experimental.

En la presente investigación, las variables independientes están constituidas por las cuatro condiciones experimentales definidas, en definitiva cada condición experimental responde a las características de la muestra sobre la que se prueba la técnica propuesta. El efecto de las distintas condiciones experimentales se verá reflejado

en la estimación de theta y en consecuencia en el error cuadrático medio, es decir, la variable dependiente objeto de interés será la estimación del valor de theta bajo cada una de las condiciones experimentales.

Tabla 38.

Generación de réplicas en cada condición experimental

Condición Experimental	Descripción	Iteraciones
1	Funcionamiento diferencial del ítem N_x=1000 grupo de referencia N_y=1000	for (i in 1:10) Incr <- 0.1*i Valor de *b* 15 ítems con DIF aumenta 0,1 en cada iter para los sujetos del grupo Y (Anexo 1)
2	Diferencia en habilidad para los grupos cuya puntuación se desea estimar. N_x=1000 N_y=1000	for (i in 1:10) theta1 <- theta1 + IncrementoOriginal El valor de Theta para el grupo Y incrementa 0,1 en cada iter (Anexo 2)
3	Dificultad de los reactivos distribución con media 0 y desviación típica 0.1. Aumento de la desviación típica en cada nueva iter	for (i in 1:10) Incr <- 0.1*i b <- rnorm(50, 0, 0.1+Incr) (Anexo 3)
4	Distribución de los parámetros b de los ítems con media 0 y desviación típica de 0.5. Incremento en la media del valor de b 0.1 en cada iter.	for (i in 1:10) Incr <- 0.1*i b <- rnorm(50, 0+Incr, 0.5) (Anexo 4)

Fuente: elaboración propia.

La selección de estos cuatro factores que configuran cada una de las condiciones experimentales (Tabla 38), cumple con el objetivo de mostrar una primera aproximación a las condiciones que más pueden incidir en el funcionamiento del procedimiento propuesto por su mayor vinculación con el posible efecto de interacción estudiado, permitiendo una justificación y valoración completa y global del mismo, siendo conscientes de la no exhaustividad de la propuesta, puesto que sería posible definir innumerables condiciones. De este modo, entiéndase el presente estudio como una primera aproximación, en la que se trabaja con una selección de las cuatro condiciones experimentales consideradas más destacadas a fin de conseguir los objetivos propuestos.

D) Generación del patrón de respuestas atendiendo a las características de cada iteración. Tal y como sucedía en la fase precedente, la generación de dicho patrón de respuestas (1/0) se realiza siguiendo el modelo de Rasch (1960).

En consecuencia, como resultado del proceso de simulación, contamos con 11 archivos de datos para cada condición experimental (para cada variable experimental objeto de interés). El archivo denominado Iter_0, en cada condición experimental, es el archivo de origen, a partir del que se generan las distintas variaciones del factor analizado, cuyo incremento se ve representado en los datos generados para las restantes iteraciones (1 a 10). En dichos archivos de datos, el tamaño de la prueba será otra de las variables constante, habiendo diseñado una prueba compuesta por una selección muestral de 50 reactivos. Tal y como sucedía en la elección del tamaño muestral, el tamaño de la prueba es un tamaño frecuente utilizado en evaluación educativa, de este modo, la simulación estadística, se ajusta con suficiente fidelidad a condiciones reales de evaluación, evitando posibles sesgos en cuanto a la validez externa del estudio propuesto.

Las hipótesis que podremos contrastar de acuerdo a la simulación de datos planteada serán las siguientes:

1. En relación a las propiedades estadísticas globales del procedimiento propuesto:

    a. Que existe efecto de interacción entre los factores principales estudiados (sujetos e ítems), efecto que no es reconocido en los procedimientos de estimación frecuentemente utilizados.

    b. Que la interacción es una importante variable a considerar pues no puede estimarse a partir de la suma de los efectos de los factores principales por separado.

    c. Que la estimación del error atendiendo al efecto de interacción es más eficiente que la realizada sin considerar tal efecto, pues presenta una información mucho más completa y precisa.

2. En relación al funcionamiento diferencial del ítem:

a. Que el procedimiento permite valorar la violación de las asunciones estadísticas del modelo empleado.

b. Que el procedimiento propuesto permite detectar funcionamiento diferencial del ítem gracias a la consideración de la interacción.

c. Que el procedimiento propuesto es sensible a la cantidad de DIF.

d. Que el procedimiento propuesto permite trabajar sin la necesidad de elaborar hipótesis previas acerca del comportamiento de los datos.

3. En relación a las diferencias entre los grupos.

a. Que la estimación del efecto de interacción permitirá detectar posibles situaciones de diferencias en nivel de habilidad entre los grupos a equiparar, situación frecuente en los estudios longitudinales.

b. Que el procedimiento presentado es sensible al grado de diferenciación en nivel de habilidad entre los grupos a equiparar.

4. En relación a la distribución del parámetro *b* (dificultad) de los ítems:

a. Que el procedimiento resulta eficiente en condiciones de variación del potencial discriminador de la prueba en cuestión.

b. Que el procedimiento resulta eficiente en diversas situaciones prácticas y atendiendo a cierta variedad en las propiedades técnicas del instrumento utilizado.

c. Que el modelo propuesto muestra estabilidad en sus estimaciones a pesar de la variación en las propiedades específicas de la prueba utilizada.

En el apartado 5.2.3, titulado "Bootstrap bidimensional: efecto de la selección de sujetos, de ítems y su interacción", se exponen, justifican y analizan las condiciones experimentales estudiadas en el presente trabajo de investigación, detallando sus características fundamentales así como las razones que justifican su inclusión en el estudio, razones que se concretan en su aparente vinculación con el efecto de interacción estudiado. Por motivos didácticos, en el presente apartado se ha ofrecido un "avance" acerca de la naturaleza de dichas condiciones experimentales a fin de exponer las características del proceso de simulación, las variables independientes utilizadas y las hipótesis objeto de interés, haciendo notar que se trata de un experimento clásico con cuatro variables independientes sometidas a estudio (analizadas en cada condición experimental) y una variable dependiente (precisión en la estimación de θ). Las características detalladas del proceso de generación de datos pueden ser analizadas a partir de las sintaxis utilizadas para la simulación (Anexos 1 a 4) y cuya explicación será completada en los siguientes apartados.

5.3.2 Bootstrap bidimensional: presentación del procedimiento de remuestreo intensivo de sujetos e ítems.

El procedimiento bootstrap (Efron, 1979; Efron & Tibshirani, 1993), puede considerarse la vía de remuestreo más general, cuya aplicación se adapta de forma más precisa a las características específicas de la investigación en ciencias sociales. Tal y como veíamos en apartados precedentes, se trata del método de remuestreo más utilizado, debido posiblemente a la claridad con la que recoge los procedimientos que subyacen al mismo, siendo el procedimiento conceptualmente más sencillo (Efron, 1990). Por otro lado, el mayor desarrollo teórico-práctico frente a otros procedimientos, se presenta como otra de las razones de su mayor uso (López Jáuregui & Elosua Oliden, 2004). En este punto, cabe destacar que se trata, más que de una técnica o modelo específico, de un método general, a partir del que es posible aproximarse a diferentes objetivos de análisis (Ledesma, 2008). La utilización de los procedimientos Jackknife o BRR (descritos en el apartado 4.2) en el ámbito de las evaluaciones internacionales a gran escala, frente al uso del bootsraping, aún cuando este puede presentar grandes

ventajas, puede deberse a que el uso de estos métodos hace más eficiente el análisis en términos computacionales (Rust, 2014, p. 147) así como a la posibilidad de replicar hasta el último decimal, hecho de gran importancia para asegurar la confiabilidad de los resultados internacionales (A. Sandoval, comunicación personal, 9 de septiembre, 2015).

La característica esencial del bootstrap es la extracción de gran número de sub-muestras (pseudoaleatorias) con reemplazamiento, de la muestra global. En esencia, se basa en una analogía entre la muestra y la población de la que es extraída dicha muestra (Mooney & Duval, 1993). Cada una de las sub-muestras, fruto del «*resampling*», contará con el mismo valor de *n* que la muestra original, el reemplazamiento implica que cada remuestra pueda tener algunos de los puntos de la muestra original representados más de una vez y otros, no representados en ninguna ocasión, cada remuestra será ligera y aleatoriamente diferente a la muestra original (Mooney & Duval, 1993).

La tesis central del bootstraping es la idea de que, en ausencia de conocimiento acerca de la distribución poblacional, la distribución de valores hallada a partir de una muestra aleatoria de tamaño n de dicha población, es la mejor guía de la distribución poblacional (Manly, 1997). De este modo, siguiendo a Mooney y Duval (1993, pp. 10-11) podemos considerar los siguientes como pasos básicos del bootstrap:

a. Construir una distribución de probabilidad empírica $\hat{F}(x)$ a partir de la muestra, asignando una probabilidad de 1/n a cada punto $x_1, x_2, x_3, x_4 \dots x_n$. De este modo obtenemos la denominada Función de Distribución Empírica (FDE) de x, siendo el estimador no paramétrico de máxima verosimilitud de la función de distribución de la población $F(X)$.

b. A partir de FDE $\hat{F}(x)$, se extrae una muestra aleatoria simple de tamaño n con reposición, constituyendo una remuestra $x^*{}_b$.

c. Se calcula el estadístico de interés, $\hat{\theta}$, utilizando dicha remuestra $\hat{\theta}_b^*$.

d. Se repiten los pasos 2 y 3 B número de veces, siendo B un número grande. La magnitud de B depende de las pruebas que se desean realizar con los datos. En general, B debería ser de entre 50 a 200 para estimar el error típico de $\hat{\theta}$, y al menos de 1000 para estimar intervalos de confianza alrededor de $\hat{\theta}$.

e. Construir una distribución de probabilidad de los $\hat{\theta}_b^*$, asignando una probabilidad de 1/B a cada punto $\hat{\theta}_1^*, \hat{\theta}_2^*, \hat{\theta}_{3,}^* \hat{\theta}_4^* \ldots . \hat{\theta}_B^*$, Esta distribución es la estimación bootsrap de la distribución muestral de $\hat{\theta}$ $\hat{F}^*(\theta^*)$. Esta distribución puede ser utilizada para hacer inferencias sobre θ.

Partiendo de la idea esencial de éste procedimiento intensivo de remuestreo, el método propuesto en la presente investigación es el que hemos denominado bootstrap bidimensional o «*bidimensional bootstraping*». Esta táctica, se presenta como una vía útil que permitirá estimar el efecto del muestreo de sujetos y de ítems de forma independiente y combinada. La consideración de ambas fuentes de error (selección de sujetos/ selección de reactivos), se presenta como un elemento esencial en el ámbito de la Evaluación Educativa, tal y como hemos podido observar a lo largo de los apartados precedentes, haciéndose necesario contar con un procedimiento como el que describimos a continuación que, a pesar de su complejidad, permite realizar estimaciones de los efectos independientes y combinados. La consideración del efecto de interacción entre dichas fuentes de error, constituye la aportación clave de la técnica propuesta.

El procedimiento de bootstrap bidimensional, puede definirse como una técnica intensiva de remuestreo en la que, las unidades de remuestreo, son dobles (filas y columnas), en nuestro caso (sujetos y reactivos). A continuación se presenta un ejemplo con 10 sujetos y 5 reactivos con el fin de ilustrar la lógica del procedimiento. La matriz *h* recoge las puntuaciones de los sujetos pertenecientes a la muestra (N=10) en los ítems que componen la prueba (5) . En nuestro caso, la matriz de origen de los datos sería una matriz de 2000 * 50, atendiendo a las características de los datos descritos en el apartado de simulación, para mejorar la claridad expositiva, a la hora de presentar la lógica del procedimiento, utilizaremos una matriz de dimensiones 10 * 5 (Tabla 39).

Tabla 39.

Matriz h respuesta de 10 sujetos a 5 reactivos

Datos a muestrear	1	2	3	4	5
1	1	0	0	0	0
2	1	1	0	0	1
3	1	0	1	1	1
4	0	1	1	0	1
5	1	0	1	1	1
6	0	0	1	0	1
7	1	0	0	0	0
8	0	1	1	0	1
9	1	1	0	1	1
10	1	0	1	0	0

Fuente: elaboración propia.

A continuación, la matriz *J* (Tabla 40) recoge 10 muestras aleatorias de los sujetos pertenecientes a la muestra, de este modo, en la columna 1 vemos como la primera submuestra aleatoria está compuesta por los sujetos 1, 3, 10, 2, 9, 8, 4, 9, 6 y 6, la segunda muestra aleatoria quedaría formada por los sujetos que aparecen en la segunda columna.

Tabla 40.

Matriz J, bootstrap de sujetos

Matriz j										
	1	2	3	4	5	6	7	8	9	10
1	1	6	3	8	6	5	4	9	8	8
2	3	9	7	6	4	1	2	4	2	2
3	10	5	2	1	1	4	6	6	6	6
4	2	7	5	4	2	6	8	8	6	4
5	9	8	10	9	6	2	5	3	9	3
6	8	9	10	10	7	5	3	4	9	1
7	4	2	7	2	1	9	3	3	6	7
8	9	8	5	3	8	5	5	10	8	2
9	6	4	8	6	6	4	2	3	5	8
10	6	10	8	4	3	7	5	2	7	6

Fuente: elaboración propia.

Nótese que, al tratarse de un procedimiento de remuestreo con reemplazamiento, algunas unidades aparecen repetidas en determinadas submuestras mientras que, otras

unidades, no quedan representadas, en consecuencia si nos fijamos en la primera submuestra, vemos como los sujetos 5 y 7 no aparecen representados en dicha submuestra y, sin embargo, los sujetos 6 y 9 aparecen en dos ocasiones.

De forma análoga, la matriz *I* (Tabla 41), representa 5 muestras aleatorias de los 5 ítems que componen la prueba. La submuestra aleatoria 1 estaría compuesta por los reactivos 1, 3, 2, 1 y 5, en la columna 2 tendríamos los reactivos que componen la segunda submuestra y así sucesivamente. De manera equivalente a lo que sucede en la matriz de datos J, al tratarse de un muestreo con reemplazamiento, algunas unidades (ítems) pueden quedar representados varias veces en una misma submuestra y, otros reactivos, no aparecer representados. De este modo, si observamos la submuestra de ítems extraída en segundo lugar (ítems 3, 2, 4, 1 y 3) podemos ver cómo el ítem 3 aparece en dos ocasiones, mientras que el ítem 5 no aparece en dicha submuestra.

Tabla 41.

Matriz I bootstrap de ítems

Matriz i					
	1	2	3	4	5
1	1	3	2	4	3
2	3	2	5	4	4
3	2	4	4	3	2
4	1	1	2	2	2
5	5	3	1	1	1

Fuente: elaboración propia.

El rasgo que caracteriza la técnica propuesta en el presente trabajo, es la combinación de éstas dos matrices de datos, es decir la consideración del efecto conjunto de los dos factores principales analizados (efecto de la selección de sujetos/ efecto de la selección de reactivos), así como el efecto de su interacción. La tesis central de la técnica presentada es precisamente esta consideración, consideración que nos permitirá analizar la interacción entre ambos factores gracias al cruce del bootstrap sobre la matriz de sujetos y reactivos, representadas en las matrices *J* e *I*. De tal modo que, en nuestro ejemplo, el muestreo con j=1 e i=1 sería el representado en la Tabla 42.

Tabla 42.

Combinación de matrices I y J, muestreo j=1 e i=1

Muestra con j=1 e i =1					
	1	3	2	1	5
1	1	0	0	1	0
3	1	1	0	1	1
10	1	1	0	1	0
2	1	0	1	1	1
9	1	0	1	1	1
8	0	1	1	0	1
4	0	1	1	0	1
9	1	0	1	1	1
6	0	1	0	0	1
6	0	1	0	0	1

Fuente: elaboración propia.

Siguiendo este procedimiento, extraeríamos 10 submuestras de sujetos y 5 submuestras de reactivos, combinando cada submuestra de sujetos con cada submuestra de reactivos y por tanto, cada submuestra de reactivos con cada submuestra de sujetos, en total se producirían 500 replicas del proceso. En definitiva se trataría de combinar j=1 con i=1, 2, 3, 4, 5, j=2 con i=1, 2, 3, 4, 5, etc.

En la investigación que presentamos, tenemos 50 reactivos y 2000 sujetos, utilizando el programa R en su versión 3.12 (R Core Team, 2013), hemos procedido a la extracción de las matrices *J* (2000 submuestas aleatorias de sujetos) e *I* (50 submuestras aleatorias de ítems), la combinación de dichas matrices, ha dado lugar a 100.000 combinaciones (para cada una de las 11 iteraciones) en cada condición experimental (4 condiciones experimentales) lo que implica un total de 4.400.000 combinaciones.

En la Tabla 43, aparece el resultado de la combinación de ambas matrices en distintos pasos sucesivos. En el primer apartado, podemos ver la generación de las matrices *J* (2000*2000) (bootstrap de sujetos) e *I* (50*50) (bootstrap de ítems). En segundo lugar, vemos la unificación de dichas matrices en la matriz que combina el procedimiento de bootstrap doble, siendo el paso clave que permite la consideración del efecto de interacción de los dos factores principales. En tercer lugar, podemos observar un panorama global de las réplicas utilizadas en la presente investigación, en la que contamos con 4 condiciones experimentales compuestas por 11 iteraciones, en las que

se produce una manipulación creciente del factor (variable independiente) objeto de interés. Como resultado de estas sucesivas fases, contamos con 4.4000.000 submuestras, tal y como puede apreciarse en la parte inferior de la Tabla 43.

Por otro lado, si analizamos la información contenida en la Tabla 43, nos damos cuenta de que el proceso de generación de datos, así como la puesta en marcha del procedimiento *bootstrap bidimensional* en las 11 bases de datos que configuran cada condición experimental, es un elemento que añade un elevado nivel de complejidad en la tesis presentada, pues los procesos de estimación de puntuaciones suelen demandar un tiempo elevado de procesamiento de datos. Téngase en cuenta que, en cada una de las submuestras extraídas (4.400.000), se han de estimar las puntuaciones theta de cada sujeto (n=2000) a partir de su respuesta a los ítems que componen la prueba (50) conforme al modelo de Rasch. No obstante, a pesar de la complejidad del trabajo aquí presentado, es preciso destacar que, en condiciones reales de evaluación, en las que contamos con una sola muestra de sujetos e ítems a partir de las que extraer las submuestras necesarias por medio del procedimiento *bootstrap bidimensional*, el procedimiento propuesto se presenta como una alternativa útil y realista, pues utilizando la sintaxis elaborada para su puesta en práctica, podremos obtener una información de gran valor al permitir tener en cuenta el efecto de interacción entre las principales fuentes de error de enlace (selección de sujetos/ selección de ítems), pudiendo detectar, sin la necesidad de elaborar hipótesis previas, problemas de Funcionamiento Diferencial del Ítem, diferencias en habilidad entre los grupos a equiparar, posibles problemas de dimensionalidad de las pruebas, efectos de posición, etc.

Tabla 43.
Combinación de matriz de remuestreo de sujetos e ítems para la realización del «bootstrap bidimensional»

1

Proceso de bootsrap matriz de sujetos.
Resultado= Matriz J de 2000*2000 que contiene 2000 submuestras aleatorias de sujetos.

	1	2	3	…	2000
1					
2					
3					
4					
5					
6					
…2000					

Proceso bootstrap matriz de ítems.
Resultado=Matriz I. De 50*50 que contiene 50 submuestras aleatorias de reactivos.

	1	2	3	…	50
1					
2					
3					
4					
5					
…					
50					

2

Combinación del proceso de bootstrap sobre la matriz de sujetos y la matriz de ítems, dando lugar a 100.000 combinaciones, cada submuestra de 2000 sujetos con cada submuestra de 50 reactivos. Cada celda de la matriz contiene el cruce de una submuestra de 2000 sujetos y 50 ítems, dando lugar a 100.000 celdas.

		Submuestras ítems				
		i=1	i=2	i=3	…	i=50
Submuestras de sujetos	j=1	j=1 i=1	j=1 i=2	j=1 i=3	j=1 i=…	j=1 i=2000
	j=2	j=2 i=1	j=2 i=2	j=2 i=3	j=2 i=…	j=2 i=2000
	j=3	j=3 i=1	j=3 i=2	j=3 i=3	j=3 i=…	j=3 i=2000
	j=4	j=4 i=1	j=4 i=2	j=4 i=3	j=4 i=…	j=4 i=2000
	j=5	j=5 i=1	j=5 i=2	j=5 i=3	j=5 i=…	j=5 i=2000
	j=…	j=… i=1	j=… i=2	j=… i=3	j=… i=…	j=… i=...
	j=2000	j=2000 i=1	j=2000 i=2	j=2000 i=3	j=2000 i=…	j=2000 i=50

3

Este procedimiento se repite desde la iteración 0 a la iteración 10 dentro de cada una de las 4 condiciones experimentales. En cada condición tendremos
1.100.000 combinaciones de 2000 sujetos y 50 reactivos.
En consecuencia, la presente investigación contendrá 4.400.000 combinaciones de sujetos e ítems.

		Iteración											T
		0	1	2	3	4	5	6	7	8	9	10	
Condición Experimental	1	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000	1100000
	2	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000	1100000
	3	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000	1100000
	4	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000	1100000
													4400000

Fuente: elaboración propia.

El procedimiento general podríamos describirlo en las fases que aparecen a continuación:

1. Selección muestral de los datos de referencia: N=2000 Número de ítems 50, población de sujetos con una distribución normal de puntuaciones Theta (0, 1), distribución de parámetros *b* de los ítems (0, 05) y (0, 0,1) (condición experimental 3).

2. Generación de las distintas condiciones experimentales, partiendo de la condición de partida, se generan 10 iteraciones adicionales, en las que la variable experimental va manipulándose de forma creciente replica a replica.

3. Extracción mediante bootstrap de una muestra aleatoria con reemplazamiento n_{xy} de la muestra N_{xy}

4. Extracción mediante bootstrap de una muestra aleatoria de ítems Z_1 con reemplazamiento de la muestra de ítems Z.

5. Estimación de la puntuación de los sujetos, a partir de los parámetros de los ítems Z_1 en las submuestras de sujetos.

6. Repetir los pasos de 0 a 3 *R* número de veces. obteniendo $\hat{e}_{y_1}(x_i)$, $\hat{e}_{y_2}(x_i)$... $\hat{e}_{y_R}(x_i)$ para cada condición experimental.

7. Análisis de varianza para analizar el efecto del muestreo de sujetos, del muestre de ítems y de su interacción.

8. Estimación del error cuadrático medio (Zu & Liu, 2010):

$$SEE(x) = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left[\hat{e}_i(x) - \bar{\hat{e}}(x)\right]^2}. \quad (56)$$

La raíz del error cuadrático medio vendría dada por:

$$RMSE(x) = \sqrt{\frac{1}{M}\sum_{i=1}^{M}[\hat{e}_i(x) - e(x)]^2}, \quad (57)$$

Donde la correspondiente media ponderada vendría dada por:

$$\sqrt{\sum_x r(x)RMSE^2\,(x)} \quad (58)$$

Lo que indica que:

$$RMSE(X) = \sqrt{[Bias(x)]^2 + [SEE(x)]^2} \quad (59)$$

Habida cuenta de la dificultad inherente al diseño presentado, y con el objetivo de ilustrar de manera apropiada el mismo, tras tratar el procedimiento de generación de datos, la selección de variables e hipótesis (Apartado 5.3.1) y presentar de forma detallada el procedimiento de bootstrap bidimensional propuesto (Apartado 5.3.2), pasamos a detenernos en la utilidad del mismo para la medida del efecto de interacción en la selección de sujetos e ítems (5.3.3).

5.3.3 Bootstrap bidimensional: efecto de la selección de sujetos, de ítems y su interacción.

Tal y como se ha expuesto en los apartados precedentes del presente capítulo, el procedimiento propuesto bootstrap bidimensional, permite estimar el error asociado a la selección de sujetos y de reactivos así como su efecto conjunto (interacción), a continuación pasamos a describir, desde un punto de vista teórico, la lógica que subyace al mismo, descripción que responde al primer objetivo de la presente investigación y que dará paso a la justificación de las condiciones experimentales seleccionadas en la misma.

En primer lugar, la Figura 22, muestra la Superficie Característica del test en una situación en la que se da una relación lineal entre el efecto de selección de sujetos y

reactivos, es decir, nos basta con conocer el efecto de los dos factores considerados sin detenernos a analizar su efecto conjunto pues nos encontraríamos ante una relación lineal y la forma que adquiriría la Superficie Característica, tal y como se puede apreciar en la Figura 22 sería la de un plano, ya que $P(\theta) = a + b\theta$, conocidos los valores de a y b, podríamos conocer sin dificultad el valor de probabilidad para cada sujeto, es decir, para conocer la altura en z (probabilidad de acertar correctamente el ítem) basta con sumar los efectos de x (dificultad del ítem) y de y (habilidad del sujeto). Por tanto, ésta sería la relación entre Theta, b y la probabilidad de acertar correctamente el ítem si ésta relación fuese lineal. Conociendo los valores de a y b podríamos situar de forma precisa la posición del sujeto en el plano representado (Figura 22).

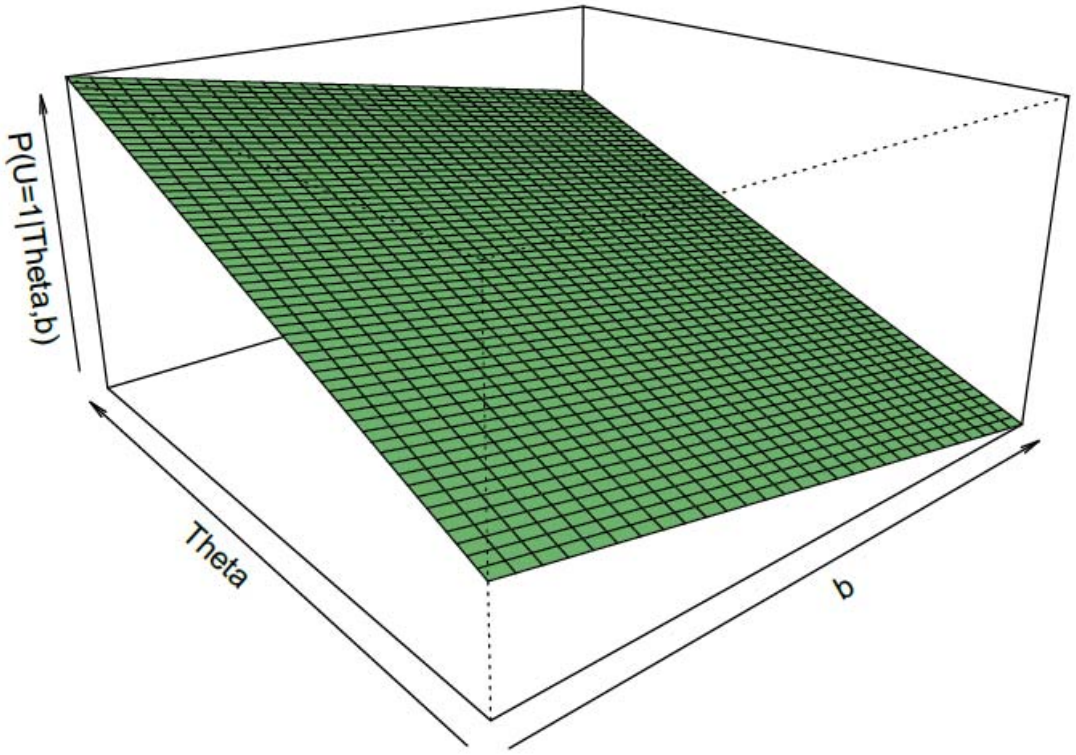


Figura 22. Superficie característica del test en la condición virtual de no interacción I
Fuente: Elaboración propia.

En el punto A (Figura 23), tendríamos a sujetos con el menor nivel de habilidad, respondiendo a los ítems más fáciles, la probabilidad de responder correctamente a éstos reactivos para los sujetos con dicho valor de theta sería inferior a 0.5. A medida que

aumenta el valor de dificultad de b, la probabilidad de dichos sujetos desciende hasta alcanzar el valor de 0 en el vértice opuesto (vértice B figura 23).

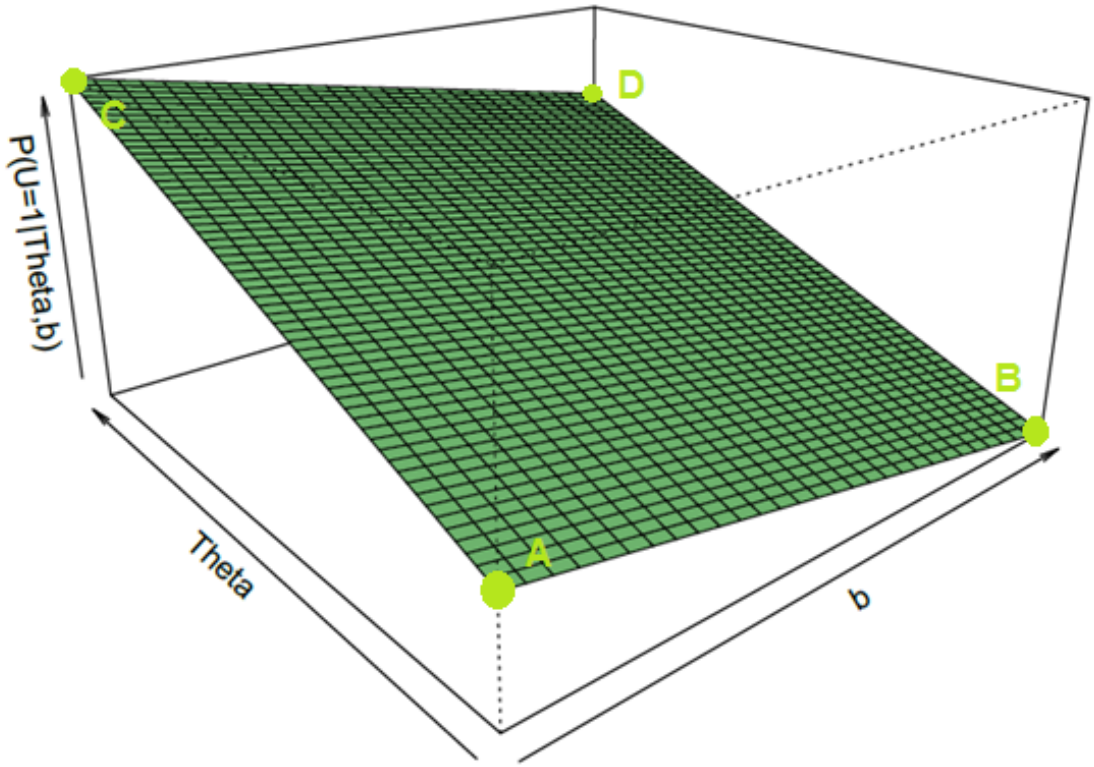


Figura 23. Superficie característica del test en la condición virtual de no interacción II
Fuente: Elaboración propia.

Del mismo modo, a medida que aumenta el nivel de habilidad del sujeto, la probabilidad de responder al ítem correctamente es mayor. Así, para el valor más alto de theta, la probabilidad de contestar correctamente a los ítems de menor nivel de dificultad, es 1 (punto C Figura 23). Si para este mismo grupo de sujetos, se produce un incremento progresivo en la dificultad de los ítems, la probabilidad de responder correctamente irá descendiendo desde el valor 1 hasta valores situados en torno al 0.67 (punto D). Si extraemos una muestra aleatoria de sujetos y reactivos, las diferencias observadas serán debidas a estos dos factores. La precisión de la medida es la misma para todo el rango de habilidades del sujeto.

Sin embargo, esta representación se aleja sustancialmente de la realidad, si representamos el funcionamiento real de la Superficie Característica del test, atendiendo a las condiciones reales de evaluación, la distribución adquiriría la forma representada en la Figura 24, gráfico en el que puede observarse cómo la suma de los efectos de b

(eje *x*) y Theta (eje *y*), no basta para conocer la altura en *z* (probabilidad de acertar el ítem correctamente), puesto que no se trata de una relación lineal.

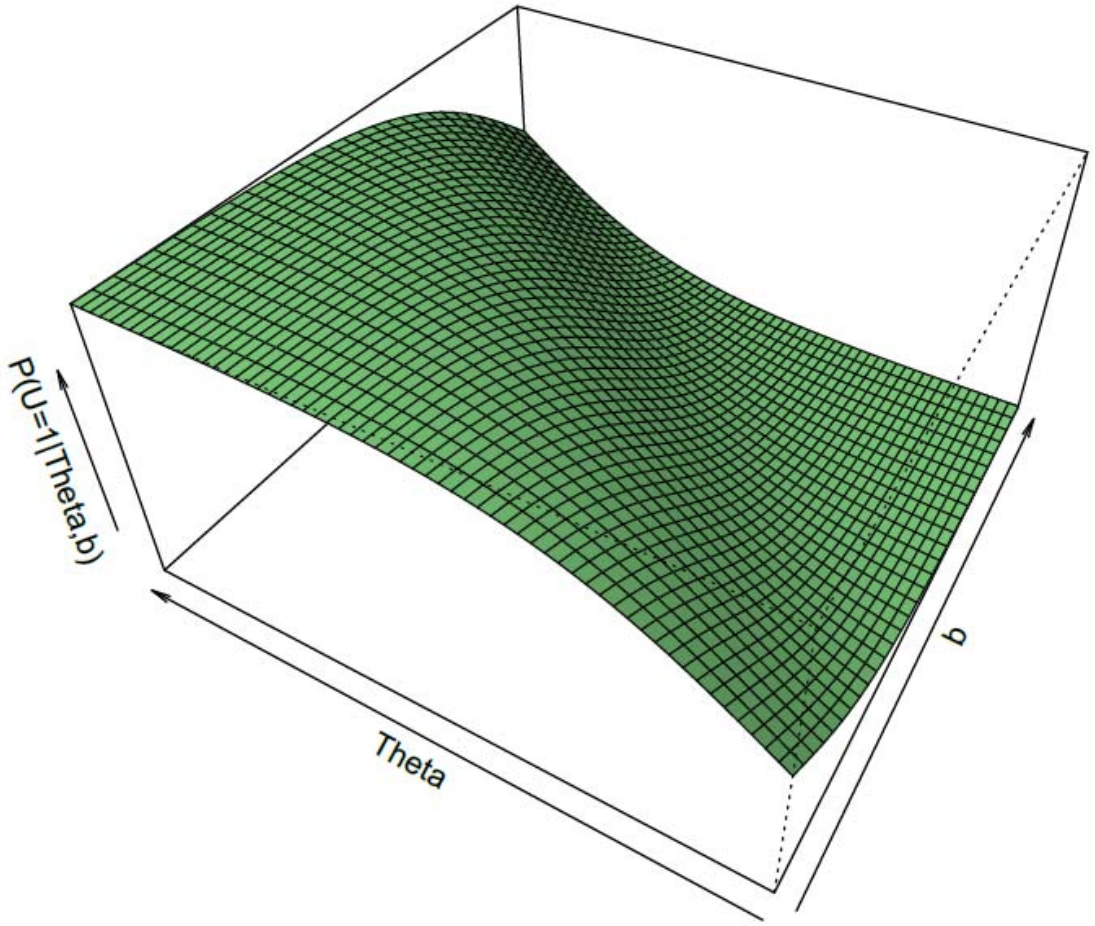


Figura 24. Superficie característica del test en condición de interacción I
Fuente: Elaboración propia.

Así, si seleccionamos un sujeto con el menor valor de theta (vértice A Figura 25), vemos como la probabilidad de acertar correctamente a los ítems de menor nivel de dificultad es inferior a 0.5, dicha probabilidad va disminuyendo progresivamente, hasta llegar un punto (punto "a" Figura 25) en el que el incremento en b no produce diferencias. Del mismo modo, si analizamos la situación de los sujetos con mayor nivel de habilidad vemos cómo, la probabilidad de responder correctamente al ítem, es igual a 1 en un amplio rango de valores de b, desde el vértice B (Figura 25) hasta el punto b en el que la probabilidad comienza a descender. Así, a pesar de producirse una situación similar en los extremos de la distribución, la realidad en el total de la distribución es muy diferente si consideramos la relación como lineal (sin efecto de interacción) o no lineal (valorando el efecto de interacción).

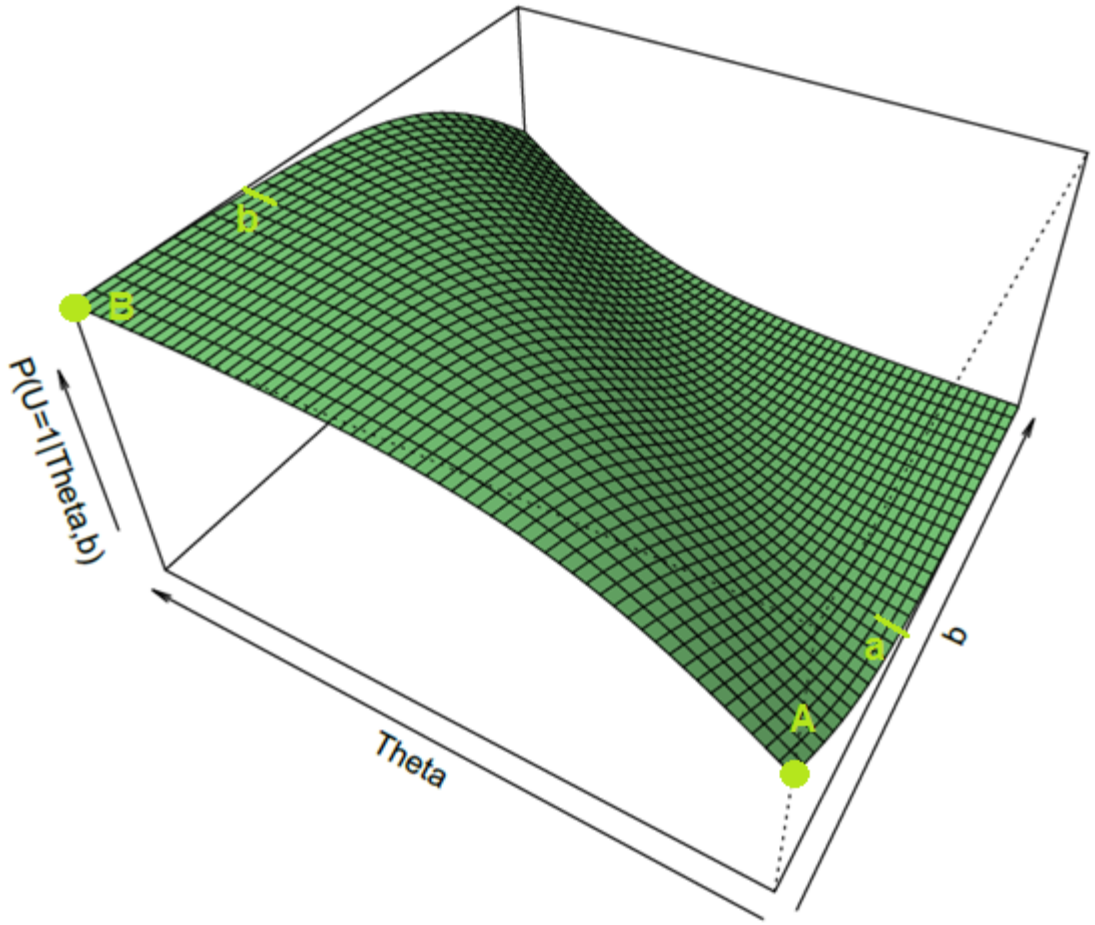


Figura 25. Superficie característica del test en condición de interacción II.
Fuente: Elaboración propia

En las Figuras 26 y 27, se representan los gráficos de las Figuras 22 y 24 pero con un cambio de perspectiva. En la figura 26 se aprecia como cuando no existe efecto de interacción, la precisión de la medida dependerá de la selección de sujetos y reactivos, así, dado un determinado valor en Theta, la altura en *z* (probabilidad de responder correctamente), sólo dependerá de b (Figura 26).

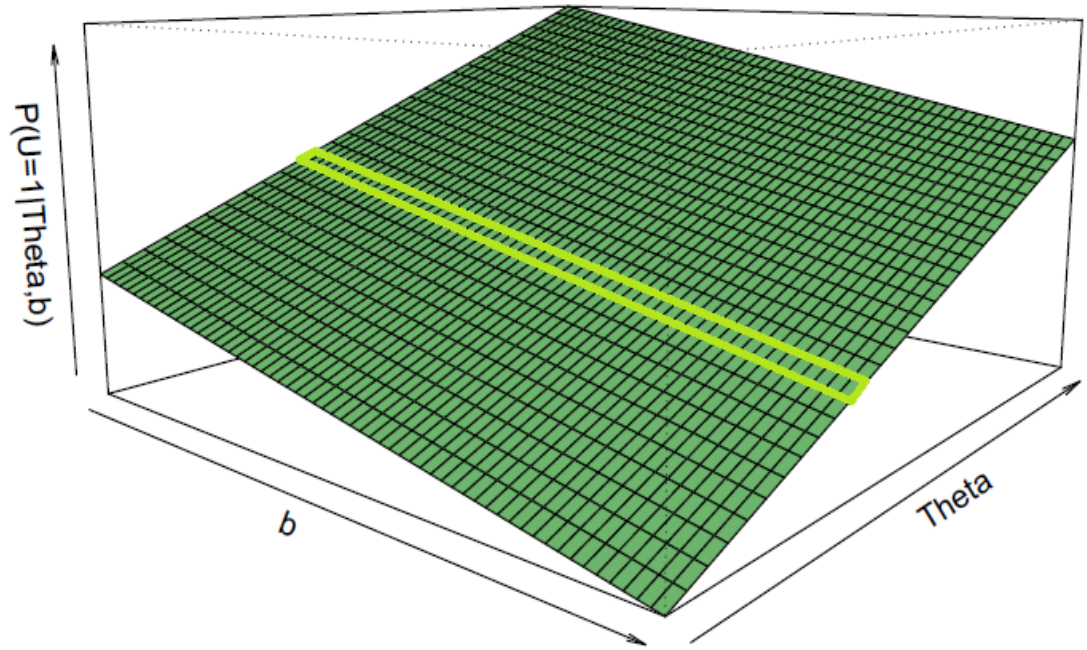


Figura 26. Superficie característica del test en la condición virtual de no interacción III
Fuente: Elaboración propia.

Sin embargo, en condiciones de interacción, dado un determinado valor de theta, la probabilidad de contestar correctamente el ítem (altura en *z*), no dependerá exclusivamente de b, pues existe un efecto de interacción no reconocido desde el punto de vista lineal (Figura 27). Así, el patrón de incremento en *z* que puede observarse en las líneas marcadas en la Figura 27, presenta una desviación considerable respecto a lo que se esperaría si se tratase de una relación lineal como la descrita con anterioridad, especialmente para determinados valores de b y theta (curvatura de las rectas trazadas entre los dos puntos), justamente si analizamos las diferencias entre las dos líneas representadas en la Figura 27, observaremos cómo la relación entre los valores de theta y b no es igual en toda la distribución, poniéndose de manifiesto la relación no lineal entre las variables y su efecto de interacción como causa de la diferenciación en las estimación de la probabilidad de acertar correctamente el ítem.

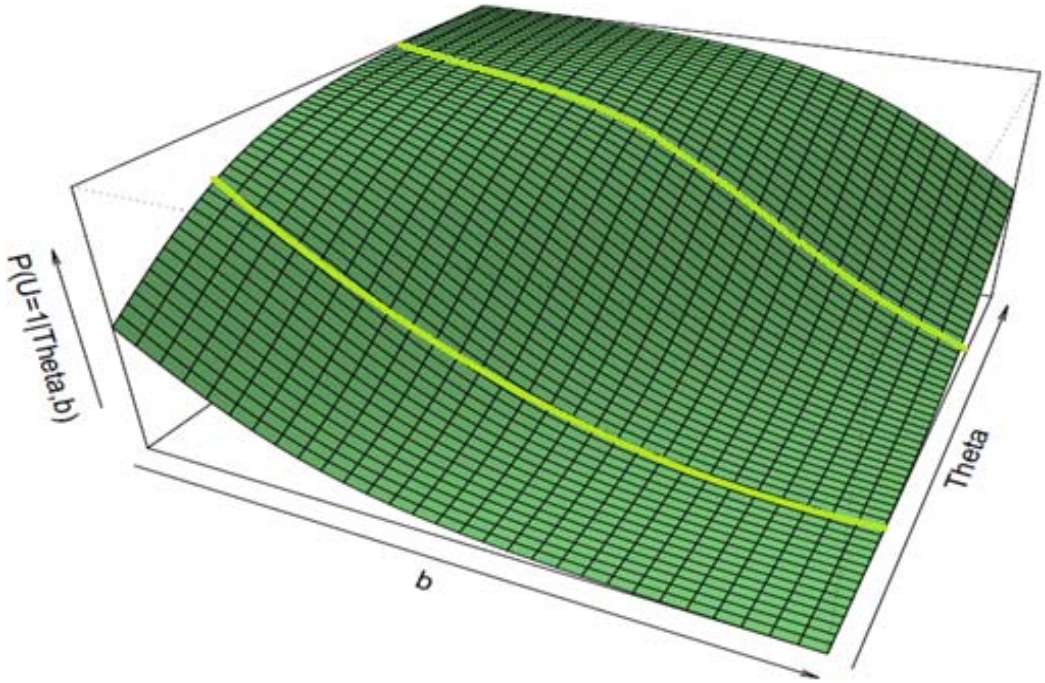


Figura 27. Superficie característica del test en condición de interacción III
Fuente: Elaboración propia.

En consecuencia, la precisión de las estimaciones en condiciones de interacción dependerá tanto de la habilidad de sujetos seleccionados (error muestral tradicional) como de los ítems seleccionados (error fruto del muestreo de ítems) así como del efecto de su interacción.

En las Figuras 28 y 29, han sido destacadas dos pequeñas áreas. En tales áreas (posiciones extremas), es donde están más próximos los valores de z teniendo en cuenta x e y en condiciones de interacción y de no interacción. De este modo, si extraemos una muestra de sujetos y reactivos que corresponda precisamente con dichas áreas, el efecto de interacción no estará presente, por la proximidad de las mismas en ambas representaciones. Sin embargo, si seleccionamos sujetos y reactivos de cualquier otro área de la distribución, observaremos como tales distancias se acentúan, pudiendo dar lugar a valores muy diferentes.

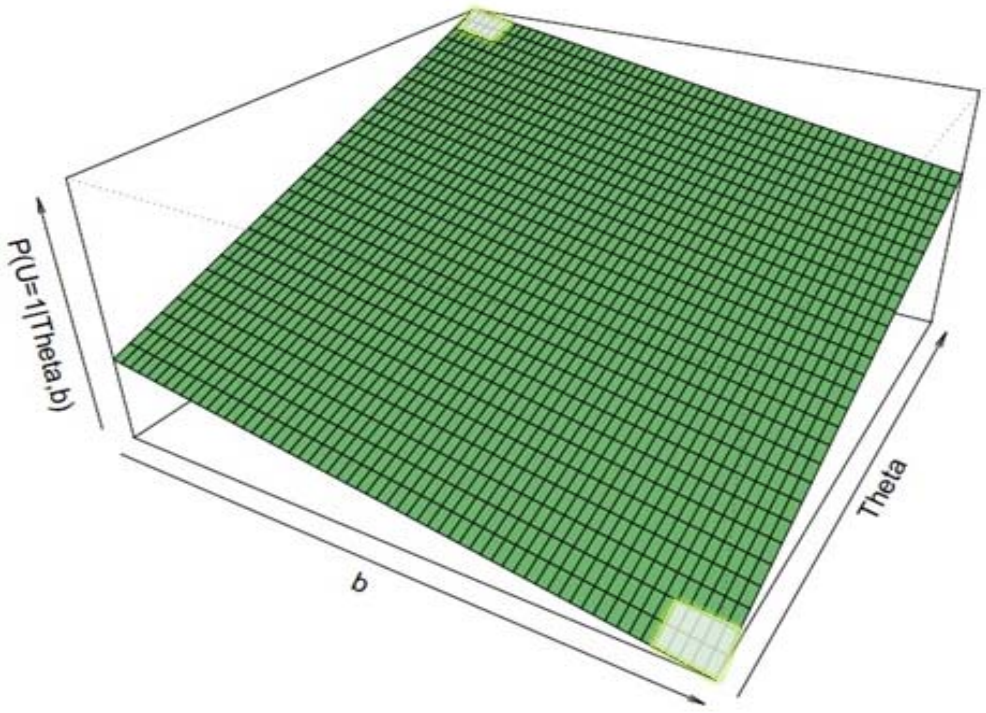


Figura 28. Superficie característica del test en la condición virtual de no interacción IV
Fuente: Elaboración propia.

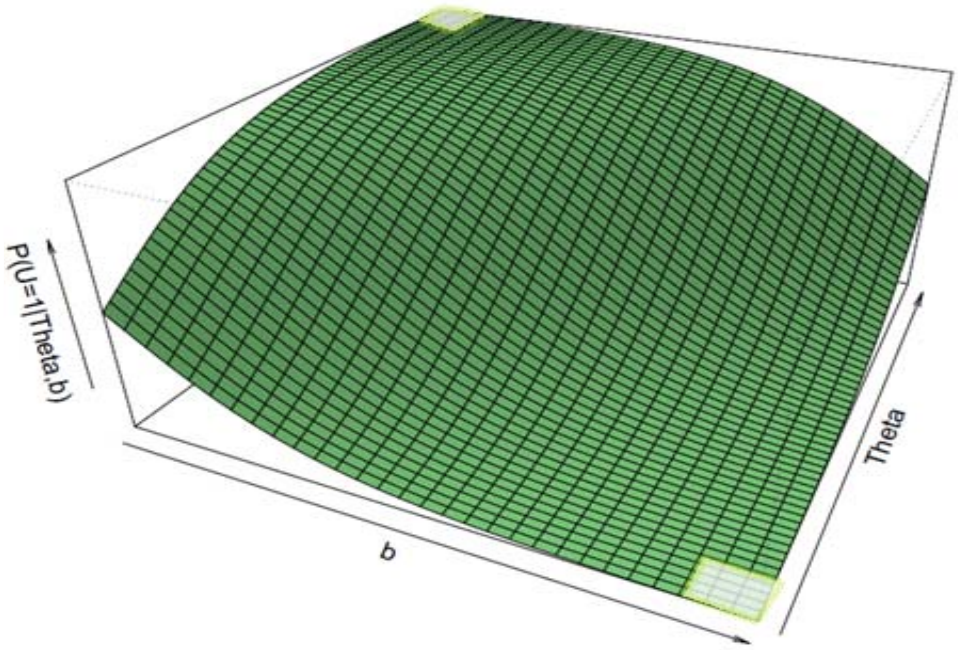


Figura 29. Superficie característica del test en condición de interacción IV
Fuente: Elaboración propia.

El gráfico presentado en la Figura 30, representa la distribución de valores de theta desde -3 a + 3 y la distribución de valores de b desde -3 a + 3. En dicha representación, observamos cómo, teniendo en cuenta la interacción entre la selección

de sujetos y reactivos, la superficie característica del test no es una superficie plana. De este modo, el error de estimación aumenta a medida que los valores de dificultad del ítem y de habilidad del sujeto se alejan, siendo la precisión de la medida óptima en el punto en el que la habilidad del sujeto y la dificultad del reactivo se aproximan.

Los sujetos situados en el vértice A de la representación, son sujetos cuyo nivel de habilidad (el más bajo de la distribución), y el nivel de dificultad del ítem (nivel más bajo de la prueba) coinciden. A medida que nos alejamos de este punto, la precisión de la medida decrece, de este modo, los ítems fáciles, no evalúan de forma adecuada a los sujetos más hábiles, y los ítems difíciles no evalúan con precisión a los sujetos cuya habilidad es menor. La diagonal formada por la distancia del vértice A a su vértice opuesto, representa el área en la que la precisión de la medida es mejor, al coincidir los niveles de habilidad y la dificultad de los ítems, atendiendo al efecto de selección de ítems, sujetos y su interacción. De este modo, si observamos el presente gráfico de forma invertida, el error de estimación alcanzaría los niveles más bajos en la diagonal que va desde el vértice A a su vértice opuesto, la escala térmica de los colores del gráfico, pretende representar precisamente esta situación, siendo las tonalidades verdes las que indican mayor precisión en la medida.

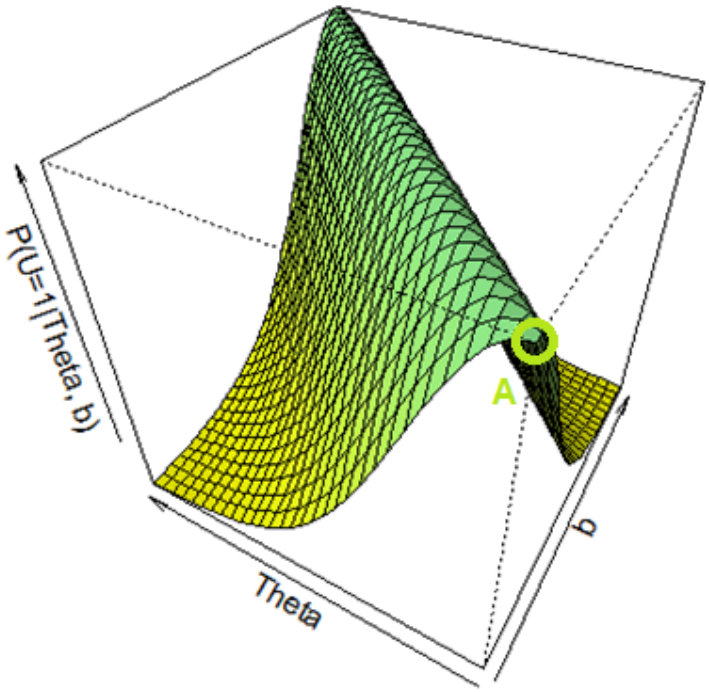


Figura 30. Superficie de información del test condición de interacción 1

Fuente: Elaboración propia.

En la representación gráfica de la Figura 31, observamos el efecto de la selección de sujetos, de reactivos y su interacción en una distribución de thetas de -3 a 3 y una distribución de valores b de -1 a 1. En este caso, la prueba no dispone de reactivos capaces de medir con precisión a los sujetos que se encuentran en los extremos de la distribución de thetas, de este modo, la prueba no dispone de reactivos óptimos para evaluar sujetos cuya habilidad es menor de -1 o mayor a 1, en consecuencia, la imprecisión de la medida se acentúa en ambos extremos de la distribución, obteniendo valores óptimos en las posiciones centrales del valor de theta. Si extraemos una muestra aleatoria de sujetos y reactivos, obtendremos un efecto debido a la selección de cada uno de los elementos así como un efecto conjunto fruto de la selección de ambos.

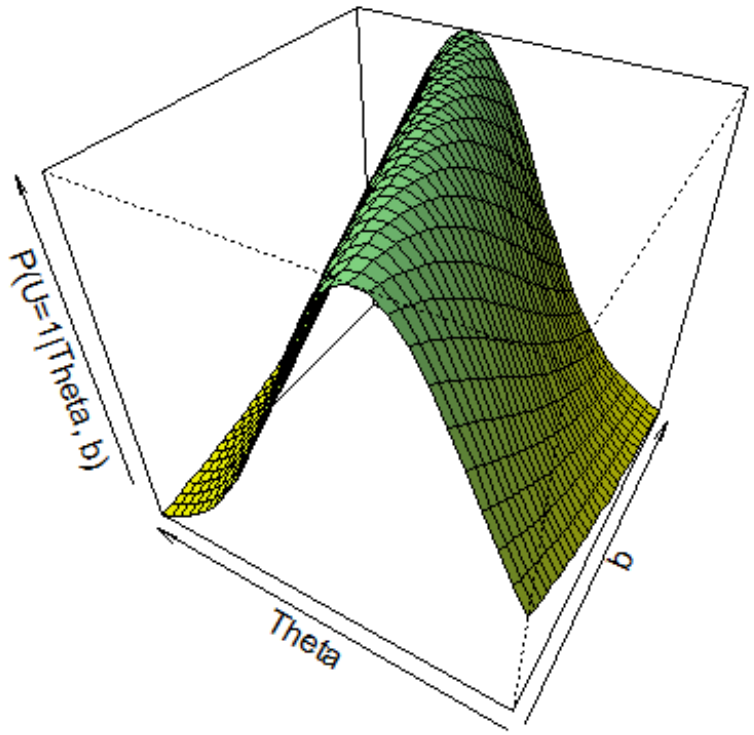


Figura 31. Superficie de información del test condición de interacción 2
Fuente: Elaboración propia.

Si representamos una distribución con los mismos valores de theta (-3 +3) pero utilizamos ítems cuyo nivel de dificultad está más próximo entre sí (0.005 a -0.005), obtendríamos una representación como la que aparece en la Figura 32. La Superficie de Información del Test, en este caso, muestra una estimación óptima para los sujetos cuyo

valor de theta se sitúa próximo a los valores de dificulta de los ítems de la prueba. Dado un valor de theta, la precisión de la medida es igual para todos los ítems seleccionados.

Los sujetos con una puntuación de theta situada en torno a 0 (valor medio de la distribución), son los que presentan una estimación más precisa, produciéndose en este punto el menor error de estimación. La prueba seleccionada no mide bien a sujetos situados en los extremos del rango de puntuación, pues no dispone de reactivos óptimos para estos grupos. De este modo, si extraemos de manera aleatoria un conjunto de sujetos y reactivos, la altura en *z* (probabilidad de responder correctamente al ítem), dependerá de forma prácticamente exclusiva de los valores de theta de los sujetos seleccionados y del efecto de interacción, siendo escaso el error debido al muestreo de ítems.

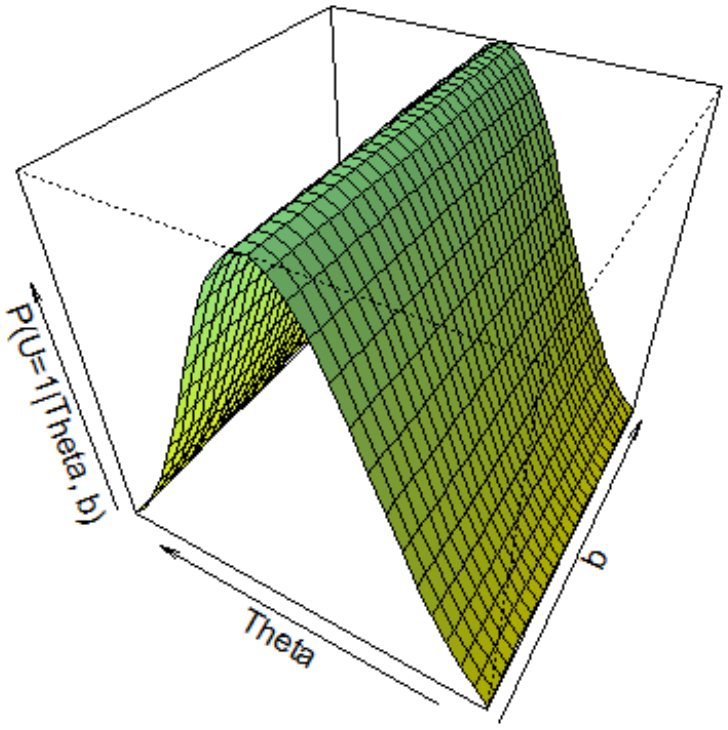


Figura 32. Superficie de información del test condición de interacción 3
Fuente: Elaboración propia.

Por el contrario, si presentamos una distribución con valores de theta más próximos entre sí (-1 + 1) y una distribución de dificultad de los ítems más amplia (-3 +3) obtendríamos una representación como la que aparece en el gráfico de la Figura 33. En este caso, la selección de reactivos con niveles medios de dificultad asegurará la

calidad de la medida. Si seleccionamos reactivos próximos a los extremos de la distribución de dificultad, obtendremos errores de estimación más elevados.

Nuestra muestra no cuenta con sujetos cuyo valores de Theta sean extremos, en consecuencia, nuestra prueba deberá construirse utilizando ítems con un rango de dificultad intermedio, situado en torno a los valores de theta de los sujetos pertenecientes a la muestra. Incluir ítems extremos, supondría una pérdida de precisión en la medida, al perder la posibilidad de incluir reactivos más acordes a nuestra distribución muestral. Con valores tan próximos de theta, el error fruto del muestreo de sujetos será menor en comparación con el error producido por el muestreo de reactivos y por la interacción.

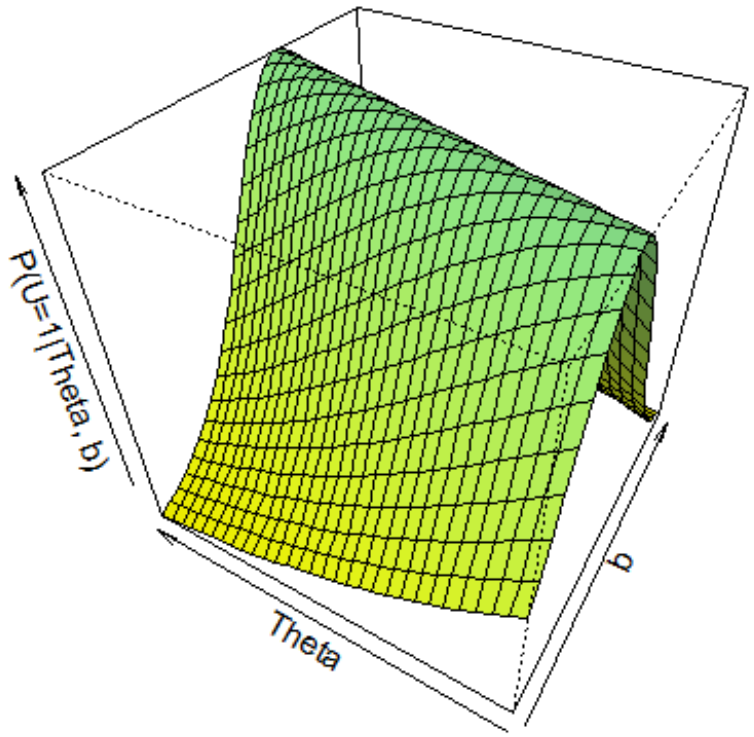


Figura 33. Superficie de información del test condición de interacción 4
Fuente: Elaboración propia.

Por último, si contamos con valores de Theta muy próximos entre sí (-0.05 a +0.05) y valores de b (-3 a 3), la precisión de la estimación dependerá prácticamente de forma total, de la selección de reactivos de nivel de dificultad medio. La selección de sujetos presentará una influencia menor en la calidad de la estimación, sin embargo, la selección de ítems será un determinante clave (ver Figura 34). Así, el error muestral

debido a la selección de sujetos será mínimo en comparación con el error fruto de la selección de reactivos.

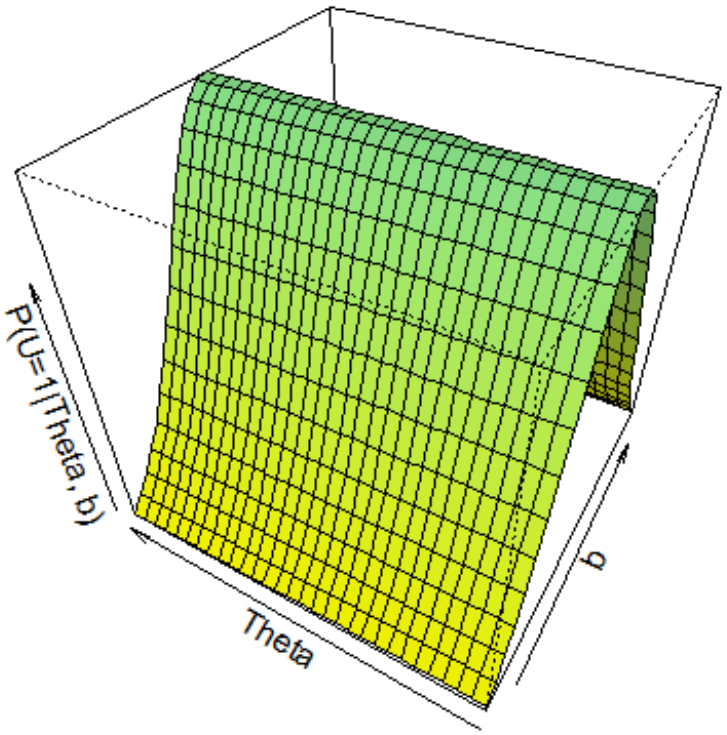


Figura 34. Superficie de información del test condición de interacción 5
Fuente: Elaboración propia.

Las condiciones experimentales seleccionadas, mostrarán los efectos destacados en el diseño, presentándose como una vía útil para contrastar las principales condiciones de funcionamiento del bootstrap bidimensional para la medida de la interacción así como los factores asociados.

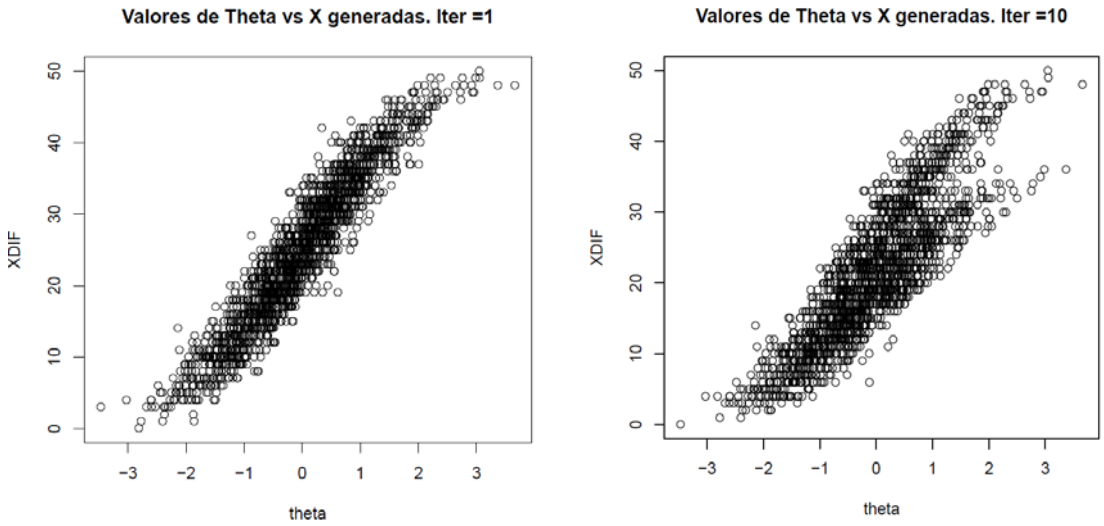
La selección de tales factores no responde a un criterio de exhaustividad sino, más bien, a uno de representatividad, habiéndose seleccionado los factores que, desde un punto de vista teórico, tienen mayor influencia potencial en el efecto de interacción objeto de estudio. Es decir, se han seleccionado aquellas condiciones que, desde un punto de vista teórico, y teniendo en cuenta lo descrito en el presente apartado, permitirán ilustrar de forma conveniente la técnica y, al mismo tiempo, analizar alguno de los factores que podrían incidir en su funcionamiento y propiedades estadísticas. Tal y como apuntábamos en el apartado de generación de datos, variables e hipótesis (5.2.1), cada condición experimental representa una variable independiente en el estudio, respondiendo a un diseño experimental clásico en el que se observa el efecto de

las variables independientes seleccionadas sobre la variable dependiente objeto de interés.

Condición experimental 1

En la primera condición experimental contamos con 50 ítems con un índice de dificultad distribuido aleatoriamente N (0,0.5) y una muestra de sujetos N= 2000 extraídos aleatoriamente de una población de sujetos con distribución N (0,1). La probabilidad de que cada sujeto conteste correctamente a cada ítem se calcula mediante el modelo de Rasch. En la Figura 20 se representa la distribución del nivel de habilidad de los sujetos y los parámetros de los ítems en la iteración 0. Partiendo de esta condición de partida, se generan 10 iteraciones adicionales, en las que se simula funcionamiento diferencial en 15 de los 50 reactivos para el grupo N_Y=1000. En cada una de las 10 iteraciones se aumenta 0.10 la dificultad de los 15 reactivos que presentan DIF respecto la iteración anterior, de este modo, la condición 0 sería la condición de partida, en la que no existe DIF, la iteración 1 sería la condición con menor grado de DIF y la condición 10 la de mayor nivel de DIF.

En las Figuras 35 y 36 puede observarse con claridad el distanciamiento de los dos grupos analizados, observándose una marcada tendencia en la diferenciación de las puntuaciones de ambos grupos entre las puntuaciones theta y el número de respuestas correctas si analizamos los datos relativos a la iteración 1 (menor cantidad de DIF) y la iteración 10 (mayor grado de DIF). El procedimiento planteado (bootstrap bidimensional) será aplicado a cada una de las muestras generadas, permitiendo contrastar las hipótesis planteadas en relación al funcionamiento diferencial del ítem.



Figuras 35 y 36. Comparación valores de theta versus X generadas en las iteraciones 1 y 10.

Fuente: elaboración propia.

De este modo, la primera variable independiente utilizada en la presente investigación será el Funcionamiento Diferencial del Ítem, ya que desde un punto de vista teórico las implicaciones del DIF en el efecto de interacción en la selección de sujetos y reactivos pueden ser notables.

Los gráficos que se representan en la Figura 37 muestran el grado creciente de DIF en cada una de las condiciones experimentales, desde la Iter_1 a la Iter_10, representación en la que se visualiza el progresivo distanciamiento de los valores de b originales y los valores de b generados en las diferentes iteraciones, en las que se incrementa 0.10 el valor de b para los 15 ítems que presentan DIF en la mitad de la muestra. Así, cada una de estas iteraciones cumple su función dentro del experimento clástico diseñado, al permitir analizar el efecto de los distintos niveles de la variable independiente sobre la variable dependiente objeto de interés.

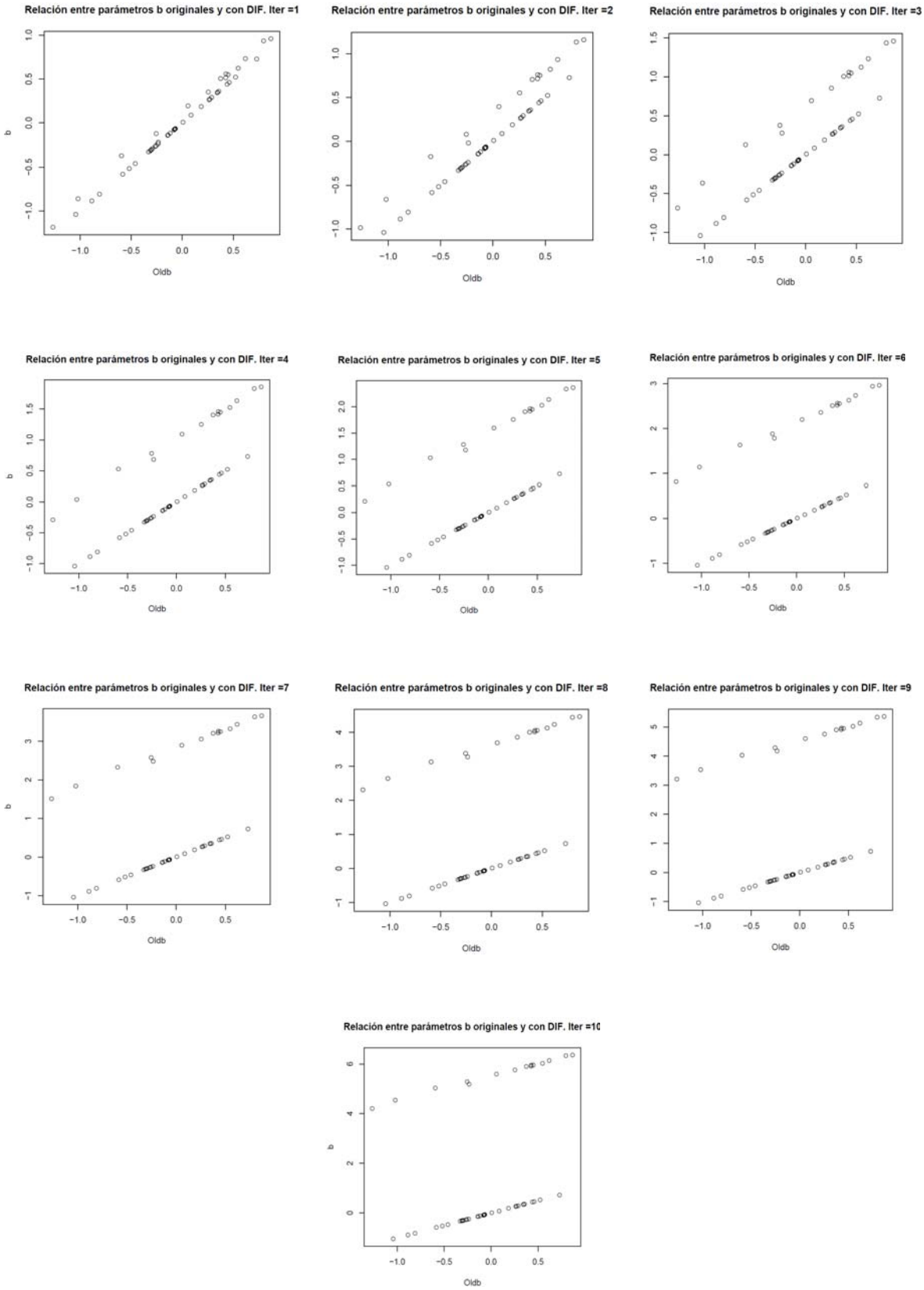


Figura 37. Comparación entre los parámetros b originales y con DIF en las diez iteraciones de la condición experimental 1.

Fuente: elaboración propia.

Condición experimental 2

La simulación se realiza contando con 50 ítems con un índice de dificultad distribuido aleatoriamente N (0,0.5) aplicados a una muestra de sujetos de tamaño N= 2000 extraídos de una población cuyo valor de Theta se ajusta a una distribución normal. La probabilidad de que cada sujeto conteste correctamente a cada ítem se calcula mediante el modelo de Rasch. En la Figura 20 se representa la distribución de los valores de habilidad de la muestra y los parámetros de los ítems para la iteración inicial. Partiendo de esta condición de partida, se generan 10 iteraciones adicionales, en las que se va aumentando la diferencia en nivel de habilidad entre el grupo N_x=1000 y N_Y=1000, aumentando la habilidad del segundo grupo 0.10 en cada iteración. Por tanto, en la condición 0 la habilidad de los dos grupos es la misma y en la condición 10 tenemos la mayor distancia entre los sujetos de ambos grupos, siendo los sujetos del grupo Y sujetos con mayor nivel de habilidad. En las representaciones gráficas que aparecen a continuación (Figura 38), se muestran los valores de theta y las x generadas en las 10 iteraciones sobre las que se aplica el procedimiento bootstrap bidimensional. Como se puede apreciar, existe un patrón creciente en el rango de puntuaciones theta, observándose un desplazamiento a valores más altos de puntuación.

La diferencia en el nivel de habilidad entre dos grupos pertenecientes a la muestra, será la segunda variable independiente utilizada en el presente diseño de investigación. Debemos destacar que, en estudios longitudinales, y en procesos de escalamiento vertical, la diferencia en el nivel de habilidad de los grupos suele ser un factor que puede condicionar enormemente los resultados del proceso. De acuerdo con lo apuntado por Kolen y Brennan (2014), la existencia de diferencias substanciales entre los grupos a equiparar supondría una de las principales fuentes de introducción de error sistemático. Desde un punto de vista operativo, trabajar con dicha fuente de error resultar problemático, sin embargo, la estimación del efecto de interacción se presenta como un elemento clave en la mejora de la calidad de la evaluación, permitiendo estimar el error asociado a la selección de sujetos y reactivos y su interacción, en relación a su efecto en sujetos con distinto nivel de habilidad. Las 11 iteraciones generadas dentro de la condición experimental 2, permitirán analizar la influencia de esta variable independiente sobre la estimación de los valores de theta, analizando la idoneidad del procedimiento presentado bootstrap bidimensional bajo estas condiciones.

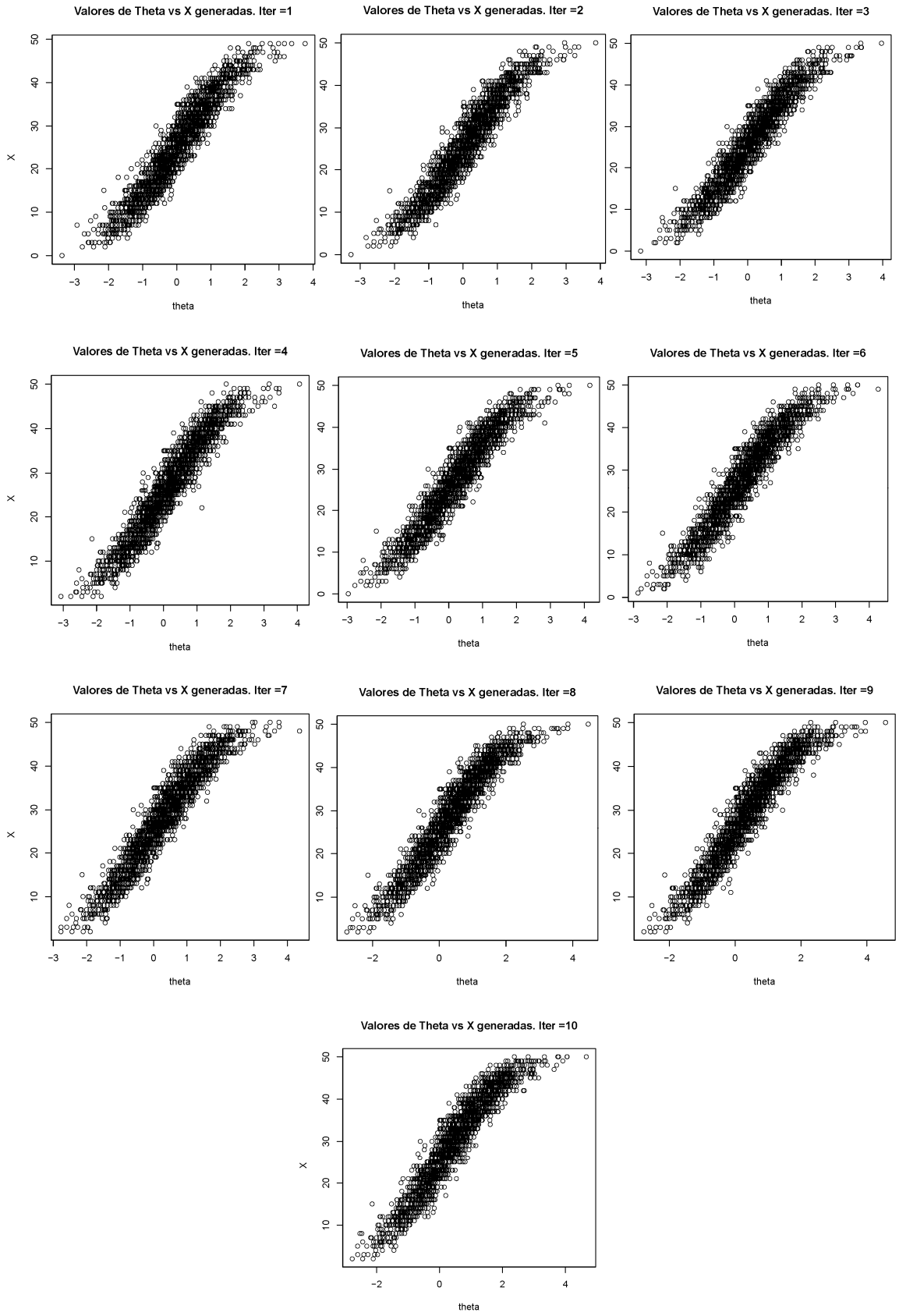


Figura 38. Comparación entre los valores de Theta y X en las diez iteraciones de la condición experimental 2.

Fuente: elaboración propia.

Condición Experimental 3

En el caso de la condición experimental tres, volvemos a contar con 50 ítems con un índice de dificultad distribuido aleatoriamente N (0, 0.1) y una muestra de sujetos de tamaño N= 2000. En la Figura 21 se puede observar la representación de dicha distribución de partida. La condición experimental 3, es la única de las cuatro condiciones en las que se modifica la distribución de los parámetros b de los ítems, pues en lugar de tener una desviación típica inicial de 0,5, es de 0,1.

La probabilidad de que cada sujeto conteste correctamente a cada ítem se calcula mediante el modelo de Rasch. Partiendo de esta condición de partida, se generan 10 iteraciones adicionales, en las que el rango de b se va incrementando progresivamente respecto al rango de theta, es decir, en la iteración 0 la desviación típica en la distribución de los valores del parámetro (b) = 0.1, dicho valor va incrementándose en cada iteración hasta alcanzar el valor de 1 en la iteracción 10. Este distanciamiento progresivo de las distribuciones de los valores de θ y b será analizado aplicando en cada iteración el procedimiento de bootstrap bidimensional propuesto. En las gráficas que aparecen a continuación (Figura 39) puede observarse la progresiva variación en el eje de abscisas, cuyos valores van de -0.4 a 0.4 en la iteración 1 y de -2 a 3 en la iteración 10.

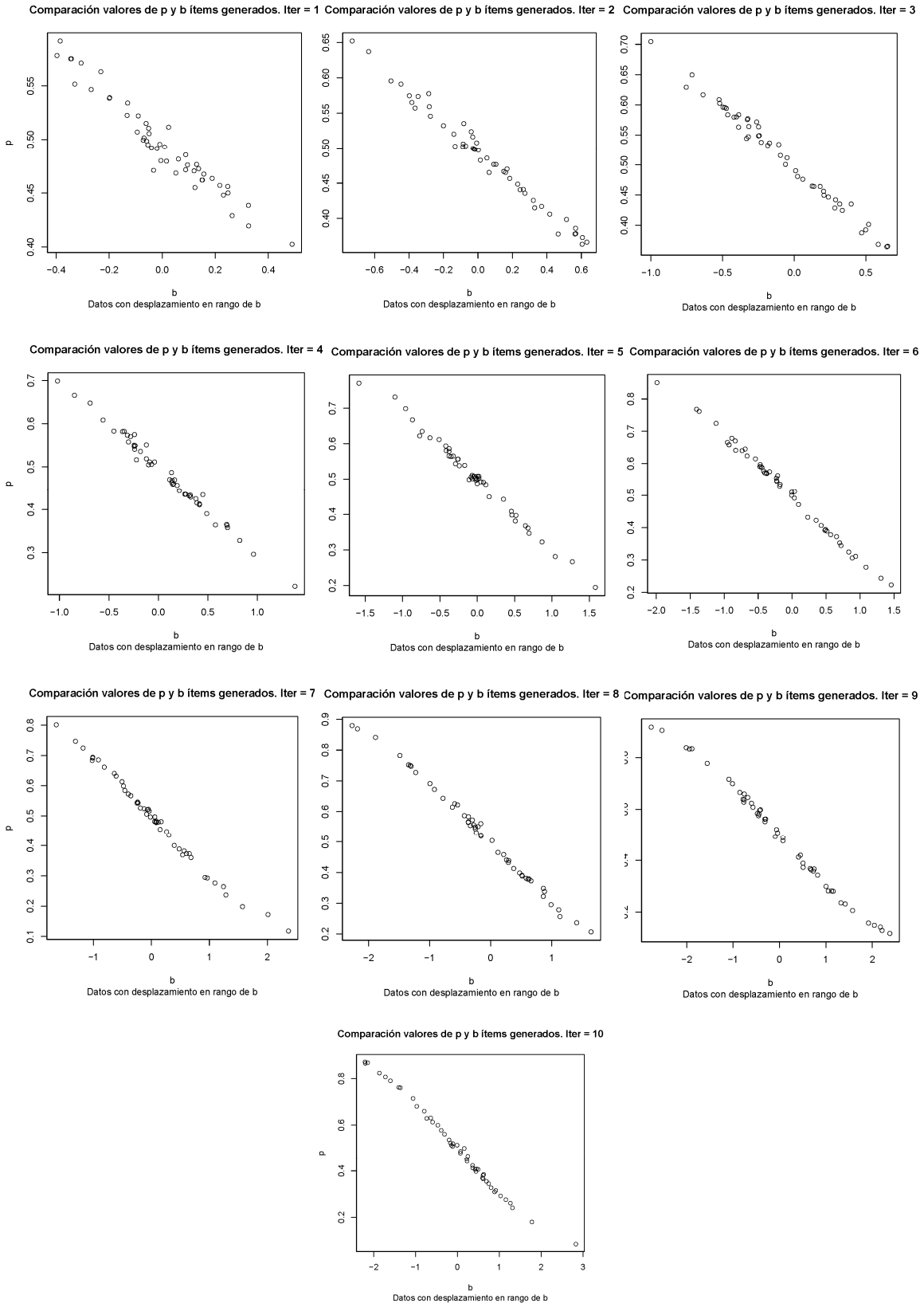


Figura 39. Comparación entre los valores p y b en los ítems generados en las diez iteraciones de la condición experimental 3.

Fuente: elaboración propia.

De manera complementaria, en las siguientes representaciones (Figura 40), se presenta un análisis comparativo entre las distribuciones de los parámetros b de los ítems y la distribución de los valores de Theta.

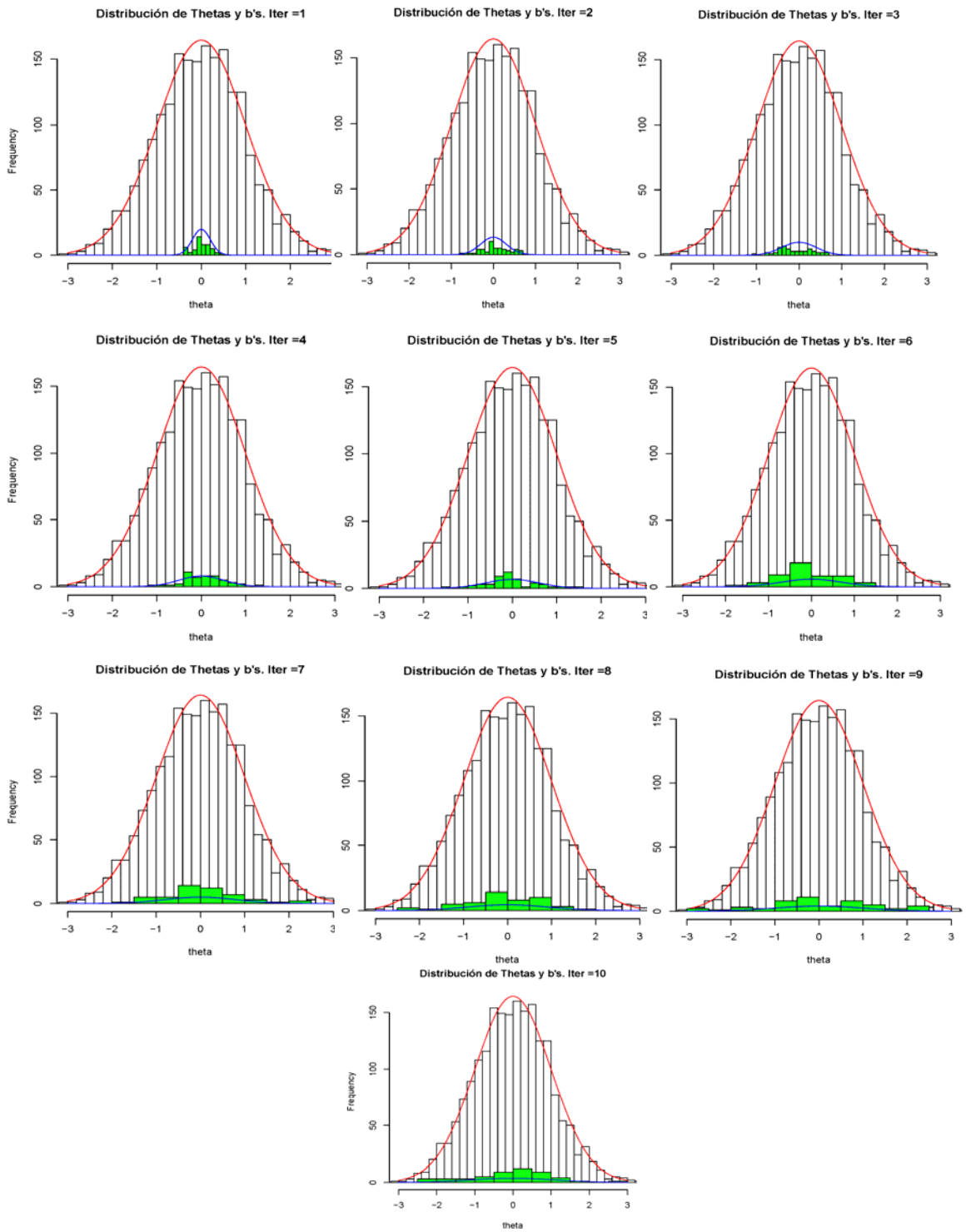


Figura 40. Comparación de las distribuciones de theta y b en las diez iteraciones de la condición experimental 3. Fuente: elaboración propia.

En las gráficas de la Figura 40, observamos el progresivo aplanamiento de la distribución de b al incrementar la desviación típica de su distribución Así, tal y como observamos en las Figuras 32, 33 y 34 (Apartado 5.3.3), la presente condición experimental permitirá analizar la influencia que la relación entre las distribuciones de Theta y b puede tener en el efecto de interacción. La variable independiente en la tercera condición experimental (distribución del parámetro b), nos permitirá analizar su influencia así como la utilidad del procedimiento bootstrap bidimensional en estas condiciones.

Condición Experimental 4

En la cuarta y última condición experimental contamos con 50 ítems con un índice de dificultad distribuido aleatoriamente N (0,0.5) y una muestra de sujetos de tamaño N= 2000 extraída de una población con un nivel de habilidad distribuido normalmente. La probabilidad de que cada sujeto conteste correctamente a cada ítem se calcula mediante el modelo de Rasch. Partiendo de esta condición inicial, se generan 10 iteraciones adicionales, en las que la media de b se va alejando progresivamente del valor medio de theta, es decir, en la iteracción 0 Rango ($\bar{b}$) ≈ Rango ($\bar{\theta}$), en la iteracion 10 el Rango ($\bar{b}$) >>Rango ($\bar{\theta}$).

La presente condición experimental nos permitirá contrastar qué sucede cuándo se produce un incremento en el nivel de dificultad de los ítems utilizados. En esta situación, si la prueba no dispone de reactivos capaces de medir con precisión a los sujetos que se encuentran en los extremos de la distribución de thetas, la precisión de la medida irá decreciendo en las puntuaciones más bajas de la distribución. De este modo, el incremento progresivo de la dificultad media de los ítems utilizados producirá un descenso en la precisión de las medidas de las puntuaciones bajas o extremadamente bajas. La aplicación del procedimiento bootstrap bidimensional permitirá analizar si el mismo es sensible a tal situación. Esta cuarta condición experimental, presenta un marcado carácter complementario respecto a la tercera condición, pues en ambos casos se analizan efectos derivados de la relación entre las distribuciones de theta y b como los descritos en las Figuras 32, 33 y 34 (Apartado 5.3.3).

Tal y como puede apreciarse en las representaciones gráficas que aparecen a continuación (Figura 41), los valores del parámetro b se desplazan progresivamente a la derecha de la distribución, siendo paulatinamente más útiles para la evaluación de sujetos cuyo nivel de habilidad es mayor.

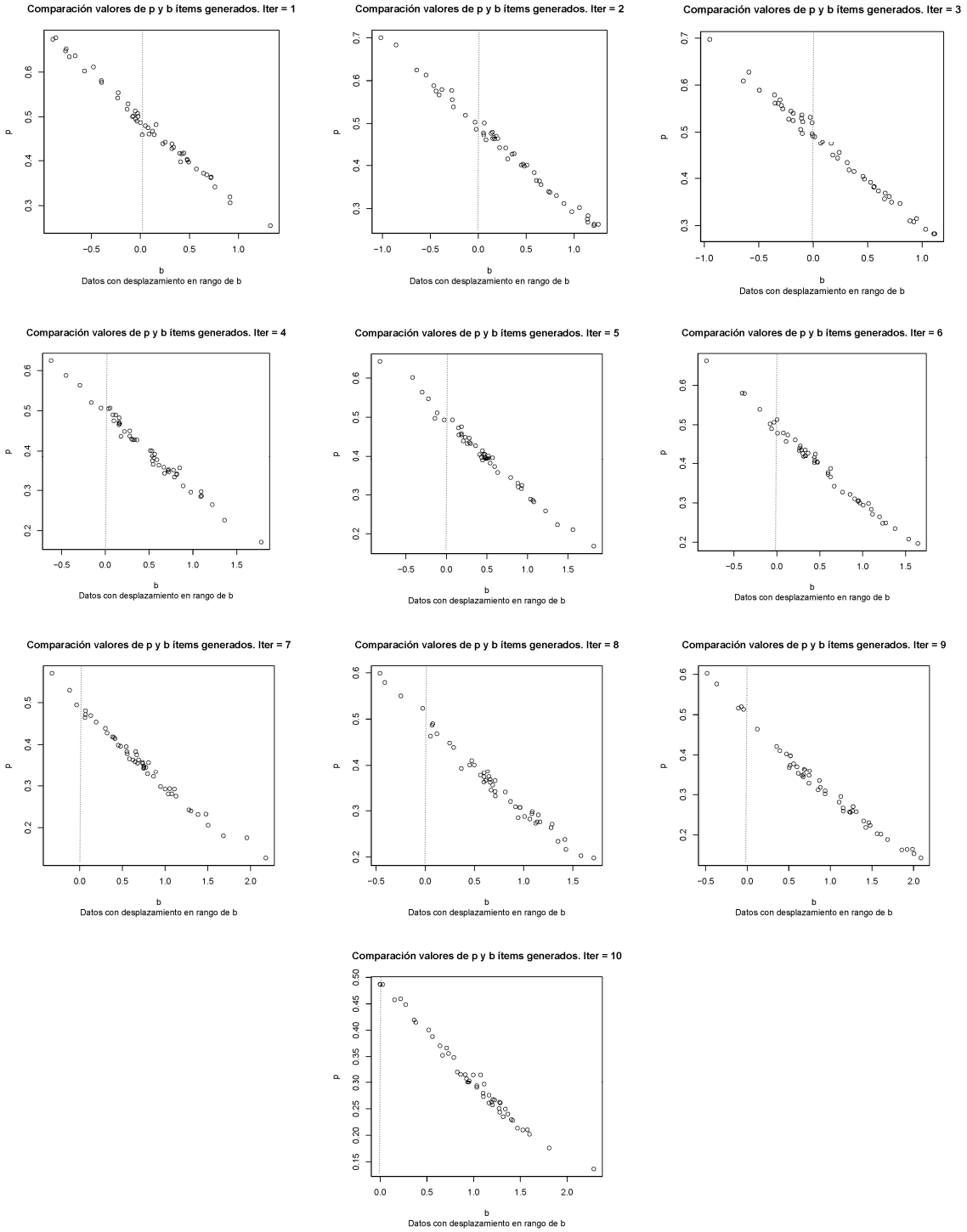


Figura 41. Comparación entre los valores p y b en los ítems generados en las diez iteraciones de la condición experimental 4.

Fuente: elaboración propia.

De manera complementaria, los gráficos recogidos en la Figura 42, muestran cómo la distribución de los valores de b, a pesar de conservar la misma forma, va desplazándose progresivamente a la derecha en relación a la distribución de las puntuaciones theta. Así, el estudio de la variable independiente que define esta cuarta condición experimental (nivel de dificultad de los ítems en relación a theta), permitirá analizar su influencia en relación a la interacción, así como contrastar el funcionamiento del procedimiento propuesto ante estas condiciones.

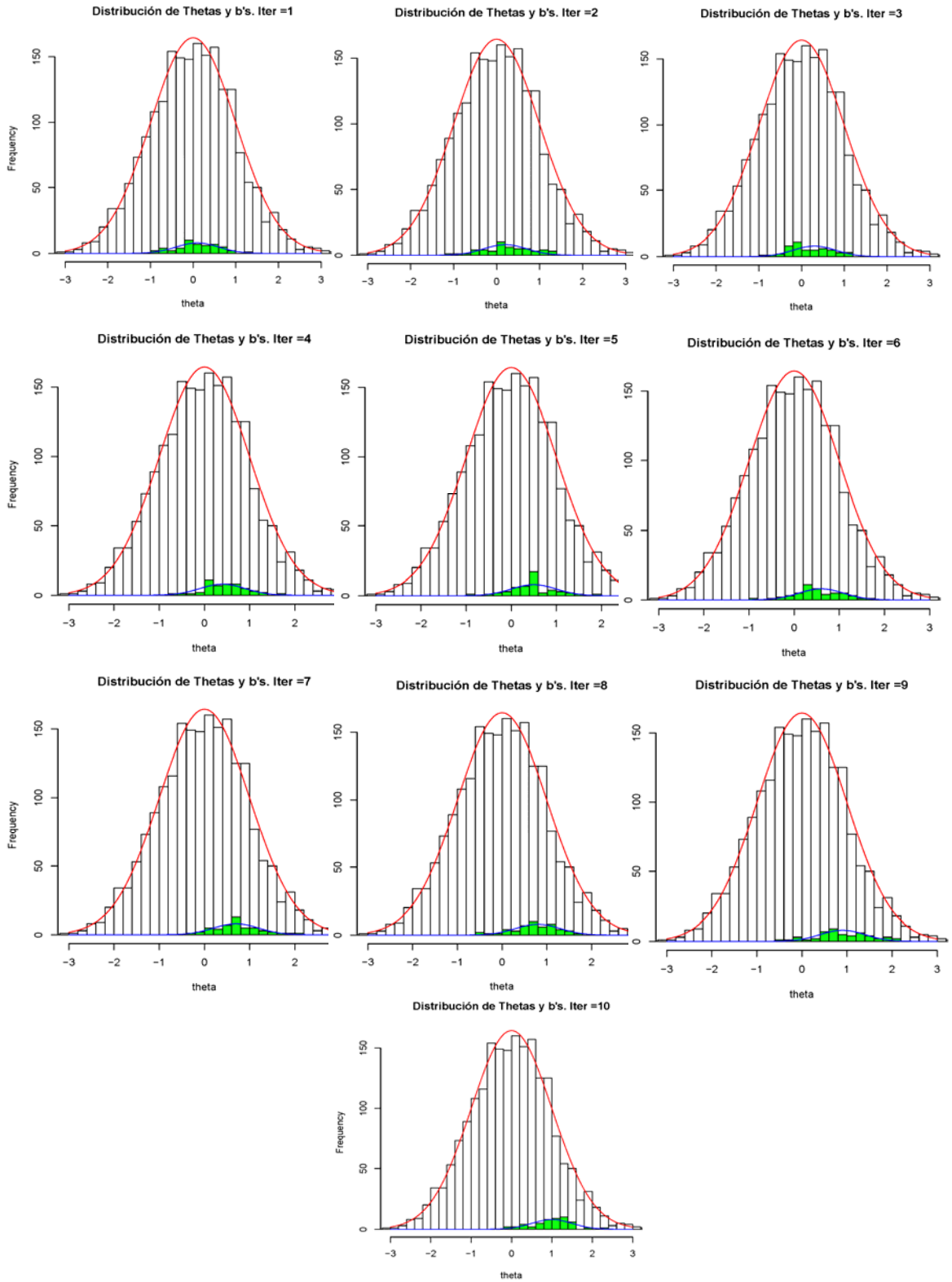


Figura 42. Comparación de las distribuciones de theta y b en los datos en las diez iteraciones de la condición experimental 4.

Fuente: elaboración propia.

CAPÍTULO 6: RESULTADOS DEL DISEÑO Y APLICACIÓN DEL PROCEDIMEINTO «BOOTSTRAP BIDIMENSIONAL» BAJO CUATRO CONDICIONES EXPERIMENTALES

En el quinto capítulo, se ha realizado una descripción en profundidad de las principales cuestiones metodológicas a considerar, detallando el procedimiento de simulación de datos utilizado así como la aproximación teórica y las características técnicas del procedimiento propuesto (bootstrap bidimensional). Del mismo modo, se ha incluido una descripción detallada de las cuatro condiciones experimentales objeto de interés, justificando con profusión los motivos que subyacen a la elección de tales condiciones experimentales.

En el presente capítulo, se presentan los detalles del procedimiento diseñado así como los resultados obtenidos bajo las distintas condiciones experimentales, resultados que nos han permitido juzgar la idoneidad del procedimiento de estimación sugerido, así como el análisis de las condiciones bajo las cuales presenta un mejor funcionamiento, analizando algunos de los factores que podrían afectar a los resultados obtenidos.

6.1 Consideraciones previas

Persiguiendo la idea de identificar el origen y cuantificar la magnitud de los errores de medida, con el fin de reducir la incertidumbre asociada a los procesos de medida y evaluación, se ha expuesto la problemática asociada a la estimación del denominado "error de enlace" (en sus componentes aleatorio y sistemático), profundizando en la necesidad de considerar tanto el error asociado a la selección de sujetos, como el error fruto de la selección de ítems, dando un paso más allá al atender, no solo al efecto de estos factores de manera independiente, sino a su efecto combinado, es decir, el efecto de interacción entre ellos.

El problema de la interacción, cobra gran importancia en cualquier proceso evaluativo y, de forma especial, en aquellos programas de evaluación que producen nuevas ediciones de sus pruebas, ya que se pretende que, el significado de las puntuaciones, sea el mismo a lo largo del tiempo. Tales pruebas, a pesar de ser construidas utilizando la misma matriz de especificaciones, podrían presentar importantes variaciones en sus propiedades psicométricas (Dorans, Moses, & Eignor, 2010) y distintos efectos de interacción con los sujetos pertenecientes a la muestra, condicionando en gran medida el proceso de escalamiento.

Tal y como se apuntaba en el capítulo 1 del presente trabajo, las pruebas aplicadas a los estudiantes año a año a lo largo de diferentes cursos escolares, pretenden medir el mismo constructo (competencia lectora, científica, matemática, etc.) en un amplio rango de dificultad, desde los cursos inferiores, con niveles menores de dificultad, a los cursos superiores, en los que es necesario aumentar la dificultad de las pruebas, puesto que se dirigen a poblaciones de distintas edades y con distinto dominio de las áreas evaluadas. El escalamiento vertical, permitirá la realización de comparaciones entre los resultados de pruebas de rendimiento realizadas en distintos grados, sin embargo, resulta imprescindible considerar el efecto de interacción, pues en estas condiciones, no es difícil apreciar la importancia del efecto combinado de la selección de sujetos y reactivos a la hora de enfrentarse a tales pruebas de evaluación. En consecuencia, medir el efecto de interacción, se presenta como un elemento imprescindible en distintas condiciones de evaluación, especialmente cuando se exige un proceso de escalamiento entre distintas pruebas. Recordemos que, una de la ironías

de la medición educativa es precisamente que, cambios en los programas de evaluación (incluso cuando dichos cambios suponen una mejora), ponen hasta cierto punto en peligro la comparabilidad de las puntuaciones (Brennan, 2007). Las características de la evaluación educativa, y los procedimientos técnicos utilizados en ella, deberán dar respuesta a la necesidad de adaptación (entendiendo la educación como una realidad en continuo cambio y evolución), sin hipotecar con ello la posibilidad de comparación con criterios de interés, la propuesta metodológica que aquí se presenta pretende la consecución de dicho objetivo.

En el ámbito de la comparabilidad, la TRI pretendía desterrar los problemas propios de la TCT, de este modo, al menos desde un punto de vista teórico, se tienen en cuenta las propiedades de simetría (independencia de los resultados respecto al test que se toma como referencia) e invariancia (independencia de los resultados respecto a la muestra utilizada) (Martínez Arias, 2005), sin embargo, en el plano operativo, dicha consideración se complejiza, Kolen y Brennan (2014) señalan que, los modelos de TRI, consiguen incrementar su flexibilidad gracias a las fuertes asunciones estadísticas del modelo pero, es probable que dichas asunciones no puedan sostenerse en condiciones reales de evaluación. Innumerables factores pueden causar variabilidad en los procesos de enlace de puntuaciones (equiparación y escalamiento) bajo el paraguas de la TRI, factores como la variabilidad entre grupos de examinados y/o ítems, la estacionalidad, diferencias regionales, diversidad en la lengua de origen, género, condiciones de aplicación, variables demográficas, etc. (Qian, Jiang, & von Davier, 2013). De acuerdo con lo apuntado anteriormente, la TRI se basa en que la medida de la habilidad latente del sujeto es obtenida a través de la interacción entre las características de los ítems y del los sujetos que contestan la prueba (von Daveir, 2011), sin embargo, desde el punto de vista aplicado, en el cálculo del error producido en el enlace de puntuaciones, parece obviarse la importancia central de dicho efecto de interacción, no quedando reconocida en los procedimientos tradicionales de cálculo utilizados.

Para dar respuesta a la necesidad de considerar el efecto de interacción, se ha propuesto el procedimiento que hemos denominado bootstrap bidimensional, caracterizado por la extracción de muestras pseudo-aleatorias de la matriz de sujetos e ítems, combinando posteriormente las matrices extraídas. Ello posibilita evaluar la existencia de interacción entre los dos factores considerados (ver apartado 5.5.3). A

partir de la variabilidad en las estimaciones fruto de cada doble submuestra (submuestra de sujetos*submuestra de ítems) es calculado el error cuadrático medio, error que incluiría el tradicional error estándar de equiparación (SEE), error aleatorio fruto del muestreo de sujetos, y el error sistemático, producto de factores como el funcionamiento diferencial del ítem, la multidimensionalidad, las diferencias entre grupos, etc. van der Linden (2013) indica que, los informes de equiparación típicos, ignoran el error de equiparación por completo, su error estándar es solo para calcular las fluctuaciones fruto del muestreo, previamente Michaelides y Hartler (2004) ya apuntaban que, la forma convencional de cálculo del error de equiparación, refleja aproximadamente la mitad del error de equiparación real (error muestral), en el marco de la presente investigación, nos atreveríamos a apuntar que, dicha proporción, se reduciría a una tercera parte, pues además del error fruto de la selección de reactivos se ha de considerar el efecto de interacción entre dichas fuentes de error.

El diseño con ítems de anclaje, es el diseño utilizado con más frecuencia en evaluación educativa, especialmente en estudios de tendencia o crecimiento, de este modo, si nos detenemos a analizar el apartado 2 del presente trabajo, observamos cómo en las evaluaciones internacionales de referencia este es el tipo de diseño más utilizado para el enlace horizontal y vertical, Martínez Arias (2005) apunta que, los procedimientos más utilizados son los basados en TRI, con el diseño con test de anclaje y especialmente con el modelo logístico de un parámetro. El número de ítems de anclaje, así como sus propiedades y características técnicas, pueden condicionar en gran medida la calidad del proceso de escalamiento. Contar con ítems que presentan variaciones en su nivel de dificultad para diferentes grupos, podría producir inestabilidad en la estimación de los parámetros y la consiguiente pérdida de precisión en la medida, produciendo procesos de enlace con serias limitaciones. Del mismo modo, asegurarse de que los grupos de examinados son razonablemente iguales en la distribución de habilidad, al menos en relación a los ítems comunes, seria otro aspecto a considerar. Por estos motivos, en la presente investigación se prueba el procedimiento sugerido en un proceso de enlace con ítems comunes, no obstante, el procedimiento descrito, podría resultar de utilidad en otros diseños.

A fin de poner a prueba la viabilidad el procedimiento reseñado (ver apartado 5.5.2), se ha diseñado un estudio de simulación Monte Carlo cuyos detalles han sido

puntualizados en el apartado 5.2.1. El estudio de simulación de datos diseñado ha permitido obtener evidencia empírica bajo 4 condiciones experimentales, contando con un total de 11 archivos de datos en cada condición, sobre cada uno de los cuales se aplicó el procedimiento de estimación propuesto. Puesto que uno de los objetivos centrales del presente trabajo ha sido la elaboración del procedimiento descrito, en el presente capítulo, dedicado a la discusión de resultados, se incluye una presentación del procedimiento propuesto así como una descripción del mismo, sus ventajas y limitaciones.

6.2 Diseño y ejecución del procedimiento bootstrap bidimensional.

El primer objetivo planteado dentro del presente proyecto de tesis doctoral es el de "Propuesta de un procedimiento para el análisis de la interacción de sujetos y reactivos en procesos de enlace «bootstrap bidimensional»", objetivo general que incluía la justificación teórica (descrita en los apartados 5.2.2 y 5.2.3), el "diseño de una propuesta eficiente de implementación" y la "elaboración de sintaxis de análisis para su puesta en práctica", el presente apartado está dedicado a abordar los resultados relativos a estos dos objetivos específicos. Recordamos que, el procedimiento de bootstrap bidimensional, ha sido definido como una técnica intensiva de remuestreo en la que, las unidades de remuestreo son dobles (filas y columnas), en nuestro caso (sujetos y reactivos), técnica que posibilita la medida del efecto de interacción entre tales fuentes de error.

La generación de los datos del estudio de Simulación Montecarlo (Anexos 1 a 4), así como la sintaxis específica para llevar a cabo el procedimiento propuesto (Anexo 5), fue diseñada dentro del programa R, en su versión 3.12 (R Core Team, 2013). A continuación pasamos a describir los detalles del procedimiento planteado.

El primer paso de la sintaxis para la ejecución del procedimiento (Cuadro 1) sería la instalación de dos paquetes requeridos para su uso como herramientas dentro de la misma, tal es el caso del paquete XLConnect (Mirai Solutions GmbH, 2014), que permite importar y exportar datos en formato Excel, y el paquete MIRT (Philip

Chalmers, 2012) (Multidimensional Ítem Response Theory), creado para el análisis de ítems dicotómicos y politómicos utilizando modelos latentes unidimensionales y multidimensionales bajo el paradigma de la Teoría de la Respuesta al Ítem. En este punto, es importante destacar la modificación de las opciones de java por medio del comando «options(java.parameters = "-Xmx4g")» (Cuadro 1), ya que al ejecutar R en una máquina virtual de java, las opciones instaladas por defecto asignan un 80% de las capacidades de memoria reales, situación que suele generar problemas en rutinas en las que el proceso de cálculo requiere mayor capacidad de memoria, con el consiguiente error de java descrito en la salida "java.lang.OutOfMemoryError: Java heap space", la utilización de dicho comando permitió aumentar la memoria disponible, mejorando el proceso de ejecución.

Cuadro 1: Sintaxis paso 1

```
#install.packages("XLConnect")
#install.packages("mirt")
Directorio <- "/Users/evaexpositocasas/Desktop/"  #
 <<<<<<<<<<<<<<<<<<<<     AQUÍ EL DIRECTORIO GENERAL

options(java.parameters = "-Xmx4g")
library(mirt)
library(XLConnect)
```

Tal y como se ha descrito en el apartado 5.2.3 del presente trabajo, el modelo de simulación Monte Carlo en el que se prueba el funcionamiento general del procedimiento implementado, así como su respuesta ante las cuatro situaciones de simulación planteadas, requiere 11 archivos de datos para cada una de las condiciones experimentales (iter 0 a iter 10) en las que se modifica la variable de interés. De este modo, el comando "for (vueltas in 0:10)" (Cuadro 2) realiza el proceso de bootsrap bidimensional en las hojas de datos mencionadas en el comando vueltas, esta opción, nos permitió dividir el procesamiento en varios ordenadores. La ejecución en distintos computadores de manera simultánea, posibilitó la consiguiente reducción en el tiempo de procesamiento de datos. En las líneas subsiguientes, se especifican las dimensiones de la matriz de datos, definiendo las filas y las columnas en las que se estructura cada archivo de datos. La carpeta de datos que contiene los archivos de la primera condición experimental se titula "EXCEL1", tras especificar la carpeta que contiene los datos, se

determinó la carpeta de salida, así como los nombres que se asignarían a los archivos de resultados, especificando la iteración a la que hace referencia, la "vuelta" o archivo de datos de la que proceden tales resultados.

Cuadro 2: Sintaxis paso 2

```
for (vueltas in 7:8  ) {

filainicial <- 1
incremento <- 1999
filafinal <- filainicial +incremento
ntotalfilas <- 2000
ntotalcolumnas <- 50

# A continuación cargamos los datos generados
DirectorioEXCEL <- paste(Directorio, "EXCEL1/", sep =
 "")       # <<<<<<<<<<<<<<<<<<<     AQUÍ EL DIRECTORIO
 DE DATOS
ArchivoXLS <- paste(DirectorioEXCEL, "DatosGenerados
 iter= ",vueltas ,".xlsx", sep = "" )
```

Posteriormente pasamos a la definición de las características del archivo de datos, el encabezado y las columnas de inicio y finalización. Los valores definitivos han de ser 2000 filas y 50 columnas. La variable incremento se generó en un primer momento para poder dividir la ejecución por filas, a fin de reducir el tiempo de ejecución. De este modo, si definiéramos la variable incremento con un valor de 0, tan solo se ejecutaría una línea, al haber definido como incremento el valor 1999 se ejecutan las filas 1 a 2000. Posteriormente se introducen los datos en x para poder operar con ellos (Cuadro 2).

Cuadro 3: Sintaxis paso 3

```
data  <- readWorksheetFromFile(ArchivoXLS , sheet = 1,
                               header = TRUE, startCol
 = 2,
                               startRow = 1, endCol =
 52,
                               endRow = 2001)
nfilas <- incremento +1
ncolumnas <- 50
x <- data
```

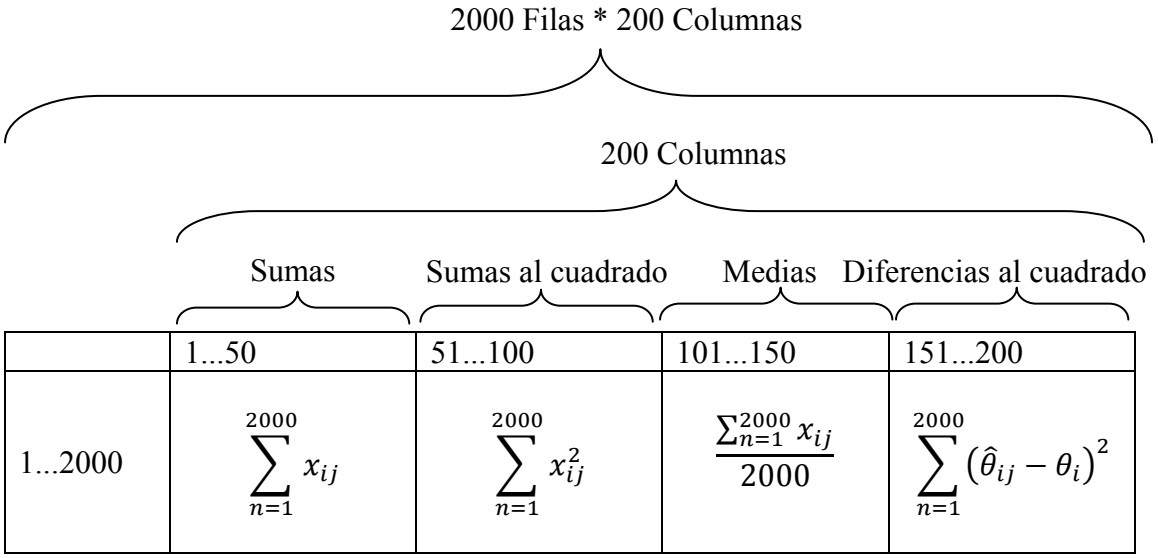
El elemento clave en el diseño del procedimiento bootstrap bidimensional, se sitúa en la ejecución del procedimiento bootstrap en las matrices de sujetos y reactivos, combinando la información procedente de dicho bootstrap en cada estimación. De este modo, es preciso definir las matrices *j* e *i* (ver apartado 5.2.2 para información detallada sobre las mismas) para su uso (Cuadros 3 y 4).

Cuadro 4: Sintaxis paso 4

```
i <-round(runif(ntotalcolumnas *ntotalcolumnas
 ,1,ntotalcolumnas ), digits = 0)
i <- matrix(i, nrow=ntotalcolumnas ,ncol=ntotalcolumnas
 )
j <-round(runif(ntotalfilas*ntotalfilas,1,ntotalfilas),
 digits = 0)
j <- matrix(j, nrow=ntotalfilas,ncol=ntotalfilas)
```

De forma análoga, uno de los factores esenciales que hacen posible el procedimiento diseñado, es la estructuración de los datos de salida, datos que representan un "resumen" de la información fruto del procesamiento que permitió las posteriores estimaciones. Debido a que cada una de las hojas de datos implica 100.000 combinaciones, resultado del bootstrap de sujetos y reactivos (vera apartado 5.2.2), resulta esencial diseñar los datos de salida, puesto que debemos tener en cuenta que el elemento que complica el procedimiento de cálculo es precisamente la estimación de las puntuaciones theta de cada sujeto en cada una de las combinaciones, situación que requiere un elevado tiempo de procesamiento.

De este modo, en cada casilla de la matriz "h" se incluyeron los estadísticos de resumen fruto de la aplicación de cada combinación de las matrices "*i*" y "*j*", es decir, el resultado de la aplicación del procedimiento bootstrap bidimensional a un archivo de datos que contiene una matriz de dimensiones 2000*50 (sujetos*ítems), será el de una nueva matiz de datos de dimensiones (2000*200), pero en cuyas casillas se incluyen los estadísticos de resumen que posibilitaron los cálculos posteriores.



	1...50	51...100	101...150	151...200
1...2000	$\sum_{n=1}^{2000} x_{ij}$	$\sum_{n=1}^{2000} x_{ij}^2$	$\frac{\sum_{n=1}^{2000} x_{ij}}{2000}$	$\sum_{n=1}^{2000} (\hat{\theta}_{ij} - \theta_i)^2$

Figura 43. Esquema de la salida de datos en la matriz "*h*".

Fuente: elaboración propia.

Esta estructura de cuatro capas implica que, el archivo de datos resultante de la aplicación del procedimiento bootstrap bidimensional a cada una de las iteraciones, sea un nuevo documento de Excel con una matriz de datos de dimensiones 2000*200 (Cuadro 5). En dicha matriz, de las columnas 1 a 50, cada casilla presenta las sumas, en las columnas 51 a 100 las sumas al cuadrado y en las columnas 101 a 150 las medias de theta de cada combinación de sujetos e ítems (Figura 43). Con los datos contenidos en estas tres capas, se procedió a los cálculos de ANOVA con dos factores (filas/ columnas). La cuarta serie de datos, que incluye las columnas 151 a 200, es la que contiene la información imprescindible para el cálculo de error cuadrático medio, puesto que en ella se realiza el cálculo de las diferencias entre las puntuaciones theta originales y theta estimadas (theta original-theta observada), sumando posteriormente los cuadrados de los residuos, información reflejada en cada casilla de la cuarta hoja de datos.

Cuadro 5: Sintaxis paso 5

```
 h <- rep(0,ntotalfilas*ntotalcolumnas*4)
dim(h) <- c(ntotalfilas, ntotalcolumnas,4)
```

Tras definir la matriz de datos de salida, se ejecutó la combinación de las matrices "*j*" e "*i*", permitiendo la combinación de los dos muestreos realizados grabando los datos de resumen del procedimiento en “h” (cuyas características se representan en

la Figura 43). La función "system.time" (Cuadro 6) se incluyó para obtener un registro del tiempo de ejecución de la rutina.

Cuadro 6: Sintaxis paso 6

```
system.time(
for (k in 1:ntotalcolumnas  ) {
 for (l in filainicial:filafinal ) {
 y <- x[j[,l],i[k,]+1]
 scores <- fscores(mirt(y, 1), full.scores = TRUE,
 method='EAP', scores.only = TRUE)
      h[l,k,1] = sum(scores)
 h[l,k,2] = sum(scores^2)
 h[l,k,3] = mean(scores)

 h[l,k,4] = sum((scores-x[j[,l],1])^2)
 }}
)
```

En este punto, la matriz *h* nos reportó los datos necesario para realizar tanto el ANOVA de dos factores (filas y columnas) como el cálculo del Error Cuadrático Medio. Tal y como veíamos en el apartado 5.2.3 para el cálculo del RMSE es necesario tener las thetas originales, restar el valor de las thetas observadas en cada condición de doble bootsrap obteniendo los residuos (residuos=thetas originales-thetas estimadas). Tras esto, para el cálculo del error cuadrático medio basta con hacer la suma de los cuadrados de los residuos ECM=sum(residuos^2). La sintaxis finaliza con los comandos que especifican la grabación de los resultados (Cuadro 7).

Cuadro 7: Sintaxis paso 7

```
Datosh <- data.frame(h)
ArchivoXLS <- paste(DirectorioEXCEL, "Resultados Iter =
 ",vueltas,".xlsx", sep = "" )
writeWorksheetToFile(ArchivoXLS , Datosh , "h")
```

La sintaxis presentada, permitió gestionar de manera eficiente la complejidad inherente al procedimiento de análisis propuesto. No obstante, es preciso realizar algunas apreciaciones relativas al mismo:

1. El tiempo estimado en el procesamiento de cada una de las hojas de datos ha consumido aproximadamente 36 horas de ejecución de datos en ordenadores convencionales, sin embargo, debemos destacar que, tras la elaboración y ejecución de la propuesta, con el fin de solucionar algunos de los problemas detectados, conseguimos depurar la rutina con el paquete "Parallel" (Urbanek, 2009), paquete que permite el trabajo con rutinas más largas de forma paralela, especialmente útil cuando se desea probar una misma función en conjuntos de datos diferentes sin la necesidad de que exista comunicación entre dichos procesos (tal y como sucede en el caso que nos ocupa).

2. Acostumbrados a la inmediatez de muchos de los análisis de datos que suelen realizarse en investigación educativa, utilizar un proceso que puede implicar un procesamiento de varias horas, puede ser estimado como “ineficiente”, sin embargo, la importancia de considerar el efecto de interacción, y la sustancial mejora en la precisión de la estimación, justifican sobradamente dicha inversión temporal. Por otro lado, la rutina ha sido ejecutada en computadores convencionales, en consecuencia, el margen de mejora en la reducción del tiempo de ejecución con el aumento de la capacidad de cálculo del ordenador utilizado puede resultar sustancial. Así mismo, téngase en cuenta que se trata de una primera versión, cuyo proceso de mejora ya ha sido iniciado. La dificultad inherente a los procesos de cálculo basados en la probabilidad, limitó en gran medida la aplicación de los procedimientos de TRI hasta la llegada del uso generalizado de ordenadores con capacidad de cálculo suficiente (Martínez Arias, 2005), el procedimiento propuesto, no es ajeno a dichas limitaciones, ya que combina la complejidad de los modelos de TRI y los procedimientos intensivos de remuestreo, sin embargo, en base a esa complejidad, la rutina propuesta, permite trabajar con computadores convencionales con tiempos de ejecución muy razonables, cuya mejora en un tiempo breve está garantizada tanto por los rápidos avances en las tecnologías de computación, como por la mejora y depuración del procedimiento propuesto en versiones subsiguientes del mismo.

3. El procedimiento propuesto se adapta con facilidad a las características de los datos que se deseen evaluar, de este modo, la adaptación tanto en número de

reactivos como en número de sujetos resultan de gran sencillez. Del mismo modo, el procedimiento propuesto presenta un carácter general, cuya adaptación a distintas condiciones de equiparación/ escalamiento o estimación bajo diferentes modelos, es posible. En este sentido, en el marco de la presente investigación, se ha optado por un procedimiento general, sin detenernos en características específicas del diseño de equiparación o escalamiento utilizado o en el modelo de estimación, por tratarse de un acercamiento preliminar a la técnica propuesta y considerando que puede tratarse de un procedimiento "marco" de análisis, que será preciso estudiar en diferentes situaciones en futuras investigaciones.

En el diseño y puesta en marcha del procedimiento propuesto, optamos por el análisis de los ítems comunes por la frecuencia de uso de este procedimiento en numerosos diseños de escalamiento, así como por su uso generalizado en las evaluaciones internacionales de referencia. El reconocimiento extendido, por parte de la comunidad científica, acerca de la importancia del análisis de la suficiencia y de las propiedades técnicas de los ítems de anclaje (Hambleton, Swaminathan & Rogers, 1991; Monseur & Berezner, 2007; Haberman, Lee & Qian, 2009), así como sus importantes implicaciones prácticas, son las principales razones que justifican la elección de dicha condición de partida.

En definitiva, uno de los resultados más destacados del presente trabajo de tesis doctoral, es la propuesta de un procedimiento para el análisis de la interacción de sujetos y reactivos en procesos de enlace (bootstrap bidimensional), cuya justificación teórica y práctica pone de manifiesto su utilidad a la hora de estimar el efecto de interacción de la selección de sujetos y reactivos en la estimación de los niveles de habilidad, situación especialmente relevante en el caso de analizar el funcionamiento de los ítems comunes en diseños de escalamiento con test de anclaje.

Debido a la complejidad del procedimiento, uno de los retos centrales a los que nos enfrentábamos, era el diseño de una propuesta eficiente de implementación así como la elaboración de una sintaxis específica para su puesta en práctica que fuese más allá de la justificación teórica realizada. El procedimiento y la sintaxis presentados, responden con fidelidad a dicho propósito, pues se trata de una propuesta práctica,

viable y útil que posibilita dar respuesta al desafío que era necesario afrontar. A pesar de las dificultades encontradas durante el diseño y programación, el resultado final es altamente satisfactorio, pues permite superar con éxito la complejidad inherente al uso de modelos de escalamiento/ equiparación bajo el paradigma de la TRI y los métodos bootstrap, en los que las numerosas réplicas, y el tiempo de procesamiento requerido para la estimación de la habilidad de cada sujeto en cada réplica, han sido elementos clave a tratar durante el diseño.

Con el fin de realizar un análisis detallado de la propuesta metodológica sugerida, el estudio de simulación Monte Carlo diseñado en el marco de la presente investigación, permitió su análisis global así como el estudio de los factores que pueden incidir en su funcionamiento.

6.3 Bootstrap bidimensional y funcionamiento diferencial del ítem.

Tal y como se ha apuntado en apartados anteriores del presente trabajo, contar con ítems que presentan variaciones en su nivel de dificultad para diferentes grupos, es una posible fuente de inestabilidad en la estimación de los parámetros, situación que implica pérdida de precisión en la medida y, en consecuencia, procesos de enlace pobres. En los diseños de evaluación con ítems comunes, gran parte del peso de la calidad del proceso de enlace recaerá en el funcionamiento de los reactivos seleccionados como ítems comunes, consecuentemente, la importancia de realizar un análisis en profundidad del funcionamiento de los reactivos que se incluyen en los tests de anclaje, ha sido reconocida por la comunidad investigadora, sin embargo, la consideración de dicha fuente de error en el proceso de enlace es todavía insuficiente. Del mismo modo, olvidar la existencia de interacción entre la selección de sujetos y reactivos puede producir una considerable infraestimación del error de enlace. En este sentido, la primera condición experimental diseñada, ha pretendido analizar la adecuación del procedimiento propuesto en condiciones de funcionamiento diferencial del ítem.

Recordemos que, en la primera condición experimental, contamos con 50 ítems comunes con un índice de dificultad distribuido aleatoriamente N (0,0.5) y una muestra

de sujetos de tamaño N= 2000. A partir de esta condición, se generaron 10 iteraciones adicionales, en las que se simuló funcionamiento diferencial en 15 de los cincuenta reactivos para el grupo N_Y=1000. En cada iteración se produjo un aumento de 0.10 puntos en la dificultad de los 15 reactivos que presentan DIF. La condición 0 es la condición de partida, en la que no existe DIF, la condición 1 es la condición con menor grado de DIF y la condición 10 la de mayor nivel de DIF. En el apartado 5.2.3 se han descrito en detalle las características de esta primera condición experimental, por ello, en el presente apartado, tan solo hemos realizado esta breve descripción a fin de ilustrar los resultados, pues el lector interesado en ello puede acudir al citado apartado si desea información más completa sobre las distintas condiciones experimentales.

De acuerdo con la rutina elaborada, la aplicación del procedimiento bootstrap bidimensional fue realizada en cada una de las iteraciones dentro de cada condición experimental, es decir, el procedimiento se aplicó sobre cada archivo de datos de la condición experimental de referencia. Ello nos ha permitido contrastar los resultados del procedimiento bajo condiciones crecientes de grado de funcionamiento diferencial del ítem. En la Figura 44, se ilustra el procedimiento, tal y como se puede apreciar, la aplicación del procedimiento bootstrap bidimensional en cada uno de los archivos de datos de las 11 iteraciones, da lugar a 11 archivos de resultado (uno para cada iteración) con los que se realiza el ANOVA y el cálculo de Error Cuadrático Medio.

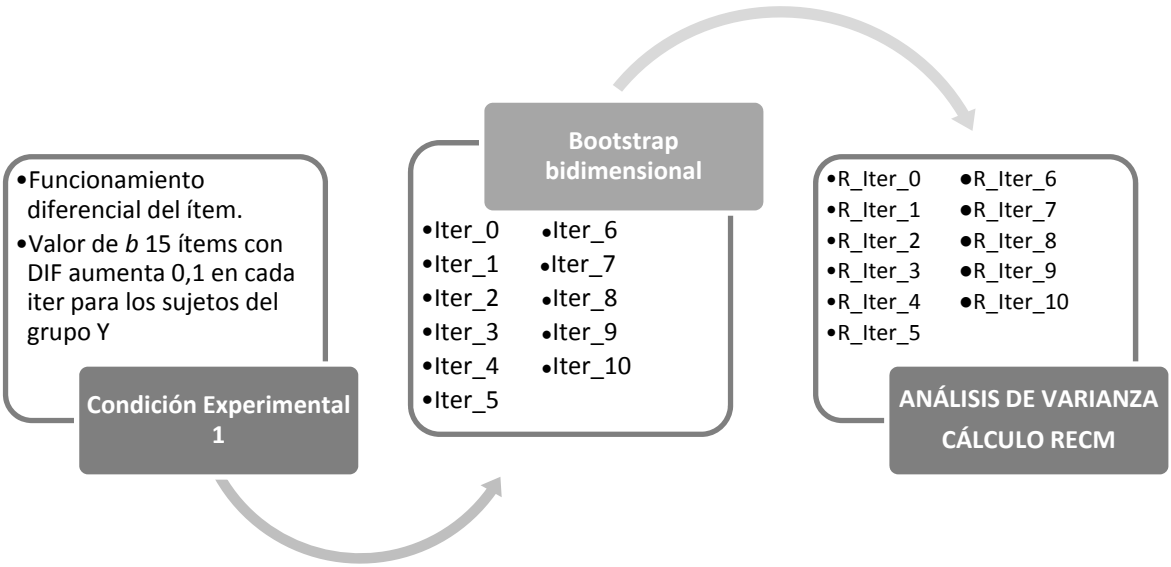


Figura 44. Diagrama que representa la lógica global del procedimiento de implementación y análisis de datos en la primera condición experimental.

Fuente: elaboración propia.

Los archivos de resultados se configuran como una estructura en cuatro capas (ver Figura 43), cada una de ellas, de dimensiones 2000*50, contiene un resumen con los estadísticos necesarios para los análisis de datos planteados. De este modo, las tres primeras capas contienen la información necesaria para realizar un Análisis de Varianza (sumas, sumas al cuadrado y medias).

El análisis de varianza (ANOVA) de dos factores, ha permitido estudiar el efecto de dos factores sobre la variable dependiente objeto de interés. En nuestro caso, el ANOVA posibilitó el análisis de la influencia de los factores filas (efecto de la selección de sujetos) y columnas (efecto de la selección de los reactivos), así como la existencia de interacción entre tales factores, es decir, el análisis del efecto combinado de dichos factores sobre la habilidad estimada. El procedimiento propuesto resulta especialmente útil al detectar dicha interacción. Los datos que aparecen en la Tabla 44 muestran los resultados del análisis de varianza para la condición experimental 1, observándose que, en todas las iteraciones, existe efecto significativo de la interacción (a un nivel de confianza del 99%), en consecuencia, la superficie característica del test en condiciones de no interacción representada por un plano inclinado (ver Figura 22 del apartado 5.2.3) se aleja de la realidad observada, aproximándose con mayor fidelidad a la representada en la Figura 24, justificándose de forma empírica a través de los resultados del Análisis de Varianza, el empleo del procedimiento sugerido.

Tabla 44.

Análisis de Varianza en las once iteraciones de la condición experimental 1

		SC	gl	MC	F	Sig
Iter 0	Entrecolumnas	39715.9182	49	810.528943	863.648337	
	Entrefilas	2411.51566	1999	1.20636101	1.28542193	
	Interacción	104685.128	97951	1.06874997	1.13879231	0.000
	Error	187604988	199900000	0.93849419		
Iter1	Entrecolumnas	73031.7612	49	1490.44411	1558.20048	
	Entrefilas	6003.54306	1999	3.00327317	3.13980355	
	Interacción	237654.983	97951	2.42626397	2.53656321	0.00
	Error	191207602	199900000	0.95651627		
Iter2	Entrecolumnas	48176.3747	49	983.19132	1033.53949	
	Entrefilas	5305.83596	1999	2.6542451	2.79016614	
	Interacción	224020.358	97951	2.28706555	2.40418372	0.000
	Error	190162008	199900000	0.95128568		
Iter3	Entrecolumnas	47853.7296	49	976.606728	1040.08951	
	Entrefilas	3711.08911	1999	1.85647279	1.97714988	
	Interacción	176774.013	97951	1.80471882	1.92203172	0.000
	Error	187698927	199900000	0.93896412		
Iter4	Entrecolumnas	36673.5151	49	748.439085	797.587826	
	Entrefilas	2631.71687	1999	1.31651669	1.40297014	
	Interacción	107419.98	97951	1.09667058	1.1686871	0.000
	Error	187581816	199900000	0.93837827		
Iter5	Entrecolumnas	67859.0592	49	1384.87876	1453.31817	
	Entrefilas	5363.52625	1999	2.68310468	2.8157012	
	Interacción	225435.156	97951	2.30150948	2.41524793	0.000
	Error	190486343	199900000	0.95290817		
Iter6	Entrecolumnas	78257.6946	49	1597.09581	1665.27896	
	Entrefilas	6460.66547	1999	3.23194871	3.36992693	
	Interacción	280596.609	97951	2.86466303	2.98696111	0.000
	Error	191715298	199900000	0.95905602		
Iter7	Entrecolumnas	66070.984	49	1348.38743	1405.03664	
	Entrefilas	5770.1031	1999	2.8864948	3.00776383	
	Interacción	214175.341	97951	2.18655594	2.27841875	0.000
	Error	191840298	199900000	0.95968133		
Iter8	Entrecolumnas	114544.083	49	2337.63435	2434.12337	
	Entrefilas	9364.38987	1999	4.6845372	4.87789783	
	Interacción	339382.163	97951	3.46481571	3.60783067	0.000
	Error	191975933	199900000	0.96035985		
Iter9	Entrecolumnas	118630.457	49	2421.02973	2555.32142	
	Entrefilas	4426.43113	1999	2.21432273	2.33714862	
	Interacción	191967.801	97951	1.95983503	2.06854479	0.000
	Error	189394508	199900000	0.94744626		
Iter10	Entrecolumnas	197618.546	49	4033.03156	4198.57257	
	Entrefilas	6088.02015	1999	3.04553284	3.17054068	
	Interacción	263790.954	97951	2.69309097	2.80363238	0.000
	Error	192018358	199900000	0.96057207		

Tras comprobar, con los resultados del Análisis de Varianza, la existencia de un efecto de interacción entre los factores analizados (sujetos y reactivos), era necesario profundizar en la utilidad del procedimiento a la hora de detectar tales interacciones. Para ello, y haciendo uso de la cuarta capa generada tras la aplicación del procedimiento bootstrap bidimensional, en la que quedaron recogidas las sumas de las diferencias al cuadrado, se procedió al cálculo de la Raíz del Error Cuadrático Medio (ver Ecuación 57), así como el intervalo de confianza de dicho error, cuyos resultados se muestran en la Tabla 45.

Tabla 45.

Error cuadrático medio e Intervalo de Confianza para las 11 iteraciones de la condición experimental 1

	RECM	LI	LS
Iter 0	0.48835469	0.48356062	0.49314876
Iter 1	0.52217825	0.51681857	0.52753793
Iter 2	0.51050731	0.50540933	0.51560530
Iter 3	0.50511251	0.50021536	0.51000965
Iter 4	0.51708709	0.51237944	0.52179475
Iter 5	0.56299583	0.55807159	0.56792008
Iter 6	0.61260003	0.60784999	0.61735008
Iter 7	0.64886648	0.64253541	0.65519754
Iter 8	0.71272900	0.70742000	0.71803800
Iter 9	0.70752456	0.70019471	0.71485441
Iter 10	0.73170722	0.72519188	0.73822256

Como puede apreciarse, los resultados muestran un progresivo incremento del valor de la raíz del error cuadrático medio, pasando de 0.48 puntos en la iteración 0 a 0.73 puntos en la iteración 10, lo que supone un incremento del 60% aproximadamente. La Figura 45 muestra esta progresión.

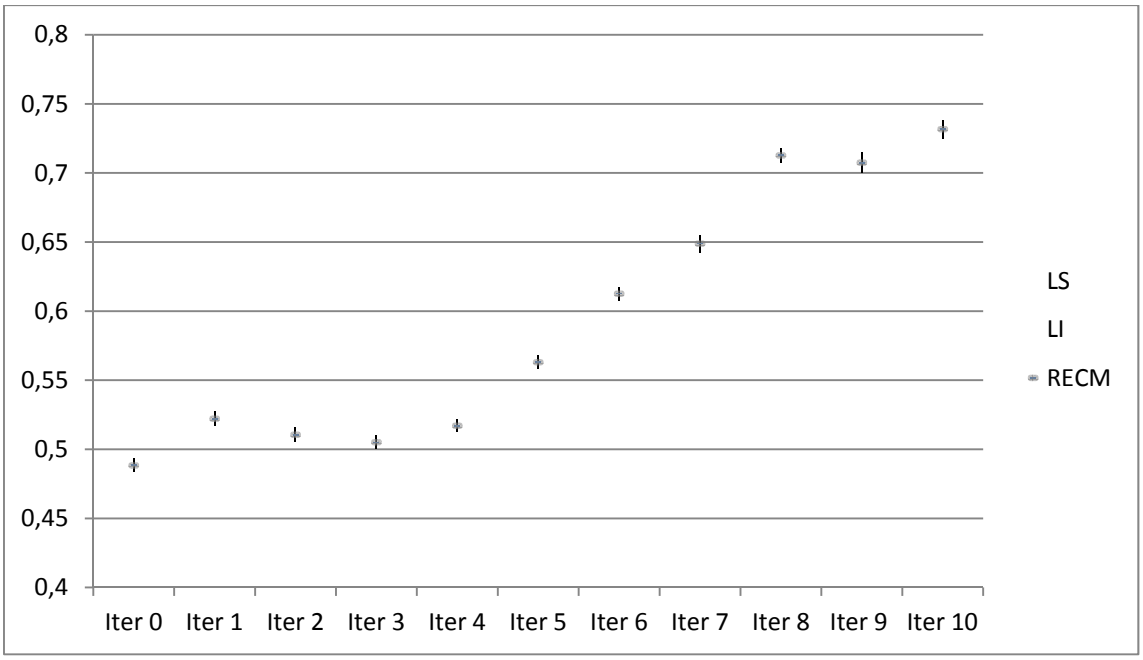


Figura 45. Error Cuadrático Medio e Intervalo de Confianza en las 11 iteraciones que forman parte de la primar condición experimental.

Fuente: elaboración propia.

La iteración 0, en la que no existe funcionamiento diferencial del ítem, presenta un valor de 0.4884 puntos. Como puede apreciarse, este es el valor más bajo de todos los datos registrados, diferencias destacadas en relación al resto de iteraciones que en ningún caso alcanzan dichos valores y cuyos IC se alejan del mismo. Es necesario tener en consideración que, tal y como indicaban los análisis de varianza previos, incluso en situaciones ideales en las que no existe DIF, existirá interacción entre la selección de sujetos y reactivos, no observándose una relación lineal entre los efectos de los factores y la habilidad estimada de los sujetos. Del mismo modo, debemos destacar que, el error cuadrático medio, también recoge la información relativa al error fruto del muestreo de sujetos y de ítems de forma independiente. A partir de la iteración 0, los valores de la RECM comienzan a incrementar de forma progresiva. Las primeras iteraciones, parecen presentar un funcionamiento más inestable, mostrando un patrón de aumento moderado entre iteraciones. Dicho patrón de aumento, podría estar justificado por la menor cantidad de DIF presente en tales condiciones y por el efecto de los factores de selección de sujetos y reactivos de forma independiente. En este sentido, el resultado más destacado sería el aumento en el valor de la RECM desde la primera iteración, iteración en la que se presenta menor grado de DIF.

De la cuarta iteración en adelante, el aumento comienza a ser progresivo, pasando de los 0,5170 puntos de la iteración 4 a los 0,7317 en la iteración 10. El procedimiento propuesto, es sensible a la cantidad de DIF presente en cada una de las iteraciones, poniendo de manifiesto su utilidad en los procesos de análisis de ítems. Así mismo, la consideración del efecto de interacción, obviada en los procedimientos clásicos, se presenta como un elemento imprescindible, a la luz de los resultados obtenidos en el análisis de la aplicación del procedimiento ante esta condición experimental.

Un aspecto destacado es que, el procedimiento de análisis sugerido, no necesita la elaboración de hipótesis previas acerca del funcionamiento de los reactivos, es decir, su utilidad con fines exploratorios radica en la necesidad de poder estudiar el funcionamiento de los reactivos más allá de las expectativas previas acerca del funcionamiento de los mismos.

De acuerdo con Monseur & Berezner (2007), los resultados muestran como, la utilización de diferentes conjuntos de ítems comunes, genera diferencias en las puntuaciones incluso cuando la muestra de sujetos es grande. Estas diferencias, son fruto tanto de la selección de reactivos en sí, como de la interacción del efecto de los mismos con los sujetos que contestan a la prueba.

En evaluación educativa, es relativamente frecuente encontrar como ítems de anclaje aquellos que están agrupados en torno a un pasaje o estímulo común. En este sentido, algunas técnicas como el "Jackknife agrupado" han abordado el efecto de selección de conjuntos de reactivos (Haberman, Lee & Qian, 2009). Sin embargo, en el marco de la presente investigación, se ha decidido considerar éstos como unidades independientes puesto que, a pesar de la complejidad añadida que supone el remuestreo de unidades independientes, consideramos que puede resultar una aportación de gran interés asegurando al mismo tiempo el cumplimento del supuesto de independencia local, esencial en los modelos de TRI.

El procedimiento implementado, podría ser puesto en práctica utilizando bloques de ítems comunes, sin embargo, en base a las razones expuestas, se ha considerado más conveniente optar por un diseño con unidades independientes que, a pesar de la

complejidad añadida que entrañan (mayor número de submuestras sobre las que estimar la puntuación) consideramos puede ajustarse de forma más apropiada a los objetivos perseguidos en el marco de la presente investigación.

Comprobar el comportamiento del procedimiento ante la manipulación de distintas condiciones de DIF (porcentaje de ítems con DIF en el test, tipo de DIF, tamaño muestral de los grupos, formato de respuesta de los ítems, etc.), así como su comparación con otros procedimientos de detección de DIF (Mantel-Haenszel, Regresión Logística, Delta-plot, etc.) se presentan como una interesante vía de análisis en futuros trabajos.

Del mismo modo, si consideramos el funcionamiento diferencial del ítem desde una perspectiva multidimensional (Kok, 1988; Ackerman, 1992), existiría una habilidad principal objeto de evaluación y una serie de habilidades espúreas que no son objeto de evaluación pero que inciden en los resultados de la misma (creando funcionamiento diferencial), en este sentido, el procedimiento propuesto parecería óptimo para el análisis de estas situaciones. Téngase en cuenta que, la multidimensionaldad de las pruebas, es otra de las condiciones que pueden afectar seriamente a procesos de escalamiento en estudios longitudinales, pues en la evaluación del sistema educativo, especialmente en materias con un marcado carácter curricular (como por ejemplo matemáticas y ciencias), pueden existir problemas de dimensionalidad que dificulten el establecimiento de adecuados procesos de escalamiento. Imaginemos que se desea evaluar la competencia matemática de estudiantes de 4º y 5º curso de Educación Primaria, a pesar de que el objetivo central de los instrumentos es la competencia matemática, competencias espúreas con vinculación específica a alguno de los dos cursos mencionados, producirían problemas en la medida del constructo de interés. En este sentido, juzgar la dimensionalidad a través del procedimiento bootstrap bidimensional, detectando el efecto de interacción, resulta una interesante vía de desarrollo. En este sentido, es preciso destacar que en el marco de la presente investigación, se ha optado por un análisis global acerca del procedimiento propuesto, en consecuencia, las perspectivas de futuro en la aplicación del mismo son muy amplias, pues será conveniente analizar de forma exhaustiva el procedimiento ante situaciones como las descritas.

En relación a los resultados obtenidos en la primera condición experimental, nos gustaría apuntar que, el procedimiento implementado permite el análisis de dos de las 4 fuentes de error sistemático que señalan Kolen y Brennan (2014), las relativas a la violación de las asunciones estadísticas del modelo (unidimensionalidad), así como problemas en la puesta en práctica del diseño de recogida de la información, diferencias en el funcionamiento de los ítems en los dos grupos a equiparar debido, por ejemplo, a efectos de posición.

6.4 Bootstrap bidimensional y diferencias en nivel de habilidad entre los grupos evaluados.

En la segunda condición experimental, estudiada a partir del modelo de simulación Monte Carlo planteado, se analizó el efecto que la diferencia en el nivel de habilidad entre los dos grupos cuyas puntaciones se desean enlazar, pudiera tener en el proceso de enlace, mediado por el efecto de interacción entre la selección de sujetos y reactivos. De este modo, tal y como se apuntaba en el capítulo 3 del presente trabajo, se recomienda que los grupos de sujetos evaluados sean razonablemente iguales en la distribución de habilidad, al menos, en los ítems comunes. El procedimiento bootstrap bidimensional se presenta como una vía útil para la detección de situaciones en las que se da dicha diferenciación entre grupos, capturando el error asociado a los principales factores considerados (sujetos e ítems) así como su efecto combinado (interacción).

Con el objetivo de probar dicho funcionamiento, en la segunda condición experimental contamos con 50 ítems con un índice de dificultad distribuido aleatoriamente N (0,0.5) y una muestra de sujetos de tamaño N= 2000. La probabilidad de que cada sujeto conteste correctamente a cada ítem fue calculada mediante el modelo de Rasch. El gráfico representado en la Figura 20 (Capítulo 5) muestra la distribución de los valores de habilidad de la muestra y los parámetros de los ítems para la iteración inicial. Partiendo de esta condición de partida, se generaron 10 iteraciones adicionales, en las que se aumentó de manera progresiva la diferencia en el nivel de habilidad entre los grupos N_x=1000 y N_Y=1000, aumentando la habilidad del segundo grupo 0.10 en cada iteración. En el apartado 5.2.3 del presente trabajo, puede consultarse información más detallada acerca de las condiciones experimentales, por motivos didácticos, hemos

incluido esta breve descripción en el apartado de resultados, a fin de facilitar al lector la interpretación de los resultados que se exponen a continuación.

De forma análoga a lo realizado en la primera condición experimental, el esquema de análisis para esta segunda condición sería similar (ver Figura 46). Partiendo de las condiciones establecidas en la segunda condición experimental, se generaron los 11 archivos de datos (ver Anexo 2), archivos sobre los que se aplicó el procedimiento propuesto y cuyos resultados fueron utilizados en una tercera fase para la realización del Análisis de Varianza y el cálculo de la Raíz del Error Cuadrático Medio.

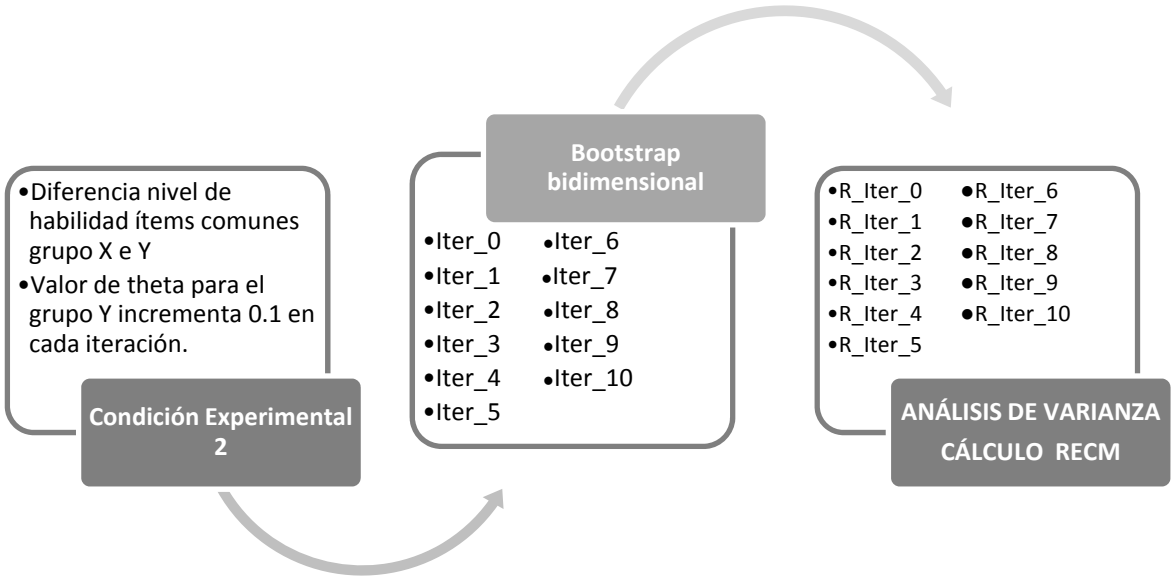


Figura 46. Diagrama que representa la lógica global del procedimiento de implementación y análisis de datos en la segunda condición experimental.

Fuente: elaboración propia.

Este diseño de trabajo, hizo posible el contar con un amplio volumen de información, que nos permitió analizar con precisión el funcionamiento de la segunda condición experimental propuesta. Téngase en cuenta que, los efectos que esta situación pudiera tener, son especialmente relevantes en estudios de tendencia o crecimiento puesto que, las diferencias en los niveles evaluados, son consustanciales a la realidad que se desea estudiar, es decir, desde el origen, en aquellas evaluaciones educativas que contemplan varios cursos o niveles, se asume que existen diferencias en la habilidad de los grupos cuyas puntuaciones se desean enlazar puesto que, de hecho, en la mayoría de los casos son sujetos pertenecientes a distintos niveles del sistema educativo. La evaluación educativa no puede modificar dicha realidad, pero deberá emplear las

estrategias metodológicas óptimas que permitan el mantenimiento de las escalas en tales condiciones. En consecuencia, asegurar un buen funcionamiento de los ítems de anclaje (como mínimo), en estas situaciones resulta un elemento esencial.

La Tabla 46 presenta los resultados del ANOVA para la segunda condición experimental. Como puede apreciarse, de forma equivalente a lo sucedido en la primera condición experimental, observamos la existencia de efecto significativo de la interacción en todas las iteraciones a un nivel de confianza del 99%, quedando ampliamente justificada la aplicación del procedimiento propuesto.

Tabla 46.

Análisis de Varianza en las once iteraciones de la condición experimental 2

		SC	gl	MC	F	Sig
Iter 0	Entrecolumnas	147068,01	49	3001,38796	3163,27642	
	Entrefilas	4355,52771	1999	2,17885328	2,29637597	
	Interacción	194174,261	97951	1,98236119	2,08928552	0,000
	Error	189669626	199900000	0,94882254		
Iter1	Entrecolumnas	36556,4402	49	746,0498	789,957117	
	Entrefilas	4159,46139	1999	2,08077108	2,20323083	
	Interacción	175854,741	97951	1,7953338	1,90099469	0,000
	Error	188789178	199900000	0,9444181		
Iter2	Entrecolumnas	78371,3174	49	1599,41464	1673,02868	
	Entrefilas	4201,35853	1999	2,10173013	2,19846354	
	Interacción	161880,568	97951	1,65266887	1,72873396	0,000
	Error	191104308	199900000	0,95599954		
Iter3	Entrecolumnas	51815,3059	49	1057,45522	1112,35996	
	Entrefilas	4492,57752	1999	2,24741247	2,36410165	
	Interacción	164485,571	97951	1,67926382	1,76645384	0,000
	Error	190033179	199900000	0,95064122		
Iter4	Entrecolumnas	138157,044	49	2819,53152	2976,35365	
	Entrefilas	2822,40147	1999	1,41190669	1,49043683	
	Interacción	105167,433	97951	1,07367391	1,13339157	0,000
	Error	189367399	199900000	0,94731065		
Iter5	Entrecolumnas	128944,998	49	2631,53056	2796,31277	
	Entrefilas	2504,02876	1999	1,2526407	1,33107904	
	Interacción	106479,054	97951	1,08706449	1,15513472	0,000
	Error	188120215	199900000	0,94107161		
Iter6	Entrecolumnas	91413,0185	49	1865,57181	1958,57356	
	Entrefilas	7965,6568	1999	3,98482081	4,18347054	
	Interacción	318510,917	97951	3,25173726	3,4138416	0,000
	Error	190407862	199900000	0,95251557		
Iter7	Entrecolumnas	89500,7815	49	1826,54656	1941,92428	
	Entrefilas	2109,22462	1999	1,05513988	1,12179005	
	Interacción	100540,824	97951	1,02644	1,09127727	0,000
	Error	188023118	199900000	0,94058588		
Iter8	Entrecolumnas	25932,0203	49	529,224904	565,41239	
	Entrefilas	5047,75302	1999	2,52513908	2,69780374	
	Interacción	158721,003	97951	1,62041228	1,73121328	0,000
	Error	187106013	199900000	0,93599807		
Iter9	Entrecolumnas	196384,086	49	4007,8385	4182,63972	
	Entrefilas	4898,89517	1999	2,45067292	2,55755862	
	Interacción	204141,111	97951	2,08411462	2,17501294	0,000
	Error	191545763	199900000	0,95820792		
Iter10	Entrecolumnas	47797,5116	49	975,4594203	1017,222226	
	Entrefilas	3495,851143	1999	1,748799972	1,823672172	
	Interacción	150446,4871	97951	1,535936204	1,601694967	0,000
	Error	191692958,7	199900000	0,958944266		

Tras analizar la pertinencia del uso del procedimiento, en base a los resultados del Análisis de Varianza, pasamos a analizar la utilidad del mismo en condiciones en las que existen diferencias en habilidad entre los dos grupos objeto de comparación. La estimación de la raíz del error cuadrático medio en cada iteración, así como el intervalo de confianza asociado a dicho valor, puede observarse en la Tabla 47.

Tabla 47.

Error cuadrático medio e Intervalo de Confianza para las 11 iteraciones de la condición experimental 2

	RECM	LI	LS
Iter 0	0,49862517	0,49322226	0,50402808
Iter 1	0,50610976	0,50113872	0,51108081
Iter 2	0,52340102	0,51797109	0,52883094
Iter 3	0,52556532	0,52031676	0,53081387
Iter 4	0,52122378	0,51616418	0,52628338
Iter 5	0,54025274	0,53591395	0,54459152
Iter 6	0,58485623	0,57995066	0,58976179
Iter 7	0,59649424	0,59238145	0,60060704
Iter 8	0,61639382	0,61285873	0,6199289
Iter 9	0,68866874	0,68362702	0,69371046
Iter 10	0,712422511	0,708410957	0,716434066

Como podemos apreciar en los datos recogidas en la Tabla 47, el valor de la Raíz del Error Cuadrático Medio es de 0.4986 puntos en la iteración de partida (0), valor muy próximo al valor de la Raíz del Error Cuadrático Medio para la iteración 0 de la primera condición experimental (0.4885), ello pone de manifiesto la igualdad en la situación de partida para ambas condiciones experimentales. Como podemos observar, el valor de la RECM en esta iteración inicial es el menor de todos los estimados, evidenciándose un incremento progresivo de dicho valor en las sucesivas iteraciones. La Figura 47 representa los datos recogidos en la Tabla 47, a partir de dicho gráfico, se puede apreciar con claridad esta evolución en los valores registrados.

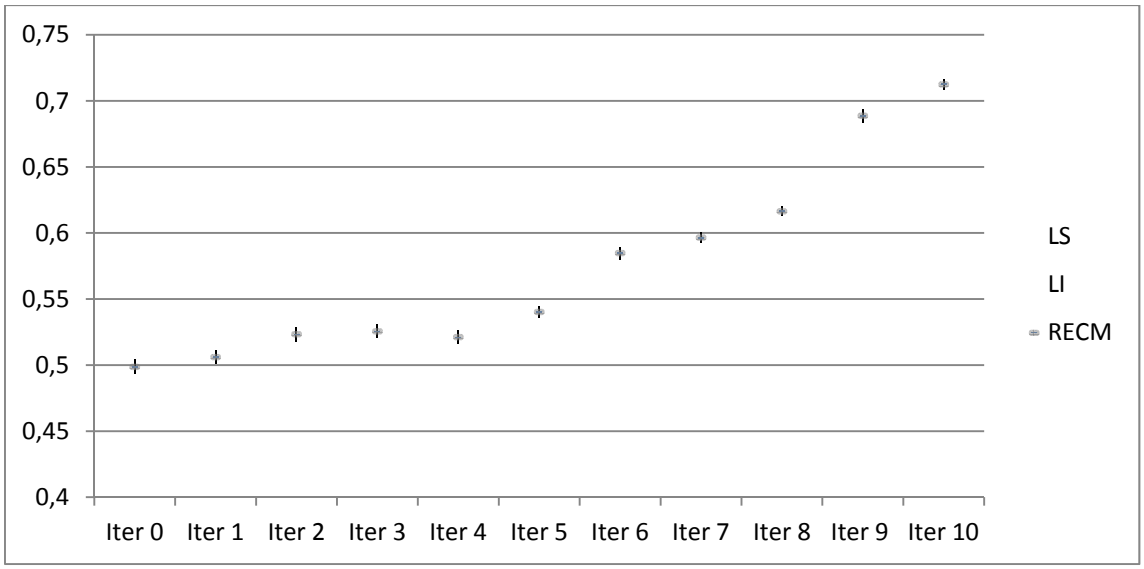


Figura 47. Error Cuadrático Medio e Intervalo de Confianza en las 11 iteraciones que forman parte de la segunda condición experimental.

Fuente: Elaboración propia.

La Figura 47 representa un marcado patrón de crecimiento en los valores de la Raíz del Error Cuadrático Medio en cada iteración, observándose un patrón más definido a partir de la iteración 5 (RECM Iter 5= 0,5402), iteración que define el punto central en lo que a diferencias entre los dos grupos se refiere. En la última iteración, el valor de la Raíz del Error Cuadrático Medio alcanza los 0.70 puntos, valor muy próximo al encontrado en el caso de la condición experimental 1, en la que se obtenía un valor de 0.73 en la décima iteración.

Por otro lado, las diferencias en el valor de la Raíz del Error Cuadrático Medio en el paso de iteración a iteración no parecen significativas en los primeros niveles, sin embargo, si realizamos comparaciones más amplias (en las que se analice la diferencia entre los valores de la RECM de iteraciones más distanciadas (4 a 6 por ejemplo), las diferencias comienzan a ser más destacadas (RECM iter 4=0.5212 y RECM Iter 6=0,5848).

El procedimiento de bootstrap bidimensional presenta un óptimo funcionamiento en estas situaciones, permitiendo el análisis de las diferencias en el nivel de habilidad entre los dos grupos objeto de interés, cuyas puntuaciones se desean comparar. Cuando existe una marcada diferencia entre los sujetos pertenecientes a dos grupos, el valor de

la RECM presenta un notable incremento. De este modo, si se desean comparar grupos extremos, tal vez resulte conveniente utilizar grupos intermedios para producir un mejor resultado en cadenas de escalamiento. Aunque la eficacia de éste diseño ha sido reconocida en anteriores ocasiones, es importante destacar la eficacia del procedimiento bootstrap bidimensional para detectar marcadas diferencias entre los grupos, con especial incidencia en los tests o ítems de anclaje, pues dichos efectos pueden tener importantes repercusiones en la validez de la medida y en consecuencia en los resultados de la evaluación y en el cumplimiento de sus propósitos.

Entre las 4 fuentes de error sistemático a las que apuntan Kolen y Brennan (2014) se encontraría la existencia de diferencias sustanciales en la conducta de los grupos a equiparar, en este sentido, el procedimiento propuesto permitiría la medida del error asociado a dicha situación. Por otro lado, Monseur y Berezner (2007), en el ámbito de las evaluaciones internacionales, proponen utilizar el procedimiento Jacknnife para el remuestreo de ítems, separando los análisis para cada país debido a que la interacción del efecto de sujetos e ítems puede alterar el cálculo del error de enlace, sin embargo nuestro doble bootstrap parece permitir la medida de este efecto de interacción, sin necesidad de elaborar hipótesis previas sobre el funcionamiento de los datos, creando divisiones a priori que no siempre se ajustarían al comportamiento real de los instrumentos en condiciones de evaluación.

6.5 Bootstrap bidimensional y variaciones en la distribución de b

A la hora de llevar a la práctica un diseño de enlace de puntuaciones con ítems de anclaje, no debemos olvidar otras dificultades de implementación que podemos encontrar en su puesta en marcha. De este modo, tal y como apunta Cook (2007), las condiciones bajo las que se debería emplear este tipo de diseño para conseguir cierta garantía en el proceso serían: similitud entre las muestras de sujetos a los que se aplican las formas a comparar, similitud entre las dos formas a equiparar y fuerte relación entre las puntuaciones a equiparar en ambas formas y el test de anclaje. El último punto, puede resultar altamente complejo en situaciones de escalamiento vertical, puesto que, los ítems incluidos en los tests de anclaje deberán poseer una fuerte relación con las

pruebas objeto de escalamiento, pruebas que posiblemente tengan cierta vinculación con el nivel de dominio evaluado. Por este motivo, las condiciones experimentales tres y cuatro tratan dos aspectos relacionados con la distribución de los parámetros b en los ítems que forman parte del test de anclaje, pues el objetivo fue observar si, variaciones en la distribución (condición experimental tres) o en la dificultad media de los reactivos utilizados (condición experimental 4), pueden incidir en un mayor o menor grado de interacción entre los efectos de la selección de sujetos y de reactivos.

En concreto, comenzando con el análisis de la condición experimental tres, iniciaremos con una breve descripción acerca de sus características generales, descritas en mayor detalle en el apartado 5.2.3 del presente trabajo. En concreto contamos con 50 ítems cuyo índice de dificultad sigue una distribución N (0, 0.1), y una muestra de sujetos de tamaño N= 2000. La condición experimental 3, es la única de las cuatro condiciones en las que se modificó el valor de la desviación típica de la distribución de los parámetros b de los ítems, pues en lugar de tener una desviación típica inicial de 0.5, su valor es de 0.1. En la Figura 48 se representan las distribución de los valores de theta y de b en tres iteraciones extremas (la iteración 1, la 5 y la 10). En el apartado 5.2.3, se incluye una representación de la distribución en cada una de las iteraciones, con fines ilustrativos, en este apartado tan solo se han destacado tres iteraciones para que el lector pueda observar con facilidad la diferencia entre las mismas. La distribución de partida (Iter 0), aparece representada en la Figura 21 del capítulo 5.

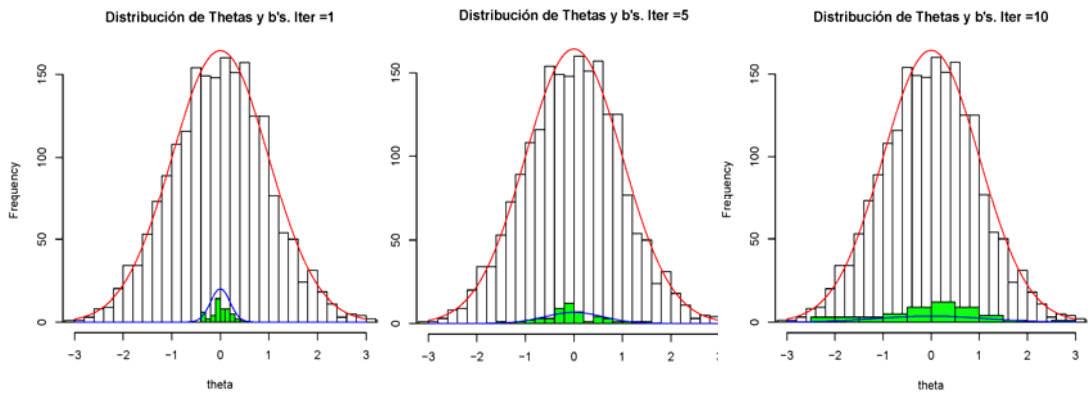


Figura 48. Comparación de las distribuciones de Theta y b en las iteraciones 1, 5 y 10 de la condición experimental 3.

Fuente: Elaboración propia

En cada una de las iteraciones generadas, el rango de b se fue incrementando progresivamente respecto al rango de theta (0.1 puntos de incremento en cada iteración), es decir, en la iteración 0 la desviación típica en la distribución de los valores del parámetro b es igual a 0.1, dicho valor fue incrementándose en cada iteración hasta alcanzar el valor de 1 en la interacción 10. Esta fluctuación en la relación entre los parámetros de θ y b pudo ser analizada aplicando en cada iteración el procedimiento de bootstrap bidimensional propuesto (Ver Figura 49).

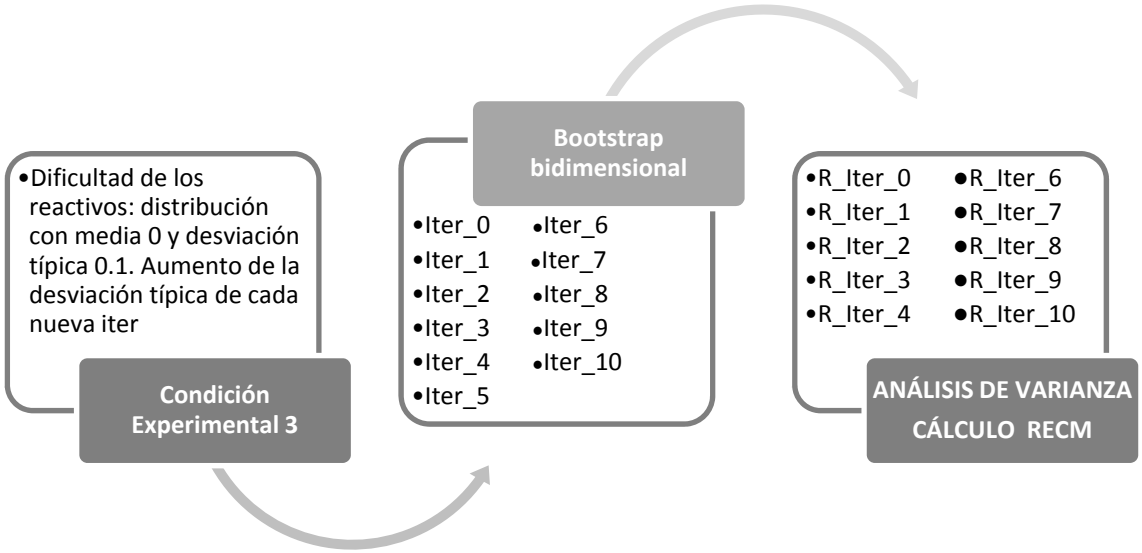


Figura 49. Diagrama que representa la lógica global del procedimiento de implementación y análisis de datos en la tercera condición experimental.

Fuente: elaboración propia.

Por tanto, el primer paso tras aplicar el procedimiento bootstrap bidimensional en cada iteración, consistió en realizar el Análisis de Varianza a partir de las tres primeras capas del archivo de resultados. Los datos expuestos en la Tabla 48, muestran los resultados del Análisis de Varianza en las once iteraciones que componen la tercera condición experimental.

Tabla 48.

Análisis de Varianza en las once iteraciones de la condición experimental 3

		SC	gl	MC	F	Sig
Iter 0	Entrecolumnas	52562,24	49	1072,69878	1131,88628	
	Entrefilas	4320,14737	1999	2,16115426	2,28039866	
	Interacción	173946,172	97951	1,77584887	1,87383356	0,000
	Error	189447023	199900000	0,94770897		
Iter1	Entrecolumnas	21050,8233	49	429,608638	455,286713	
	Entrefilas	3244,89365	1999	1,62325845	1,720282	
	Interacción	138675,801	97951	1,41576707	1,50038868	0,000
	Error	188625682	199900000	0,94360021		
Iter2	Entrecolumnas	33741,6001	49	688,604083	733,148422	
	Entrefilas	2868,56929	1999	1,43500215	1,52782939	
	Interacción	110395,814	97951	1,12705142	1,19995799	0,000
	Error	187754556	199900000	0,9392424		
Iter3	Entrecolumnas	35752,2246	49	729,637237	771,710101	
	Entrefilas	3903,20626	1999	1,95257942	2,06517045	
	Interacción	177604,307	97951	1,81319545	1,91774922	0,000
	Error	189001652	199900000	0,945481		
Iter4	Entrecolumnas	28701,858	49	585,752205	616,823814	
	Entrefilas	5903,0283	1999	2,95299064	3,10963397	
	Interacción	254137,945	97951	2,59454161	2,73217077	0,000
	Error	189830326	199900000	0,94962644		
Iter5	Entrecolumnas	69185,5252	49	1411,94949	1501,57744	
	Entrefilas	3376,24518	1999	1,68896707	1,79617958	
	Interacción	160433,724	97951	1,63789777	1,74186849	0,000
	Error	187968131	199900000	0,94031081		
Iter6	Entrecolumnas	71726,3592	49	1463,80325	1534,12568	
	Entrefilas	4934,69511	1999	2,46858185	2,58717475	
	Interacción	201583,063	97951	2,05799903	2,15686716	0,000
	Error	190736831	199900000	0,95416123		
Iter7	Entrecolumnas	102574,679	49	2093,3608	2206,09673	
	Entrefilas	3911,53904	1999	1,95674789	2,06212667	
	Interacción	112394,076	97951	1,14745206	1,20924698	0,000
	Error	189684713	199900000	0,94889801		
Iter8	Entrecolumnas	123673,834	49	2523,95579	2625,10116	
	Entrefilas	5647,14253	1999	2,82498376	2,93819257	
	Interacción	240994,636	97951	2,46035912	2,55895591	0,000
	Error	192197836	199900000	0,96146991		
Iter9	Entrecolumnas	135907,347	49	2773,61933	2897,55782	
	Entrefilas	7908,41479	1999	3,95618549	4,13296665	
	Interacción	335282,452	97951	3,42296099	3,57591514	0,000
	Error	191349591	199900000	0,95722657		
Iter10	Entrecolumnas	157695,001	49	3218,26533	3340,29669	
	Entrefilas	7961,83245	1999	3,98290768	4,13393304	
	Interacción	345715,116	97951	3,52947001	3,66330175	0,000
	Error	192597035	199900000	0,96346691		

Los resultados del análisis de varianza muestran un efecto significativo de la interacción en las diez iteraciones, de este modo, no atender a dicho efecto en la estimación supondría una considerable pérdida de precisión en los resultados, aconsejándose plenamente el procedimiento de análisis presentado.

En un segundo momento, si nos detenemos a analizar los resultados de la Raíz del Error Cuadrático Medio en las 11 iteraciones de la condición experimental 3, observamos un aumento moderado de dicho valor (Tabla 49). Los valores para la RECM en las iteraciones 0 a 5, son valores muy similares a los observados en la iteración 0 de las condiciones experimentales 1 (RECM iter 0= 0.4883), 2 (RECM iter 0= 0.4986) y 4 (RECM iter 0= 0.49580708). Téngase en cuenta que, en dichas condiciones experimentales, la iteración de partida correspondería con la iteración 5 de la presente condición, pues haría referencia a una situación en la que no existen diferencias entre los grupos, no hay funcionamiento diferencial del ítem, y la distribución de los valores de b corresponde con una media de 0 puntos y una desviación típica de 0.5 (iteración 4 de la presente condición experimental).

Tabla 49.

Error cuadrático medio e Intervalo de Confianza para las 11 iteraciones de la condición experimental 3

	RECM	LI	LS
Iter 0	0,49432841	0,48905615	0,49960066
Iter 1	0,49444223	0,48938394	0,49950051
Iter 2	0,48254949	0,47776329	0,48733569
Iter 3	0,51381536	0,5087154	0,51891532
Iter 4	0,50995145	0,50506623	0,51483666
Iter 5	0,50088581	0,49577593	0,50599568
Iter 6	0,52532338	0,52002841	0,53061836
Iter 7	0,53403998	0,52854496	0,539535
Iter 8	0,54656336	0,54106576	0,55206096
Iter 9	0,56814855	0,56294334	0,57335377
Iter 10	0,55059656	0,54557075	0,55562238

El incremento en el valor de la Raíz del Error Cuadrático Medio es muy leve, si estudiamos la tendencia de dicho valor a través de la Figura 50, observamos que tal incremento es apenas perceptible entre iteraciones, pues el intervalo de confianza observado indica que dichas diferencias no son estadísticamente destacables. No obstante, si realizamos un análisis comparativo entre iteraciones más alejadas,

observamos que dicho incremento si parece ser más destacado, de este modo, entre las iteraciones 2 (RECM = 0.4825) y 9 (RECM = 0.5681), las diferencias son más notables.

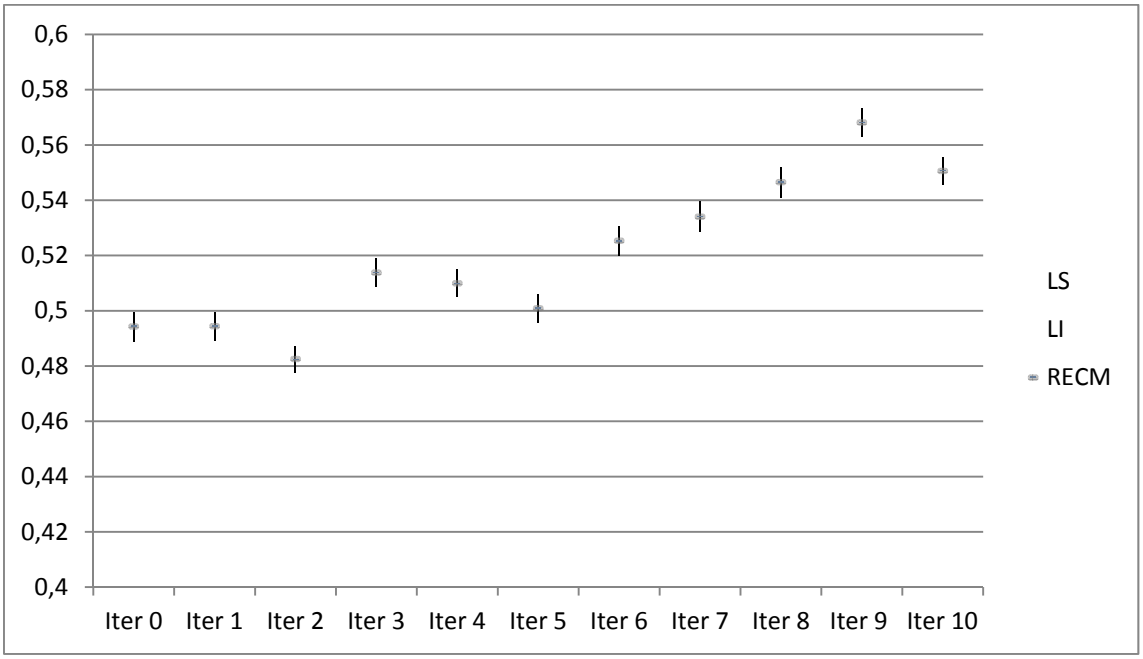


Figura 50. Error Cuadrático Medio e Intervalo de Confianza en las 11 iteraciones que forman parte de la tercera condición experimental.

Fuente: Elaboración propia.

Una distribución de los valores de b, que abarque un rango de habilidad similar al rango de habilidad de los sujetos evaluados (desviación típica de 0.1 a 0.4), parece producir mejores resultados. Esta idea, apoyaría la recomendación de incluir ítems de anclaje que abarquen el rango de habilidad de los sujetos evaluados, mostrando mejor funcionamiento cuando las distribuciones de theta y b son más próximas. A pesar de haber encontrado cierto incremento en la RECM, las diferencias son muy sutiles. Consideramos que, si unido a esta variación en los valores de la desviación típica, se da una diferenciación entre los dos grupos (condición frecuente en procesos de escalamiento vertical), el incremento en el valor de la RECM sería mucho más destacado.

De forma complementaria, la condición experimental 4 nos ha permitido analizar de manera pormenorizada el funcionamiento del procedimiento en condiciones de variación de la distribución de los valores de b respecto a la distribución de valores de theta, en concreto, el funcionamiento del procedimiento cuando la media de los

valores de dificultad incrementa progresivamente. De este modo, en la cuarta y última condición experimental partimos de 50 ítems con un índice de dificultad distribuido aleatoriamente N (0, 0.5) y un tamaño muestral de N= 2000 sujetos. La probabilidad de que cada sujeto conteste correctamente a cada ítem se estimó mediante el modelo de Rasch. Las 10 iteraciones adicionales creadas a partir de esta condición de partida, presentan un valor en la media de b que se aleja progresivamente del valor medio de theta, es decir, en la iteración 0 Rango ($\bar{b}$) ≈ Rango ($\bar{\theta}$), en la iteracion 10 el Rango ($\bar{b}$) >> Rango ($\bar{\theta}$).

En la representación gráfica de la Figura 51, puede observarse el desplazamiento del valor medio de b en iteraciones extremas (1, 5 y 10). La información detallada acerca de la distribución de puntuaciones y valores de b en cada una de las iteraciones puede observarse en el apartado 5.2.3 del presente trabajo.

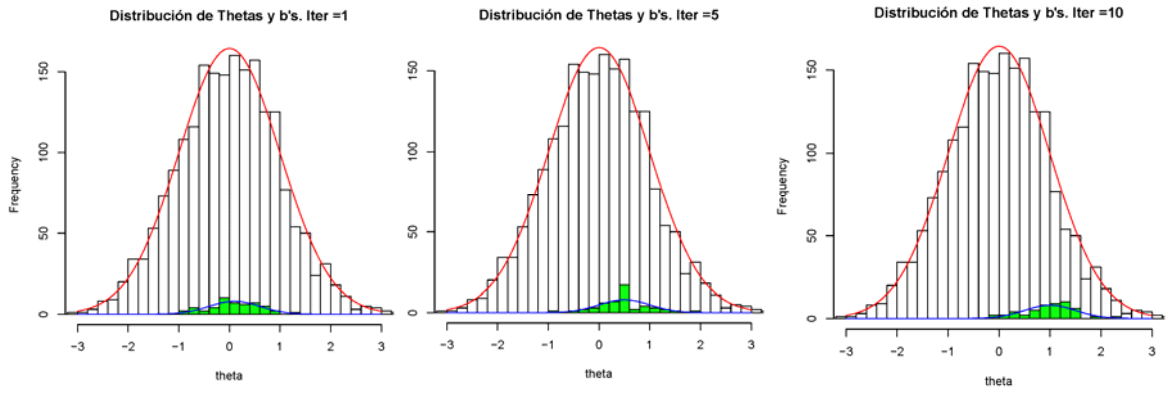


Figura 51. Comparación de las distribuciones de theta y b en las iteraciones 1, 5 y 10 de la condición experimental 4.

Fuente: Elaboración propia.

De forma análoga a lo realizado en las condiciones experimentales precedentes, el esquema de trabajo sobre la condición experimental 4 vendría representado por la Figura 52, en la que se aprecia cómo tras el primer paso de definición de la condición experimental, y la generación de los 11 archivos de datos fruto de la simulación Monte Carlo, se aplicó el procedimiento de bootstrap bidimensional para analizar su funcionamiento en condiciones en las que la media de b va aumentando de manera progresiva. El Análisis de Varianza, realizado a través de las tres primeras capas del

archivo de resultados, hizo posible el análisis de la pertinencia de aplicar una técnica como la propuesta.

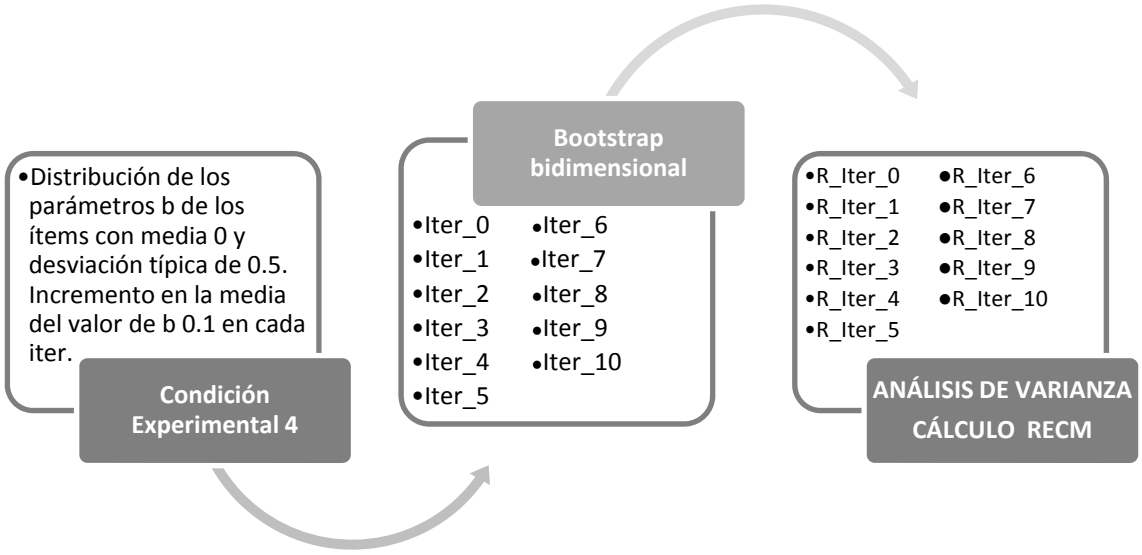


Figura 52. Diagrama que representa la lógica global del procedimiento de implementación y análisis de datos en la cuarta condición experimental.

Fuente: elaboración propia.

La Tabla 50 recoge los resultados del Análisis de Varianza para la cuarta condición experimental. Los resultados del análisis ponen de manifiesto la adecuación del uso de la técnica sugerida a la luz de la confirmación de efecto de interacción entre los factores selección de sujetos e ítems en todas las iteraciones de la condición experimental analizada.

Tabla 50.

Análisis de Varianza en las once iteraciones de la condición experimental 4

		SC	gl	MC	F	Sig
Iter 0	Entrecolumnas	52562,24	49	1072,69878	1131,88628	
	Entrefilas	4320,14737	1999	2,16115426	2,28039866	
	Interacción	173946,172	97951	1,77584887	1,87383356	0,000
	Error	189447023	199900000	0,94770897		
Iter1	Entrecolumnas	21050,8233	49	429,608638	455,286713	
	Entrefilas	3244,89365	1999	1,62325845	1,720282	
	Interacción	138675,801	97951	1,41576707	1,50038868	0,000
	Error	188625682	199900000	0,94360021		
Iter2	Entrecolumnas	33741,6001	49	688,604083	733,148422	
	Entrefilas	2868,56929	1999	1,43500215	1,52782939	
	Interacción	110395,814	97951	1,12705142	1,19995799	0,000
	Error	187754556	199900000	0,9392424		
Iter3	Entrecolumnas	127810,163	49	2608,37067	2735,50651	
	Entrefilas	4631,75272	1999	2,31703488	2,42997059	
	Interacción	208407,808	97951	2,12767412	2,2313801	0,000
	Error	190609415	199900000	0,95352384		
Iter4	Entrecolumnas	28701,858	49	585,752205	616,823814	
	Entrefilas	5903,0283	1999	2,95299064	3,10963397	
	Interacción	254137,945	97951	2,59454161	2,73217077	0,000
	Error	189830326	199900000	0,94962644		
Iter5	Entrecolumnas	69185,5252	49	1411,94949	1501,57744	
	Entrefilas	3376,24518	1999	1,68896707	1,79617958	
	Interacción	160433,724	97951	1,63789777	1,74186849	0,000
	Error	187968131	199900000	0,94031081		
Iter6	Entrecolumnas	71726,3592	49	1463,80325	1534,12568	
	Entrefilas	4934,69511	1999	2,46858185	2,58717475	
	Interacción	201583,063	97951	2,05799903	2,15686716	0,000
	Error	190736831	199900000	0,95416123		
Iter7	Entrecolumnas	102574,679	49	2093,3608	2206,09673	
	Entrefilas	3911,53904	1999	1,95674789	2,06212667	
	Interacción	112394,076	97951	1,14745206	1,20924698	0,000
	Error	189684713	199900000	0,94889801		
Iter8	Entrecolumnas	123673,834	49	2523,95579	2625,10116	
	Entrefilas	5647,14253	1999	2,82498376	2,93819257	
	Interacción	240994,636	97951	2,46035912	2,55895591	0,000
	Error	192197836	199900000	0,96146991		
Iter9	Entrecolumnas	135907,347	49	2773,61933	2897,55782	
	Entrefilas	7908,41479	1999	3,95618549	4,13296665	
	Interacción	335282,452	97951	3,42296099	3,57591514	0,000
	Error	191349591	199900000	0,95722657		
Iter10	Entrecolumnas	157695,001	49	3218,26533	3340,29669	
	Entrefilas	7961,83245	1999	3,98290768	4,13393304	
	Interacción	345715,116	97951	3,52947001	3,66330175	0,000
	Error	192597035	199900000	0,96346691		

El efecto de la selección de sujetos, de reactivos y su combinación (interacción), está presente en todas las iteraciones. El cálculo de la Raíz del Error Cuadrático Medio muestra un patrón de ligero incremento. De este modo, si observamos los valores de las primeras iteraciones 0 (RECM iter 0= 0.4958) y 1 (RECM iter 1=0.5019), podemos comprobar la similitud de los mismos con los valores iniciales en las restantes condiciones experimentales (1, 2 y 3).

Tabla 51.

Error cuadrático medio e Intervalo de Confianza para las iteraciones de la condición experimental 4

	RECM	LI	LS
Iter 0	0,49580708	0,49066938	0,50094477
Iter 1	0,50188132	0,4967672	0,50699543
Iter 2	0,52419406	0,51887756	0,52951055
Iter 3	0,51788342	0,51256718	0,52319965
Iter 4	0,53499914	0,52999551	0,54000278
Iter 5	0,50941761	0,50430616	0,51452905
Iter 6	0,53055924	0,52495065	0,53616784
Iter 7	0,5408397	0,53562457	0,54605484
Iter 8	0,54175477	0,53628762	0,54722192
Iter 9	0,56754528	0,56268482	0,57240574
Iter 10	0,53967911	0,53475119	0,54460702

En la Figura 53, se muestra la evolución general de la tendencia en aumento del valor de la RECM en la cuarta condición experimental. De forma semejante a lo observado en la condición experimental 3, el incremento observado es muy ligero, sin embargo, si paramos a analizar iteraciones más distanciadas, comienzan a observarse algunas diferencias más importantes (RECM iter 1=0.5019 y RECM iter 9= 0.5675).

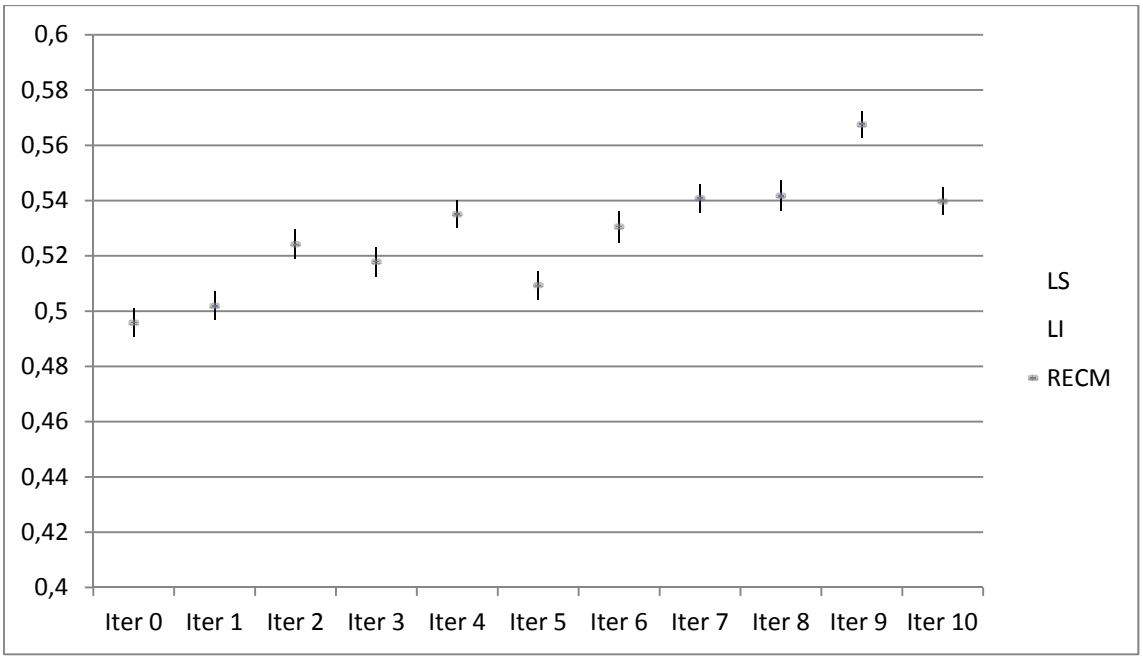


Figura 53. Error Cuadrático Medio e Intervalo de Confianza en las 11 iteraciones que forman parte de la cuarta condición experimental.

Fuente: Elaboración propia.

A pesar de tratarse de valores muy próximos, podemos observar cierto patrón de aumento, en este sentido, nos gustaría destacar, igual que hicimos en el análisis de resultados relativos a la condición experimental 3, que en la simulación de datos que nos ocupa se mantuvo el mismo nivel de habilidad para los dos grupos de sujetos evaluados, de este modo, las diferencias encontradas pueden ser fuertemente atenuadas por dicho factor. Sin embargo, en condiciones reales de aplicación, en las que el uso de ítems de anclaje se justifica por la necesidad de llevar a cabo un proceso de escalamiento vertical, cabría esperar que las diferencias entre los dos grupos en nivel de habilidad sean un elemento a considerar. En esta situación, se prevé una pauta de incremento más marcada, al contar con ítems con distinto nivel de precisión para los dos grupos que se desean comparar.

En cualquier caso, la recomendación esencial en relación a los ítems de anclaje que deben ser incluidos, iría en la línea de lo apuntado en el apartado 2.6 del presente trabajo, pues contar con reactivos representativos de todos los rangos de dificultad ha de ser uno de los criterios clave que guíen el proceso de elaboración de las pruebas. En el contexto de simulación analizado, cuando los reactivos exigen progresivamente mayor nivel de habilidad de los sujetos, se observa un ligero incremento en el efecto de

interacción. Este problema, puede verse acentuado en situaciones de escalamiento vertical en las que se dan diferencias de habilidad en los sujetos pertenecientes a los grupos a evaluar.

El cuidadoso análisis de la distribución de los parámetros de los ítems de anclaje, se presenta como un elemento decisivo en la mejora de los procesos de enlace. Las consecuencias prácticas de utilizar ítems cuya distribución no ha sido analizada en detalle, puede producir gran inestabilidad en los procesos de medida vinculados a la comparabilidad de puntuaciones.

6.6 Consideraciones generales.

Por último, nos gustaría concluir este apartado con un breve análisis comparativo del procedimiento bajo las cuatro condiciones experimentales expuestas, destacando una vez más que, la selección de los cuatro factores que definen cada una de las condiciones experimentales trabajadas no responde a un propósito de exhaustividad. Lejos de ello, cumple con el objetivo de mostrar una primera aproximación al procedimiento, atendiendo a los problemas más destacados a los que nos podemos enfrentar en los procesos de enlace de puntuaciones, problemas con especial vinculación con el efecto de interacción y que constituyen, por este motivo, variables independientes de especial relevancia en el marco de la investigación presentada.

Los resultados obtenidos, ponen de manifiesto la posible efectividad del procedimiento, no obstante, existen destacadas diferencias en relación a cada una de las condiciones. El gráfico de la Figura 54, representa los valores de la RECM en las condiciones experimentales 1 y 2. Como puede apreciarse, el patrón de crecimiento es muy similar en estas condiciones, observándose un incremento progresivo en el valor de la RECM similar para ambos casos. En las primeras iteraciones, en las que se presentan niveles más bajos de la variable independiente objeto de estudio (DIF/ diferencia entre grupos) los valores son muy próximos, sin embargo, a partir de la quinta iteración el Funcionamiento Diferencial del Ítem y las diferencias entre los grupos a comparar comienzan a tener una importancia algo más destacada. Las graves consecuencias de problemas de Funcionamiento Diferencial de de los ítems de anclaje se ven claramente

representadas en el gráfico puesto qué, el efecto de 15 reactivos con DIF en un test de 50 ítems, puede llevar a establecer verdaderas diferencias entre los grupos objeto de estudio. Las consecuencias del incumplimiento de los supuestos del modelo (unidimensionalidad/ DIF) así como las diferencias destacadas en el nivel de habilidad de los grupos que se desean evaluar, requieren especial atención en el diseño de pruebas y en la interpretación de los resultados. El bootstrap bidimensional nos permitiría llegar a estimar el componente de interacción "clave" en la detección de estas situaciones, pues si analizamos en exclusiva el error fruto del muestreo de sujetos o de reactivos, estaremos olvidando este componente crucial y, en consecuencia, pasando por alto situaciones que pueden invalidar el objetivo central perseguido en la evaluación.

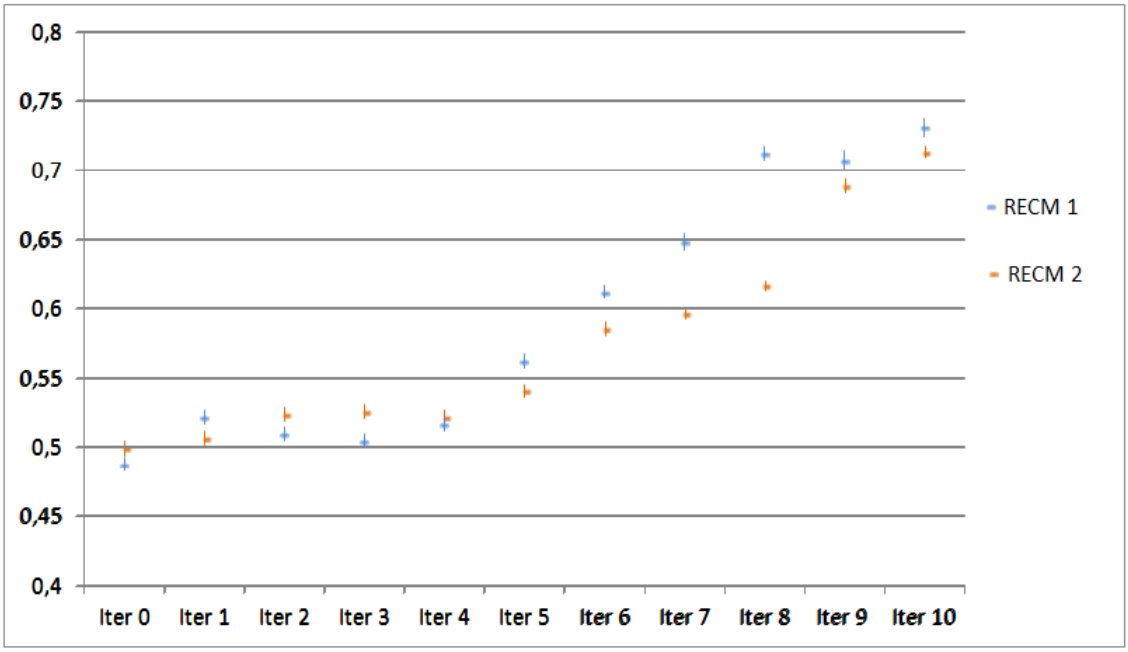


Figura 54. Error Cuadrático Medio e Intervalo de Confianza en las 22 iteraciones que forman parte de las condiciones experimentales 1 y 2.

Fuente: Elaboración propia.

Por otro lado, las condiciones experimentales 3 y 4 presentan un alto grado de similitud entre ellas, tal y como puede apreciarse en la Figura 55, el patrón de incremento en el valor de la RECM en ambas condiciones es muy leve, observándose un ligero incremento apenas perceptible. En ambos casos, la comparación entre iteraciones más distanciadas, nos permite observar el funcionamiento del procedimiento, puesto que, a pesar de tratarse de leves incrementos, el patrón de ascenso es un aspecto a destacar. La variación en las pautas distribucionales de b, tienen

el mismo efecto de forma independiente a si se trata de una modificación en sus valores promedios o en su dispersión. A pesar de que el incremento pudiera parecer poco reseñable, es importante advertir que, en condiciones reales de aplicación este efecto puede verse multiplicado si a ello se suman ciertas diferencias entre los grupos que se desean comparar.

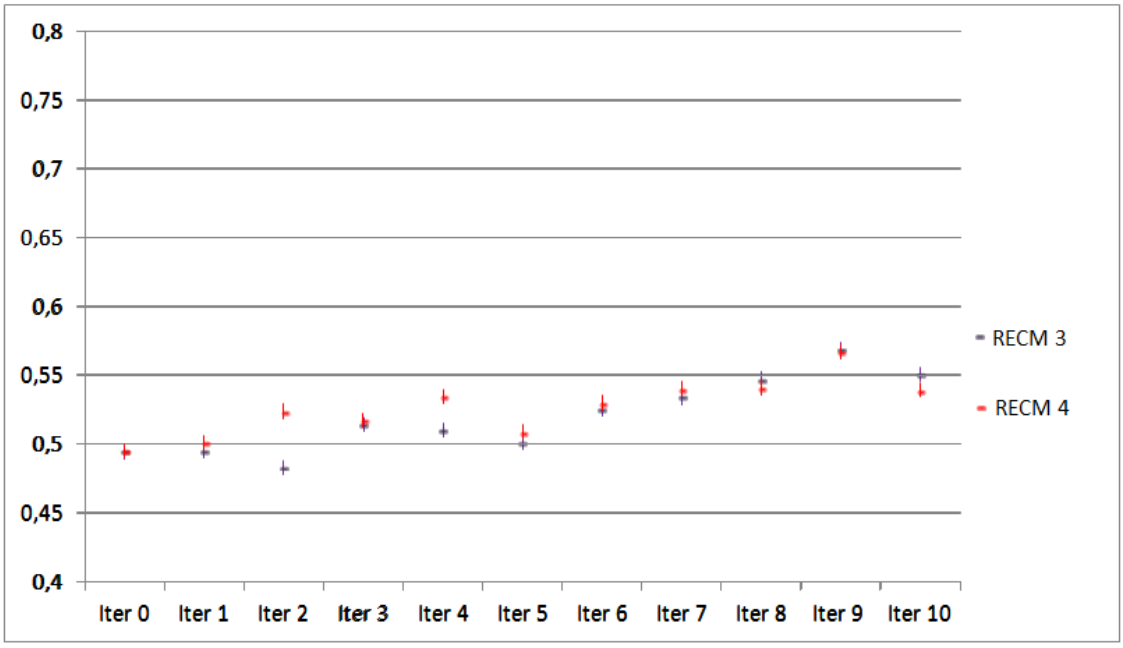


Figura 55. Error Cuadrático Medio e Intervalo de Confianza en las 22 iteraciones que forman parte de las condiciones experimentales 3 y 4.

Fuente: Elaboración propia.

La Figura 56, que agrupa los valores en las 4 condiciones experimentales, nos muestra una panorámica general del funcionamiento del procedimiento propuesto. En la Iteración 0 (iteración de partida en las cuatro condiciones experimentales) el valor de la RECM es prácticamente el mismo en todas ellas. A partir de la iteración 5, el funcionamiento entre condiciones comienza a distanciarse, observándose los patrones de incremento antes reseñados. En todas las condiciones experimentales, el funcionamiento del procedimiento en las primeras iteraciones es menos perceptible, presentando una mejora en su funcionamiento a partir de niveles medios del factor analizado.

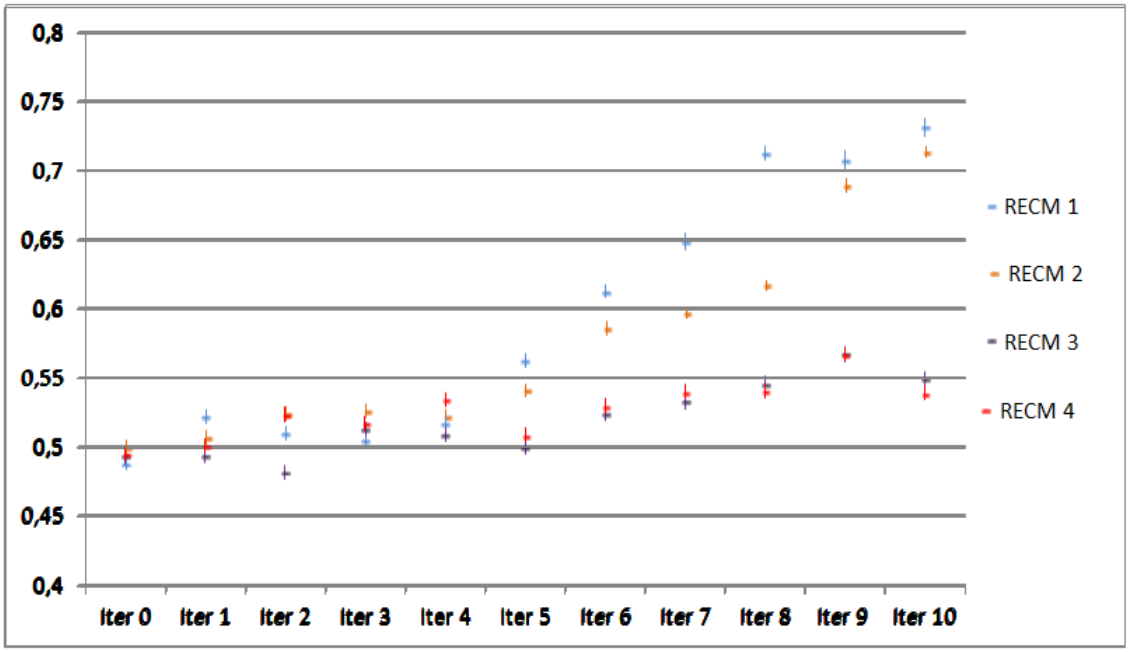


Figura 56. Error Cuadrático Medio e Intervalo de Confianza en las 44 iteraciones que forman parte de las condiciones experimentales 1, 2 3 y 4.

Fuente: Elaboración propia.

Tras este análisis pormenorizado de los resultados obtenidos, pasamos a analizar las hipótesis que pretendíamos contrastar de acuerdo al diseño de simulación de datos y análisis propuesto.

6.7 Contraste de las hipótesis de investigación.

En relación a las propiedades estadísticas globales del procedimiento propuesto, los resultados obtenidos en las cuatro condiciones experimentales, muestran la existencia de un efecto de interacción entre los factores principales estudiados (sujetos e ítems) efecto que no es reconocido en los procedimientos de estimación frecuentemente utilizados, los resultados del ANOVA muestran un efecto significativo de la interacción más allá del efecto significativo de los factores filas (selección de sujetos) y columnas (selección de reactivos). Podemos afirmar que, la interacción es una importante variable a considerar, al no poder estimarse a partir de la suma de los efectos de los factores principales por separado. La estimación del error, atendiendo al efecto de interacción, constituye una alternativa eficiente y realista, pues presenta una información mucho más completa y precisa.

En cuanto al Funcionamiento Diferencial del Ítem, los resultados muestran una respuesta óptima del procedimiento a la hora de valorar la violación de las asunciones estadísticas del modelo empleado. Gracias a la consideración del efecto de interacción, el procedimiento permite detectar funcionamiento diferencial, siendo sensible a la cantidad de DIF que presentan los datos. Otro resultado destacado es la confirmación de la utilidad del procedimiento, al funcionar sin la necesidad de elaborar hipótesis previas acerca del comportamiento de los datos, es decir, lejos de la necesidad de elaborar o definir hipótesis acerca de cuál será el comportamiento de los reactivos en relación a las muestras de sujetos participantes, el bootstrap bidimensional funciona sin la necesidad de establecer dichas hipótesis, alternativa que se ajusta de manera más realista a las condiciones reales de evaluación, en las que sería imposible definir todas las posibles fuentes de DIF que podrían estar presentes.

Si consideramos el funcionamiento del procedimiento en relación a las diferencias en el nivel de habilidad de los grupos, podemos afirmar que la estimación de la interacción permitirá detectar posibles situaciones de diferencias en nivel de habilidad entre los grupos a equiparar, situación frecuente en los estudios longitudinales, siendo un procedimiento sensible al grado de diferenciación en nivel de habilidad entre los grupos a equiparar.

Por último, en relación a las hipótesis relativas a la distribución del parámetro *b* de los ítems, podemos afirmar que el procedimiento muestra cierta eficiencia en condiciones de variación del potencial discriminador de la prueba en cuestión, atendiendo a cierta variedad en las propiedades técnicas del instrumento utilizado. Del mismo modo, muestra estabilidad en sus estimaciones a pesar de la variación en las propiedades específicas de la prueba utilizada. En relación a este conjunto de hipótesis debemos destacar que, los resultados obtenidos, pueden presentarse fuertemente atenuados por la simulación de grupos del mismo nivel de habilidad.

CONCLUSIONES, LIMITACIONES Y PROSPECTIVA

En este último apartado se recogen las principales conclusiones derivadas del presente trabajo de tesis doctoral, cuyo objetivo central ha sido el diseño, puesta en práctica y evaluación de un procedimiento para el análisis del efecto de interacción de la selección de sujetos y reactivos en los procesos de enlace de puntuaciones, denominado bootstrap bidimensional. A fin de estructurar de manera clara este apartado, y con el objetivo de dar debida respuesta a los interrogantes inicialmente planteados, las conclusiones aquí expuestas se estructurarán en torno a los mismos, sintetizando de un modo global el análisis detallado recogido en los capítulos que integran este trabajo.

En segundo lugar, se presentan las principales limitaciones detectadas así como las futuras líneas de investigación, pues se trata de un novedoso ámbito de trabajo que demanda nuevos proyectos que profundicen en los hallazgos más destacados y atiendan a las principales necesidades detectadas.

¿Cuál es la importancia de la evaluación en el sistema educativo actual?

El presente trabajo de tesis doctoral nos ha permitido dibujar un panorama "global" de la evaluación educativa, comprobando su fuerte influjo en la sociedad actual a muy distintos niveles. En el ámbito nacional, una revisión de las últimas leyes educativas pone de manifiesto el «*boom»* evaluativo ante el que nos encontramos, observándose una mención explícita creciente acerca de la importancia de la evaluación en la legislación educativa, con especial incidencia en las leyes de la última década, desde la Ley Orgánica de Educación (2002) hasta la Ley Orgánica para la Mejora de la Calidad Educativa (2013). Precisamente, es en este último mandato legislativo (LOMCE, 8/2013), donde encontramos importantes alusiones a la evaluación de aprendizajes y competencias; De este modo, desde su preámbulo, se apunta a la justificación de esta nueva normativa en base a los resultados de España en las evaluaciones internacionales en las que ha participado, destacando las evaluaciones externas de final de etapa como una de las novedades más importantes destinadas a mejorar la calidad del sistema educativo.

Más allá de nuestras fronteras, el horizonte internacional también muestra un interés generalizado y creciente por lograr una evaluación educativa de calidad dirigida a distintos propósitos. De este modo, desde mediados del siglo XX, la atención internacional en el ámbito de la evaluación educativa ha ido en aumento, consolidándose una compleja red evaluativa auspiciada por distintos organismos e instituciones internacionales (privadas, públicas, intergubernamentales, de cobertura mundial o regional), entre las que destacan las llevadas a cabo por la «*International Association for the Evaluation of Educational Achievement»* (IEA), la «*Organization for Economic Cooperation and Development»* (OECD), el «*Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación»* (LLECE) o el «*Southern and Eastern Africa Consortium for Monitoring Educational Quality»* (SACMEQ).

Comprender, analizar y mejorar las condiciones en las que se realizan estas evaluaciones, supone una necesidad a la que ha de dar respuesta la investigación educativa actual, enfrentándose a la creciente demanda de contar con evaluaciones realizadas con precisión, rigurosidad y cuyos resultados informen fielmente de la

situación y progreso de los sistemas educativos en general y de los estudiantes en particular.

¿Cuáles son las principales exigencias a las que se enfrenta la evaluación educativa?

Son numerosos los retos a los que se enfrenta la evaluación educativa en nuestros días, entre las primeras necesidades detectadas se encontrarían las derivadas del denominado proceso de evaluación por competencias, proceso que implica tener en cuenta que la finalidad es evaluar el desarrollo de conocimientos, actitudes y destrezas a distintos niveles, e informar a la administración educativa, a profesores, a padres y a alumnos del grado de consecución de los objetivos educativos propuestos. Este enfoque por competencias lleva implícita la consideración de su carácter continuo, con diversos niveles o estadios de desarrollo, su contextualización y su interdisciplinariedad. Todo ello exige la adopción de procedimientos de evaluación que permitan considerar tales características, dichos procedimientos implican modificaciones relativas a la construcción de los ítems, diseño de las pruebas, diseños de recogida de información, construcción de escalas, procedimientos de análisis de datos, etc. La complejidad inherente al constructo "competencias" implica procesos evaluativos eminentemente más complejos.

En esta misma línea, el amplio dominio competencial que las diferentes evaluaciones educativas pretenden abarcar, supone un importante desafío. En las evaluaciones educativas, la intención es evaluar un amplio rango del dominio de interés, por consiguiente deberían incluirse gran cantidad de reactivos en una misma prueba si se pretende abarcar dicho dominio de forma exhaustiva. La aplicación de tales pruebas sería compleja, puesto que los estudiantes deberían contestar a instrumentos excesivamente largos, poniendo en riesgo las propiedades psicométricas de los mismos. En consecuencia, evaluaciones educativas como TIMSS, PIRLS o PISA implementan un sistema de "bloques" que permiten abarcar un amplio rango del dominio evaluado sin que sea necesario construir formas excesivamente largas, de este modo, bloques de anclaje rotativos permiten la comparación entre los cuadernillos de un mismo curso o nivel, el reto de la comparabilidad o enlace de puntuaciones tendría en este apartado una de sus principales implicaciones.

Unido a ello, la seguridad en el proceso de evaluación exige la elaboración de diferentes formas de un mismo instrumento, formas que han de ser debidamente enlazadas si se pretende garantizar tanto la seguridad como la calidad del proceso. Por otro lado, cambios curriculares o cambios en el dominio evaluado dificultarían la posibilidad de comparación entre distintas ediciones de una misma prueba, exigiendo la puesta en marcha de diferentes técnicas de enlace de puntuaciones pues, tal y como apunta Brennan (2007), una de la ironías de la medición educativa es que estos cambios en los programas de evaluación (incluso cuando dichos cambios suponen una mejora) ponen hasta cierto punto en peligro la comparabilidad de las puntuaciones.

Por otro lado, una única toma de datos no nos informará del progreso en el aprendizaje, siendo necesario contar con medidas repetidas que nos permitan un análisis del cambio, lo que implica el reconocimiento del aprendizaje como una realidad en continua transformación y evolución, en la que no caben concepciones simplistas de carácter estático.

En último lugar, en el presente trabajo ha quedado reconocida la importancia que en las evaluaciones a gran escala tiene la posibilidad de comparación con referentes de interés tales como escuelas, ediciones, distritos, estados, países, etc., pues es necesario poder detectar y analizar las mejores prácticas educativas identificando líneas de cambio y mejora exitosas.

En conclusión, la comparabilidad se postula como una exigencia transversal que va más allá de consideraciones particulares para erguirse como un reto central en la evaluación educativa, permitiendo afrontar gran variedad de desafíos asociados.

¿Qué implica la comparabilidad?

A pesar de su importancia y sus fuertes repercusiones en la calidad de los procesos de medición educativa y psicológica, la comparabilidad de puntuaciones, no ha sido un tema tratado en profundidad desde la psicometría clásica. De este modo, el presente trabajo ha perseguido aportar una clarificación conceptual en torno a los principales términos utilizados en el ámbito de la comparabilidad o enlace de

puntuaciones, creando un marco de trabajo adecuado que permita un punto de partida óptimo para su investigación en profundidad.

Como aspecto central dentro de esta clarificación conceptual, se encuentra la definición y distinción de los términos equiparación y escalamiento, atendiendo a sus características particulares y a su papel dentro de la evaluación educativa, entendiendo ambos conceptos como parte de la comparabilidad o enlace de puntuaciones pero con características y objetivos claramente diferenciados. De este modo, el exhaustivo análisis presentado en el capítulo 2 del presente trabajo, ha permitido afrontar uno de los problemas centrales en torno al enlace de puntuaciones, la ausencia de una teoría de base satisfactoria.

Del mismo modo, el amplio abanico de procedimientos estadísticos que permiten la obtención de puntuaciones comparables, produce cierta confusión en la investigación sobre esta temática. El capítulo 3 de la presente tesis doctoral, ha realizado un completo recorrido por estos procedimientos estadístico, observando cómo ciertamente las mayores diferencias entre los diferentes tipos de enlace de puntuaciones, son de tipo interpretativo y no procedimental. Por tanto, la información recogida en los capítulos 2 y 3 ha permitido atender a uno de los desafíos más acuciantes en el ámbito de la comparabilidad, disponer de una teoría de base satisfactoria en la que se integren conceptos, procedimientos y aplicación.

¿Qué factores pueden incidir en los procesos de enlace de puntuaciones?

La posibilidad de comparar puntuaciones, componente esencial en evaluación educativa, exige fuertes asunciones metodológicas, así como complejos diseños de recogida y análisis de datos. En función de los objetivos perseguidos, cobrarán mayor importancia fuentes particulares de error.

El componente de error por el que se ve afectado cualquier proceso de enlace, puede descomponerse en dos elementos: *error sistemático* y *error aleatorio*. El error sistemático se produce normalmente cuando existen fallos en las asunciones del modelo o método seleccionado y, el error aleatorio, está causado por las características muestrales.

De este modo, algunos factores que pueden causar variabilidad en los procesos de escalamiento y equiparación basados en TRI, podrían ser diferencias entre grupos de examinados, diferencias en ítems, diferencias culturales, dimensionalidad, problemas de aplicación, problemas en el diseño, insuficiencia de los ítems de anclaje utilizados, definición de las escalas verticales, etc. De manera independiente a las fuentes particulares de error que podamos enumerar, el objetivo de este trabajo ha ido más allá, al considerar la necesidad de estimar el error de enlace utilizando procedimientos adecuados que trasciendan las fuentes particulares que puedan ser causa de ello.

¿Cuál es la perspectiva tradicional en la estimación del error de enlace?

Hasta la fecha, evaluar los resultados de un proceso de enlace de puntuaciones resulta difícil, al no estar disponible ningún criterio estándar que permita investigar los sesgos potenciales de los resultados de dicho enlace (Jiang, von Davier, & Chen, 2012). A pesar de los avances desarrollados en los últimos años, con un volumen creciente de investigación al respecto, los trabajos desarrollados son todavía escasos, especialmente en lo relativo al error sistemático. La forma convencional de calcular el denominado error estándar de equiparación tan solo captura la varianza fruto del muestreo de sujetos (error aleatorio).

Por tanto, nos encontramos ante una perspectiva tradicional limitada a una estimación del error de enlace basada en la selección de los sujetos que forman parte de la muestra. Entre los estudios más novedosos, destacan aquellos que consideran el error asociado a la selección de los ítems (Monseur & Berezner, 2007; Haberman, Lee, & Qian, 2009; Zu & Liu, 2010). En cualquier caso, la aportación esencial del presente trabajo ha sido la consideración de ambos factores de manera combinada, aludiendo a la repercusión de la interacción entre la selección de sujetos y reactivos y su importancia en la cuantificación del error. Los procesos actuales de estimación se limitan a detectar una pequeña proporción del error de enlace real, proporción relativa a la selección muestral de sujetos o reactivos, olvidando la existencia de una posible interacción ente ambos elementos. De este modo, si consideramos los trabajos realizados en el ámbito de la estimación del error, observamos cómo estos son enfocados de manera mayoritaria a la medida del error debido al muestreo de sujetos y, en menor medida, al error debido al

muestreo de ítems, dejando a un lado el elemento clave considerado en éste trabajo "la interacción" entre ambos factores.

¿Existe efecto de interacción en la selección de sujetos y reactivos?

Tal y como se ha representado en el apartado 5.2.3 de la memoria que nos ocupa, la superficie característica del test en una situación de no interacción entre la selección de sujetos y reactivos, quedaría representada por un plano inclinado (ver Figura 22), ya que $P(\theta) = a + b\theta$. Sin embargo, los resultados del Análisis de Varianza en las cuatro condiciones experimentales, muestran que, en todas las iteraciones (44 en total), existe efecto significativo de la interacción (a un nivel de confianza del 99%). En consecuencia, si representamos la superficie característica del test, esta se aproximaría con mayor fidelidad a la reflejada en la Figura 24. Ello supone una evidencia de la imposibilidad de que exista una relación lineal entre la variable latente (-∞ a +∞) y la probabilidad (0-1).

El planteamiento tradicional en el que se analizan por separado el efecto de la selección de sujetos y reactivos, no se ajusta a situaciones reales de evaluación, pues existe una combinación de los efectos de ambos factores que difiere de la suma de los mismos de forma independiente. Por tanto, el presente trabajo ha permitido justificar desde un punto de vista teórico y empírico la necesidad de considerar el efecto de interacción.

Cabe destacar que, a pesar de que el procedimiento implementado ha surgido en el marco del estudio de los procedimientos de enlace de puntuaciones, tras el desarrollo completo de la investigación se observa que, ésta situación, estaría presente en cualquier evaluación que exija una selección muestral de sujetos y de ítems, no estando estrictamente vinculada a los procesos de enlace. Así, el estudio de un diseño con ítems comunes, nos ha permitido precisamente avalar esta realidad, apuntando a su utilidad para la estimación del error en contextos más amplios de evaluación, y no sólo en el ámbito del enlace de puntuaciones. Este hallazgo inesperado, sugiere futuras vías de análisis.

¿Es posible mejorar los procesos de enlace de puntuaciones teniendo en cuenta el efecto de interacción en la selección de sujetos e ítems?

Entendiendo que la mejora de los procesos de enlace, es uno de los principales retos a los que nos enfrentamos en la denominada "era de la evaluación", podemos considerar que, uno de los resultados más destacados del presente trabajo de tesis doctoral, es la propuesta de un procedimiento para el análisis de la interacción de sujetos y reactivos en procesos de equiparación y escalamiento (generalizable a otras situaciones de evaluación), procedimiento que hemos denominado «bootstrap bidimensional», y cuya justificación teórica y práctica pone de manifiesto su utilidad a la hora de mejorar la precisión en la medida de la habilidad de los sujetos, permitiendo una estimación más ajustada del error en general y del error de enlace en particular. Esta cuantificación, ayudará tanto a la valoración de los procesos generales de enlace como a la toma de decisiones acerca del diseño de recogida de datos, el tamaño de la muestra, los ítems a incluir en los test de anclaje, etc.

¿Cómo estimar la interacción entre dichas fuentes de error?

Precisamente, el marco general del problema de investigación trabajado en la presente tesis doctoral, ha sido definido como el planteamiento y evaluación de una propuesta metodológica que permita la estimación del efecto de interacción entre la selección de sujetos y reactivos en procesos de enlace de puntuaciones. En concreto, el método propuesto es el que hemos denominado bootstrap bidimensional, técnica intensiva de remuestreo en la que, las unidades de remuestreo son dobles (filas y columnas), en nuestro caso sujetos y reactivos.

La aplicación de este doble procedimiento de remuestreo a la matriz de datos, y la estimación de la habilidad de cada sujeto en cada una de las submuestras resultantes, nos ha permitido analizar, a través del Análisis de Varianza, la existencia del efecto de cada factor (sujetos y reactivos), así como el efecto de interacción entre ambos factores. Al mismo tiempo, la estimación de la Raíz del Error Cuadrático Medio, a partir de los datos extraídos de la aplicación del bootstrap bidimensional, ha permitido combinar los errores sistemático y aleatorio de enlace, errores tratados de manera independiente desde una perspectiva clásica.

La sintaxis para la puesta en práctica del procedimiento, ha sido elaborada en el programa R, bajo su versión 3.12. Dicha sintaxis destaca por el tratamiento óptimo de los datos de entrada, por posibilitar la combinación de las matrices *i* y *j,* así como por la claridad en el diseño de salida de los resultados, pues la compleja combinación de un procedimiento de remuestreo doble, en el ámbito de la TRI, hace imprescindible trabajar con estrategias de análisis eficientes, que minimicen el coste temporal del procedimiento de análisis frente a las ganancias que los resultados obtenidos reportan.

¿Resulta de utilidad el estudio de la interacción en el análisis de ítems comunes?

El estudio de la interacción se presenta como un elemento esencial en el análisis de los ítems comunes o ítems de anclaje. Tal y como se recoge en diversos apartados del presente trabajo, el diseño con ítems comunes es el diseño de enlace más frecuente, tanto en procesos de escalamiento como de equiparación. De este modo, hemos querido poner especial énfasis en la importancia de analizar el efecto de interacción en estos reactivos, pues las propiedades específicas de los mismos, su calidad, su suficiencia, su estabilidad, etc. pueden condicionar el proceso de enlace de puntuaciones. La utilización de diferentes conjuntos de ítems comunes genera diferencias en las puntuaciones, incluso cuando la muestra de sujetos es grande (Monseur & Berezner, 2007). Estas diferencias, son fruto tanto de la selección de reactivos en sí, como de la interacción del efecto de los mismos con los sujetos que contestan a la prueba.

En consecuencia, estudiar el efecto de interacción precisamente en estos reactivos tiene un importante valor informativo, que nos ayudará a tomar decisiones cruciales acerca de la calidad del proceso de enlace llevado a cabo, así como a elaborar interpretaciones más ajustadas a la realidad.

¿En qué condiciones muestra su efectividad el procedimiento propuesto?

Numerosos factores podrían estar relacionados con el efecto de interacción, nótese que, la presente tesis doctoral no ha perseguido analizar de manera exhaustiva todos los posibles factores relacionados con la interacción, sin embargo, se ha presentado una propuesta óptima para su detección así como su funcionamiento bajo

ciertas condiciones, atendiendo a determinados factores que podrían tener una importancia destacada.

En relación a los factores analizados, el procedimiento implementado nos ha permitido analizar 3 de las 4 fuentes de error sistemático que señalan Kolen y Brennan (2014), las relativas a la violación de las asunciones estadísticas del modelo (unidimensionalidad), problemas en la puesta en práctica del diseño de recogida de la información (diferencias en el funcionamiento de los ítems en los dos grupos a comparar debido, por ejemplo, a efectos de posición) y la existencia de diferencias sustanciales en la conducta de los grupos a equiparar. El procedimiento propuesto, ha presentado un funcionamiento muy adecuado en tales situaciones, presentándose como un eficiente detector de DIF o de diferencias entre los grupos que se desean enlazar.

Por otro lado, en el ámbito de las evaluaciones internacionales, autores como Monseur y Berezner (2007), sugieren la utilización del procedimiento Jackknife para el remuestreo de ítems realizando análisis diferenciados por país, debido a posibles diferencias en el funcionamiento de los ítems en cada país que pudieran alterar el cálculo del error de enlace. El procedimiento bootstrap bidimensional permitiría la medida de este efecto de interacción, sin necesidad de formular hipótesis previas sobre el funcionamiento de los datos, creando divisiones a priori que no siempre se ajustarían a la realidad evaluada.

El alto grado de similitud detectado en el funcionamiento del procedimiento bajo las condiciones experimentales 3 y 4, muestra un efecto moderado de la variación en los patrones distribucionales de b. Cabe destacar que el efecto es similar de forma independiente a si se trata de una modificación en sus valores promedio o en su dispersión. A pesar de tratarse de un incremento poco reseñable, es importante apuntar que, en condiciones reales de evaluación, este efecto puede verse multiplicado si a ello se suman ciertas diferencias entre los grupos que se desean comparar.

Entre los factores más destacados, se encontraría el funcionamiento diferencial del ítem. El procedimiento presentado, además de permitir su detección, se ha mostrado sensible a la cantidad de DIF presente en cada una de las iteraciones. Unido a ello, uno de los aspectos más relevantes, es la utilidad del procedimiento al no depender de la

formulación de hipótesis previas acerca del funcionamiento de los reactivos, es decir, su eficacia práctica se ve incrementada ante la posibilidad de poder estudiar el funcionamiento de los reactivos más allá de las expectativas previas acerca del comportamiento de los mismos.

En consecuencia, la consideración del efecto de interacción, obviada en los procedimientos clásicos, se presenta como un componente imprescindible, a la luz de los resultados obtenidos en el análisis de la aplicación del procedimiento ante las cuatro condiciones experimentales trabajadas en la presente tesis doctoral. El procedimiento aquí presentado, supone una interesante propuesta para la estimación del efecto de interacción, abordando un terreno escasamente explorado y con fuertes implicaciones en la práctica evaluativa actual. La versatilidad de la propuesta, permite su aplicación ante situaciones muy variadas de enlace de puntuaciones, con diferentes diseños de recogida de información, procedimientos de enlace, métodos de estimación, tamaños muestrales, etc.

LIMITACIONES Y PROSPECTIVA

El presente estudio supone una aproximación inicial al procedimiento propuesto y, lejos de concluir la misma con una respuesta completa y definitiva a los interrogantes suscitados durante su desarrollo, finalizamos apuntando las principales limitaciones encontradas así como las futuras líneas de trabajo.

En primer lugar, la ausencia de una teoría de base satisfactoria en el ámbito del enlace de puntuaciones, ha supuesto un gran reto, especialmente si consideramos su análisis desde el punto de vista de la investigación educativa. Las particulares necesidades de la educación, exigen una atención específica desde dicho ámbito, pues es necesario dar respuesta a la compleja realidad que entraña la medida en educación. Por tanto, una perspectiva integrada, que aúne los esfuerzos de diversas áreas, será la más conveniente a la hora de afrontar nuevos desafíos.

A pesar de haber conseguido una solución satisfactoria ante la complejidad inherente al procedimiento de análisis propuesto, existen algunas limitaciones sobre las que sería preciso trabajar. En concreto, el tiempo de ejecución en ordenadores

convencionales sería una de las principales áreas de mejora, puesto que, a pesar de las últimas modificaciones para su reducción, consideramos que es posible optimizar los procesos de cálculo, implementando procedimientos verdaderamente eficientes que permitan la estimación en espacios temporales más razonables, sin necesidad de dividir el proceso y con la ventaja de disponer de forma rápida de los resultados. Los apresurados avances en las tecnologías de computación, así como la depuración del procedimiento propuesto, permitirán sustanciales mejoras en futuras versiones.

Puesto que uno de los hallazgos más destacados de la presente investigación es la importancia de la estimación del efecto de interacción en el Funcionamiento Diferencial del Ítem, analizar el comportamiento del procedimiento ante la manipulación de distintas condiciones de DIF, como el porcentaje de ítems con DIF, el tipo de DIF, el tamaño muestral de los grupos, el formato de respuesta de los ítems, etc., así como la comparación de este procedimiento con otros métodos de detección (Mantel Haenszel, Regresión logística, etc.), constituyen futuras líneas de investigación que es preciso destacar. No cabe duda en relación a la necesidad de seguir investigando de cara a valorar la idoneidad y ajuste del procedimiento implementado.

Del mismo modo, teniendo en cuenta que la multidimensionalidad de las pruebas es otra de las condiciones que pueden atentar seriamente contra la calidad de los procesos de enlace de puntuaciones, y con especial incidencia en el escalamiento vertical, sería necesario estudiar de forma más completa el funcionamiento del bootstrap bidimensional en la detección de tales situaciones pues, en la presente investigación, tan solo se ha realizado una pequeña aproximación al mismo desde la perspectiva de una concepción multidimensional del DIF.

A la luz de los resultados obtenidos en relación a la existencia de diferencias en el nivel de habilidad de los sujetos cuyas puntuaciones se desean comparar, sería interesante estudiar cómo funciona este procedimiento en el análisis de cadenas de escalamiento, puesto que, las marcadas diferencias en el nivel de habilidad de los grupos implicados, suelen ser una dificultad frecuente que puede suponer un incremento exponencial en la estimación del error de enlace, incremento más destacado cuanto mayores sean las diferencias en nivel de habilidad dentro de la cadena de escalamiento.

El moderado efecto detectado en relación a la variación en los patrones distribucionales de b (dispersión/media), hace necesario pensar en análisis más ajustados a condiciones reales de evaluación, en los que se estudie este efecto combinado con variables como las posibles diferencias entre los grupos a evaluar, pues ciertamente puede existir un efecto multiplicativo que sitúe la importancia de la distribución de los valores de b en condiciones de aplicación más realistas.

El procedimiento de estimación presentado en estas páginas, puede adaptarse a innumerables diseños y condiciones de evaluación. A pesar de haber optado en este trabajo de investigación por el análisis de los ítems comunes, una interesante vía de desarrollo sería el estudio de su funcionamiento ante diferentes situaciones de enlace (diseños, procedimientos, número de sujetos, número de ítems, tipo de anclajes, etc.). Del mismo, los resultados obtenidos evidencian la importancia de considerar el efecto de interacción en la estimación del error en el marco de cualquier evaluación en la que sea necesaria la selección muestral de sujetos o de reactivos, avalando su uso más extendido, no limitado a condiciones particulares de escalamiento o equiparación. En esta línea, evaluar el procedimiento en situaciones de evaluación que no impliquen enlace de puntuaciones, se evidencia como una interesante cuestión para futuros trabajos.

Por último, tal y como manifestábamos en la descripción del método, los estudios de simulación, pretenden reproducir características y comportamientos reales y, al mismo tiempo, ejercer un control sobre factores que pueden resultar objeto de interés, gracias a la determinación a priori de la situación a analizar. La realización de un estudio de simulación queda plenamente justificada de acuerdo al objetivo central del presente trabajo, no obstante, el riesgo derivado de que las condiciones del estudio de simulación sean poco realistas y, en consecuencia, se produzca una limitación en la generalización de sus resultados (validez externa) (Revuelta & Ponsoda, 2003), haría recomendable su aplicación a datos reales tras su diseño y justificación en entornos controlados de simulación. En esta línea, las Thetas originales para el cálculo del Error Cuadrático Medio, podrían ser consideradas a partir de la respuesta de los sujetos a la forma completa (sin contar con los ítems de anclaje). Es decir, el valor del ECM podría calcularse en relación a la diferencia de la Theta original (entendida como la theta estimada a partir de los resultados de la respuesta de los sujetos a los ítems que

componen la prueba) y la Theta estimada a partir de las puntuaciones de los sujetos en los ítems que componen el test de anclaje.

CONCLUSIONS, LIMITATIONS AND OUTLOOK

This final section contains the main conclusions drawn from this PhD thesis study, whose core objective was to design, put into practice and evaluate a procedure for analysing the effect of interaction of the selection of subjects and tasks in score linking processes, known as "two-dimensional bootstrap". In order to structure this section clearly the conclusions discussed here are structured around the issues initially outlined in all overall summary of the detailed analysis contained in the chapters making up this study.

Secondly, the main limitations detected are presented, along with future lines of research, as this is a novel field of work that demands new studies that delve further into the major findings and address the main needs detected.

What is the importance of evaluation in the current education system?

This PhD thesis study has enabled us to sketch a "global" panorama of educational evaluation, confirming its strong influence on today's society at very different levels. In the national scene, a look at the latest education laws highlights the evaluation boom facing us and there is a growing explicit mention of the importance of evaluation in education legislation, with a special emphasis on laws passed during the

last decade, from the Organic Law on Education (LOE, 2002) to the Organic Law for the Improvement of Education Quality (LOMCE, 2013). It is precisely in the latter legislative change (LOMCE, 8/2013) where we find important allusions to evaluating learning and competencies; in fact, from its preface it talks about justifying this new law based on Spain's results in the international evaluations it has taken part in, highlighting end of stage external evaluations as one of the most important new features aimed at improving the quality of the education system.

Beyond our borders, the international community is also showing a generalised and growing interest in achieving quality educational evaluation directed at different purposes. So, from the mid-20th century, international attention in the field of Educational Evaluation has been on the rise, consolidating a complex evaluation network backed by various international bodies and institutions (private, public, intergovernmental, with worldwide or regional scope), including those set up by the *International Association for the Evaluation of Educational Achievement* (IEA), the *Organization for Economic Cooperation and Development* (OECD), the *Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación* (LLECE) and the *Southern and Eastern Africa Consortium for Monitoring Educational Quality* (SACMEQ).

Understanding, analysing and improving the conditions under which these evaluations are carried out is a need to which today's education research must respond, tackling the growing demand for evaluations carried out with precision, rigour and with results informing accurately of the situation and progress of education systems in general and of students in particular.

What are the main demands facing Educational Evaluation?

Educational evaluation is facing many challenges at the moment, and the top needs detected are those arising from the so-called competency evaluation process, which involves taking into account that the end purpose is to evaluate the development of knowledge, attitudes and skills at different levels, and inform the education administration, teachers, parents and students of the extent to which the proposed educational objectives have been met. This focus on competencies implicitly entails

considering its continuous nature, with various levels or phases of development, its contextualization and its interdisciplinarity. All this requires the adoption of evaluation procedures that allow these features to be considered. These procedures involve changes relating to constructing items, designing the tests, designs for information collection, constructing scales, data analysis procedures, etc. The inherent complexity of the "competencies" construct involves eminently more complex evaluation processes.

In this same line, the broad areas of competency that the various educational evaluations aim to address present a major challenge. In educational evaluations, the intention is to evaluate a wide range of the domain of interest, so a large quantity of tasks should be included in the same test if this domain is to be thoroughly dealt with. Applying these tests would be a complex exercise, as students would have to answer excessively long tests, thereby putting the psychometric properties of these tests at risk. Consequently, educational evaluations such as TIMSS, PIRLS and PISA implement a system of "blocks" allowing a wide range of the domain being evaluated to be covered without the need to construct excessively long forms, using rotating anchor blocks to enable comparisons to be drawn between the exercise books of the same year or level. In this case, the challenge of compatibility or score linking would be one of the main implications.

Together with this, security in the evaluation process requires producing different versions of the same instrument, versions that have to be duly linked together if both the security and the quality of the process are to be guaranteed. On the other hand, changes to the curriculum or changes in the domain evaluated would hinder the possibility of drawing comparisons between different editions of the same test, requiring various score linking techniques to be set up; as Brennan (2007) noted, one of the ironies of educational measurement is that these changes to evaluation programmes (even when such changes lead to improvement) endanger the comparability of scores to a certain extent.

However, a single data capture will not tell us about learning progress. Repeated measurements are necessary that enable us to analyse change, which involves recognising learning as a continual process of change and development, in which there is no room for simplistic and static conceptions.

Lastly, this study recognises the importance in large scale evaluations of the possibility of comparison based on factors of interest such as schools, editions, districts, states, countries, etc.. as we need to be able to detect and analyse best educational practices by identifying successful cases of change and improvement.

In conclusion, comparability is proposed as a transversal requirement that goes beyond particular considerations to stand as a central challenge in educational evaluation, enabling a huge variety of associated challenges to be tackled.

What does comparability involve?

Despite its importance and its major repercussions on the quality of educational and psychological measuring processes, the topic of score comparability has not been addressed in any detail from the point of view of classic psychometrics. This study has pursued a conceptual clarification of the main terms used in the sphere of comparability or score linking, creating a suitable framework that serves as an ideal starting point for in-depth research of this issue.

A central feature of this conceptual clarification is the definition of the terms equating and scaling, according to their particular characteristics and their role within educational evaluation, both understood as part of comparability or score linking but with clearly differentiated characteristics and objectives. The thorough analysis presented in chapter 2 of this study has enabled one of the main problems with score linking to be addressed, that of the absence of a satisfactory base theory.

Likewise, the wide range of statistical procedures that allow comparable scores to be obtained causes a certain amount of confusion in research on this topic. Chapter 3 of this PhD thesis goes through all these statistical procedures, observing how the greatest differences between the different types of score linking are actually interpretative rather than procedural. Therefore, the information contained in chapters 2 and 3 has enabled one of the most pressing challenges in the sphere of comparability to be addressed, that of having a satisfactory base theory that integrates concepts, procedures and application.

What factors can affect the score linking process?

The possibility of comparing scores, an essential component in educational evaluation, requires strong methodological assumptions plus complex designs for data collection and analysis. Depending on the objectives pursued, particular error sources will acquire greater importance.

The error component by which any linking process is affected can be broken down into two elements: *systematic error* and *random error*. Systematic error usually arises when there are mistakes in the selected model or method and random error is caused by sample characteristics.

So, some factors that can cause variability in the scaling and equating processes based on IRT may be differences between groups of examinees, differences in items, cultural differences, dimensionality, application problems, design problems, failure of the anchoring items used, definition of vertical scales, etc. Apart from the particular sources of error that might be listed, the objective of this study was to go further, to consider the need to estimate the scaling error by using suitable procedures that transcend the particular sources that can cause it.

What is the traditional view on estimating linking error?

To date, evaluating the results of a score linking process has been difficult, as there are no standard criteria available enabling the potential biases of the results of that link to be investigated (Jiang, von Davier, & Chen, 2012). In spite of progress made over recent years, with a growing volume of research on the topic, there is still a shortage of work being produced, especially on the subject of systematic error. The conventional way of calculating the so-called standard equating error only captures the variance resulting from subject sampling (random error).

Therefore, this is a traditional view limited to an estimation of linking error based on the selection of subjects making up the sample. The most innovative studies include the ones that discuss the error associated with item selection (Monseur & Berezner, 2007; Haberman, Lee, & Qian, 2009; Zu & Liu, 2010). In any case, the

essential contribution of this study is the combined consideration of both factors, alluding to the repercussion of the interaction between the selection of subjects and items and its importance for quantifying error. Current estimation processes are limited to detecting a small proportion of the real linking error, a proportion relative to sample selection of subjects and items, forgetting the existence of a possible interaction between both elements. If we consider the studies done in the field of error estimation, we can see how they mostly focus on measuring the error caused by subject sampling and, to a lesser extent, the error caused by item sampling, neglecting the key element of "interaction" discussed in this study.

Is there an interaction effect in the selection of subjects and items?

As conveyed in section 5.2.3 of the report under discussion, the characteristic surface of the test in a situation of non-interaction between subject and items selection, would be represented by an inclined plane (see Figure 22), as $P(\theta) = a + b\theta$. However, the results of the Variance Analysis in the four experimental conditions show that in all iterations (44 in total) there is a significant effect of the interaction (at a confidence level of 99%). Consequently, if we represent the characteristic surface of the test, it would be closer to the one reflected in Figure 24. This is evidence of the impossibility of a linear relationship between the latent variable (-∞ to +∞) and probability (0-1).

The traditional approach in which the effect of subject and items selection is analysed separately is not suitable for real evaluation situations, as there is a combination of the effects of both factors that differs from their sum obtained independently. This study has therefore shown the need to consider the effect of interaction to be justified from a theoretical and empirical point of view.

It should be highlighted that, despite the fact that the procedure implemented has arisen as part of the study of score linking procedures, after the research had been completed it was observed that this situation would be found in any evaluation requiring a sample selection of subjects and items, not strictly related to equating processes. Thus, the study of a design with common items has enabled us to precisely establish this fact,

pointing to its usefulness for error estimation in wider evaluation contexts, not only in the score linking sphere. This unexpected finding suggests future lines of analysis.

Is it possible to improve the score linking process taking into account the interaction effect in the selection of subjects and items?

By understanding that improving linking processes is one of the main challenges we face in the so-called "evaluation era", we can consider that one of the most striking results of this PhD thesis study is the proposal for a procedure for analysing the interaction of subjects and items in equating and scaling processes (generalizable to other evaluation situations). We have called the procedure "two-dimensional bootstrap" and its theoretical and practical justification highlights its usefulness for improving accuracy in measuring the skills of subjects, enabling a tighter estimation of error in general and of scaling error in particular. This quantification will help both in the valuation of general linking processes and in the decisions taken about data collection design, sample size, items to be included in the anchoring test, etc.

How can the interaction between these sources of error be estimated?

The general framework of the research problem addressed in this PhD thesis has been defined as the proposal and evaluation of a methodological approach that enables an estimation to be made of the effect of interaction between the selection of subjects and tasks in score linking processes. Specifically, the method proposed is what we have called "two-dimensional bootstrap", an intensive resampling technique in which the resampling units are double (rows and columns), in our case (subjects and tasks or items).

The application of this double resampling procedure to the data matrix, and the estimation of the skill of each subject in each of the resulting sub-samples, has enabled us to analyse, using Analysis of Variance, the existence of the effect of each factor (subjects and items), as well as the effect of interaction between the two factors. At the same time, the estimation of the Root Mean Square Error (RMSE), based on data extracted from the application of the two-dimensional bootstrap procedure, has allowed

systematic and random linking errors to be combined, rather than the classic approach of handling these errors independently of each other.

The syntax for putting this procedure into practice has been produced in the R programme, version 3.12. This syntax is notable for its optimal handling of entry data, for the possibility of combining the *i* and *j* matrices and for the clarity in the output of the results obtained, because the complex combination of a double sampling procedure, in the sphere of IRT, means it is essential to work with efficient analysis strategies that minimise the time invested in the analysis procedure compared to the benefits gleaned from the results obtained.

Is it useful to study interaction in the analysis of common items?

The study of interaction is an essential component of the analysis of common items or anchoring items. As discussed in several sections of this study, the design of common items is the most frequent linking design, both in scaling and equating processes. We have thus opted for placing special emphasis on the importance of analysing the effect of interaction on these tasks, as their specific properties, their quality, their adequacy, their stability, etc., can determine the score linking process. The use of various sets of common items produces differences in scores, even when the subject sample size is large (Monseur & Berezner, 2007). These differences are the result both of the selection of tasks and of the interaction of their effect with the subjects undertaking the test.

Consequently, studying the effect of interaction precisely in these tasks has an important informative value, which will help us to make crucial decisions about the quality of the linking process carried out, as well as to produce interpretations that are closer to reality.

Under what conditions is the proposed procedure effective?

Numerous factors could be connected with the effect of interaction; note that this PhD thesis has not set out to analyse thoroughly all the possible factors related to interaction, however, an optimum proposal is put forward for their detection as well as

their functioning under certain conditions, taking into account certain factors that may have particular importance.

In respect of the factors analysed, the procedure implemented has enabled us to analyse 3 of the 4 sources of systematic error identified by Kolen and Brennan (2014), those relative to the violation of the statistical assumptions of the model (unidimensionality), problems with putting the information collection design in place (differences in the functioning of the items in the two groups being compared owing, for example, to position effects) and the existence of substantial differences in the behaviour of the groups being equated. The procedure proposed has been shown to function very adequately in such situations, proving to be an efficient detector of DIF or of differences between the groups to be linked.

In the field of international evaluations, authors like Monseur and Berezner (2007), suggest the use of the Jackknife procedure for resampling items when carrying out analyses differentiated by country, due to possible differences in the functioning of the items in each country that could affect linking error calculation. The two-dimensional bootstrap procedure would allow this effect of interaction to be measured, without the need to produce preliminary hypotheses on data functioning, creating a priori divisions that do not always reflect the reality evaluated.

The high degree of similarity detected in the functioning of the procedure under experimental conditions 3 and 4, shows a moderate effect of variation in the distribution patterns of *b*. It should be highlighted that the effect is similar irrespective of whether there is a change in mean values or in the dispersion. Despite not being a very notable increase, it is important to point out that, in real evaluation conditions this effect can be multiplied if certain differences between the groups being compared are added.

Amongst the most prominent factors would be the differential functioning of the item. The procedure presented, in addition to enabling its detection, has been shown to be sensitive to the quantity of DIF present in each of the iterations. Together with this, one of the most important aspects is that the procedure works without the need to draw up prior hypotheses about how the items work. Specifically, its practical usefulness

increases because it makes it possible to study the how the items function without taking into consideration prior expectations about their behaviour.

Consequently, the effect of interaction, ignored in classic procedures, is presented as an essential component, in the light of the results obtained in the analysis of the application of the procedure under the four experimental conditions handled in this PhD thesis. The procedure presented here is an interesting proposal for estimating the interaction effect, addressing under-explored territory and with major implications for current evaluation practice. The versatility of the proposal enables it to be applied in widely diverse score linking situations, with different designs for information gathering, linking procedures, estimation methods, sample sizes, etc.

LIMITATIONS AND OUTLOOK

This study is an initial approach to the proposed procedure and, far from concluding with a full and definitive answer to the tasks arising during its course, it ends with a discussion of the main limitations found plus future lines of work.

Firstly, the absence of a satisfactory base theory in the sphere of score linking has proved to be a major challenge, especially if we consider its analysis from the point of view of education research. The particular needs of education require specific attention in this sphere, as the complex reality involved in measurement in education needs a response. Therefore, an integrated approach that combines efforts in various areas is the best way forward when tackling new challenges.

In spite of having achieved a satisfactory solution in the face of the complexity inherent to the proposed analysis procedure, there are some limitations and more work is needed. Specifically, the execution time on conventional computers is one of the areas for improvement, as despite latest modifications to reduce this time, we think it is possible to optimise calculation processes by implementing truly efficient procedures that enable estimations to be made in more reasonable time-scales, without the need to split the process and with the advantage of being able to obtain results faster. Fast-paced advances in computational technology, as well as streamlining the proposed procedure, will enable substantial improvements to be made in future versions.

Since one of the major findings of this research is the importance of the estimation of the effect of interaction in Differential Item Functioning (DIF), analysing the behaviour of the procedure when manipulating different DIF conditions, like the percentage of items with DIF, the type of DIF, the sample size of the groups, the response format of items, etc., as well as comparing this procedure with other methods of detection (Mantel Haenszel, logistical regression, etc.) are all future lines of research that must be highlighted. There is no doubt about the need to continue carrying out research to assess the suitability and accuracy of the procedure implemented.

Likewise, taking into account that the multidimensionality of the tests is another of the conditions that can seriously damage the quality of score linking processes, and with special emphasis on vertical scaling, it would be necessary to study more fully how the two-dimensional bootstrap procedure works in the detection of these situations. During this study only a brief look was taken at the procedure from the perspective of a multidimensional concept of DIF.

In the light of results obtained in relation to the differences observed in the skill level of subjects whose scores are being compared, it would be interesting to study how this procedure works in the analysis of scaling chains, as the marked differences in the skill level of the groups involved is usually a frequent problem that can result in an exponential increase in linking error estimation, becoming further increased as the differences in skill level within the scaling chain become greater.

The moderate effect observed in relation to the variation in distributional patterns of b (dispersion / mean) makes it necessary to think about analyses that are more closely linked to real evaluation conditions, in order to study this combined effect with variables such as the possible differences between the groups being evaluated, as there can certainly be a multiplier effect that places the importance of the distribution of b values in more realistic conditions of application.

The estimation procedure presented in this study can be adapted to countless designs and evaluation conditions. Despite having opted in this research study for the analysis of common items, an interesting line of development would be to study how it works in different linking situations (designs, procedures, number of subjects, number

of items, type of anchoring, etc.). The results obtained show the importance of considering the effect of interaction on the estimation of error in the context of any evaluation in which a sample selection of subjects or tasks is required, supporting its more widespread use, not limited to particular scaling or equating conditions. In this sense, evaluating the procedure in evaluation situations that do not involve score linking is clearly an interesting issue for future work.

Lastly, as stated in the method description, simulation studies seek to reproduce real features and behaviours and, at the same time, exercise control over factors that can turn out to be objects of interest, because the situation to be analysed has been decided beforehand. Performing a simulation study is fully justified in accordance with the central aim of this study, however, the risk arises when the simulation study conditions are not very realistic and, consequently, there is a limit to how generalised the results can be (external validation) (Revuelta & Ponsoda, 2003). This means that it would be advisable for it to be applied to real data following its design and justification in controlled simulation environments. On this point, the original thetas for calculating the Mean Squared Error could be considered based on the responses of subjects to the complete form (without counting the anchoring items). That is, the value of the RMSE could be calculated in relation to the difference of the original theta (understood as the theta estimated on the basis of the results of subjects' response to the items making up the test) and the theta estimated on the basis of subjects' scores for the items that comprise the anchoring test.

REFERENCIAS BIBLIOGRÁFICAS

Acevedo Díaz, J. A. (2005). TIMSS y PISA. Dos proyectos internacionales de evaluación del aprendizaje escolar en ciencias. *Revista Eureka sobre Enseñanza y Divulgación de las Ciencia* , 2(3), 282-301.

Adams, R. J., Wilson, M. R., & Wang, W. (1997). The multidimensional random coefficients. *Applied Psychological Measurement*, (21), 1-24.

Aitkin, M., & Longford, M. (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society*, 149, 1-43.

Ackerman, T.A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*(1), 67-91.

Álvarez-Méndez, J. M. (2001). *Evaluar para conocer, examinar para excluir.* Madrid: Morata.

Alvaro Page, M. (1990). *Hacia un modelo causal del rendimiento académico.* Madrid: Centro de Publicaciones del Ministerio de Educación y Ciencia (CIDE).

Amadeo, J., Torney-Purta, J., Lehmann, R., Husfeldt, V., & Nikolova, R. (2002). *Civic Knowledge and Engagement An IEA Study of Upper Secondary Students in Sixteen Countries.* Amsterdam: The International Association for the Evaluation of Educational Achievement.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1985). *Standards for educational and psychological testing*. Washington , DC: American Psychological Association.

American Educational Research Association, American Psychological Association y National Council onMeasurement in Education. (1990). *Standards for Educational and Psychological Testing.* Washington, DC, EEUU: American Psychological Asociation.

American Edcational Research Association, Américan Psicological Association & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington: American Educational Research Association.

American Edcational Research Association, Américan Psicological Association & National Council on Measurement in Education (2014). *Standards for educational and psychological testing.* Washington: American Edcational Research Association, Américan Psicological Association & National Council on Measurement in Education.

Angoff, W. H. (1971). Scales, norms and equivalent scores (2nd ed.). En R. L. Thorndike, *Educational Measurement* (pp. 508-600). Washington, DC: American Council on Education.

Angoff, W. H. (1987). Technical and practical issues in equating: a discussion of four papers. *Applied Psychological Measurement*, *11,* 291-300.

Antoni Adell, M. (2002). *Estrategias para mejorar el rendimiento académico de los adolescentes.* Madrid: Pirámide.

Baker, F. B. (1992). Equating Tests Under the Graded Response Model. *Applied Psychological Measurement, 16*(1), 87-96.

Baker, F. B. (1993). Equating test under the nominal response model. *Applied Psychological Measurement, 17*(3), 239-251.

Beaton, A. E. (1999). *The benefits and limitations of international educational achivement studies.* París: International Institute of Educational Planning, International Academy of Education. UNESCO, París.

Beaton, A., Martin, M., Mullis, I., González, E., Smith, T., & Kelly, D. (1996). *Science Achievement in the Middle School Years. IEA's Third International Mathematics and Science Study.* Chestnut Hill, MA: Lynch School of Education. Boston College.

Bock, R. D. (1972). Estimating ítem parameters and latent ability when responses are scored in two o more nominal categories. *Psychometrika*, *37*(1), 29-51.

Bolívar, A. (2010). *Competencias básicas*. Bizkaia: Wolters Kluwer España.

Braun, H. I., & Holland, P. W. (1982). Observed- Score Test Equating: A Mathematical Analysis of Some ETS Equating Procedures. En P. W. Holland, & D. B. Rubin, *Test Equating* (pp. 9-49). New York: Academic Press.

Brennan, R. L. (2007). Tests in Transition: Discussion and Synthesis. In N. J. Dorans, M. Pommerich, & P. W. Holland, *Linking and Aligning Scores and Scales* (pp. 161-175). New York: Springer.

Brennan, R. L., & Kolen, M. (1987a). Some practical issues in equating. *Applied Psychological Measurement, 11,* 279-290.

Brennan, R. L., & Kolen, M. (1987b). A reply to Angoff. *Applied Psychological Measurement, 11,* 301-306.

Bruner, J. J., & Elacqua, G. (2004). Factores que inciden en una educación efectiva. Evidencia internacional. *Revista Virtual La educación,* (139-140), 1-11.

Camilli, G., & Sephard, L. A. (1994). *Mthods for Identifying Biased Test Items*. Thousand Oaks, California: SAGE publications, Inc.

Campbell, J., Kelly, D., Mullis, I., Martin, M., & Sainsbury, M. (2001). *Farmework and Specifications for PIRLS Assesment 2001.* Chestnut Hill, MA: International Study Center. Lynch School of Education, Boston College.

Carabaña, J. (1987). Origen social, inteligencia y rendimiento académico al final de la EGB. En C. Lerena (ed.), *Educación y Sociología en España,* (pp. 262-289). Madrid: Akal.

Carlson, J. E. (2011). Statistical Models for Vertical Linking. En A. A. von Davier, *Statistical Models for Test Equating, Scaling, and Linking,* (pp. 59-70). New York: Springer.

Casassus, J., Cusato, S., Froemel, J. E., & Palafox, J. C. (2001). *Primer estudio internacional comparativo sobre lenguaje, matemática y factores asociados, para alumnos del tercer y cuarto grado de la Educación Básica. Informe técnico.* Santiago de Chile: Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación.

Casassus, J., Froemel, J. E., Palafox, J. C., & Cusato, S. (1998). *Primer Estudio Internacional Comparativo sobre Lenguaje, Matemática y Factores Asociados en Tercero y Cuarto Grado.* Santiago de Chile: Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación UNESCO-SANTIAGO.

Castillo Arredondo, S., & Cabrerizo Diago, J. (2010). *Evaluación educativa de aprendizajes y competencias.* Madrid: Pearson.

Castro Morera, M. (1997). *Meta-Análisis. Aportaciones metodológicas a la síntesis cuantitativa de la evidencia.Uun estudio de simulación Monte Carlo.* Madrid, Madrid: Tesis doctoral.

Cervini, R. (2003). Relaciones entre composición estudiantil, proceso escolar y el logro en matemáticas en la educación secundaria en Argentina. *Revista Electrónica de Investigación Educativa*, *5*(1), 72-98.

Childs, R., & Jaciw, A. P. (2003). Matrix sampling of items in large-scale assessments. *Practical Assessment, Research & Evaluation, 16* (8), 16-28.

Chong, H. J., & Osborn Poopp, S. E. (2005). Test Equating by Common Items and Common Subjects: Concepts and Applications. *Practical Assessment, Research & Evaluation*, *10*(4), 1-19.

Coleman, J. E., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A.M., Weinfield, F.D., & York, R.L. (1966). *Equality of educational opportunity.* Washington. D.C.: US Government Printing Office.

Coll, C. (2007). Las competencias en la educación escolar: algo más que una moda y mucho menos que un remedio. *Aula de Innovación Educativa*, 34-39.

Cook, L. L. (2007). Practical Problems in Equating Test Scores: A particioner´s perspective. En N. J. Dorans, M. Pommerich, & P. W. Holland, *Linking and Aligning Scores and Scales*, (pp.73-88). New York: Springer.

Cook, L. L., & Petersen, N.S. (1987). Problems Related to the Use of Conventional and Item Response Theory Equating Methods in Less Than Optimal Circumstances. *Applied Psychological Measurement*, *11,* 225-244.

Cortina, K. S. (2015). PIAAC and PISA: Pedagogically Paradoxical Parallels. *Zeitschrift fur Padagogik,* 61, 223 - 242.

Crocker, L., & Algina, J. (1986). *Introduction to Classical & Modern test theory.* New York: Holt, Rinehart and Winston.

Cronbach, L. (1963). Course improvement through evalation. *Teachers College Record, 64*, 672-683.

Cui, Z., & Kolen, M. J. (2008). Comparison of Parametric and Nonparametric Bootstrap Methods for Estimating. *Applied Psychological Measurement, 32*(4), 334-347.

De la Garza Vizcaya, E. (2004). La Evaluación educativa. *Revista Mexicana de Investigación Educativa*, *9*(23), 807-816.

De la Orden, A. (1972). Evaluación de los conocimientos. *Bordón*, (187), 201-222.

De la Orden, A. (2000). La función optimizante de la evaluación de programas evaluativos. *Revista de Investigación Educativa*, *18*(2), 381-389.

De la Orden, A., Oliveros, L., MafoKozi, J., & Gonzalez, C. (2001). Modelos de investigación del bajo rendimiento. *Revista Complutense de Educación, 12*(1), 159-178.

Deakin Crick, R. (2008). Key Competencies for Education in a European Context: narratives of accountability or care. *European Educational Research Journal, 7* (3), 311-318.

Delors, J. (1996). *Learning: the treasure within.* Paris: UNESCO.

Denyer, M., Furnémount, J., Poulain, R., & Vanloubbeeck, G. (2007). *Las competencias en educación. Un balance.* México: Fondo de cultura económica.

Dolata, S. (2005). *Construction and validation of pupil socioeconomic status index for SACMEQ Education Systems*. Comunicación presentada en SACMEQ Conference. París.

Dorans, N. J. (2000). *Distinctions among classes of linkages*. College Board Research Notes (RN-11). New York: The College Board.

Dorans, N. J. (2004). Equating, Concordance, and Expectation. *Applied Psychological Measurement*, *28*, 227-246.

Dorans, N. J. (2012). The Contestant Perspective on Taking Tests: Emanations From the Statue Within. *Educational Measurement: Issues and Practice, 31*(4), 20-37.

Dorans, N. J. (2013). On Attempting to do what Lord said was impossible: commentary on van der Linden`s "Some conceptual issues in Observed- Score Equating". *Journal of Educational Measurement, 50*(3), 304-314.

Dorans, N. J., & Holland, P. W. (2000). Population Invariance and the Equatability of Tests: Basic Theory and The Linear Case. *Journal of Educational Measurement, 37*(4), 281-306.

Dorans, N. J., Moses, T. P., & Eignor, D. (2010). *Principles and Practices of Test Score Equating.* Educational Testing Service. Princeton, NJ: ETS.

Dorans, N. J., Moses, T. P., & Eignor, D. (2011). Equating Test Scores: Toward Best Practices. En A. A. von Davier, *Statistical Models for Test Equating, Scaling, and Linking,* (pp. 21-42). New York: Springer.

Dorans, N. J., Pommerich, M., & Holland, P. W. (2007). *Linking and Aligning Scores and Scales.* New York: Springer.

Edmonds, R. (1979). Effective schools for the urban poor. *Educational Leadership, 1*(37), 15-24.

Efron, B. (1979). Bootsrap methods: another look at the jackknife. *Annals of statistics*, *7*(1), 1-26.

Efron, B. (1990). *The Jackknife, the Bootstrap and other resampling plans* (5 ed.). Philadelphia: Society for Industrial and Applied Mathematics.

Escamilla, A. (2008). *Las competencias básicas. Claves y propuestas para su desarrollo en los centros.* Barcelona: Graó.

Eurydice. (2002). *Key competencies. A developing concept in general compulsory education.* Belgium: Eurydice.

Fairbank, B. A. (1987). The Use of Presmoothing and Postsmoothing to Increase the Precision of Equipercentile Equating. *Applied Psychological Measurement*, *11*, 245-262.

Ferrer, G., & Arregui, P. (2003). *Las pruebas internacionales de aprendizaje en América Latina y su impacto en la calidad de la educación: criterios para guiar futuras aplicaciones.* Programa de Promoción de la Reforma Educativa en América Latina y el Caribe.

Feuer, M. H., Holland, P.W., Green, B.F., Bertenthal, M.W., & Hemphill, F.D. (1999). *Uncommon measures: Equivalence and linkage among educational tests.* Washington: National Acadey Press.

Flanagan, J. C. (1951). Units, socres, and norms. En E. F. Lindquist, *Educational Measurement* (pp. 695-763). Washington, DC: American Council on Education.

Galton, F. (1988). Regression towards mediocrity in hereditary stature. *Anthropological miscellanea*, 246-263.

García Ramos, J. (1999). Investigación y evaluación. Implicaciones y efectos. Algunas reflexiones metodológicas sobre la investigación y evaluación educativas. *Revista Complutense de Educación, 10*(2), 189-214.

Garden, R., Lie, S., Robitaille, D., Angell, C., Martin, M., Mullis, I., Foy, P., & Arora, A. (2006). *TIMSS Advanced 2008 Assessment Frameworks.* Chestnut Hill: TIMSS & PIRLS International Study Center Lynch School of Education, Boston College.

Gempp, R. (2010). Equiparacion, alineamiento y predicion de puntuaciones en medicion educativa. *Revista Iberoamericana de Evaluacion Educativa*, *3*(2), 103-126.

Gimeno-Sacristán, J., Pérez Gómez, A., Torres Santomé, J., Angulo Rasco, F., & Álvarez-Méndez, J. M. (2008). *Educar por competencias ¿qué hay de nuevo?* Madrid: Morata.

Goldstein, J. (1997). Methods in school effectiveness research. *School Effectiveness and school improvement, 8*, 369-395.

Goldstein, H. (2004). The Evaluation World Cup: international comparisons of student achievement. *Plenary talk to Association for Educational Assessment,* Budapest.

González Fernández, D. (1975). Procesos escolares inexplicables. *Aula Abierta, 11*, 12.

Haberman, S. J., Lee, Y.-H., & Qian, J. (2009). *Jackknifing Techniques for Evaluation of Equating Accuracy.* Princeton, New Jersey: Educational Testing Service.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22*(3), 144-149.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory. Principles and Applications.* New York: Springer.

Hambleton, R., Swaminathan, H., & Rogers, H. (1991). *Fundamentals of item response theory.* Newbury Park, California: Sage.

Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26*(1), 3-24.

Hanushek, E.A., Schwerdt, G., Wiederhold, S., & Woessmann, L. (2015). Returns to Skills around the World: Evidence from PIAAC. *European Economic Review,* 73, pp. 103-130.

Harris, D. J. (2007). Practical Issues in Vertical Scaling. En N. J. Dorans, M. Pommerich, & P. W. Holland, *Linking and Aligning Scores and Scales* (pp. 233-251). New York: Springer.

Harris, D. J., & Crouse, J. D. (1993). A Study of Criteria Used In Equating. *Applied measurement in education, 6*(3), 95 - 240.

Holland, P. W. (2007). A framework and history for score linking. En N. J. Dorans, M. Pommerich, P. W. Holland, N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and Aligning Scores and Scales* (pp. 5-30). New York, NY: Springer-Verlag.

Holland, P. W. (2013). Comments on van der Linden`s critique and proposal for equating. *Journal of Educational Measurement, 50*(3), 286-294.

Holland, P. W., & Dorans, N. J. (2006). Linking and Equating. En R. L. Brennan, & R. L. Brennan (Ed.), *Educational Measurement* (4 ed., pp. 187-220). Westport, CT: Praeger Publishers.

Holland, P. W., Dorans, N. J., & Petersen, N. S. (2007). Equating Test Scores. In C. R. Rao, & S. Sinharay, *Handbook of statistics 26. Psychometrics* (pp. 169-203). Amsterdam: North Holland.

Holland, P. W., & Hoskens, M. (2003). Clasical test theory as a first-order item response theory: application to true-scores prediction from a possibly nonparalel test. *Psychometrika, 68* (1), 123-149.

Holland, P., & Rubin, D. (1982). *Test Equating.* New York: Academic Press.

Holland, P. W., & Thayer, D. T. (1988). Differential Item Performance and the Mantel Haenszel Procedure. En H. Wainer, & H. I. Braun, *Test Validity* (pp. 129-145). Hillsdale, New Jersey: Lawewnce Erlbaum Associates.

Hungi, N., Makuwa, D., Ross, K., Saito, M., Dolata, S., van Cappelle, F., Paviot, L., & Vellien, J. (2010). *SACMEQ III: Project Results. Pupil Achivement levels in reading and mathematics.* SACMEQ.

Husén, T. (1987). Policy impact of IEA research. *Comparative Educational Review, 1* (31), 29-46.

Instituto de Evaluación. (2009). *Evaluación General de Diagnóstico 2009. Marco de la Evaluación.* Madrid: Ministerio de Educación.

Instituto de Evaluación. (2010). *ICCS 2009. Estudio Internacional de Civismo y Ciudadanía. IEA. Informe Español.* Madrid: Ministerio de Educación.

Instituto Nacional de Ciencias de la Educación. (1976). *Determinantes del Rendimiento Académico.* Madrid: Servicio de Publicaciones del Ministerio de Educación y Ciencia. Departamento de Prospeccción Educativa.

Jencks, C., Smith, M. S., Ackland, H., Bane, M. J., Cohen, D., Grintlis, H., Heynes, B., & Michelson, S (1972). *Inequality: A reassessment of the effect of family and schooling in America.* New York: Harper & Row.

Jiang, Y., von Davier, A. A., & Chen, H. (2012). Evaluating Equating Results: Percent Relative Error for Chained Kernel Equating. *Journal of Educational Measurement*, *49*(1), 39-58.

Kaczynska, M. (1965). *El rendimiento escolar y la inteligencia.* Madrid: Espasa Calpe.

Kelley, T. L. (1927). *The interpretation of educational measurements.* New York: World Book.

Kendall, M., & Stuart, A. (1977). *The advanced theory of statistics* (4 ed., Vol. 1). New York: Macmillan.

Kim, S. (2010). A comparative study of IRT Fixed Parameter Calibration Meethods. *Journal of Educational Measurement, 4*(43), 355-381.

Kim, S. (2010). An Extension of Least Squares Estimation of IRT Linking Coefficients for the Graded Response Model. *Applied Psychological Measurement, 34*, 505-520.

Kim, S., & Cohen, A. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement, 26*, 25-41.

Kim, S., & Kolen, M. J. (2007). Effects on Scale Linking of Different Definitions of Criterion Functions for the Irt Characteristic Curve Methods. *Journal of Educational and Behavioral Statistics, 32*(4), 371-397.

Kirsch, I. (2001). *The International Adult Literacy Survey (IALS): Understanding What Was Measured.* Princeton: Educational Testing Service.

Kish, L. & Frankel, M.R. (1970). Balanced Repeated Replications for Standard Errors. *Journal of the American Statistical Association, 65*(331), 1071-1094.

Klitgaard, R. E., & Hall, G. R. (1974). Are there unusually effective schools? *Journal of Human Resources, 74*, 90–106.

Kok, F.G. (1988). Item bias and test multidimensionality. En R. Langeheine & J. Rost (Eds.), Latent trait and latent class models (pp. 263-274). New York: Plenum.

Kolen, M. J. (1988). An NCME instructional module on Traditional Equating Methodology. *Educational Measurement, 7*(4), 29-37.

Kolen, M. J. (2001). Linking assessments effectively: purpose and design. *Educational Measurement: Issues and Practice, 20*(1), 5-19.

Kolen, M. J. (2004). Linking Assessments: concept and history. *Applied Psychological Measurement, 28*(4), 219-226.

Kolen, M. J. (2007). Data Collection Designs and Linking Procedures. En N. J. Dorans, M. Pommerich, & P. W. Holland, *Linking and Aligning Scores and Scales* (pp. 31- 55). New York: Springer.

Kolen, M. J., & Brennan, D. (1995). *Test Equating. Methods and practices.* New York: Springer- Verlag.

Kolen, M. J., & Brennan, D. (2004). *Test Equatin, Scaling and Linking. Methods and practice* (2 ed.). New York: Springer- Verlag.

Kolen, M. J., & Brennan, R. L. (2014). *Test Equating, Scaling, and Linking. Methods and Practices.Third Edition* (3 ed.). New York: Springer.

Kolen, M. J., & Jarjoura, R. L. (1987). Analytic smoothing for equipercentile equating under the common item nonequivalent populations design. *Psychometrica,52*, 43-59.

Kolen, M. J., Tong, Y., & Brennan, R. L. (2011). Scoring and Scaling Educational Tests. In A. A. von Davier, *Statistical Models for Test Equating, Scaling, and Linking* (pp. 43-58). New York: Springer.

Lapointe, A., Mead, N., & Philips, G. (1989). *Un mundo de diferencias. Un estudio internacional de Evaluación de las Matemáticas y las Ciencias*. (S. Marcos Pérez, Trad.) Madrid: Centro de publicaciones del ministerio de Educación y Ciencia. CIDE.

Lee, W.C., & Ban, J.C. (2010). A Comparison of IRT Linking Procedures. *Applied Measurement in Education, 23*(1), 23-48.

Ley 14/1970, de 4 de agosto, General de Educación y Financiamiento de la Reforma Educativa. BOE núm. 187 (6 de agosto de 1970), 12525-12546.

Ley Orgánica 1/1990, de 3 de octubre, de Ordenación General del Sistema Educativo. BOE núm. 238 (4 de octubre de 1990), 28927-28942.

Ley Orgánica 10/2002, de 23 de diciembre, de Calidad de la Educación. BOE núm. 307 (24 de diciembre de 2002), 45188-45220.

Ley Orgánica 2/2006, de 3 de mayo, de Educación. BOE num. 106 (4 de mayo 2006) 17158-17207.

Ley Orgánica 8/2013, de 9 de diciembre, para la mejora de la calidad educativa. BOE num. 295 (10 de diciembre 2013) 97858 - 97921.

Li, D., Jiang, Y., & von Davier, A. A. (2012). The Accuracy and Consistency of a Series of IRT True Score Equatings. *Journal of Educational Measurement, 49*(2), 167–189.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement* Education, 6(1), 83-102.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1980). *An investigation of item bias in a test of reading comprehension*. Washington, National Institute of Education.

Linn, R. L., Levine, M. V., Hastings, N. C., & Wardrop, J. (1981). Item Bias in a Test of Reading Comprehension. *Applied Psychological Measurement, 5*(2), 159-173.

Liou, M., & Cheng, P. E. (1995). Asymptotic Standard Error of Equipercentile Equating. *Journal of Educational and Behavioral Statistics, 20*(3), 259-286.

Liu, J., & Walker, M. E. (2007). Score Linking Issues Related to Test Content Changes. En N. J. Dorans, M. Pommerich, & P. W. Holland, *Linking and Aligning Scores and Scales* (págs. 109-134). New York: Springer.

LLECE. (2001). *Informe Técnico. Primer Estudio Internacional Comparativo sobre lenguaje, matemática y factores asociados para alumnos del tercer y cuarto grado de la Educación Básica.* Santiago: Publicado por el Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación.

Lomax, R. G., & Hahs-Vaughn, D. L. (2012). *An introduction to Statistical Concepts. Third edition.* New York: Taylor & Francis Group.

López Jáuregui, A., & Elosua Oliden, P. (2004). Estimaciones bootstrap para el coeficiente de determinación: un estudio de simulación. *Revista Electrónica de Metodología Aplicada , 9*(2), 1-14.

Lord, F. M. (1975). *A survey of equating methods based on item characteristic curve theory.* Reserach Bulletin Nª. 75-13, Princeton,. New Jersey: Educational Testing Service.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* New Jersey: Lawrwnce Erlbaum associates, publishers.

Lord, F. M. (1982a). The Standard Error of Equipercentile Equating. *Journal of Educational Statistics, 7*(3), 65-174.

Lord, F. M. (1982b). *Standard error of an equating by item response theory.* Princeton: Educational Testing Service.

Lord, F. M., & Novick, M. R. (2008). *Statistical theories of mental test scores.With contributions by A.Brinbaum.* New Jersey: Information Age Publishing.

Manly, B. F. (1997). *Randomization, Bootsrap and Monte Carlo methods in biology* (2 ed.). Florida: Chapman&Hall/CRC.

Martín, E. (2009). Curriculo y evaluación estandarizada: cooperación o tensión. En E. Martín, & F. Martíenz Rizo, *Avances y desafíos en la evaluación educativa* (pp. 89-97). Madrid: OEI y Fundación Santillana.

Martín, E., & Coll, C. (2003). *Aprender contenidos, desarrollar capacidades. Intenciones educativas y planificación de la enseñanza.* Barcelona: Edebé.

Martin, M., Mullis, I., González, E., Gregory, K., Smith, T., Chorostowski, S., Garden, R.A., & O´Connor, K..M. (2000). *TIMSS 1999 International Mathematics Report: Findings from IEA´S Repeat of the Third International Mathematics and Science Study at the Eight grade.* Chestnut Hill, MA: International Study Center. Lynch school of Education. Boston College.

Martin, M., Mullis, I., González, E., & Kennedy , A. (2003). *Trends in children´s reading literacy achievement 1991-2001: IEA´s Repeat in nine countries of the 1991 Reading Literacy Study.* Chestnut Hill, MA: International Study Center. Lynch School of Education. Boston College.

Martínez Arias, R. (2005). *Psicometría: teoría de los tests psicológicos y educativos.* Madrid: Síntesis.

Martinez Arias, R., Gaviria Soto, J. L., & Castro Morera, M. (2009). Concepto y evolución de los modelos de valor añadido en educación. *Revista de Educación*, *348*, 15-47.

Martínez Arias, R., Hernández Lloreda, M. J., & Hernández Lloreda, M. V. (2006). *Psicometría.* Madrid: Alianza Editorial.

Martínez Rizo, F. (2004). *La Comparabilidad de los Resultados de las Pruebas Nacionales de Lectura y Matemáticas, 1998-2003.* Instituto Nacional para la Evaluación de la Educación. México: Colección Cuadernos de Investigación.

Martinez Rizo, F. (2009). La evaluación de la calidad de los sistemas educativos: propuesta de un modelo. En E. Martín, & F. Martínez Rizo, *Avances y desafíos en la evaluación educativa*, (pp. 27-39). Madrid: OEI y Fundación Santillana.

Martinez-Otero, V. (1997). *Los adolescentes ante el estudio. Causas y consecuencias del rendimiento académico.* Fundamentos.

Marzano, R. J. (2000). *A new era of school reform: going where the research takes us*. Aurora, CO: Mid-continent Research for Education and Learning.

Masters, G. (1982). A Rasch Model for Partial Credit Scoring. *Psychometrika, 47*(2), 149-174.

Matas Terrón, A. (2003). Estudio diferencial de indicadores de rendimiento en pruebas objetivas. *Revista Electrónica de Investigación y Evaluación Educativa. RINACE, 9*(2), 184-197.

McArdle, J. J., & Grimm, K. J. (2011). An Empirical Example of Change Analysis by Linking Longitudinal Item Response Data From Multiple Tests. En A. A. von Davier, *Statistical Models for Test Equating, Scaling and Linking* (pp. 71-88). New York: Springer.

McClelland, D. C. (1973). Testing for Competence Rather Than for "Intelligence", *Amerincan Psychologist*, *28*, 1-14.

Meroni, E.C., Vera-Toscano, E., & Costa, P. (2015). Can low skill teachers make good students? Empirical evidence from PIAAC and PISA. *Journal of Policy Modeling,* 37, 308–323.

Michaelides, M. P., & Haertel, E. H. (2004). *Sampling of Common Items: An Unrecognized Source of Error in Test Equating.* Los Angeles: Center for the Study of Evaluation (CSE) National Center for Research on Evaluation,Standards, and Student Testing (CRESST). Graduate School of Education & Information Studies University of California.

Millet, G., & Sánchez, P. (2007). *Las competencias básicas. Cultura imprescindible de la ciudadanía.* Madrid: Proyecto Atlántida.

Mirai Solutions GmbH (2014). XLConnect: Excel Connector for R. R package version 0.2-9. http://CRAN.R-project.org/package=XLConnect

Mislevy, R. J. (1992). Linking educational assessments: Concepts, issues, methods, and prospects. Princeton, NJ: ETS Policy Information Center.

Mislevy, R. J. (1995a). What can we learn from International Assessments? *Educational Evaluation and Policy Analysis, 21*(17), *419-437.*

Mislevy, R. J. (1995b). *Linking Adult Literacy Assessments.* Princeton, NJ: Educational Testing Service.

Monereo, C. (2005). *Internet y competencias básicas. Aprender a colaborar, a comunicarse, a participar, a aprender.* Barcelona: Graó.

Monereo, C., & Castelló, M. (2009). La evaluación como herramienta de cambio educativo: evaluar las evaluaciones. En C. Monereo (Coord.), *PISA como excusa repensar la educación para cambiar la enseñanza.* Barcelona: Graó.

Monseur, C., & Berezner, A. (2007). The computation of Equating Errors in International Surveys in Education, *Journal of applied Measurement*, *8*(3), 323-335.

Monseur, C., Sibberns, H., & Hastedt, D. (2006). Equating Errors in International surveys in education. *The Second IEA International Research Conference: Proceedings of the IRC-2006. Volume II* (pp. 61-65). Amsterdam: International Association for Evaluation of Educational Achievement (IEA).

Mooney, C. Z., & Duval, R. D. (1993). *Bootstraping a nonparametric approach to statistical inference.* London: SAGE University Paper.

Morris, C. N. (1982). On the Foundations of Test Equating. En P. Holland, & D. Rubin, *Test Equating* (pp. 169-191). Princeton, New Jersey: Academic Press.

Mortimore, P., Sammons, P., Stoll, L., Lewis, D., & Ecob, R. (1988). *School Matters.* Gran Bretaña: A.Wheaton & Co.

Moser, C. (1951). Problems and Designs of Cross-Validation. *Educational and Psychological*, *11*, 5-11.

Moya Otero, J. (2009). *Proyecto Atlántida de las competencias básicas al currículo integrado.* Madrid: Proyecto Atlántida.

Mullis, I. V., & Martin, M. O. (Eds) (2013). *Timss 2015 Assessment Frameworks.* Boston, USA: International Association for the Evaluation of Educational Achievement (IEA) & Lynch School of Education.

Mullis, I. V., & Martin, M. O.(Eds) (2015). *Pirls 2016 Assessment Frameworks.* Boston, USA: International Association for the Evaluation of Educational Achievement (IEA) & Lynch School of Education.

Mullis, I., Martin, M., González, E., & Chrostowski, S. (2004). *TIMSS 2003 International Mathematics Report. Findings From IEA´s Trends in International Mathematics and Science Study at the Fourth and Eighth Grades.* Chestnut Hill, MA, USA: International Study Center. Lynch School of Education.

Mullis, I., Martin, M., & Foy, P. (2008). *TIMSS 2007 International Mathematics Report. Findings from IEA´s Trends in International Mathematics and Science Study at the Fourth and Eighth Grades.* Chestnut Hill, MA, USA: International Study Center. Lynch School of Education. Boston College.

Mullis, I., Martin, M., González, E., & Kennedy, A. (2003). *PIRLS 2001 International Report: IEA's Study of Reading Literacy Achievement in Primary Schools.* MA, USA: Boston College: Chestnut Hill.

Mullis, I., Martin, M., Kennedy, A., & Foy, M. (2007). *PIRLS 2006 International Report. IEA`s Progress in International Reading Literacy Study in Primary chools in 40 countries.* Chestnut Hill, MA, USA: International Study Center. Lynch School of Education. Boston College.

Mullis, I., Martin, M., Kennedy, A., Trong, K. L., & Sainsbury, M. (2009). *PIRLS 2011 Assessment Framework.* MA, USA: Boston: Boston College.

Mullis, I., Martin, M., Ruddock, G., O´Sulivan, C., & Preuschoff, C. (2009). *TIMSS 2011 Assessment Framewoks.* Chestnut Hill, MA, USA: International Study Center Lynch School of Education, Boston College.

Muñiz, J. (1997). *Introducción a la teoría de respuesta a los ítems.* Madrid: Psicología Piramide.

Muñiz, J. (2003). *Teoría clásica de los tests.* Madrid: Ediciones Pirámide.

Muñiz, J., & Hambleton, R. K. (1992). Medio siglo de Teoría de Respuesta a los Ítems, *Anuario de psicología, 52*, 41-66.

Muñoz Arroyo, A. (1977). *Valoración del rendimiento de centros docentes de EGB. VI Plan de Investigación Educativa.* CIDE (MEC). ICE, Universidad de Extremadura.

Murillo, F. J. (2003). El movimiento de investigación de Eficacia Escolar. In F. J. Murillo, *La investigación sobre Eficacia Escolar en Iberoamérica. Revisión internacional del estado del arte.* (pp. 1-36). Bogotá: Convenio Andrés Bello-Centro de Investigación y Documentación Educativa.

National Commission on Excellence in Education. (1983). *A nation at risk.* Washington DC: Government Printing Office.

Navas, M. J. (1996). Equiparación de Puntuaciones. En J. Muñiz, *Psicometría* (pp. 295-369). Madrid: Universitas.

Navas, M. J. (2000). Equiparación de puntuaciones: exigencias actuales y retos de cara al futuro. *Metodología de las Ciencias del Comportamiento*, *2*(2), 151-165.

Noreen, E. W. (1989). *Computer inensive methods for testing hypotheses. An introduction.* New York: John Wiley & Sons, Inc.

North Central Regional Educational Laboratory and Metiri Group (2003). enGauge® 21st Century Skills: Literacy in the Digital Age. Naperville, IL: North Central Regional Educational Laboratory and the Metiri Group. Url: http://pict.sdsu.edu/engauge21st.pdf

OECD. (2002). *Conocimientos y aptitudes para la vida. Primeros resultados del programa Internacional de Evaluación de Estudiantes (PISA) 2000 de la OCDE.* París: Santillana.

OECD. (2005a). *The definition and selection of key competencias. Executive summary.* París: Organisation for Economic Co-operation and Development.

OECD. (2005b). *PISA 2003 Data Analysys Manual: SPSS Users.* París: Organisation for Economic Co-operation and Development

OECD. (2006). *Assessing Scientific, Reading and Mathematical Literacy A Framework for PISA 2006.* París: Organisation for Economic Co-operation and Development.

OECD. (2009). *PISA 2009 Assessment Framework Key competencies in reading, mathematics and science.* París: Organization for the Economic Cooperation and Development.

OECD. (2010a). *PISA 2009 Results: What Students Know and Can Do. Student Performance in Reading, Mathematics and Science (Volume I).* París: Organization for the Economic Cooperation and Development.

OECD. (2010b). *TALIS 2008 Technical Report.* París: Organisation for Economic Co-Operation and Development.

OECD.(2012a).The Organisation for Economic Co-operation and Development (OECD) [sitio web]. París: OECD [Consulta 17 de Agosto de 2012]. Disponible en: http://www.oecd.org/about/

OECD. (2012b). *PISA 2012 Assessment and Analytical Framework Mathematics, Reading, Science,Problem Solving and Financial Literacy.* París: Organization for the Economic Co-operation and Development.

OECD. (2012c). *PISA 2009 Technical Report.* París: Organization for the Economic Co-operation and Development.

OECD (2013). OECD Skills Outlook 2013: First Results from the Survey of Adult Skills, OECD Publishing. http://dx.doi.org/10.1787/9789264204256-en

OECD. (2014). *TALIS 2013 Results An International Perspective on Teaching and Learning.* OECD Publishing. http://dx.doi.org/10.1787/9789264196261-en.

Ogasawara, H. (2001). Item Response Theory True Score Equatings and Their Standard Errors. *Journal of Educational and Behavioral Statistics, 26*(1), 31-50.

Olson, J., Martin, M., & Mullis, I. (2008). *TIMSS 2007 Technical Report.* Boston, MA: TIMSS & PIRLS International Study Center. Lynch Scholl of Education. Boston College.

Orden ECD/65/2015, de 21 de enero, del Ministerio de Educación, Cultura y Deporte, por la que se describen las relaciones entre las competencias, los contenidos y los criterios de evaluación de la educación primaria, la educación secundaria obligatoria y el bachillerato. BOE núm. 25 (29 de enero de 2015), 6986-7003.

Pacheco-Villamil, J. (2007). La equiparación de puntuaciones en procesos de comparación de pruebas diferentes. *Avances en Medición, 5*, 153-157.

Pajares, R., Sanz, A., & Rico, L. (2004). *Aproximación a un modelo de evaluación: el proyecto PISA 2000.* Madrid: Ministerio de Educación Cultura y Deporte.Secretaría General técnica, Subdirección General de Información y Publicaciones.

Pardo, A., & San Martín, R. (2010). *Análisis de datos en ciencias sociales y de la salud II*. Madrid: Síntesis.

Parlamento Europeo y Consejo de la Unión Europea. (2006). Recomendación del Parlamento Europeo y del Consejo, de 18 de diciembre de 2006, sobre las competencias clave para el aprendizaje permanente. *Diario Oficial de la Unión Europea,* pp. L394/10-18 (30 de diciembre de 2006).

Patz, R. J., & Yao, L. (2007). Methods and Models for Vertical Scaling. En N. J. Dorans, M. Pommerich, & P. W. Holland, *Linking and Aligning Scores and Scales* (pp. 253- 272). New York: Springer.

Pedraza Daza, F. P., Mantilla Cárdenas, W., Duarte Agudelo, P., García Conde, M., Ortiz Rojas, H. M., Martínez, R., Reyes, C., García, D., & Soarez,M.T. (2013). *Tercer estudio regional comparativo y explicativo TERCE: Análisis curricular*. Santiago, Chile: Oficina Regional de Educación para América Latina y el Caribe (OREALC/UNESCO).

Pérez, G. (1981). *Origen social y rendimiento escolar*. Madrid: Centro de investigaciones sociológicas.

Perrenoud, P. (1999). *Dix nouvelles compètences pour enseigner. Diez nuevas competencias para enseñar*. París: ESF.

Petersen, N. S. (2007). Equating: Best Practices and Challenges to Best Practices. En N. J. Dorans, M. Pommerich, & P. W. Holland, *Linking and Aligning Scores and Scale,* (págs. 59-72). New York: Springer.

Philip Chalmers, R. (2012). mirt: A Multidimensional Item ResponseTheory Package for the R Environment. Journal of Statistical Software, *48*(6), 1-29. URL http://www.jstatsoft.org/v48/i06/

Piñeros, L. J., & Rodríguez, A. (1998). Los Insumos Escolares en la Educación Secundaria y su efecto sobre el Rendimiento Académico de los estudiantes: Un estudio en Colombia. *Human Development Department LCSHD paper series*, 36.

Plowden, C. (1967). *Children and their primary Schools.* London: HMSO.

Pommerich, M. (2007). Concordance: The Good, the Bad, and the Ugly. En N. J. Dorans, M. Pommerich, & P. W. Holland, *Linking and Aligning Scores and Scales*, (pp. 199-216). New York: Springer.

Pommerich, M., & Dorans, N. J. (2004). Linking socres via concordance: introduction to the special issue. *Applied Psychological Measurement, 28*(4), 216-218.

Pommerich, M., Hanson, B. A., Harris, D. J., & Sconing, J. A. (2004). Issues in Creating and Reporting Concordance Results Based on Equipercentile Methods. *Applied Psychological Measurement , 28*(4), 247-273.

Prellezo García, J. M. (2009). *Diccionario de Ciencias de la Educación* . Alcalá de Henares, Madrid: CCS.

Prieto Adánez, G., & Dias Velasco, A. (2003). Uso del modelo de Rasch para poner en la misma escala las puntuaciones de distintos tests. *Actualidades en psicología , 19*(106), 5-23.

Qian, J., Jiang, Y., & von Davier, A. A. (2013). *Weighting Test Samples in IRT Linking and Equating: Toward an Improved Sampling Design for Complex Equating.* Princeton, N. J: Educational Testing Service.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Real Decreto 1513/2006, de 7 de diciembre, por el que se establecen las enseñanzas mínimas de la Educación primaria. BOE núm. 293 (8 de diciembre de 2006), 43053-43102

Real Decreto 1631/2006, de 29 de diciembre, por el que se establecen las enseñanzas mínimas correspondientes a la Educación Secundaria Obligatoria. BOE núm. 5 (5 de enero de 2007), 677-773.

Revuelta, J., & Ponsoda, V. (2003). *Simulación de modelos estadísticos en ciencias sociales.* Madrid: La Muralla.

Rodríguez Espinar, S. (1982). *Factores de Rendimiento Escolar.* Barcelona: Oikos-tau.

Ruiz De Miguel, 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.

Rutter, M., Maughan, B., Mortimer, P., & Ouston, J. (1979). *Fifteen thousand hours: Secondary schools and their effects on children.* London: Open Books.

Rust, K. (2014). Sampling, weighting, and variance estimation in international large-scale assessments.En Rutkowski, L., von Davier, M., Rutkowski D. (eds) (2014). *Handbook of international large-scale assessment*. New York: CRC Press.

Southern and Eastern Africa Consortium for Monitoring Educational Quality (2010). [sitio web]. París: UNESCO [Consulta 20 de Septiembre de 2012]. Disponible en http://www.sacmeq.org/

Salvador Mata, F., Rodríguez Dieguez, J. L., & Bolivar Botia, A. (2004). *Diccionario Enciclopedico de Didáctica* (Vol. II). Málaga: Aljibe.

Samejima, F. (1969). Estimation of Latent Ability Using a Response Pattern of graded scores. *Psychometric Monograph , 34* (17), 1-100.

Scheerens, J. (2000). *Improving school effectivenness.* París: International Institute for Educational Planning.

Schulz, W., & Sibberns, H. (2004). *IEA Civic Education Study. Thecnical Report.* Amsterdam: International Association for the Evaluation of Educational Achievement (IEA).

Schulz, W., Ainley, J., & Fraillon, J. (2011). *ICCS 2009 Technical Report.* Amsterdam: International Association for the Evaluation of Educational Achievement (IEA).

Schulz, W., Ainley, J., Fraillon, J., Kerr, D., & Losito, B. (2010). *Initial Findings from the IEA International Civic an Citizenship Education Study.* Amsterdam: International Association for the Evaluation of Educational Achievement (IEA).

Schulz, W., Fraillon, J., Ainley, J., Losito, B., & Kerr, D. (2008). *International Civic and Citizenship Education Study Assessment Farmework.* Amsterdam: International Association for Evaluation of Educational Achievement.

Scriven, M. (1967). The methodology of evaluation. *Perspectives of curriculum evaluation*, *1*, 39-83.

Sheehan, K. M., & Mislevy, R. J. (1988). *Some Consequences of the Uncertainty in IRT Linking Procedures.* Princeton, N.J: Educational Testing Service.

Skaggs, G., & Lissitz, R. (1982). Test Equating: Relevant Issues and a Review of recent Research. *Anual Meeting of the American Educational Research Association.* Los Ángeles.

Skaggs, G., & Lissitz, R. W. (1986). IRT Test Equating: relevant issues and review of recent research. *Review of Educational Research, 56*(4), 495-529.

Solanas, A., & Sierra, V. (1992). Bootstrap: fundamentos e introducción a sus aplicaciones. *Anuario de psicología, 55*, 143-154.

Stocking, M. L., & Lord, F. M. (1983). Developing a Common Metric in Item Response Theory. *Applied psychological measurement, 7*(2), 201-210.

Stufflebeam, D. (1973). *Toward a science of educational evaluation.* Englewood Cliffs, NJ: Educational Thechnology Publications.

Svensson, A. (1971). *Relative Achievement School performance in relation to inteligence, sex and home enviroment.* Stockholm: Almqvist & Wiksell.

Tatto, M., Schwille, J., Senk, S., Ingvarson, L., Peck, R., & Rowley, G. (2008). *Teacher Education and Development Study in Mathematics (TEDS-M) Policy, Practice, and Readiness to Teach Primary and Secondary Mathematics Conceptual Framework.* Amsterdam: International Asociation for the Evaluation of Educational Achievement (IEA).

Tatto, M., Schwille, J., Senk, S., Ingvarson, L., Rowley, G., Peck, R., Bankov, K., Rodríguez, M., & Reckase, M (2012). *Policy, practice, and readiness to teach primary and secondary mathematics in 17 countries: Findings from the IEA Teacher Education and Development Study in Mathematics (TEDS-M).* Amsterdam: International Association for Evaluation of Educational Achievement (IEA).

Teresi, J. A., & Jones, R. N. (2013). Bias in psychological assessment and other measures. En K. F. Geisinger, *APA Handbok of Testing and Assessment in Psychology* (págs. 139-164). Washington D.C: American Psychological Asociation.

Theule, S. (2006). Examining Instruction, Achievement, And Equity with NAEP Mathematics Data. *Education Policy Analysis Archives*, *14* (14).

Tiana, A. (2009). Evaluación y cambio educativo: los debates ectuales sobre las ventajas y los riesgos de la evaluación. En E. Martín, & F. Martínez Rizo, *Avances y desafíos en la evaluación educativa* (pp. 17-26). Madrid: OEI y Fundación Santillana.

Tong, Y., & Kolen, M. J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education, 20*(2), 227-253.

Torney-Purta, J., Lehmann, R, Oswald, H., & Schulz, W. (2001). *Citizenship and Education in Twenty- eight Countries. Civic Konowledge and Engagement at age fourteen.* Amsterdam: International Asociation for the Evaluation of Educational Achievement (IEA).

Tourón, J. (1985). La predicción del rendimiento académico: procedimientos, resultados e implicaciones. *Revista Española de Pedagogía*, *169-170*(43), 473-495.

Tremblay, K., Lalancette, D., & Roseveare, D. (2012). *Assessment of higher education Learning Outcomes AHELO. Feasibility Study Report. Volume 1 – Design and Implementation.* París: Organization for the Economic Cooperation and Development.

Tremblay, K., Lalancette, D., & Roseveare, D. (2013). *Assessment of Higher Education Learning Outcomes. Feasibility Study Report. Volume 2 – Data Analysis and National Experiences*. París: Organization for the Economic Cooperation and Development.

Treviño, E., Katherine, P., Gempp, R., & Donoso Rivas, F. (2013). *Factores asociados al aprendizaje en el SERCE: análisis de los factores latentes y su vínculo con los resultados académicos de los niños.* Santiago, Chile: oficina Regional de Educación para América Latina y el Caribe (OREALC/UNESCO).

Tsai, T.-H., Hanson, B. A., Kolen, M. J., & Forsyth, R. A. (2001). A Comparison of Bootstrap Standard Errors of IRT Equating Methods for the Common-Item Nonequivalent Groups Design. *Applied Measurement in Education, 14*(1), 17-30.

Tukey, J. W. (1958). Bias and confidence in not quite large samples. *Annals of Mathematical Statistics, 29* (2).

Tyler, R. (1950). *Basic principles of curriculum and instruction.* Chicago: University of Chicago Press.

UNESCO. (1998). *Declaración mundial sobre la educación superior en el siglo XXI: visión y acción*. Conferencia Mundial sobre la Educación Superior, 9 de octubre de 1998, París. ED98/CONF202/CLD.40.

Urbanek, S (2009). Multicore: Parallel processing of R code on machines with multiple cores or CPUs. URL http://cran.r-project.org/package=multicore.

Valdés, H., Treviño, E., Acevedo, C. G., Castro, M., Carrillo, S., Costilla, R., Bogoya, D., & Pardo, C. (2008). *Los aprendizajes de los estudiantes de América Latina y el Caribe Primer reporte de los resultados del Segundo Estudio Regional Comparativo y Explicativo.* Santiago, Chile: Oficina Regional de Educación de la UNESCO para América Latina y el Caribe OREALC/UNESCO.

van der Linden, W. J. (2013). Some Conceptual Issues in Observed-Score Equating. *Journal of Educational Measurement, 50*(3), 249-285.

von Davier, A. A. (2007). Potential Solutions to Practical Equating Issues. En N. J. Dorans, M. Pommerich, & P. W. Holland, *Linking and Aligning Scpres and Scales* (pp. 89-106). New York: Springer.

von Davier, A. A. (2011). *Statistical Models for Test Equating, Scaling and Linking.* New York: Springer.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The Kernel Method of Test Equating.* New York: Springer.

Walberg, H. J. (1984). Improving the productivity of America's schools. *Educational Leadership*, *41(*8), 19-27.

Wall, W., & Varma, V. (1975). *Avances en psicología de la educación.* Madrid: Ediciones Morata.

Wang, T. (2006). Standard Errors of Equating for Equipercentile Equating with Log-Linear Pre- Smoothing using the Delta Method. *Center for Advanced Studies in Measurement and Assessment. CASMA Research Report,* (14).

Weber, G. (1971). *Inner-city children can be taught to read: four successful schools*.Washington, DC: Council for Basic Education.

Weeks, J., & Domingue, B. (2013). IRT-Based Test Linking in R. *Trainging Session at National Council on Measurement in Education,* 26 de Abril de 2013, San Francisco.

World Conference on Education for All. (1990). *World Declaration on Education for All and Framework for Action to meet Basic Learning Needs.* New York: UNESCO.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, *31*, 114-128.

Wu, M. (2010). Measurement, Sampling, and Equating Errors in Large-scale Assessments. *Educational Measurement: Issues and Practice*, *29*(4), 15-27.

Wu, M., Adams, R. J., Wilson, M. R., & Haldane, S. (2009). *ConQuest (Version 2.0,)* [Computer Software]. Camberwell: ACER.

Xu, X., & von Davier, M. (2010). *Linking Errors in Trend estimation in Large-Scale Surveys: a case study.* Princeton, NJ: ETS: Educational Testing Service.

Yang, W.-L., & Houang, R. (1996). The Effect of Anchor Length and Equating Method on the Accuracy of Test Equating: Comparisons of Linear and IRT-Based Equating Using an Anchor-Item Design. *Annual Meeting of the American Educational Research Association.* New York.

Yin, R., Brennan, R., & Kolen, M. (2004). Concordance Between ACT and ITED Scores From Different Populations. *Applied Psychological Measurement, 28*(4), 274-289.

Zabala, A., & Arnau, L. (2007). *Cómo aprender y enseñar competencias.* Barcelona: Graó.

Zeng, L. (1991). *Standard Errors of Linear Equating for the Single-Group Design.* Iowa City: American College Testing.

Zu, J., & Liu, J. (2010). Observed Score Equating Using Discrete and Passage-Based Anchor Items. *Journal of Educational Measurement, 47*(4), 395-412.

ANEXOS

ANEXO 1
Sintaxis de R para la generación de datos de simulación
Condición experimental 1

```
# Generación de Thetas y parámetros b de los ítems y su representación gráfica. También se generan los datos con distintas cantidades de DIF y se guardan

#################################################################

# Variable  IncrementoOriginal =  cantidad que se añadirá a los parámetros b en cada iteración con el fin de generar sucesivas condiciones de DIF.
IncrementoOriginal <- 0.1

# Variable directorio idem donde queremos los resultados
Directorio <- "/Users/Eva/Desktop/"

# Carga del paquete XLConnect para poder guardar datos en EXCEL
install.packages("XLConnect")

# Creación de subdirectorios para resultados
DirectorioGráficos <- paste(Directorio, "Gráficos/", sep = "")
DirectorioEXCEL <- paste(Directorio, "EXCEL/", sep = "")

# Creación de los subdirectorios donde se pondrán los gráficos y los resultados de EXCEL
# ShowWarnings = FALSE permite que en caso de que ya exista el directorio no avise y simplemente continúe ejecutando el programa
dir.create(DirectorioGráficos, showWarnings = FALSE)
dir.create(DirectorioEXCEL, showWarnings = FALSE)

# El comando sink envía la salida de la consola a un archivo (Al final se restaura)
# La opción sep = "" se pone para que no incluya ningún carácter entre los dos literales (si no, pone un blanco)
Archivo <- paste(Directorio, "Resultados.lis", sep = "")
sink(Archivo, append=TRUE)

# Se fija la semilla para que en todas las ejecuciones los valores generados sean los mismos y sea posible realizar comparaciones (De una iteración a la siguiente cambian los valores generados,  pero cuando se empieza a ejecutar, se repiten en el mismo orden)
set.seed(222)

# Generación de los valores de theta y b y representación gráfica
theta <- rnorm(2000, mean = 0, sd = 1)
b <- rnorm(50, 0, 0.5)
```

```
hist(theta, freq=T, breaks= "Scott", xlim = c(-3,
3),main="Distribución de Thetas y b's")
curve(dnorm(x,0,1)*412,col="red",add=T)
hist(b, freq=T, breaks= 12, col="green", add= T)
curve(dnorm(x,0,0.5)*10,col="blue",add=T)

# Se guarda el gráfico de resultados
Archivo <- paste(DirectorioGráficos, "Distribución Theta y
b.pdf", sep = "" )
dev.print(pdf, file=Archivo, width=5.7, height=5.7,
pointsize=12)

# Generación de grupo focal(1)y grupo de referencia (0)
set.seed(111)
Grupo <- rbinom(2000*1, size=1, prob=0.5)
DatosGeneradosBase <-transform(Grupo, Theta = theta)
rownames(DatosGeneradosBase) <- paste("Sujeto ", 1:2000, sep="")
colnames(DatosGeneradosBase) <- c("Grupo", "Theta")

# Comprobación de normalidad
shapiro.test(DatosGeneradosBase$Theta)
Theta como b
shapiro.test(b)

# Definición de la variable Grupo como factor, para poderla usar
en una prueba t como VI
DatosGeneradosBase$Grupo = factor(DatosGeneradosBase$Grupo )

# Prueba t para dos colas
t.test(theta~Grupo, alternative='two.sided', conf.level=.95,
var.equal=TRUE, data=DatosGeneradosBase)

# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #

# Generación del patrón de 1's y 0's.
# Primero se crean valores True False, luego se convierten a 1's
y 0s
respuestas <- outer(theta, b, function(theta, b) 1/(1+exp(-
(theta-b)))>(runif(100000,0,1)))
respuestas <- respuestas + 0 # Con este truco se pasa de matriz
lógica a numérica

# Combinación de los datos generados con las respuestas
DatosGenerados <- data.frame(DatosGeneradosBase, respuestas)

# Se guardan los datos en formato EXCEL
Archivo <- paste(DirectorioEXCEL, "DatosGenerados.csv", sep = ""
)
ArchivoXLS <- paste(DirectorioEXCEL, "DatosGenerados.xlsx", sep
= "" )

# En caso de no tener instalado el paquete XLConnect, se usaría
la instrucción
# write.csv y se guardaría en formato csv
# write.csv(DatosGenerados, file = Archivo, dec = ",",
fileEncoding = "UTF-16LE")
```

```
writeWorksheetToFile(ArchivoXLS , DatosGenerados,
"DatosGenerados")

# Comprobación de si los p de los ítems generados se
# comportan como los valores de b
p <- apply(respuestas, 2, mean)
plot(b,p, main="Comparación valores de p y b ítems generados")

# Se guarda el gráfico de resultado
Archivo <- paste(DirectorioGráficos, "Compara b y p.pdf", sep =
"" )
dev.print(pdf, file=Archivo, width=5.7, height=5.7,
pointsize=12)

# Calculo de los valores de X observadas
X <- apply(respuestas, 1, sum)
plot(theta, X, main="Comparación de los valores de Theta con las
X generadas.")
Archivo <- paste(DirectorioGráficos, "Compara X y Theta.pdf",
sep = "" )
dev.print(pdf, file=Archivo, width=5.7, height=5.7,
pointsize=12)

# Se guardan los valores originales de b en Oldb y le sumamos a
b el incremento
Oldb <- b
# Aquí generamos los valores de DIF
Incrb <- rnorm(15, 0, 0.05)
Nb <- rep(0, times=35)
Incrb <- c(Nb, Incrb)

# A continuación aparece la secuencia que hay que repetir para
cada valor distinto de incremento de las dificultades del DIF

# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #

for (i in 1:10) {

# Generación de los incrementos adicionales de b para los 15
últimos ítems
Incr <- 0.1*i
Nb <- rep(0, times=35)
Nb2 <- rep(Incr, times=15)
Nb <- c(Nb, Nb2)

# Generación del vector de incrementos
Incrb <- Incrb + Nb

# Generación de los nuevos valores de b sumándole a Oldb el
incremento
b <- Oldb +Incrb

# Generación del patrón de 1's y 0's para los datos con DIF.
# En esta matriz todos los sujetos contestan a los 15 últimos
ítems con los valores de dificultad más altos
```

```
# Primero se crean valores True False, luego se convierten a 1's
y 0s
respuestasDIF <- outer(theta, b, function(theta, b) 1/(1+exp(-
(theta-b)))>(runif(100000,0,1)))
respuestasDIF <- respuestasDIF + 0 # Con este truco se pasa de
matriz lógica a numérica

# En la matriz RespuestasT ponemos a los sujetos del grupo 0(
referencia) el valor de la matriz respuestas, y a los sujetos
del grupo 1 (focal) los valores de la matriz RespuestasDIF
# Primero se crea la matriz RespuestasT y se dimensiona
RespuestasT <- rep(0, times= 100000)
dim(RespuestasT) <- c(2000,50)
RespuestasT[which(Grupo==0),] <- respuestas[which(Grupo==0),]
RespuestasT[which(Grupo==1),] <- respuestasDIF
[which(Grupo==1),]

# Comprobación de los resultados de la generación
# Se comprueba si los p de los ítems generados se
# comportan como los valores de b
Título <- paste("Comparación valores de p y b ítems generados.
Iter = ", i, sep = "")
pDIF <- apply(RespuestasT, 2, mean)
plot(b,pDIF , main=Título , sub="Datos con DIF")


# Se guarda el gráfico de resultado
Archivo <- paste(DirectorioGráficos, "Compara b y p (Datos con
DIF) ", "Iter =",i,".pdf", sep = "" )
dev.print(pdf, file=Archivo , width=5.7, height=5.7,
pointsize=12)

XDIF <- apply(RespuestasT, 1, sum)
Título <- paste("Valores de Theta vs X generadas. Iter =", i,
sep = "")
plot(theta, XDIF , main=Título , sub="Datos con DIF")
Archivo <- paste(DirectorioGráficos, "Compara X y Theta (Datos
con DIF) ", "Iter =",i,".pdf", sep = "" )
dev.print(pdf, file=Archivo , width=5.7, height=5.7,
pointsize=12)

# Contraste de hipótesis
Título <- paste("Comparación de XDIF para los dos grupos iter
=", i, sep = "")
print(Título)
tstat <- t.test(XDIF ~Grupo, alternative='two.sided',
conf.level=.95, var.equal=FALSE)
print(tstat)

# Se comparan los datos de los ítems con y sin DIF
Título <- paste("Relación entre parámetros b originales y con
DIF. Iter =", i, sep = "")
plot(Oldb,b, main=Título )
Archivo <- paste(DirectorioGráficos, "Compara Oldb y b (Datos
con DIF) ", "Iter =",i,".pdf", sep = "" )
```

```
dev.print(pdf, file=Archivo , width=5.7, height=5.7,
pointsize=12)

# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #

Título <- paste("Relación entre parámetros b originales y p con
DIF. Iter =", i, sep = "")
plot(Oldb, pDIF,xlim=c(-1.5,1.5),ylim=c(0.3,0.8), main=Título)
Título <- paste("Relación entre parámetros b y p sin DIF. Iter
=", i, sep = "")
plot(b,p,xlim=c(-1.5,1.5),ylim=c(0.3,0.8),col="red",main=Título)

# También al haber DIF tiene que aumentar la varianza observada,
ya que aumenta la de error
Título <- paste("Comparación de varianza original con observada
(posible cambio en la varianza) iter =", i, sep = "")
print(Título)
print(var(X))
print(var(XDIF))

# Se guardan en EXCEL los datos generados en esta iteración
# Primero se crea el data frame con los datos que incorporan el
DIF
# En cada iteración varía el contenido de la matriz
'respuestas', pero la base es la misma
DatosGenerados <- data.frame(DatosGeneradosBase, RespuestasT)
ArchivoXLS <- paste(DirectorioEXCEL, "DatosGenerados iter=
",i,".xlsx", sep = "" )

# En caso de no tener instalado el paquete XLConnect, se usaría
la instrucción
# write.csv y se guardaría en formato csv
# Archivo <- paste(DirectorioEXCEL, "DatosGenerados.csv", sep =
"" )
# write.csv(DatosGenerados, file = Archivo, dec = ",",
fileEncoding = "UTF-16LE")
writeWorksheetToFile(ArchivoXLS , DatosGenerados,
"DatosGenerados")

}
# Hasta aquí llega lo que se repite para cada valor del
incremento
sink()

# Fin
```

ANEXO 2
Sintaxis de R para la generación de datos de simulación
Condición experimental 2

```
# Generación de Thetas y parámetros b de los ítems y su
representación gráfica. También se generan los datos con
distintas medias para los grupos

#################################################################

IncrementoOriginal <- 0.1

# En la variable Directorio se pone el idem donde queremos los
resultados
Directorio <- "c:/Users/Eva/Desktop/"    #
<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<     AQUÍ

# Se carga el paquete XLConnect para poder guardar datos en
EXCEL Si no tenemos instalado el paquete, hay que ejecutar el
comando
install.packages("XLConnect")

# library(XLConnect)

#################################################################

# Creación de subdirectorios para resultados
DirectorioGráficos <- paste(Directorio, "Gráficos2/", sep = "")
DirectorioEXCEL <- paste(Directorio, "EXCEL2/", sep = "")

# Creación de los subdirectorios donde se pondrán los gráficos y
los resultados de EXCEL
# La opción showWarnings = FALSE se utiliza para que en caso de
que ya exista el directorio no avise y simplemente continúe
ejecutando el programa
dir.create(DirectorioGráficos, showWarnings = FALSE)
dir.create(DirectorioEXCEL, showWarnings = FALSE)

# El comando sink envía la salida de la consola a un archivo (Al
final se restaura)
# La opción sep = "" se pone para que no incluya ningún carácter
entre los dos literales (si no, pone un blanco)
Archivo <- paste(Directorio, "Resultados.lis", sep = "")
sink(Archivo, append=TRUE)

# Se fija la semilla para que en todas las ejecuciones los
valores generados sean los mismos y sea posible hacer
comparaciones ((De una iteración a la siguiente cambian los
valores generados, pero cuando se empieza a ejecutar, se repiten
en el mismo orden)

set.seed(222)
```

```
# Generación de los valores de theta y b y representación
gráfica de valores de theta para los grupos 0 y 1 por separado
theta0 <- rnorm(1000, mean = 0, sd = 1)
theta1 <- rnorm(1000, mean = 0, sd = 1)
theta <- c(theta0 ,theta1 )
b <- rnorm(50, 0, 0.5)
hist(theta, freq=T, breaks= "Scott", xlim = c(-3,
3),main="Distribución de Thetas y b's")
curve(dnorm(x,0,1)*412,col="red",add=T)
hist(b, freq=T, breaks= 12, col="green", add= T)
curve(dnorm(x,0,0.5)*10,col="blue",add=T)

# Se guarda el gráfico de resultado
Archivo <- paste(DirectorioGráficos, "Distribución Theta y
b.pdf", sep = "" )
dev.print(pdf, file=Archivo, width=5.7, height=5.7,
pointsize=12)

# Generación de grupo focal(1)y grupo de referencia (0)
set.seed(111)
Grupo <- rep(0:1, each = 1000)
DatosGeneradosBase <-transform(Grupo, Theta = theta)
rownames(DatosGeneradosBase) <- paste("Sujeto ", 1:2000, sep="")
colnames(DatosGeneradosBase) <- c("Grupo", "Theta")

# Comprobación de normalidad
shapiro.test(theta0)
shapiro.test(theta1)
shapiro.test(DatosGeneradosBase$Theta)
shapiro.test(b)

# La variable Grupo se define como factor, para poderla usar en
una prueba t como VI
DatosGeneradosBase$Grupo = factor(DatosGeneradosBase$Grupo )

# Hacemos la prueba t para dos colas
t.test(theta~Grupo, alternative='less', conf.level=.95,
var.equal=TRUE, data=DatosGeneradosBase)

# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #

# Generación del patrón de 1's y 0's.
# Primero se crean valores True False, luego se convierten a 1's
y 0s
respuestas <- outer(theta, b, function(theta, b) 1/(1+exp(-
(theta-b)))>(runif(100000,0,1)))
respuestas <- respuestas + 0 # Con este truco se pasa de matriz
lógica a numérica

# Combinación de los datos generados con las respuestas
DatosGenerados <- data.frame(DatosGeneradosBase, respuestas)

# Se guardan los datos en formato EXCEL
Archivo <- paste(DirectorioEXCEL, "DatosGenerados.csv", sep = ""
)
ArchivoXLS <- paste(DirectorioEXCEL, "DatosGenerados.xlsx", sep
```

```
= "" )

# En caso de no tener instalado el paquete XLConnect, se usaría
la instrucción
# write.csv y se guardaría en formato csv
# write.csv(DatosGenerados, file = Archivo, dec = ",",
fileEncoding = "UTF-16LE")
writeWorksheetToFile(ArchivoXLS , DatosGenerados,
"DatosGenerados")

# Comprobación del comportamiento de los p de los ítems
generados (si comportan como los valores de b)
p <- apply(respuestas, 2, mean)
plot(b,p, main="Comparación valores de p y b ítems generados")

# Se guarda el gráfico de resultado
Archivo <- paste(DirectorioGráficos, "Compara b y p.pdf", sep =
"" )
dev.print(pdf, file=Archivo, width=5.7, height=5.7,
pointsize=12)

# Cálculo de los valores de X observadas
X <- apply(respuestas, 1, sum)
plot(theta, X, main="Comparación de los valores de Theta con las
X generadas.")
Archivo <- paste(DirectorioGráficos, "Compara X y Theta.pdf",
sep = "" )
dev.print(pdf, file=Archivo, width=5.7, height=5.7,
pointsize=12)


# A continuación aparece la secuencia que hay que repetir para
cada valor distinto de incremento

# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #

for (i in 1:10) {

# Generación de los incrementos adicionales en habilidad para el
grupo 1

theta1 <-  theta1 + IncrementoOriginal
theta <- c(theta0 ,theta1 )

DatosGeneradosBase <-transform(Grupo, Theta = theta)
rownames(DatosGeneradosBase) <- paste("Sujeto ", 1:2000, sep="")
colnames(DatosGeneradosBase) <- c("Grupo", "Theta")


# Generación del patrón de 1's y 0's para los datos.
# Primero se crean valores True False, luego se convierten a 1's
y 0s
respuestasDIF <- outer(theta, b, function(theta, b) 1/(1+exp(-
(theta-b)))>(runif(100000,0,1)))
respuestasDIF <- respuestasDIF + 0 # Con este truco se pasa de
matriz lógica a numérica
```

```
# En la matriz RespuestasT ponemos a los sujetos del grupo 0(
referencia) el valor de la matriz respuestas, y a los sujetos
del grupo 1 (focal) los valores de la matriz RespuestasDIF

# Creación de la matriz RespuestasT y se dimensiona
RespuestasT <- rep(0, times= 100000)
dim(RespuestasT) <- c(2000,50)
RespuestasT[which(Grupo==0),] <- respuestas[which(Grupo==0),]
RespuestasT[which(Grupo==1),] <- respuestasDIF
[which(Grupo==1),]

XDIF <- apply(RespuestasT, 1, sum)
Título <- paste("Valores de Theta vs X generadas. Iter =", i,
sep = "")
plot(theta, XDIF , main=Título , sub="Datos con diferencias en
Theta media")
Archivo <- paste(DirectorioGráficos, "Compara X y Theta
(diferencias en Theta media) ", "Iter =",i,".pdf", sep = "" )
dev.print(pdf, file=Archivo , width=5.7, height=5.7,
pointsize=12)

# Contraste de hipótesis
Título <- paste("Comparación de XDIF para los dos grupos iter
=", i, sep = "")
print(Título)
tstat <- t.test(XDIF ~Grupo, alternative='less', conf.level=.95,
var.equal=FALSE)
print(tstat)

# Comparación de los datos de los ítems con y sin diferencias en
Theta media

# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #

Título <- paste("Comparación de varianza original con observada
(posible cambio en la varianza) iter =", i, sep = "")
print(Título)
print(var(X))
print(var(XDIF))

# Se guardan en EXCEL los datos generados en esta iteración
# Primero creamos el data frame con los datos que incorporan la
diferencia
# En cada iteración varía el contenido de la matriz
'respuestas', pero la base es la misma
DatosGenerados <- data.frame(DatosGeneradosBase, RespuestasT)
ArchivoXLS <- paste(DirectorioEXCEL, "DatosGenerados iter=
",i,".xlsx", sep = "" )

# En caso de no tener instalado el paquete XLConnect, se usaría
la instrucción
# write.csv y se guardaría en formato csv
# Archivo <- paste(DirectorioEXCEL, "DatosGenerados.csv", sep =
"" )
# write.csv(DatosGenerados, file = Archivo, dec = ",",
```

```
fileEncoding = "UTF-16LE")
writeWorksheetToFile(ArchivoXLS , DatosGenerados,
"DatosGenerados")

}
# Hasta aquí llega lo que se repite para cada valor del
incremento
sink()

# Fin
```

ANEXO 3
Sintaxis de R para la generación de datos de simulación
Condición experimental 3

```
# Generación de Thetas y parámetros b de los ítems y su
representación gráfica. También se generan los datos con un
incremento en la varianza de b y se representan gráficamente.

#################################################################
##

# En la variable IncrementoOriginal se indicca la cantidad que
se añadirá a los parámetros b en cada iteración. Esto es lo que
tenemos que cambiar hasta que tengamos un resultado razonable.
IncrementoOriginal <- 0.1

# En la variable Directorio se pone el idem donde queremos los
resultados
    Directorio <- "/Users/Eva/Desktop/"    #
<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<     AQUÍ

# Se carga el paquete XLConnect para poder guardar datos en
EXCEL
# Si el paquete no está instalado, hay que ejecutar el comando
install.packages("XLConnect")
# desde la consola. También se puede hacer desde el menú
library(XLConnect)

#################################################################
##

# Creación de subdirectorios para poner resultados
DirectorioGráficos <- paste(Directorio, "Gráficos3/", sep = "")
DirectorioEXCEL <- paste(Directorio, "EXCEL3/", sep = "")
# Creación de los subdirectorios donde se pondrán los gráficos y
los resultados de EXCEL
# La opción showWarnings = FALSE es utilizada para que en caso
de que ya exista el directorio no avise y simplemente continúe
ejecutando el programa
dir.create(DirectorioGráficos, showWarnings = FALSE)
dir.create(DirectorioEXCEL, showWarnings = FALSE)

# El comando sink envía la salida de la consola a un archivo (Al
final se restaura)
# La opción sep = "" se pone para que no incluya ningún carácter
entre los dos literales (si no, pone un blanco)
Archivo <- paste(Directorio, "Resultados.lis", sep = "")
sink(Archivo, append=TRUE)


# Se fija la semilla para que en todas las ejecuciones los
valores generados sean los mismos y sea posible hacer
```

```
comparaciones ((De una iteración a la siguiente cambian los
valores generados, pero cuando se empieza a ejecutar, se repiten
en el mismo orden)
set.seed(222)

# Generación de los valores de theta y b y representación
gráfica
theta <- rnorm(2000, mean = 0, sd = 1)
b <- rnorm(50, 0, 0.1)
hist(theta, freq=T, breaks= "Scott", xlim = c(-3,
3),main="Distribución de Thetas y b's")
curve(dnorm(x,0,1)*412,col="red",add=T)
hist(b, freq=T, breaks= 12, col="green", add= T)
curve(dnorm(x,0 ,0.1)*10,col="blue",add=T)

# Se guarda el gráfico de resultado
Archivo <- paste(DirectorioGráficos, "Distribución Theta y
b.pdf", sep = "" )
dev.print(pdf, file=Archivo, width=5.7, height=5.7,
pointsize=12)

# Generación de grupo
set.seed(111)
Grupo <- rbinom(2000*1, size=1, prob=0.5)
DatosGeneradosBase <-transform(Grupo, Theta = theta)
rownames(DatosGeneradosBase) <- paste("Sujeto ", 1:2000, sep="")
colnames(DatosGeneradosBase) <- c("Grupo", "Theta")

# Comprobación de normalidad
shapiro.test(DatosGeneradosBase$Theta) # Salen normales tanto
Theta como b
shapiro.test(b)

# La variable Grupo se define como factor, para poderla usar en
una prueba t como VI
DatosGeneradosBase$Grupo = factor(DatosGeneradosBase$Grupo )

# Prueba t para dos colas
t.test(theta~Grupo, alternative='two.sided', conf.level=.95,
var.equal=TRUE, data=DatosGeneradosBase)

# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #
#

# Generación del patrón de 1's y 0's.
# Creación de los valores True False, luego se convierten a 1's
y 0s
respuestas <- outer(theta, b, function(theta, b) 1/(1+exp(-
(theta-b)))>(runif(100000,0,1)))
respuestas <- respuestas + 0 # Con este truco se pasa de matriz
lógica a numérica

# Se combinan los datos generados con las respuestas
DatosGenerados <- data.frame(DatosGeneradosBase, respuestas)

# Se guardan los datos en formato EXCEL
```

```
Archivo <- paste(DirectorioEXCEL, "DatosGenerados.csv", sep = ""
)
ArchivoXLS <- paste(DirectorioEXCEL, "DatosGenerados.xlsx", sep
= "" )

# En caso de no tener instalado el paquete XLConnect, se usaría
la instrucción
# write.csv y se guardaría en formato csv
# write.csv(DatosGenerados, file = Archivo, dec = ",",
fileEncoding = "UTF-16LE")
writeWorksheetToFile(ArchivoXLS , DatosGenerados,
"DatosGenerados")

# Comprobación del comportamiento de los p de los ítems
generados (en relación a los valores de b)
p <- apply(respuestas, 2, mean)
plot(b,p, main="Comparación valores de p y b ítems generados")

# Se guarda el gráfico de resultado
Archivo <- paste(DirectorioGráficos, "Compara b y p.pdf", sep =
"" )
dev.print(pdf, file=Archivo, width=5.7, height=5.7,
pointsize=12)

# Cálcujlo de los valores de X observadas
X <- apply(respuestas, 1, sum)
plot(theta, X, main="Comparación de los valores de Theta con las
X generadas.")
Archivo <- paste(DirectorioGráficos, "Compara X y Theta.pdf",
sep = "" )
dev.print(pdf, file=Archivo, width=5.7, height=5.7,
pointsize=12)


# A continuación aparece la secuencia que hay que repetir para
cada valor distinto de varianza de b

# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #
#

for (i in 1:10) {

# Generación de los incrementos adicionales de b
Incr <- 0.1*i
b <- rnorm(50, 0, 0.1+Incr)

# Generación del patrón de 1's y 0's para los datos.
# Primero se crean valores True False, luego se convierten a 1's
y 0s
respuestas <- outer(theta, b, function(theta, b) 1/(1+exp(-
(theta-b)))>(runif(100000,0,1)))
respuestas <- respuestas + 0 # Con este truco se pasa de matriz
lógica a numérica


# Comprobación de los resultados de la generación
```

```
# Se comprueba si los p de los ítems generados se comportan como
los valores de b
Título <- paste("Comparación valores de p y b ítems generados.
Iter = ", i, sep = "")
p <- apply(respuestas , 2, mean)
plot(b,p , main=Título , sub="Datos con desplazamiento en rango
de b")


# Se guarda el gráfico de resultado
Archivo <- paste(DirectorioGráficos, "Compara b y p diferencias
en rango de b ", "Iter =",i,".pdf", sep = "" )
dev.print(pdf, file=Archivo , width=5.7, height=5.7,
pointsize=12)
X <- apply(respuestas, 1, sum)
Título <- paste("Valores de Theta vs X generadas. Iter =", i,
sep = "")
plot(theta, X, main=Título)
Archivo <- paste(DirectorioGráficos, "Compara X y Theta ", "Iter
=",i,".pdf", sep = "" )
dev.print(pdf, file=Archivo, width=5.7, height=5.7,
pointsize=12)
Título <- paste("Distribución de Thetas y b's. Iter =", i, sep =
"")
hist(theta, freq=T, breaks= "Scott", xlim = c(-3,
3),main=Título)
curve(dnorm(x,0,1)*412,col="red",add=T)
hist(b, freq=T, breaks= 12, col="green", add= T)
curve(dnorm(x,0 ,0.1+Incr)*10,col="blue",add=T)

# Se guarda el gráfico de resultado
Archivo <- paste(DirectorioGráficos, "Distribución de Thetas y
b's ", "Iter =",i,".pdf", sep = "")
dev.print(pdf, file=Archivo, width=5.7, height=5.7,
pointsize=12)


# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #
#

# Varianza original y observada
Título <- paste("Comparación de varianza original con observada
(posible cambio en la varianza) iter =", i, sep = "")
print(Título)
print(var(X))
print(var(XDIF))

# Se guardan en EXCEL los datos generados en esta iteración
# Primero se crea el data frame con los datos
# En cada iteración varía el contenido de la matriz
'respuestas', pero la base es la misma
DatosGenerados <- data.frame(DatosGeneradosBase, respuestas)
ArchivoXLS <- paste(DirectorioEXCEL, "DatosGenerados iter=
",i,".xlsx", sep = "" )

# En caso de no tener instalado el paquete XLConnect, se usaría
```

```
la instrucción
# write.csv y se guardaría en formato csv
# Archivo <- paste(DirectorioEXCEL, "DatosGenerados.csv", sep =
"" )
# write.csv(DatosGenerados, file = Archivo, dec = ",",
fileEncoding = "UTF-16LE")
writeWorksheetToFile(ArchivoXLS , DatosGenerados,
"DatosGenerados")

}
# Hasta aquí llega lo que se repite para cada valor del
incremento
sink()

# Fin
```

ANEXO 4
Sintaxis de R para la generación de datos de simulación
Condición experimental 4

```
# Generación de Thetas y parámetros b de los ítems y su
representación gráfica. Se generan distintos valores para
la distribución de los valores de b, desplazando su media
en cada iteración.

##############################################################
##

# En la variable Directorio se indica el idem donde
queremos los resultados
Directorio <- "/Users/Eva/Desktop/"    #
<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<     AQUÍ

# Carga del paquete XLConnect para poder guardar datos en
EXCEL
# Si no esta instalado el paquete, hay que ejecutar el
comando install.packages("XLConnect")
# desde la consola. También se puede hacer desde el menú
library(XLConnect)

##############################################################
#

# Cración de subdirectorios para poner resultados
DirectorioGráficos <- paste(Directorio, "Gráficos4/", sep =
"")
DirectorioEXCEL <- paste(Directorio, "EXCEL4/", sep = "")

# Creación de los subdirectorios donde se pondrán los
gráficos y los resultados de EXCEL
# La opción showWarnings = FALSE se utilizapara que en caso
de que ya exista el directorio no avise y simplemente
continúe ejecutando el programa
dir.create(DirectorioGráficos, showWarnings = FALSE)
dir.create(DirectorioEXCEL, showWarnings = FALSE)

# El comando sink envía la salida de la consola a un
archivo (Al final se restaura)
# La opción sep = "" se pone para que no incluya ningún
carácter entre los dos literales (si no, pone un blanco)
Archivo <- paste(Directorio, "Resultados.lis", sep = "")
sink(Archivo, append=TRUE)


# Se fija la semilla para que en todas las ejecuciones los
```

```
valores generados sean los mismos y sea posible realizar
comparaciones ((De una iteración a la siguiente cambian los
valores generados,pero cuando se empieza a ejecutar, se
repiten en el mismo orden)
set.seed(222)

# Se generan los valores de theta y b y los representamos
gráficamente
theta <- rnorm(2000, mean = 0, sd = 1)
b <- rnorm(50, 0, 0.5)
hist(theta, freq=T, breaks= "Scott", xlim = c(-3,
3),main="Distribución de Thetas y b's")
curve(dnorm(x,0,1)*412,col="red",add=T)
hist(b, freq=T, breaks= 12, col="green", add= T)
curve(dnorm(x,0,0.5)*10,col="blue",add=T)

# Se guarda el gráfico de resultado
Archivo <- paste(DirectorioGráficos, "Distribución Theta y
b.pdf", sep = "" )
dev.print(pdf, file=Archivo, width=5.7, height=5.7,
pointsize=12)

# Generación de grupos
set.seed(111)
Grupo <- rbinom(2000*1, size=1, prob=0.5)
DatosGeneradosBase <-transform(Grupo, Theta = theta)
rownames(DatosGeneradosBase) <- paste("Sujeto ", 1:2000,
sep="")
colnames(DatosGeneradosBase) <- c("Grupo", "Theta")

# Comprobación de normalidad
shapiro.test(DatosGeneradosBase$Theta) # Salen normales
tanto Theta como b
shapiro.test(b)

# La variable Grupo se define como factor, para poderla
usar en una prueba t como VI
DatosGeneradosBase$Grupo = factor(DatosGeneradosBase$Grupo
)

# Prueba t para dos colas
t.test(theta~Grupo, alternative='two.sided',
conf.level=.95, var.equal=TRUE, data=DatosGeneradosBase)

# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #
#


# Generación del patrón de 1's y 0's.
# Primero se crean valores True False, luego se convierten
a 1's y 0s
```

```
respuestas <- outer(theta, b, function(theta, b) 1/(1+exp(-
(theta-b)))>(runif(100000,0,1)))
respuestas <- respuestas + 0 # Con este truco se pasa de
matriz lógica a numérica

# Combinación de los datos generados con las respuestas
DatosGenerados <- data.frame(DatosGeneradosBase,
respuestas)

# Se guardan los datos en formato EXCEL
Archivo <- paste(DirectorioEXCEL, "DatosGenerados.csv", sep
= "" )
ArchivoXLS <- paste(DirectorioEXCEL, "DatosGenerados.xlsx",
sep = "" )

# En caso de no tener instalado el paquete XLConnect, se
usaría la instrucción
# write.csv y se guardaría en formato csv
# write.csv(DatosGenerados, file = Archivo, dec = ",",
fileEncoding = "UTF-16LE")
writeWorksheetToFile(ArchivoXLS , DatosGenerados,
"DatosGenerados")

# Se comprueba si los p de los ítems generados se comportan
como los valores de b
p <- apply(respuestas, 2, mean)
plot(b,p, main="Comparación valores de p y b ítems
generados")

# Se guarda el gráfico de resultado
Archivo <- paste(DirectorioGráficos, "Compara b y p.pdf",
sep = "" )
dev.print(pdf, file=Archivo, width=5.7, height=5.7,
pointsize=12)

# Cálculo de los valores de X observadas
X <- apply(respuestas, 1, sum)
plot(theta, X, main="Comparación de los valores de Theta
con las X generadas.")
Archivo <- paste(DirectorioGráficos, "Compara X y
Theta.pdf", sep = "" )
dev.print(pdf, file=Archivo, width=5.7, height=5.7,
pointsize=12)


# A continuación aparece la secuencia que hay que repetir para
cada valor distinto de media en los parámetros de b

# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #
#
```

```
for (i in 1:10) {

# Se generan los incrementos adicionales en la media de los
valores de b
Incr <- 0.1*i
b <- rnorm(50, 0+Incr, 0.5 )

# Se genera el patrón de 1's y 0's para los datos con DIF.
# En esta matriz todos los sujetos contestan a los 15
últimos ítems con los valores de dificultad más altos
# Primero se crean valores True False, luego se convierten
a 1's y 0s
respuestas <- outer(theta, b, function(theta, b) 1/(1+exp(-
(theta-b)))>(runif(100000,0,1)))
respuestas <- respuestas + 0 # Con este truco se pasa de
matriz lógica a numérica


# Comprobación de los resultados de la generación
# Se comprueba si los p de los ítems generados se comportan
como los valores de b
Título <- paste("Comparación valores de p y b ítems
generados. Iter = ", i, sep = "")
p <- apply(respuestas , 2, mean)
plot(b,p , main=Título , sub="Datos con desplazamiento en
rango de b")


# Se guarda el gráfico de resultado
Archivo <- paste(DirectorioGráficos, "Compara b y p
desplazamiento en rango de b ", "Iter =",i,".pdf", sep = ""
)
dev.print(pdf, file=Archivo , width=5.7, height=5.7,
pointsize=12)
X <- apply(respuestas, 1, sum)
Título <- paste("Valores de Theta vs X generadas. Iter =",
i, sep = "")
plot(theta, X, main=Título)
Archivo <- paste(DirectorioGráficos, "Compara X y Theta ",
"Iter =",i,".pdf", sep = "" )
dev.print(pdf, file=Archivo, width=5.7, height=5.7,
pointsize=12)
Título <- paste("Distribución de Thetas y b's. Iter =", i,
sep = "")
hist(theta, freq=T, breaks= "Scott", xlim = c(-3,
3),main=Título , sep = "")
curve(dnorm(x,0,1)*412,col="red",add=T)
hist(b, freq=T, breaks= 12, col="green", add= T)
curve(dnorm(x,0+Incr,0.5)*10,col="blue",add=T)

# Se guarda el gráfico de resultado
```

```
Archivo <- paste(DirectorioGráficos, "Distribución de
Thetas y b's ", "Iter =",i,".pdf", sep = "")
dev.print(pdf, file=Archivo, width=5.7, height=5.7,
pointsize=12)

# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #
#

# Representación aumento de la varianza observada
Título <- paste("Comparación de varianza original con
observada (posible cambio en la varianza) iter =", i, sep =
"")
print(Título)
print(var(X))
print(var(XDIF))

# Se guarda en EXCEL los datos generados en esta iteración
Primero se crea el data frame con los datos que incorporan
incremento en la media de b. En cada iteración varía el
contenido de la matriz 'respuestas', pero la base es la
misma
DatosGenerados <- data.frame(DatosGeneradosBase,
respuestas)
ArchivoXLS <- paste(DirectorioEXCEL, "DatosGenerados iter=
",i,".xlsx", sep = "" )

# En caso de no tener instalado el paquete XLConnect,se
usaría la instrucción
# write.csv y se guardaría en formato csv
# Archivo <- paste(DirectorioEXCEL, "DatosGenerados.csv",
sep = "" )
# write.csv(DatosGenerados, file = Archivo, dec = ",",
fileEncoding = "UTF-16LE")
writeWorksheetToFile(ArchivoXLS , DatosGenerados,
"DatosGenerados")

}
# Hasta aquí llega lo que se repite para cada valor del
incremento
sink()
# Fin
```

ANEXO 5
Sintaxis de R para la ejecución del procedimiento «bootstrap bidimensional»

```
#install.packages("XLConnect")
#install.packages("mirt")
Directorio <- "/Users/evaexpositocasas/Desktop/"  #
<<<<<<<<<<<<<<<<<<     AQUÍ EL DIRECTORIO GENERAL
options(java.parameters = "-Xmx4g")
library(mirt)
library(XLConnect)

for (vueltas in 0:10  ) {

filainicial <- 1
incremento <- 1999
filafinal <- filainicial +incremento
ntotalfilas <- 2000
ntotalcolumnas <- 50

# A continuación se cargan los datos generados
DirectorioEXCEL <- paste(Directorio, "EXCEL1/", sep = "")
# <<<<<<<<<<<<<<<<<<     AQUÍ EL DIRECTORIO DE DATOS
ArchivoXLS <- paste(DirectorioEXCEL, "DatosGenerados iter=
",vueltas ,".xlsx", sep = "" )

# El comando sink envía la salida de la consola a un
archivo (Al final se restaura)
# La opción sep = "" se pone para que no incluya ningún
carácter entre los dos literales (si no, pone un blanco)
# Para que la ejecución sea más rápida, se deja esto con
comentarios, Si se desea ver la salida paso a paso, se
quita el comentario de la linea siguiente
#sink(Archivo, append=TRUE)

data  <- readWorksheetFromFile(ArchivoXLS , sheet = 1,
                               header = TRUE, startCol =
2,
                               startRow = 1, endCol = 52,
                               endRow = 2001)

# Se definen las filas y columnas de subíndices. Los
valores definitivos tienen que ser nfilas = 2000 y
ncolumnas = 50.
nfilas <- incremento +1
ncolumnas <- 50

# Datos en x para operar con ellos
x <- data
# Definición de las matrices cuadradas de subíndices que
```

```
vamos se utilizarán, para las filas y las columnas
i <-round(runif(ntotalcolumnas *ntotalcolumnas
,1,ntotalcolumnas ), digits = 0)
i <- matrix(i, nrow=ntotalcolumnas ,ncol=ntotalcolumnas )
j <-round(runif(ntotalfilas*ntotalfilas,1,ntotalfilas),
digits = 0)
j <- matrix(j, nrow=ntotalfilas,ncol=ntotalfilas)

# En h se guardan los resultados de cada casilla, la suma,
la suma de cuadrados y la media (ANOVA) y las diferencias
entre theta y theta original (error cuadrático medio).
# Estructura de cuatro capas, cada una de ellas con nfilas
X ncolumnas casillas.
# En la primera capa se guardan las sumas de scores, en la
segunda las sumas de scores cuadradas, en la tercera la
media de las scores y en la cuarta la diferencia entre
theta originales y estimadas.
h <- rep(0,ntotalfilas*ntotalcolumnas*4)
dim(h) <- c(ntotalfilas, ntotalcolumnas,4)

# Interacción del muestreo y resultados en h
# La función system.time() indica cuánto tiempo se ha
tardado en ejecutar la rutina
system.time(
for (k in 1:ntotalcolumnas  ) {
    for (l in filainicial:filafinal ) {
    y <- x[j[,l],i[k,]+1]
    scores <- fscores(mirt(y, 1), full.scores = TRUE,
method='EAP', scores.only = TRUE)
      h[l,k,1] = sum(scores)
    h[l,k,2] = sum(scores^2)
    h[l,k,3] = mean(scores)

    h[l,k,4] = sum((scores-x[j[,l],1])^2)
    }}
)

# Se guardan en EXCEL los datos generados en esta iteración
# Primero se crea el data frame. En cada iteración varía el
contenido de la matriz 'respuestas', pero la base es la
misma
Datosh <- data.frame(h)
ArchivoXLS <- paste(DirectorioEXCEL, "Resultados Iter =
",vueltas,".xlsx", sep = "" )
writeWorksheetToFile(ArchivoXLS , Datosh , "h")
}
```