UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE PSICOLOGÍA

Departamento de Metodología de las Ciencias del Comportamiento

TESIS DOCTORAL

RECUPERACIÓN DE ESTRUCTURAS LATENTES Y ESTIMACIÓN DE PARÁMETROS EN TEST MULTIDIMENSIONALES ORDINALES: ANALISIS COMPARADO ENTRE LA TRI LA TCT Y EL ANÁLISIS FACTORIAL

MEMORIA PARA OPTAR ALGRADO DE DOCTOR PRESENTADA POR

Rodrigo A. Asún Inostroza

Director

Jesús Mª Alvarado Izquierdo

Madrid, 2015

Dedicado a mi madre, María Cristina Inostroza,
a mi padre Anselmo Domingo Asún,
a mi abuelo Domingo Asún,
y a mi abuela María Salazar,
sin cuyas enseñanzas de cariño y esfuerzo
no sólo esta tesis, sino yo mismo no hubiera existido.

y también

Dedicado a Johannes Kepler,
por enseñar pasión y rigor en tiempos más oscuros.

AGRADECIMIENTOS

En primer lugar, agradezco a mi profesor guía, Jesús Alvarado, no sólo por brindarme el espacio para aprender, discutir y cuestionar, sino por su amistad y apoyo constantes durante los largos años en que se produjo esta tesis.

En segundo lugar, agradezco a la Facultad de Ciencias Sociales de la Universidad de Chile y, específicamente a su Departamento de Sociología, por haberme brindado el financiamiento y tiempo para poder desarrollar esta tesis.

En tercer lugar, agradezco a la Comisión Nacional de Investigación Científica y Tecnológica de Chile (CONICYT), y al programa Becas Chile (beca N º 26081114FIC) sin cuyo apoyo no habría sido posible disponer de la oportunidad y los recursos para la terminar este producto.

Finalmente, debo agradecer Karina Rodríguez Navarro, mi pareja, no sólo por su apoyo afectivo en el difícil proceso de construir una tesis de doctorado, sino también por la generosidad con que me ofreció su capacidad académica, que está en cada una de las páginas de esta tesis, tan suya como mía.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

LISTADO ALFABÉTICO DE ACRÓNIMOS

1P	Modelo dicotómico unidimensional de un parámetro
2P	Modelo dicotómico unidimensional de dos parámetros
3P	Modelo dicotómico unidimensional de tres parámetros
SAC	Sesgo absoluto de estimación del parámetro de correlación
ADF	Distribución asintóticamente libre
AF	Análisis factorial
AFI	Análisis factorial de ítems
CAT	Test adaptativo informatizado
CCI	Curva característica de los ítems
CCT	Curva característica del test
CFA	Análisis factorial confirmatorio
DEE	Desviación estándar de la estimación
DL	Distorsión de la linealidad
DTF	Funcionamiento diferencial del test
DIF	Funcionamiento diferencial de los ítems
DWLS	Estimación por mínimos cuadrados diagonalmente ponderados
$DWLS_{PO}$	Estimación por mínimos cuadrados diagonalmente ponderados basada en las correlaciones policóricas
E	Error aleatorio de las respuestas observadas
EAP	Estimación esperada a posteriori
FI	Estimación por información completa
GRM	Modelo de respuesta graduada
LI	Estimación por información limitada
MGRM	Modelo multidimensional de respuesta graduada

ML	Estimación por máxima verosimilitud
ML_{PE}	Estimación por máxima verosimilitud basada en las correlaciones de Pearson
MLR	Estimación máximo verosimilitud robusta
MML	Estimación por máxima probabilidad marginal
MTRI	Modelo multidimensional de TRI
NOMGRM	Modelo multidimensional de respuesta graduada de ojiva normal
PB	Puntuaciones brutas
PE	Procedimiento de estimación
PV	Puntuaciones verdaderas
SRL	Sesgo relativo de estimación de los parámetros lambda
RMSE	Raíz de la media cuadrática de los errores de estimación
DE	Desviación estándar
DSC	Desviación estándar de las estimaciones de las correlaciones
DSL	Desviación estándar de estimación de los parámetros lambda
SR	Sesgo relativo de estimación
TC	Tasa de convergencia
TCT	Teoría clásica del test
TRI	Teoría de respuesta al ítem
ULS	Estimación por mínimos cuadrados no ponderados
ULS_{PO}	Estimación por mínimos cuadrados no ponderados basada en las correlaciones policóricas
WLS	Estimación por mínimos cuadrados ponderados
WLS_{PE}	Estimación por mínimos cuadrados ponderados basada en las correlaciones de Pearson
X	Puntuación observada

RESUMEN

A través de la realización de cuatro estudios basados en datos simulados y unidos por una preocupación común, aunque parcialmente independientes en su metodología y foco específico, esta tesis se plantea como objetivo general poner a prueba algunas de las ventajas que podría obtener un investigador aplicado al utilizar modelos clásicos de TRI frente a otras opciones, para de esta forma evaluar la posibilidad de obtener resultados relativamente equivalentes utilizando procedimientos alternativos.

Específicamente, en esta tesis se pone a prueba: a) la calidad diferencial de las estimaciones de las propiedades de los ítems obtenidas por los procedimientos TRI frente a los factoriales; b) la validez de los procedimientos de escalamiento TRI frente a los de la TCT; y c) los límites de la propiedad de invarianza de las estimaciones de sujetos e ítems obtenidas con procedimientos TRI. La hipótesis general de esta tesis es que, en el tipo de condiciones en que se ha enfocado el estudio, las ventajas de utilizar modelos clásicos de TRI no serán demasiado amplias, con lo que existirán condiciones en las que emplear procedimientos alternativos será recomendable, dada su simplicidad relativa y su equivalencia de resultados.

La primera investigación utiliza datos simulados para comparar las estimaciones de discriminación de ítems politómicos provenientes del AFI, frente a las posibles de obtener con distintas estrategias relacionadas con el análisis factorial clásico. La segunda investigación, también utiliza datos simulados para poner a prueba la equivalencia de las estimaciones de discriminación de los ítems obtenidas con los procedimientos TRI y AFI. La tercera investigación, utiliza dos estudios basados en datos simulados tanto para comparar los escalamientos de los sujetos obtenidos por la TRI y la TCT, como para aclarar las condiciones en que es posible esperar la presencia similitudes o diferencias entre ambos. Finalmente, la cuarta investigación, emplea varios ejercicios basados en datos simulados,

para tratar de precisar los límites y consecuencias aplicadas de la propiedad de invarianza de la TRI.

Los resultados de la tesis indican que: a) los procedimientos AFI de estimación de parámetros de discriminación de ítems politómicos, obtienen productos notoriamente más precisos e insesgados que los procedimientos derivados del análisis factorial clásico; b) los procedimientos TRI y AFI son altamente equivalentes al estimar los parámetros de discriminación de ítems politómicos en la mayor parte de las condiciones; c) los escalamientos de sujetos obtenidos por procedimientos TRI y TCT son similares en la mayoría de las condiciones, aunque los escalamientos TRI tienden a ser más válidos cuando se dispone de instrumentos compuestos por ítems más discriminadores y mejores condiciones para realizar la estimación, mientras que lo contrario ocurre en presencia de ítems menos discriminadores y peores condiciones de estimación; d) La propiedad de invarianza de la TRI sólo se confirma estrictamente para los conjuntos de respuestas e ítems para los que se ha demostrado ajuste del modelo a los datos, por lo que resulta justificado atribuir a la TRI la propiedad de "invarianza interna", pero no la de "invarianza externa", como a veces parece suponerse en la literatura.

La discusión y conclusiones generales de la tesis tratan de resituar los méritos de la TRI en la perspectiva del investigador aplicado, estableciendo que si bien la TRI posee ventajas innegables para este tipo de usuario, estas son de un alcance más limitado que las que habitualmente se le atribuyen, por lo que es posible proponer alternativas como el AFI, que permite obtener estimaciones de modelos de medida de similar calidad a los alcanzados por la TRI, pero con un menor nivel de complejidad de cálculo e interpretación.

ABSTRACT

Throughout four research studies using simulated data, this thesis aimed to test some of the potential benefits applied researchers might get from using classic item response theory models (IRT) over other available options, and assessed the possibility of obtaining relatively equivalent results using alternative methods designed for the same purpose.

Specifically in this thesis tested: (a) the accuracy item property estimates achieved with TRI models compared to factor analytic procedures; (b) the validity of IRT subject scaling procedures compared to the classical test theory (CTT) scaling procedure; and (c) the limits the property of invariance of subject and item estimates obtained with IRT modeling procedures. The general hypothesis of this thesis is that, benefits derived from using classic IRT models might not be large in conditions assessed here; therefore, there will be circumstances in which the use of alternative procedures might be recommended, given its relative simplicity and its equivalent outcomes.

The first research study assessed the item factor analysis (IFA) estimates for polytomous data and compared these estimations to those possible to achieve with classic factor analysis techniques. The second research assessed the equivalence between IRT and AFI estimates for item discrimination parameters. The third research compared IRT and CTT scaling procedures to clarify the situation where we might expect to find larger similarities and differences between them. Lastly, the fourth research aims to delimitate the limits and empirical consequences of the property of invariance of IRT models.

Research findings presented in the current thesis show that: (a) AFI discrimination parameter estimates for polytomous items are notoriously more precise and unbiased than parameter estimates from classic factor analysis procedures; (b) IRT and AFI procedures are highly equivalent to estimate discrimination parameters of polytomous items in most conditions; (c)

IRT and CTT subject scaling procedures yield similar results in most situations, although IRT subject scaling tend to be more valid when tests or scales are comprised by items with large discrimination parameters in combination with large samples and large number of items, whereas the CTT subject scaling supersede IRT subject scaling when items have small discrimination parameters, specially in combination with small samples and small test length; (d) the property of invariance of IRT models is restricted to the set of responses and items for which model fit has been demonstrated, therefore, it is justified to assume the existence of internal invariance but not the existence of external invariance.

In the discussion and conclusion section of this thesis, it is attempted to resituate the merits of using IRT models in applied research by establishing that, even though using IRT models have important advantages for this type of user, this advantages are more restricted than previously thought and other procedures, such as IFA which allows achieving results that are more easy to calculate and interpret and, in addition, have similar levels of accuracy to IRT models.

PRÓLOGO: MOTIVACIÓN PERSONAL QUE DA ORIGEN A ESTA TESIS

La bióloga y filósofa Donna Haraway (1995) postula que el conocimiento es situado, es decir, que pese a los legítimos esfuerzos de los científicos por lograr la objetividad y apegarse a la evidencia empírica, nunca es posible lograr una posición completamente neutral y despegarse totalmente de los sesgos disciplinares y biográficos de los investigadores y de las comunidades científicas en las que son socializados. Desde esa perspectiva, para comprender el sentido de esta tesis es importante saber que su autor es sociólogo, por lo que proviene de una comunidad distinta a aquellas que nutren habitualmente a la psicometría, y se siente tan atraído por la medición de constructos sociológicos y psicosociales, como por la docencia universitaria y el estudio aplicado de movimientos sociales.

En ese marco, la andadura que dio origen a esta tesis se inició hace ya 15 años, cuando su autor ingresó con 20 minutos de retraso a su primera clase del Diploma de Estudios Avanzados en Metodología de las Ciencias del Comportamiento y de la Salud, en la Facultad de Psicología de la Universidad Autónoma de Madrid.

Al entrar al aula, los profesores Vicente Ponsoda y Julio Olea se encontraban explicando muy pedagógicamente algunos aspectos de una teoría psicométrica hasta entonces ignorada por completo por el autor de estas líneas: los fundamentos de la teoría de respuesta al ítem (TRI). Perplejo ante su completo desconocimiento del tema, quien escribe supuso que, por descuido, había leído mal la fecha de inicio del curso y se había perdido un par de semanas de clase. Confiado en esa interpretación, esperó al final de la lección para preguntar a los docentes como ponerse al día. Resultó un poco intranquilizante la preocupación de ambos profesores cuando supieron que tenían un estudiante que no sabía nada de la TRI, pero le recomendaron una lectura para comenzar su puesta al día: *Introducción a la Teoría de Respuesta a los Ítems* de José Muñiz (1997).

Al comenzar a leer el libro, el autor de esta tesis descubrió con sorpresa que su desconocimiento no se debía a 20 minutos de atraso ni a dos semanas de ausencia a clases, sino a 20 años de desactualización: todo lo que sabía de psicometría -básicamente, teoría clásica del test (TCT) y construcción de escalas tipo Likert- aparecía en un capítulo de historia de la psicometría que culminaba en los años 80.

Pasado este shock, quien escribe leyó éste y otros textos para tratar de aprender lo máximo posible de esta nueva y prometedora herramienta, portadora de muy interesantes posibilidades, como el permitir modelar las respuestas de los sujetos a los ítems en función de distintos modelos estadísticos, obtener mejores y variadas estimaciones de las propiedades de los ítems, obtener propiedades invariantes de sujetos e ítems, evaluar el ajuste de los datos a los modelos o diseñar tests adaptados a cada sujeto, entre otras.

Tomando en cuenta que su -en ese tiempo- débil conocimiento matemático le impedía adentrarse en las complejidades de la TRI, el autor de esta tesis pensó que su aporte a la disciplina podría consistir en difundir esta nueva teoría de los tests en el mundo que conocía: la sociología y el estudio de las actitudes sociales con escalas tipo Likert de respuesta politómica.

Sin embargo, una vez finalizado su diplomado y a su regreso a Chile, quien escribe se reencontró con una comunidad de sociólogos y psicólogos sociales Latinoamericanos que, enmarcados en una tradición altamente crítica, cuestionaron la relación costo-eficiencia de emplear la TRI en sus actividades habituales de construcción y aplicación de instrumentos de medida, es decir para: (a) escalar a sus sujetos; (b) determinar las propiedades de los ítems; (c) estimar la calidad de la medida lograda por un instrumento en un determinado grupo; o para (d) determinar la relación entre el constructo latente foco del estudio y otros constructos. En este contexto, el autor de esta tesis fue consultado en reiteradas ocasiones, con genuino escepticismo de parte de sus interlocutores, respecto de los beneficios de emplear una

herramienta mucho más sofisticada y que requería esfuerzo adicional de comprensión, para realizar operaciones psicométricas habituales en investigación aplicada que implican baja o media complejidad relativa. En síntesis, se le preguntó explícitamente, ¿vale la pena invertir esfuerzos en aprender TRI?

Evidentemente, la TRI ofrece algunas ventajas notorias para la realización de operaciones psicométricas habituales en investigación aplicada; ventajas que no requieren mayor prueba, como por ejemplo, permite disponer de errores de medida contextuales al nivel de rasgo de los sujetos o curvas características de los ítems y funciones de información tanto de ítems y como de tests, que permiten una caracterización más fina de las características de un instrumento. No obstante, en relación a otros temas (escalamiento, propiedades de los ítems, correlación con otras variables, invarianza de las estimaciones) el autor de esta tesis notó que no tenía la posibilidad de proveer de una respuesta clara a sus interlocutores. Es por ello que aprovechó la realización de su trabajo de fin de master para poner a prueba la capacidad de la TRI para escalar a sujetos reales y para determinar el grado de relación entre el constructo medido y otros constructos de interés pues, de acuerdo a su razonamiento, las ventajas de la TRI sobre la TCT, deberían evidenciarse en ambos tipos de actividades.

En dicha investigación, se aplicó diversos modelos politómicos de TRI a las respuestas de una muestra representativa de chilenos frente a una escala tipo Likert, diseñada para medir intolerancia y discriminación (Asún y Zúñiga, 2008). Para sorpresa de quien escribe, los escalamientos TRI y TCT tuvieron una correlación lineal muy alta entre sí y evidenciaron una casi igual relación con todas las variables criterio, por lo que se debía concluir que, al menos en este caso particular, escalar a las personas usando modelos de TRI no parecía ofrecer ventajas respecto de la simple suma de puntuaciones. En cambio, el diagnóstico de la calidad y características de los ítems que se podía lograr utilizando dichos

modelos era mucho más claro que el que se podía obtener con la TCT. No obstante, el estudio se había realizado con datos reales, por lo que no era clara la posibilidad de generalizar sus resultados.

Años más tarde, cuando el autor de esta tesis regresó a España para integrarse al programa de Doctorado en Metodología de las Ciencias del Comportamiento y de la Salud, en la Universidad Complutense de Madrid, bajo la guía del profesor Jesús Alvarado, pudo profundizar en estas preocupaciones, pues encontró un espacio de diálogo estimulante del pensamiento autónomo que comenzó con la lectura -recomendada por su tutor- de dos textos: *Test Theory* de Roderick McDonald (1999) y *Is Psychometrics Pathological Science?* de Joel Michell (2008).

Estos y otros textos afines, le hicieron ver a quien escribe que existían múltiples conexiones, integraciones o reinterpretaciones de los distintos modelos psicométricos, y que lo que parecía ser un conjunto de procedimientos estadísticos estancos, podían constituir un conjunto más integrado de propuestas analíticas. Así, quedó claro que a pesar de los innegables aportes de la psicometría y su evidente progreso en los últimos años, no siempre los fundamentos de cada procedimiento eran interpretados de manera uniforme, pudiendo encontrarse interesantes debates respecto de ellos y sus alcances.

Durante todo su proceso de formación doctoral, las preocupaciones personales del autor de esta tesis continuaron ligadas al aporte que podía hacer la psicometría moderna al avance de los estudios de los investigadores aplicados, por lo que las conexiones entre procedimientos y los debates respecto a sus fundamentos, enriquecieron sus inquietudes orientadas a responder preguntas como: ¿cuáles son las ventajas que puede ofrecer el empleo de procedimientos TRI a los investigadores aplicados?, o ¿qué alternativas existen a la TRI que permitan obtener algunas de sus ventajas, pero con un menor costo en dominio

matemático o que deriven de técnicas más familiares al investigador no experto en psicometría?

Estas preguntas condujeron a la realización de una serie de estudios independientes, pero interconectados a nivel de su preocupación global, que son los que componen la presente tesis. Para poder abordarlas, en el primer capítulo se desarrollarán algunos conceptos y debates que permiten situar el resto de la tesis en un marco general, como es i) intentar ponderar el impacto y evolución de la TRI hasta nuestros días; ii) delimitar lo que constituye y cuáles son las principales funciones de una teoría de los test; iii) elaborar una propuesta respecto de en que consisten la TRI y la TCT como teorías de los test; iv) desarrollar las potenciales ventajas de la TRI, distinguiendo especialmente aquellas que pueden ser de mayor interés para los investigadores aplicados.

INTRODUCCIÓN

IMPACTO Y EVOLUCIÓN DE LA TEORÍA DE RESPUESTA AL ÍTEM

Desde su aparición, la teoría de respuesta al ítem (TRI) ha tenido un fuerte y positivo impacto en la psicometría, abriendo nuevas posibilidades conceptuales y operativas en el terreno del diseño y evaluación de instrumentos de medición de variables latentes. Desde el punto de vista conceptual la TRI ha permitido, por ejemplo, pensar en la elaboración de modelos probabilísticos que explican las respuestas que los sujetos dan a los ítems (a veces incluso utilizando modelos cognitivos, como hacen Gorin y Embretson, 2006), ha asentado la idea de someter a prueba empírica los supuestos de los modelos psicométricos, y ha permitido establecer conexiones matemáticas entre la psicometría y otros procedimientos estadísticos para datos observados o latentes (Mellenbergh, 1994a). Desde el punto de vista operativo, la TRI ha permitido un diagnóstico más fino de las propiedades de los ítems y tests, ha complejizado el diagnóstico de los errores de medida producidos al aplicar un instrumento a un conjunto de sujetos (e.g., estimando el error de medida contextual a cada patrón de respuestas) y ha permitido el desarrollo de tests óptimos para cada sujeto, entre otras muchas posibilidades. Por estas razones, la gran mayoría de las investigaciones académicas en las principales revistas del campo tienen como foco esta teoría de los tests (Abad, Olea, Ponsoda, y García, 2011). No obstante, la consolidación y evolución de la TRI ha implicado procesos de difusión y complejización que en ocasiones han dificultado la comprensión cabal de sus beneficios y limitaciones, especialmente para las operaciones de medición que más habitualmente desarrollan los investigadores aplicados.

Con respecto a la difusión de la TRI, una lectura de sus principales manuales (e.g., Embretson y Reise, 2000; Hambleton, Swaminathan, y Rogers, 1991; Muñiz, 1997; Reise y

Haviland, 2005; Reise y Henson, 2003) permite hipotetizar que, pese a sus potencialidades, no siempre ha sido fácil popularizar la TRI entre los investigadores no expertos en psicometría. Las razones de esta dificultad pueden ser múltiples, encontrándose argumentos que destacan las fuertes barreras de entrada que impone la mayor complejidad matemática de la TRI (Abad et al., 2011), la carencia de software amigable con el usuario no experto (Muñiz, 1997), las mayores demandas que hace la TRI respecto de tamaños de muestra requeridos para lograr estimaciones precisas (Martínez-Arias, Hernández-Lloreda, y Hernández-Lloreda, 2006), el carácter restrictivo de los supuestos de los modelos de TRI más difundidos, más simples y que exigen menos muestra (Reise y Henson, 2003), e incluso se ha señalado que el que la TRI disponga de índices de ajuste puede ser visto como una dificultad por los investigadores aplicados, pues el desajuste de los datos al modelo puede ser un obstáculo para el avance de sus investigaciones, lo que puede llevar a que opten por una teoría de los tests que no dispone de esos indicadores (Borsboom, 2006).

Quizá producto de esas dificultades para popularizar la TRI fuera del campo de los expertos en medición, una parte de los textos de psicometría focalizados en la TRI y dirigidos al público no experto (e.g., Embretson y Reise, 2000; Hambleton, Swaminathan, y Rogers, 1991; Reise y Haviland, 2005; Reise y Henson, 2003), han enfatizado tanto sus atractivos y beneficios, así como sus diferencias frente a la tradicional Teoría Clásica del Test (TCT), desarrollando con menos profundidad algunas de sus limitaciones o sus similitudes y equivalencias con la TCT tal como ésta formulada en su forma original (Gulliksen, 1950/1987) o en sus desarrollos más recientes (Raykov y Marcoulides, 2015); o sus similitudes y equivalencias con otros procedimientos (e.g., Christoffersson, 1975).

Por otro lado, desde su aparición, la TRI ha experimentado un aumento de complejidad, entendida ésta como una creciente diversificación y aumento de sus interconexiones internas y externas. Lo que nació como un conjunto de modelos centrados en

modelar las respuestas dicotómicas de las personas a los ítems a partir de funciones monotónicas crecientes y un único rasgo latente (Lord y Novick, 1968), hoy en día se ha extendido a distintos formatos de respuesta (Embretson y Reise, 2000), al empleo de funciones no monotónicamente crecientes (Mokken, 1997), a la estimación de modelos que contemplan múltiples rasgos latentes (Reckase, 2009), etc. En relación al aumento de sus interconexiones internas y externas, cada vez es más frecuente encontrar literatura centrada en generar enfoques integrados que permiten deducir modelos específicos o casos anidados en modelos más generales (Mellenbergh, 1994a) y evidenciar las conexiones entre la TRI y otras técnicas de análisis estadístico u otras teorías de los tests (Raykov y Marcoulides, 2015).

Este aumento de complejidad ha permitido ampliar el campo de aplicabilidad de la TRI y el desarrollo de miradas más integradas, no obstante, hipotetizamos al mismo tiempo ha generado algunas dificultades, especialmente para el investigador aplicado no experto en psicometría, para quien se han tornado borrosas las fronteras entre la TRI y algunos procedimientos alternativos y, por tanto, le es difícil comprender el contenido esencial de la TRI, sus límites y ventajas comparativas. En este contexto, esta tesis ha buscado contribuir a precisar algunas de las ventajas y limitaciones prácticas de la TRI, estableciendo condiciones bajo las cuales es posible obtener resultados de calidad equivalente (o incluso superior) a ella empleando estrategias o teorías de los tests alternativas. Para lograr este objetivo, es necesario comenzar por definir o clarificar el significado de algunos conceptos clave involucrados en esta tarea: especificar qué son las teorías de los tests, cuáles son los aspectos esenciales que caracterizan a la TRI y a la TCT y cuáles son las potenciales ventajas que se podrían obtener de la aplicación de la TRI frente a sus alternativas.

¿QUÉ ES UNA TEORÍA DE LOS TESTS?

La literatura psicométrica señala que las teorías de los tests están compuestas por modelos matemáticos de carácter probabilístico que relacionan las respuestas a los tests con la asignación puntuaciones a los sujetos medidos y estiman la calidad de esas puntuaciones (Martínez-Arias et al., 2006; Muñiz, 2001), lo que permite hacerse cargo de los problemas que implica la medición de constructos latentes, diseñar procedimientos para disminuirlos, construir mejores instrumentos de medida y comparar distintas poblaciones (Crocker y Algina, 1986).

En base a esta definición, creemos que una teoría de los tests, para ser considerada como tal, debe contener al menos cuatro tipos de componentes o procedimientos esenciales: (a) procedimientos de estimación de las puntuaciones de los sujetos en el o los constructos medidos, a partir de las respuestas que ellos han dado al test; (b) procedimientos de estimación del error de medida de las puntuaciones asignadas a los sujetos (de manera de saber en que proporción esa puntuación es producto de un error de medida o corresponde a una propiedad estable de los sujetos); (c) procedimientos que permitan estimar las propiedades de los ítems que componen el test (de manera de tener criterios para la eliminación de ítems inadecuados, para la generación de nuevos ítems y para la construcción definitiva del test); y (d) procedimientos de transformación, estandarización y equiparación de puntuaciones (que permitan clasificar a los sujetos en categorías, hacer comparaciones válidas entre grupos, aplicaciones, etc.), así como procedimientos que permitan detectar sesgos (e.g., funcionamiento diferencial del ítem o del test) en las puntuaciones en desmedro o beneficio de algún grupo en particular.

Desde la psicometría actualmente sólo se reconoce la existencia de dos teorías de los test[1] (i.e., la TCT y la TRI), por lo que a continuación trataremos de definir en qué consiste y cómo ha evolucionado cada una de ellas. Esto permitirá introducir el Análisis Factorial de Ítems (AFI; Wirth y Edwards, 2007), como procedimiento alternativo para la estimación de variables latentes y posteriormente situar en mejor medida cuáles son las principales potenciales ventajas de la TRI frente a ambas alternativas.

LA TCT Y SU EVOLUCIÓN

Habitualmente se reconoce que la teoría clásica de los tests tiene su origen en los trabajos de Spearman (1904), cuyo interés era generar un procedimiento que permitiese 'separar' el error de medida aleatorio (E_i) de la puntuación verdadera (PV_i), que se define como el valor esperado de las respuestas del sujeto, cuando se cuenta con mediciones observadas (X_i) obtenidas a partir de la aplicación de un test a un conjunto de sujetos. En consecuencia, la conocida ecuación fundamental de la TCT que define $X_i = PV_i + E_i$, expresa dos características centrales de la TCT que la diferencian de la TRI: su focalización en el test total (y no en sus ítems componentes o ítems) y en la estimación del error de medida global.

Si bien desde la TCT existen otras formas de estimación de las puntuaciones de los sujetos que implican un mayor grado de complejidad y sofisticación (e.g., ponderar las respuestas a los ítems según su índice de validez), la forma más habitual de obtener la puntuación observada de los sujetos al test consiste simplemente sumar todas las respuestas correctas de las personas a ítems de aptitud dicotómicos o sumar las puntuaciones obtenidas por las personas en sus respuestas a preguntas con formato politómico de respuesta ordenada empleando enteros sucesivos (i.e., 0, 1, 2, 3, …, *k*) (Gulliksen, 1950/1987). Por su parte, si

[1] Ya que la teoría de la generalizabilidad es frecuentemente considerada una extensión de la TCT.

bien se han desarrollado múltiples estadísticos para estimar el error de medida (Gulliksen, 1950/1987) la mayor parte de ellos son dependientes directa o indirectamente de la noción de formas paralelas y tienden a producir una medida global de la calidad de la medida. A pesar de las controversias desarrolladas en el campo (Sijtsma, 2009), en la actualidad, el procedimiento más empleado con la finalidad de establecer el error de medida es el coeficiente α de Cronbach y/o sus versiones corregidas adaptadas a ítems ordinales (Elosua y Zumbo, 2008). Además, si bien en sus comienzos la TCT proponía numerosos índices para evaluar las propiedades de los ítems (Gulliksen, 1950/1987), en general se han tendido a consolidar principalmente dos: el índice *p* de dificultad de los ítems, que se define como la proporción de acierto a ítems dicotómicos (que en las versiones politómicas es reemplazado por la media de puntuación en un ítem tipo Likert) y la correlación ítem-escala corregida, que indicaría el grado en que cada ítem mide el constructo general evaluado por el test completo (Muñiz, 2001). Dentro de las críticas que se formulan a estos estadísticos, frecuentemente se señala que éstos no son invariantes respecto de los sujetos a los que se aplica el tests (Gulliksen, 1950/1987). Finalmente, dentro de la TCT se han desarrollado procedimientos para estandarizar puntuaciones de los sujetos (e.g., transformación a percentiles o a puntuaciones *Z*) y equiparar puntuaciones obtenidas en distintos grupos (e.g., igualación equipercentil), no obstante, su utilidad práctica no pareciera del todo satisfactoria pues involucran supuestos no siempre posibles de cumplir (Crocker y Algina, 1986).

Las definiciones y estadísticos presentados en los párrafos precedentes corresponden a la caracterización que se hace de la TCT en la literatura psicométrica tradicional. Sin embargo, es importante destacar que en años recientes ha surgido una corriente de autores que emplea el modelo congenérico (Jöreskog, 1971; Steyer, 1989) o el supuesto de la existencia de una variable normal subyacente a los ítems categóricos (Raykov y Marcoulides, 2011), para redefinir a la TCT como un modelo aplicable a ítems de tests que midan un

mismo rasgo latente unidimensional, aunque lo hagan con distinta varianza error y unidad de medida (Raykov y Marcoulides, 2011). Esta reformulación de la TCT no es una simple derivación de la TCT tradicional, pues requiere supuestos adicionales que no se encuentran en la versión inicial (Kohli, Koran, y Henn, 2014), modifica el significado de la puntuación verdadera y tiende a asimilar la TCT al AFI (e.g., Raykov y Marcoulides, 2015). Pese a ello, este reciente desarrollo sirve para mostrar las interesantes conexiones que existen entre la TCT, el análisis factorial y algunos modelos de la TRI.

LA TRI Y SU EVOLUCIÓN

La creciente complejización de la TRI genera un cierto grado de dificultad para definir sus características esenciales en tanto teoría de los tests, pues los progresivos avances en investigación psicométrica han tendido a sobrepasar muchas definiciones anteriormente aceptadas, ampliando continuamente sus alcances. Así por ejemplo, es tradicional señalar que la TRI es un conjunto de modelos probabilísticos que tienen por finalidad relacionar las respuestas discontinuas (dicotómicas o politómicas, nominales u ordinales) de los sujetos a los ítems que componen un test (Embretson y Reise, 2000) con una variable latente unidimensional o multidimensional, que se supone continua, diferenciando así los modelos de TRI de los modelos de clase latente (De Ayala, 2009; Muñiz, 1997; Reckase, 2009). De este modo, la TRI consistiría en un conjunto de modelos de medida de variable latente caracterizados por el modelamiento de una curva característica del ítem (CCI), que relacionaría la probabilidad de respuesta al ítem con el nivel de rasgo o rasgos latentes de los sujetos a través de una función monotónica creciente definida de antemano por el modelo (De Ayala, 2009; Embretson y Reise, 2001; Hambleton et al., 1991; Muñiz, 1997). Por otro lado, si bien no es habitual que se considere explícitamente a los procedimientos de estimación

como parte de la delimitación de la TRI, frecuentemente se señala que en la TRI se emplean procedimientos de información completa para la estimación de parámetros, es decir, procedimientos que emplean la totalidad de los patrones de respuesta de los sujetos a los ítems (Abad et. al., 2011; Embretson y Reise, 2000; Martínez-Arias et al., 2006). En base lo anterior, se podría afirmar que tradicionalmente se ha sostenido que la TRI consistiría en un conjunto de modelos de medida, que proponen modelar las probabilidades de respuesta de los sujetos a un conjunto de ítems de respuesta discontinua, empleando funciones sigmoides definidas de antemano por cada modelo, que relacionan esas probabilidades con uno o más rasgos latentes y cuyos parámetros son estimados empleando procedimientos de información completa.

Pese a la aparente claridad de esta definición, la mayor parte de los límites que establece han sido superados por la investigación psicométrica. Al respecto, es importante destacar cinco ejemplos de ampliaciones relevantes:

1. Han sido propuestos modelos de TRI para respuestas continuas (Samejima, 1973; Mellenbergh, 1994b), con lo que el trabajar con datos discontinuos no sería en si mismo un elemento distintivo de la TRI.
2. Se han desarrollado modelos de punto ideal o *unfolding* (Drasgow, Chernyshenko, y Stark, 2010; Roberts, Donoghue, y Laughlin, 2000) que, a diferencia de los modelos tradicionales de TRI que asumen una CCI monotónicamente creciente, estiman CCI que contienen un punto de inflexión a partir del cual son decrecientes; con lo cual ha superado la caracterización de los modelos de TRI en función de una CCI creciente.
3. Se han desarrollado modelos no paramétricos de TRI (e.g., Mokken, 1997; Sijtsma y Molenaar, 2002), los cuales no proponen ninguna función a priori para la CCI, por lo que incluso pareciera no ser prerrequisito de la TRI el definir previamente dicha función.

4. Respecto a los procedimientos de estimación, algunos autores plantean que es posible estimar modelos de TRI con procedimientos heurísticos (Embretson y Reise, 2000) o con procedimientos de información limitada (Forero y Maydeu-Olivares, 2009), con lo que el tipo de procedimiento de estimación tampoco sería un elemento distintivo de este tipo de modelos.
5. Finalmente, se han propuesto modelos de mixturas (*mixture models*) que permiten emplear la TRI para ajustar modelos de medida de clases latentes (Maij-de Meij, Kelderman, y van der Flier, 2008) y no sólo rasgos latentes, ampliando así el horizonte de posibles aplicaciones de la TRI.

De acuerdo a estos ejemplos, resulta evidente que intentar delimitar un objeto complejo como es la TRI, cuyo desarrollo ha tendido a multiplicar sus modelos con mucha rapidez (cf. e.g., van der Linden y Hambleton, 1997) es una empresa difícil y pudiera ser considerada innecesaria. No obstante, no intentarlo podría por ejemplo: (a) contaminar la discusión sobre las ventajas comparativas de la TRI (u otra teoría de los test) con la definición de los objetos que estamos comparando; (b) disponer de una definición tan amplia de la TRI que incluya prácticamente todo (pues la única distinción que parece quedar en pie es afirmar que esta teoría de los tests modela probabilidades de respuesta a ítems suponiendo constructos latentes) y no permita la emergencia de teorías competidoras con la cual contrastarla; (c) dificultar la comunicación interdisciplinaria, pues un mismo procedimiento puede tener dos o más denominaciones dependiendo de la comunidad que lo aplique; por ejemplo, en algunas publicaciones se indica que se utilizaron modelos de TRI para los análisis pero al revisar en detalle, es posible notar que se emplearon procedimientos y enfoques tradicionalmente asociados al análisis factorial (e.g., Finch, 2010) o inversamente, en algunas publicaciones se denomina análisis factorial a perspectivas y procedimientos que

tradicionalmente se han considerado TRI (e.g., Galbraith, Moustaki, Bartholomew, y Steele, 2002).

Con la finalidad de saldar este problema, en esta tesis propongo una delimitación más estricta de lo que constituye la TRI, retornando a cinco de los elementos centrales de la definición original que hemos señalado en las páginas precedentes. En otras palabras, se propone reservar la etiqueta de TRI para: (a) los modelos de medida (b) que se propongan modelar las probabilidades de respuesta de los sujetos a un conjunto de ítems de respuesta discontinua (c) empleando funciones continuas (definidas previamente u obtenidas a partir de los datos) (d) asumiendo que las respuestas se pueden explicar a través de uno o más rasgos latentes y (e) que estimen sus parámetros a través de procedimientos de información completa. Por tanto, propongo para efectos de esta tesis que aquellos modelos que no establezcan a priori o no produzcan CCI, que estimen los parámetros por información limitada, que supongan la existencia de clases latentes o que excedan los límites de un modelo de medida, reciban otras denominaciones, y que se distinga con claridad cuando se está estimando propiamente un modelo de TRI usando sus reglas internas, de aquellas ocasiones en que se emplea un enfoque distinto como medio alternativo para obtener una aproximación razonable a sus parámetros.

Esta definición dejaría fuera de la TRI a los modelos de clase latente, a los modelos para respuestas continuas y a los modelos AFI (Wirth y Edwards, 2007). Interesa destacar esta última distinción, pues la conocida equivalencia matemática entre algunos modelos de TRI y el modelo de factor común (McDonald, 1999) bajo algunos supuestos y condiciones (Christoffersson, 1975; Muñiz, 1997; Takane y De Leeuw, 1987), no debería borrar la distinción entre el AFI y la TRI. Si se acepta esta distinción, se podría concluir que la equivalencia entre ambos modelos implica que el AFI permitiría obtener (para aquellos modelos en que hay equivalencia) parámetros de los ítems y sujetos similares a los que se

estiman con algunos modelos de TRI -que incluyen a los modelos uni y multidimensionales de ojiva normal o logísticos, dicotómicos de uno o dos parámetros y politómico de respuesta graduada (Samejima, 1969)-, haciendo la salvedad que la diferencia entre ambos tipos de modelos (i.e., AFI y TRI) consistiría en que, al emplear procedimientos de información limitada, los parámetros estimados por el AFI potencialmente pueden tener mayor sesgo y error de estimación que aquellos que estima la TRI empleando procedimientos de información completa, es decir, empleando toda la información contenida en el patrón de respuestas de los sujetos.

Desde una perspectiva TCT en su versión tradicional, también es posible encontrar conexiones con algunos modelos de TRI. Por ejemplo, las CCI de los modelos clásicos de TRI (e.g., dicotómicos de uno y dos parámetros) se modelan con funciones no lineales, lo que ha llevado a calificar a este procedimiento como no-lineal en contraste con la linealidad de la TCT (Muñiz, 1997; Santisteban, 2009); no obstante, de Lord (1980) ha mostrado que, si bien estos modelos de TRI emplean una función sigmoidea para relacionar el rasgo con la probabilidad de respuesta a un ítem discontinuo, esto es equivalente a suponer una relación lineal entre el rasgo y la respuesta subyacente que el sujeto habría dado de haber dispuesto de alternativas de respuesta continuas. En consecuencia, se podría considerar a estos modelos de TRI como esencialmente lineales, donde su no linealidad sería producto de la adaptación a estar modelando datos discontinuos (Muthén, 1993). En ese marco, la diferencia entre estos modelos de TRI y la TCT consistiría en que la primera distingue entre una linealidad subyacente y la no linealidad producto de disponer de respuestas observadas discontinuas a ítems (y por añadidura tests), mientras que la segunda mantiene un modelo lineal para las respuestas observadas al test.

Finalmente, las reformulaciones actuales de la TCT derivadas del modelo congenérico (Kohli, Koran, y Henn, 2014; Raykov y Marcoulides, 2015), también tienden a difuminar las

distinciones entre la TCT y la TRI, planteando que la primera es un caso especial de algunos modelos de la segunda. Pese al mérito de evidenciar que no hay tantas diferencias entre ambas teorías de los tests como se creía, esta propuesta desde un punto de vista técnico y práctico homologa la TCT al AFI, sin que quede claro el aporte del cambio de denominación, por lo que se ha desestimado su utilización en la presente tesis.

VENTAJAS POTENCIALES DE LA TRI COMO TEORÍA DE LOS TESTS

En este contexto de equivalencias teóricas y confluencias empíricas entre la TRI y el AFI, y entre la TRI y la TCT, es pertinente preguntarse cuáles son las potenciales ventajas que se pueden obtener del empleo de la TRI frente a sus alternativas. Para responder a esta pregunta, se ordenará la descripción de las ventajas de la TRI en función de los elementos que se han propuesto en las páginas previas como constitutivos de una teoría de los tests (i.e., procedimientos de estimación de puntuaciones sujeto, de error de medida, de propiedades de los ítems, y de estandarización de puntuaciones), distinguiendo también entre las ventajas asociadas a operaciones psicométricas complejas, de aquellas que pueden ser más aplicables por investigadores aplicados no expertos en psicometría.

Estimaciones de las Puntuaciones de los Sujetos

En la literatura psicométrica actual no es muy frecuente encontrar afirmaciones respecto de las ventajas comparativas de la estimación del rasgo (i.e., $\hat{\theta}$) que se realiza con la TRI, frente a los escalamientos posibles de lograr con otros procedimientos. No obstante, ello no implica que no existan algunos estudios que si lo argumentan. Por ejemplo, Lord (1980) demuestra que, si se ha especificado adecuadamente el modelo, la estimación de θ que realizan los modelos unidimensionales de TRI clásicos (logísticos o de ojiva normal de uno, dos y tres parámetros) tiene una relación lineal con el rasgo latente, mientras que la puntuación

verdadera o *true score* de la TCT tiene una relación no lineal con dicho rasgo, por lo que resulta un estimador menos adecuado.

Pese a ello, la diferencia entre ambos tipos de escalamiento para los citados modelos clásicos de TRI (pues para modelos más complejos las diferencias son evidentes ya que la TCT no tiene posibilidades de emular modelos multidimensionales o de punto ideal, entre otros) no ha sido completamente aclarada en la literatura psicométrica actual, siendo posible encontrar investigaciones que reportan de una alta semejanza en los escalamientos producidos por la TRI y la TCT y/o similares correlaciones con una variable criterio (Ferrando y Chico, 2007; Xu y Stone, 2012), así como también es posible encontrar investigaciones que reportan altas tasas de error Tipo I en la detección de efectos de interacción cuando se emplea la TCT para escalar a los sujetos, problema que no ocurre en la misma magnitud cuando se emplea la TRI (Embretson, 1996; Kang y Waller, 2005; Morse, Johanson, y Griffeth, 2012).

En relación a la diferencia entre los escalamientos TRI y AFI, la evidencia indica que ellos son altamente semejantes (Kohli, Koran, y Henn, 2014), al menos para el conjunto de modelos en que existe equivalencia matemática entre ambos procedimientos. Ante esto, podría afirmarse que la principal ventaja de aplicar la TRI en este contexto es más bien teórica, dado que el modelo de análisis factorial se basa únicamente en la matriz de varianza-covarianza o la matriz de correlaciones entre los ítems, por lo que, en rigor, las puntuaciones factoriales o *factor scores* que se pueden obtener a través de análisis factorial no son parte integral del modelo, sino agregados posteriores más o menos arbitrarios (DiStefano, Zhu, y Mindrila, 2009) pues el modelo no contempla a los sujetos, sino sólo la relación entre los ítems.

Dentro de las ventajas más citadas de las estimaciones TRI (tanto de sujetos como de ítems) está la propiedad de invarianza (De Ayala, 2009). La propiedad de invarianza

facilitaría, por ejemplo, implementar procedimientos de equiparación de puntuaciones, el análisis del funcionamiento diferencial de los ítems (DIF, por sus siglas en inglés), la comparación entre grupos y el diseño de bancos de ítems y test adaptativos, entre otras posibilidades (Muñiz y Hambleton, 1992).

A su vez, dentro de las ventajas que la estimación de las puntuaciones de sujetos que permite la TRI estaría el hecho de considerar las propiedades de los ítems y la CCI del modelo (DeMars, 2010) para las estimaciones, lo que genera que: (a) las estimaciones obtenidas estén más determinadas por los ítems más discriminadores que por los de menor calidad; (b) que en las estimaciones de θ se considere la dificultad de los ítems respondidos; (c) se pueda incluir sin mayores problemas ítems con distinto número de alternativas de respuesta en un mismo test analizado (Embretson y Reise, 2000); y, (d) que se pueda estimar puntuaciones empleando modelos de relación no lineal entre las respuestas subyacentes a los ítems y el rasgo latente, como ocurre en los modelos de punto ideal (Roberts, Donoghue, y Laughlin, 2000) o modelos con efectos cuadráticos del rasgo sobre los ítems (Rizopoulos y Moustaki, 2008). Los puntos a y b deberían ser de gran interés para el investigador aplicado, mientras que los puntos c y d parecen más pertinentes para el investigador experto en psicometría por la complejidad que implican.

Por otro lado, la existencia de indicadores de ajuste para los modelos de TRI permite que el investigador pueda tener una mayor confianza en los escalamientos obtenidos por la TRI que los obtenidos por la TCT, pues en estos últimos el investigador se ve obligado a confiar en la validez de sus supuestos, sin ponerlos a prueba. Pese a ello, hay autores que señalan que la posibilidad de poner a prueba los supuestos (y rechazarlos) en la TRI pudiera ser visto como un problema por los investigadores aplicados (Borsboom, 2006) quienes, ante el desconocimiento sobre qué hacer ante la violación de supuestos, podrían tender a utilizar procedimientos más toscos que no entreguen información sobre el ajuste.

Finalmente, se ha señalado que las estimaciones de θ son más fáciles de interpretar que las de la TCT pues es posible referirlas al contenido de los ítems, al estar en la TRI los parámetros de sujetos e ítems en la misma métrica (Embretson y Reise, 2000). No obstante, con la excepción de la investigación educativa (e.g., Rittle-Johnson, Matthews, Taylor, y McEldoon, 2011), sobre todo de gran escala, la interpretación de puntuaciones basadas en procedimientos similares al Mapa de Wright no parece ser muy habitual en el mundo de la investigación aplicada donde los investigadores parecieran continuar prefiriendo otros métodos de interpretación (e.g., establecimiento de baremos).

En consecuencia, postulamos que, desde el punto de vista del investigador aplicado, las principales ventajas de emplear los escalamientos producidos por los modelos de TRI que podrían ser mayor interés para el investigador aplicado (modelos uni y multidimensionales de ojiva normal o logísticos, dicotómicos de uno a tres parámetros y politómico de respuesta graduada), serían la posibilidad de: (a) obtener estimaciones más precisas del constructo latente; (b) que las estimaciones obtenidas tengan mejor capacidad de estimar la correcta relación entre la variable medida y otras variables de interés, (c) obtener mediciones más robustas producto de la propiedad de invarianza.

Determinación de la Calidad de la Estimación de las Puntuaciones de los Sujetos

En este campo, la TRI pareciera tener ventajas evidentes frente a la TCT y el AFI, pues sólo con la TRI es posible obtener una estimación del error de medida para cada patrón de respuestas, y no suponer que aquel es una propiedad global de la medición realizada (Embretson y Reise, 2000; Muñiz, 1997) tal como ocurre en la TCT.

Como se ha señalado anteriormente, la manera más tradicional de estimar la fiabilidad de una medida desde la TCT es utilizar el coeficiente α de Cronbach, práctica que no ha estado exenta de importantes cuestionamientos (Sijtsma, 2009) y que sólo permite obtener una medida global de la fiabilidad. Si bien desde el AFI hay alternativas que parecen más

adecuadas para la estimación de la fiabilidad ante datos categóricos (Elosua y Zumbo, 2008), desde el modelo factorial tampoco es posible obtener medidas de fiabilidad para cada patrón de respuestas.

Por su parte, disponer de la función de información del test y de los ítems en la TRI facilita un diagnóstico claro de las debilidades y potencialidades métricas del instrumento que no está al alcance de la TCT (Asún y Zúñiga, 2009).

Evidentemente, conocer la calidad de las estimaciones de las puntuaciones de los sujetos debiera ser una ventaja de mucho interés para los investigadores aplicados, y en este campo la ventaja de la TRI resulta evidente.

Estimaciones de las Propiedades de los Ítems

En este tema hay pocas investigaciones que comparen sistemáticamente la calidad de las estimaciones de la dificultad y discriminación de los ítems entre la TCT y la TRI. Destacan, sin embargo, los estudios de Fan (1998) y de MacDonald y Paunonen (2002) quienes, pese a las ventajas teóricas de la TRI sobre la TCT, no detectan diferencias sustanciales. Ese resultado quizá no es del todo sorprendente considerando que, bajo ciertas condiciones, existe equivalencia relativa entre los parámetros de dificultad y discriminación de la TRI y la TCT (Muñiz, 1997). No obstante, también se ha demostrado indirectamente que existe una relación no lineal entre los parámetros de los ítems estimados con la TRI versus los estimados con la TCT (Embretson y Reise, 2000).

En cualquier caso, en este campo, la ventaja de la TRI se ha argumentado desde dos perspectivas diferentes. En primer lugar, se ha señalado acertadamente que la mayor versatilidad de los modelos de TRI permiten modelar numerosas propiedades de los ítems (e.g., el parámetro c de pseudo azar) que desde la TCT no resultan estimables, lo que permite un diagnóstico mucho más preciso del funcionamiento de los ítems, pudiendo modelarse incluso los procesos cognitivos que orientaron la respuesta (Gorin y Embretson, 2006).

Además, disponer de una CCI permite al investigador formarse una imagen mucho más precisa del funcionamiento de cada ítem que lo alcanzable con la TCT. En segundo lugar, se ha argumentado que la propiedad de invarianza de la TRI (Embretson y Reise, 2000) daría mayor validez a los parámetros estimados para los ítems, los que no serían dependientes de la muestra concreta que fue utilizada para estimarlos (Reise, Ainsworth, y Haviland, 2005).

Respecto a las diferencias y similitudes entre las propiedades de los ítems estimadas por la TRI y el AFI existe mayor número de estudios comparativos (e.g., De Mars, 2012; Reiser y VanderBerg, 1994). En términos generales, se ha mostrado que el AFI es capaz de obtener parámetros de discriminación y dificultad muy similares a los que se obtienen con la TRI, pero la TRI tiende a obtener parámetros de mejor calidad en condiciones más difíciles (e.g., ítems con alta asimetría), especialmente ante ítems dicotómicos. Frente a ítems politómicos la ventaja de la TRI no es tan clara pues el único estudio existente (Forero y Maydeu-Olivares, 2009) no distingue claramente resultados para datos dicotómicos y politómicos. No obstante, una ventaja incontestable de la TRI frente al AFI para la evaluación de las propiedades de los ítems es que el AFI sólo permite estimar los parámetros de dificultad y discriminación de los ítems, mientras que la versatilidad de los modelos de TRI permiten modelar otras muchas propiedades de éstos.

Si bien algunas de esas propiedades de los ítems no parecen de mucho interés para el investigador aplicado no experto en psicometría, el poder estimar por ejemplo el parámetro de seudo azar en test de aptitud si sería una ventaja incontestable de la TRI en esta comparación. Evidentemente, también debería ser de interés para el investigador aplicado el poder estimar los parámetros clásicos de discriminación y dificultad de manera más precisa con la TRI, y que ellos sean invariantes.

Transformación, Estandarización y Equiparación de Puntuaciones

Si bien se reconoce que la TCT dispone de procedimientos de transformación, estandarización y equiparación que permiten comparar puntuaciones a lo largo del tiempo o entre grupos (Kolen y Brennan, 2004), la propiedad de invarianza de la TRI facilitaría en gran medida estas actividades (Muñiz, 1997) y brindaría una solución más elegante (Muñiz, 2010), lo que la ha transformado en una teoría de los tests muy empleada cuando se quiere comparar grupos o culturas distintas en un mismo rasgo, o cuando se desea obtener evidencia de que un instrumento no presenta funcionamiento diferencial en sus ítems o test (Abad et. al, 2011).

En cualquier caso, en el contexto de esta tesis no se someterá a prueba las ventajas de la TRI en lo que a transformación, estandarización y equiparación de puntuaciones se refiere, pues este tipo de actividades no son utilizadas frecuentemente por el investigador aplicado, quedando usualmente en manos de diseñadores de tests con mayor dominio psicométrico.

PREGUNTA E HIPÓTESIS DE INVESTIGACIÓN DE LA PRESENTE TESIS

Como se ha señalado anteriormente, el objetivo de esta tesis es someter contraste a algunas de las ventajas que podría obtener un investigador aplicado de la utilización de la TRI y evaluar la posibilidad de obtener resultados similares empleando procedimientos alternativos. De la revisión presentada en las páginas precedentes es posible señalar que esas ventajas consistirían en obtener estimaciones:

1. Más precisas y válidas de él o los constructos latentes que se intentan medir.
2. Con propiedades más robustas, es decir, invariantes.
3. Más precisas del error de medida con que se está evaluando a los sujetos.

4. Más exactas e insesgadas de las propiedades básicas de los ítems (discriminación y dificultad).
5. De otras propiedades simples de los ítems, como la propensión al acierto al azar.

Esta tesis no pondrá a prueba los puntos 3 y 5, pues en ellos las ventajas de la TRI son auto-evidentes, por lo que se centrará en someter a prueba la capacidad de la TRI de obtener estimaciones de los sujetos y de los parámetros de los ítems más válidos que los alcanzables por otros procedimientos alternativos (TCT y AFI), y en la propiedad de invarianza que poseerían dichas estimaciones. En consecuencia, la pregunta de investigación que guiará esta tesis será: ¿es posible obtener estimaciones de parámetros de los ítems y de los sujetos de calidad y con propiedades similares a las de la TRI empleando procedimientos alternativos en condiciones cercanas a la investigación aplicada? La hipótesis de esta investigación es que, en el tipo de condiciones en que se ha enfocado el estudio, las ventajas de la TRI no serán demasiado notorias, con lo que existirán condiciones en las que emplear procedimientos TCT o AFI será recomendable, dada su simplicidad relativa y su equivalencia de resultados.

La tesis se ha diseñado de manera que cada capítulo tenga la forma de un artículo, autocontenido y relativamente independiente, para así facilitar el envío de cada uno de ellos a diferentes revistas de metodología. A continuación se explica el sentido y aporte de cada capítulo o artículo para la pregunta de investigación de la tesis.

El primer capítulo presenta la primera investigación de la tesis, en la que, partiendo de la demostrada equivalencia matemática entre algunos modelos de TRI y el modelo AFI para datos ordinales, se desarrolla una investigación que muestra que el AFI logra estimaciones de las cargas factoriales de los ítems más precisas e insesgadas que otras alternativas frecuentemente empleadas en el análisis factorial clásico.

En el segundo capítulo, se profundiza en las fuertes conexiones entre la TRI y el AFI y se presentan los resultados de una investigación que muestra que, en algunas condiciones y

frente a ítems politómicos, con el AFI es posible obtener estimaciones de parámetros de los ítems de similar, e incluso mayor precisión y ausencia de sesgo que con la TRI, aunque esta última habitualmente emplee la información completa contenida en la base de respuestas y el AFI comúnmente sólo información limitada.

El capítulo tres se centra en las potenciales ventajas de la TRI sobre la TCT en el escalamiento de los sujetos. En este capítulo se presentan los resultados de dos estudios de simulación en los que se muestra la capacidad de la TCT de producir parámetros de los sujetos similares a los de la TRI en algunas condiciones, constatándose la fuerte relación entre ambos escalamientos en la mayor parte de las condiciones, pero evidenciando y discutiendo también la ventaja relativa del escalamiento TRI en otras.

El capítulo cuarto presenta un estudio en el que se pone a prueba la capacidad de la TRI de obtener estimaciones invariantes de los ítems y sujetos, encontrándose que si bien es una pretensión correcta, su alcance es limitado para el investigador aplicado.

Finalmente, la discusión y conclusiones de esta tesis tratan de establecer nociones generales respecto de los méritos relativos de la TRI para el investigador aplicado y la posibilidad de proponer alternativas de calidad a esta. En conjunto, se trata de resituar las ventajas de la TRI sobre un terreno más sólido y generar un marco que le permita al investigador no experto en psicometría comprender qué es, qué se puede ganar y en qué condiciones, se deberían obtener beneficios de la utilización de esta teoría de los tests frente a aproximaciones alternativas.

Cada uno de los capítulos de esta tesis está basado en artículos de investigación publicados o que están actualmente en una primera o segunda ronda de revisión en revistas indexadas en la *Web of Science* (anteriormente ISI Web of Knowledge) y tienen índices de impacto que las coloca en el primer o segundo cuartil. Dichos capítulos han sido traducidos del inglés y han sido moderadamente modificados para mejorar la coherencia de la tesis.

En consecuencia, cada uno de los capítulos esta basado en los siguientes artículos:

Capítulo 1: Asún, R. A., Rdz-Navarro, K., y Alvarado, J. M. (2015). Developing multidimensional Likert scales using item factor analysis: the case of four-point items. *Sociological Methods & Research*. Publicada en línea antes de prensa. http://dx.doi.org/10.1177/0049124114566716

Capítulo 2: Asún, R. A., Rdz-Navarro, K., y Alvarado, J. M. (bajo segunda ronda de revisión). Comparing modern procedures to assess Likert-type scales: Polytomous IRT and Confirmatory Item Factor Analysis estimations.

Capítulo 3: Asún, R. A., Rdz-Navarro, K., y Alvarado, J. M. (bajo segunda ronda de revisión). Revisiting subject scaling procedures: Classical test theory raw scores versus item response theory theta estimates.

Capítulo 4: Asún, R. A., Rdz-Navarro, K., y Alvarado, J. M. (esperando decisión final). The sirens' call in psychometrics: The invariance of IRT models.

Nota técnica: En todos estos artículos se han realizado estudios Monte Carlo basados en los generadores de números aleatorios que por defecto incluyen los softwares que hemos empleado y que son de uso habitual en investigación psicométrica. De esta forma, el software MPlus (Muthén y Muthén, 2011) emplea ran2 (Press, Teukolsky, Vetterling, y Flannery, 1992), que es un generador de números aleatorios de precisión doble. Por su parte, el software Lisrel (Jöreskog y Sörbom, 2006) también emplea un generador de números aleatorios de precisión doble (Schrage, 1979). Finalmente, el software R (R Development Core Team, 2012) emplea el procedimiento Mersenne-Twister (Matsumoto y Nishimura, 1998) para generar números aleatorios de precision simple.

CAPITULO 1

DESARROLLANDO ESCALAS LIKERT MULTIDIMENSIONALES EMPLEANDO ANÁLISIS FACTORIAL DE ÍTEM: EL CASO DE LOS ÍTEMS DE 4 ALTERNATIVAS DE RESPUESTA

RESUMEN

En esta investigación se compara el desempeño de dos enfoques posibles de utilizar cuando se analizan escalas de clasificación tipo Likert de cuatro puntos empleando procedimientos factoriales: el análisis factorial (AF) clásico y el análisis factorial de ítem (AFI). Para el AF, se compara la efectividad de los procedimientos de estimación por máxima verosimilitud (ML) y por mínimos cuadrados ponderados (WLS) empleando como información las matrices de correlación de Pearson entre los ítems. Para el AFI, se compra la efectividad de los procedimientos de estimación por mínimos cuadrados diagonalmente ponderados (DWLS) y por mínimos cuadrados no ponderados (ULS), empleando como información las matrices policóricas de correlación de los ítems. A través de un estudio Monte Carlo, se simularon 210 condiciones, considerando de uno a tres factores latentes (independientes y correlacionados en dos niveles), ítems de mediana o baja calidad, tres diferentes niveles asimetría de los ítems y cinco tamaños de muestras. Los resultados muestran que los dos procedimientos AFI logran equivalentes y precisas estimaciones de parámetros, mientras que los dos procedimientos AF obtienen estimaciones sesgadas de los parámetros. Por lo tanto, no recomendamos emplear el AF clásico bajo las condiciones consideradas. Se discuten en profundidad los requerimientos mínimos para lograr resultados precisos usando procedimientos AFI.

INTRODUCCIÓN

El mecanismo de construcción de escalas desarrollado por Likert (Likert, 1932; Likert, Roslow, y Murphy, 1934) constituye una técnica simple y rápida para generar instrumentos de medición, por lo que es ampliamente usado por los científicos sociales para medir una importante variedad de constructos. Por ello, se han desarrollado procedimientos estadísticos meticulosos para diseñar y validar este tipo de escalas (DeVellis, 1991; Spector, 1992). Sin embargo, muchos de estos procedimientos ignoran la naturaleza ordinal de las respuestas dadas a los ítems y asumen la presencia de variables observadas continuas, medidas a nivel de intervalo. Aunque en la literatura aún es posible encontrar mucho debate sobre la robustez de las técnicas estadísticas paramétricas para desarrollar escalas Likert disponiendo de datos ordinales (Jamieson, 2004; Carifio y Perla, 2007; Norman, 2010), la evidencia muestra que, bajo circunstancias relativamente comunes, en estas condiciones el análisis factorial (FA) clásico produce resultados imprecisos al caracterizar la estructura interna de la escala o seleccionar los ítems más informativos dentro de cada factor (Berstein y Teng, 1989; DiStefano, 2002; Holgado-Tello, Chacón-Moscoso, Barbero-García, y Vila-Abad, 2010).

Afortunadamente, el análisis factorial de ítems (AFI) provee una alternativa que evita esos problemas (Wirth y Edwards, 2007) porque reconoce la naturaleza ordinal de las variables observadas.

Aunque la relevancia del análisis AFI para desarrollar Escalas Likert ha sido reconocido (Flora y Curran, 2004), aún existe debate respecto a los procedimientos de estimación específicos que habría que emplear, especialmente en el caso de ítems politómicos (Savalei y Rhemtulla, 2012), y no ha sido descartado que algún procedimiento de estimación alternativo podría permitir el uso del AF en vez del AFI.

Así, este capítulo apunta a llenar ese vacío, presentando los resultados de un estudio de simulación que compara el desempeño de los procedimientos de estimación AFI más recomendados, frente a algunas alternativas desarrolladas dentro del AF clásico. Dado que el desempeño de los procedimientos de estimación dependen del número de categorías de respuesta de los ítems (Beauducel y Herzberg, 2006; Dolan, 1994; Savalei y Rhemtulla, 2013), esta investigación se focalizará en el caso de los ítems de 4 puntos, dado que esa estructura de respuestas ha sido poco investigada, a pesar de ser el formato más empleado en las escalas Likert cuando se sospecha que la categoría intermedia es inadecuada.

El Número de Categorías de Respuesta en las Escalas Tipo Likert

Desde que por primera vez Rensis Likert sugirió el procedimiento de construcción de escalas que ahora lleva su nombre, ha habido un considerable debate acerca del número óptimo de categorías que se debe presentar a los sujetos que responden el cuestionario. Interesantemente, la evidencia encontrada en la literatura apoya posiciones altamente contrastantes: algunos investigadores sugieren que un mayor número de categorías de respuesta logra niveles de confiabilidad más altos (Garner, 1960) y validez (Hancock y Klockars, 1991; Loken, Pirie, Virnig, Hinkle, y Salmon, 1987), mientras que otros sugieren que el número de categorías de respuesta no está relacionado con la confiabilidad de la escala (Boote, 1981; Brown, Wilding, y Coulter, 1991) o su validez (Chang, 1994; Matell y Jacoby, 1971). En general, la evidencia tiende a indicar que: a) los investigadores deberían evitar presentar pocas categorías de respuestas (dos o tres), ya que podría decrecer la validez de la escala y los sujetos pueden sentir que no son capaces de expresar su verdadera opinión cuando responden el cuestionario (Preston y Colman, 2000); y b) los beneficios de aumentar el número de las categorías de respuesta desaparecen si se presenta a los sujetos más de siete alternativas, porque los sujetos serán incapaces de discriminar entre ellas (Miller, 1956).

Por esas razones, la mayor parte de las escalas Likert emplean de cuatro a siete categorías de respuestas, siendo cinco y siete el formato más comúnmente usado en investigación aplicada (Cox, 1980). La preferencia por un número impar de categorías de respuesta refleja la tendencia a emplear ítems que permitan a los sujetos definirse como neutrales respecto al constructo medido (Preston y Colman, 2000).

No obstante, la categoría intermedia puede afectar la validez de los resultados porque: (a) los sujetos podrían usar esta categoría por razones diferentes a tener una opinión intermedia, como por ejemplo: no tener opinión, no querer expresar su verdadera opinión, no entender la pregunta, que la pregunta no sea pertinente para la persona, entre otras posibilidades (Kulas, Stachowski, y Haynes, 2008, Raaijmakers, van Hoof, Hart, Verbogt, y Wollebergh, 2000); (b) se ha reportado la existencia de relación entre la deseabilidad social y el empleo de la categoría intermedia (Garland, 1991); (c) en algunos idiomas –como el Castellano- es una tarea engorrosa expresar semánticamente la idea de neutralidad de forma coherente con el resto de categorías de respuesta (González-Romá y Espejo, 2003); y (d) en ciertas ocasiones la información aportada por una categoría intermedia es casi nula (Andrich, 1978).

Por lo tanto, un formato de respuesta de cuatro puntos es altamente atractivo cuando se sospecha que la deseabilidad social afecta a la población diana, los sujetos son heterogéneos en sus capacidades para discriminar entre las categorías (i.e., la muestra es obtenida de una población general) o cuando el método de administración de la escala hace difícil emplear un mayor número de categorías de respuesta (i.e., cara a cara).

Sin embargo, al considerar utilizar un formato de respuesta de cuatro puntos, los investigadores debieran tener en cuenta que cuando el número de categorías de respuesta decrece, los ítems no tendrán un nivel de medición de intervalo, con lo que cualquier análisis

estadístico que presuponga dicha propiedad (como el AF clásico) posiblemente produzca resultados imprecisos.

Las Escalas Tipo Likert y el Análisis Factorial Clásico

El análisis factorial ha sido ampliamente reconocido como un procedimiento central para desarrollar escalas Likert (Nunnally, 1978). Así, el enfoque tradicional indica que cuando se desea producir una escala unidimensional o se sospecha que esa es la naturaleza del constructor latente medido, se debería seleccionar a aquellos ítems que maximicen la consistencia interna de la escala, empleando ya sea las correlaciones de Pearson entre el ítem y la escala total, y/o el α de Cronbach (DeVellis, 1991), procedimiento que permanece popular a pesar de la crítica que ha recibido (Sijtsma, 2009). El análisis factorial podría ser posteriormente empleado para chequear que efectivamente se haya producido una escala unidimensional. Por otro lado, si se supone que se está midiendo un constructo multidimensional, los investigadores deberían comenzar el proceso de construcción de la escala empleando el análisis factorial para evaluar la estructura interna de la base de datos, para luego proceder a seleccionar aquellos ítems que miden mejor cada factor, empleando las cargas factoriales de los ítems o los mismos análisis estadísticos utilizados en el caso de una escala unidimensional, pero individualmente para cada dimensión (Spector, 1992).

Uno de los problemas de esta perspectiva es que el AF clásico asume que los ítems son variables continuas medidas a nivel de intervalo, y los procedimientos de estimación más usados dentro de ese enfoque, tales como la estimación por máxima verosimilitud (ML, por sus siglas en inglés), funcionan más eficientemente cuando las respuestas observadas tienen una distribución multivariada normal. En contraste, los ítems en una escala Likert se codifican usando un procedimiento conocido como puntaje entero (González-Romá y Espejo, 2003), el cual asigna sucesivos enteros a cada categoría de respuesta (1, 2, 3,…, *n*). En

consecuencia, los ítems pueden ser considerados solo como mediciones ordinales, en el mejor de los casos.

Varios autores han señalado que la validez de los análisis estadísticos no depende de los niveles de la medición (Gaito, 1980; Lord, 1953a; Velleman y Wilkinson, 1993), que los análisis estadísticos paramétricos son robustos a disponer de datos ordinales (Norman, 2010) y, más aún, que las escalas Likert producen mediciones de nivel de intervalo (Carifio y Perla, 2007). Sin embargo, la teoría de la medición claramente establece que no es posible inferir cantidades de atributos ordinales (Michell, 2009). Esto implica que, aunque en algunos casos utilizar estadística que asume nivel de medición de intervalo sobre datos ordinales podría funcionar bien, saltarse este supuesto podría ser altamente problemático en otras condiciones, especialmente cuando los datos se alejen de la normalidad multivariante.

Esta situación es particularmente problemática para el AF clásico, porque cuando éste es aplicado a datos discontinuos, la correlación entre las variables observadas depende tanto de la cantidad real de asociación entre ellas, como de la distribución de frecuencias relativa de cada par de variables, de manera tal que aquellos ítems que tengan muy diferentes frecuencias de respuesta, mostrarán correlaciones artificialmente atenuadas (McDonald, 1999), lo que podría generar: (a) la aparición de factores espurios debido a correlaciones artificialmente más altas entre ítems con frecuencias de respuesta más similares, aumentando la complejidad dimensional del instrumento (Berstein y Teng, 1989) y; (b) la subestimación de las cargas factoriales de ítems con frecuencia de respuesta asimétrica (DiStefano, 2002), lo que podría producir una selección no óptima de ítems para el instrumento final.

Aunque se han propuesto algunas soluciones para este problema, tales como crear parcelas de ítems para así obtener un mayor número de categorías de respuesta en cada variable observada y así emular el disponer de variables continuas (Hau y Marsh, 2004),

actualmente se considera que el AFI es la alternativa que mejor preserva la lógica de análisis factorial aplicado a los ítems, tratando cada uno de ellos como indicador independiente.

El Análisis Factorial de Ítems

Durante los últimos cuarenta años, diversos investigadores han estado desarrollando métodos que permitan al AF clásico tratar con variables dicotómicas y ordinales (Christofferson, 1975; Christofferson, 1977; McDonald, 1982; Muthén, 1978, 1984, 1989). Muchas de las propuestas que se han desarrollado están basadas en una metodología de tres pasos.

Primero, se asume que cada variable categórica observada obtenida de la respuesta a un ítem es sólo un registro aproximado y tosco de la verdadera variable continua con distribución normal subyacente que habría sido lograda si los sujetos no hubieran tenido que restringir su respuesta a un número limitado de alternativas ordinales. En consecuencia, se estiman puntuaciones umbrales que representan el valor que habría permitido la ordinalización de las variables continuas subyacentes.

Formalmente, si un ítem tiene una serie de categorías de respuesta ordenadas (1, 2, 3, …, *m*), *z* es la respuesta ordinal dada por el sujeto en el ítem y *z** es el verdadero puntaje subyacente que el sujeto debiera haber contestado; la conexión entre *z* y *z** será:

$$Si... \quad \tau_{i-1} < z^* < \tau_i \quad \rightarrow \quad z = i \tag{1.1}$$

Donde *m*-1 parámetros umbrales fragmentan la escala de z*:

$$-\infty < \tau_1 < \tau_2 < \ldots < \tau_{m-1} < +\infty \tag{1.2}$$

Segundo, usando parámetros umbrales y la distribución bivariada entre las variables, se estiman las correlaciones tetracóricas o policóricas entre los ítems (en el caso de variables observadas dicotómicas o politómicas, respectivamente), para reflejar la asociación entre las variables continuas subyacentes.

Por último, se ajusta un modelo factorial y se estiman cargas factoriales –lambdas (λ)- para cada ítem empleando procedimientos que minimizan las diferencias entre las matrices de correlación tetra o policóricas observadas y la matriz reproducida por el modelo.

Tres procedimientos de estimación han sido aconsejados para este tipo de datos: (a) cuadrados mínimos ponderados (WLS, por sus siglas en inglés; Muthén, 1984) el cual minimiza la matriz residual ponderada por la varianza-covarianza de la matriz de estimaciones de las correlaciones tetra o policóricas; (b) cuadrados mínimos diagonalmente ponderados (DWLS, por sus siglas en inglés; Muthén, du Toit, y Spisic, 1997) el cual minimiza la matriz residual ponderada por las varianzas de las estimaciones de correlación tetra o policóricas y; (c) mínimos cuadrados no ponderados (ULS, por sus siglas en inglés; Muthén, 1993) el cual minimiza la matriz residual no ponderada.

Estudios previos han mostrado que el AFI tiende a producir estimaciones más precisas comparadas con el AF clásico (este último estimado empleando ML) en datos dicotómicos u ordinales con pocas alternativas de respuesta, y que ambos procedimientos tienden a converger cuando se dispone de cinco o más alternativas de respuesta (Beauducel y Herzberg, 2006; DiStefano, 2002; Dolan, 1994; Holgado–Tello et al., 2010; Rhemtulla, Brosseau-Liard, y Savalei, 2012).

Además, cuando se usa AFI, los diferentes procedimientos de estimación tienen diferentes desempeños; por ejemplo, aunque WLS tiene sobresalientes propiedades asintóticas, cuando se aplica a datos ordinales, requiere muestras muy grandes para obtener estimaciones insesgadas y en muestras pequeñas presenta problemas de convergencia y obtiene estimaciones de parámetros sesgados e inestables (Flora y Curran, 2004).

Con respecto a ULS y DWLS, la información actualmente es escasa y en cierta medida inconsistente; por ejemplo, Rigdon y Ferguson (1991) no hallaron diferencias entre estos dos procedimientos, resultado casi igual al reportado por Yang-Wallentin, Jöreskog, y

Luo (2010), mientras que Rhemtulla et al. (2012) señalan que ambos procedimientos obtuvieron apropiadas estimaciones de parámetros y tasas de convergencia equivalentes, aunque ULS logró más bajas tasas de error Tipo I, al tiempo que Forero, Maydeu-Olivares, y Gallardo-Pujol (2009) reportaron que DWLS muestra tasas más altas de convergencia que ULS, pero ULS es más robusto a las condiciones más difíciles (muestras pequeñas, distribuciones asimétricas y respuestas dicotómicas). Debe notarse que esta última investigación es la única que emplea datos politómicos, aunque no diferencia por tipo de datos en sus resultados, por lo que no es posible saber cuál producirá mejores resultados en escalas Likert con más de dos categorías de respuesta.

Así, considerando la información disponible hoy en día, no es posible definir cuál es el mejor procedimiento de estimación para analizar escalas tipo Likert de cuatro alternativas de respuesta porque, aunque la mayoría de las investigaciones concluyen que el número de categorías de respuesta afecta la efectividad de los procedimientos de estimación (Beauducel y Herzberg, 2006; Dolan, 1994; Savalei y Rhemtulla, 2013), solamente unos pocos estudios han evaluado este formato de respuesta, mientras que la mayor parte evaluó el caso dicotómico o el disponer de un número impar de categorías de respuesta (tres o cinco).

Además, pese a que WLS no parece ser una opción para estimar modelos AFI, es conocido que fue desarrollado como una alternativa a ML cuando no existe normalidad multivariante en el AF clásico basado en correlaciones Pearson (por ello WLS también es conocido como distribución asintóticamente libre –ADF, por sus siglas en inglés–; Browne, 1984), por lo que podría ser robusto a trabajar con datos ordinales. Pese a ello, su desempeño no ha sido testeado en el contexto de este tipo de datos, es decir, asumiendo que las respuestas ordinales son medidas a nivel de intervalo y empleando como información las correlaciones Pearson entre los ítems. Considerando que WLS está disponible en varios programas de estimación de modelos factoriales bien conocidos, tales como AMOS

(Arbuckle, 2010) y LISREL (Jöreskog y Sörbom, 2006), su desempeño es de gran interés, porque podría constituir una alternativa más sencilla para el investigador aplicado, poco familiarizado con las correlaciones tetra y policóricas.

Por lo tanto, con el objetivo de proveer de pautas para la investigación aplicada interesada en analizar o construir escalas Likert compuestas por ítems con cuatro alternativas de respuesta, se realizó un estudio Monte Carlo que comparó el desempeño de dos procedimientos de estimación AFI -concretamente DWLS y ULS (de aquí en adelante $DWLS_{PO}$ y ULS_{PO} para indicar que las estimaciones son realizadas a partir de la matriz de correlaciones policóricas)- con procedimientos de AF clásico –concretamente WLS y ML (de ahora en adelante WLS_{PE} y ML_{PE} para indicar que las estimaciones son realizadas a partir de las correlaciones de Pearson entre los ítems), donde ML_{PE} será considerado la línea base para comparar las mejoras posibles de obtener con los otros tres procedimientos.

Con este estudio esperamos aportar información que clarifique las consecuencias de utilizar uno u otro procedimiento de estimación al aplicar modelos factoriales y de esta forma ayudar a los investigadores aplicados a mejorar sus prácticas, para lograr instrumentos más confiables y válidos.

MÉTODO

Procedimiento de Simulación

Los datos fueron generados usando el software PRELIS 2 (Jöreskog y Sörbom, 2002) según el siguiente modelo factorial multidimensional:

$$X_{ij} = \sum_{k=1}^{k} \lambda_{jk} \times F_k + \left(1 - \sum_{k=1}^{k} \lambda_{jk}^2\right)^{0.5} \times e_j \qquad (1.3)$$

Donde X_{ij} es la respuesta simulada del sujeto *i* al ítem *j*, λ_{ik} es la carga factorial del ítem *i* en factor *k* (se simuló estructuras simples sin cargas cruzadas, así $\lambda_{jk} = 0$ para la

relación entre cada ítem y los factores pertenecientes a un factor distinto al asignado para él), F_k son factores latentes creados a partir de una distribución normal (dichos los factores fueron simulados independientes o linealmente asociados entre sí), y e_j es el error de medición aleatorio de cada ítem simulado a partir de una distribución estándar normal.

Dado que las variables X_j fueron generadas continuas, ellas fueron posteriormente recodificadas en cuatro categorías de respuesta de acuerdo a la proporción deseada de respuestas en cada categoría (este proceso será explicado más tarde) para representar ítems tipo Likert con cuatro alternativas de respuesta.

Condiciones Simuladas

Los datos fueron generados para estructuras factoriales con una, dos y tres dimensiones, dada la frecuencia de ese tipo de estructuras en la investigación aplicada. Para las condiciones multidimensionales, se simularon tres grados de correlación entre los factores para representar situaciones habituales: correlación nula ($\rho = 0$), baja ($\rho = .3$) y alta ($\rho = .6$).

Para aumentar la probabilidad de obtener factores bien especificados (Fabrigar, Wegener, MacCallum, y Strahan, 1999), se crearon seis ítems para cada dimensión; de esta forma, fueron simulados seis, doce y dieciocho ítems para las condiciones unidimensionales, bidimensionales y tridimensionales respectivamente.

Para evaluar la robustez de cada procedimiento de estimación en función de la calidad de la escala, las cargas factoriales fueron adaptadas para representar ítems con cargas bajas ($\lambda = .3$) y medianas ($\lambda = .6$). Si bien es razonable argumentar que aceptar saturaciones de .3 implica trabajar con ítems de muy baja calidad (pues sólo el 9% de la varianza del indicador se puede atribuir al factor) y es posible encontrar textos que recomiendan sólo interpretar ítems con cargas iguales o mayores a .4 (Stevens, 1992), esa recomendación no es universalmente aceptada, lo que quizá explique que diversos estudios relativamente recientes han mostrado que, pese a que .4 parece ser el punto de corte más frecuentemente empleado

(Henson y Roberts, 2006; Peterson, 2000), una proporción importante de los investigadores emplean .3 como el valor mínimo para considerar que un ítem puede ser retenido como indicador de un factor (Hair, Anderson, Tatham, y Black, 1998; Henson y Roberts, 2006; Merenda, 1997; Peterson, 2000). La intensión de esta investigación es chequear el impacto de trabajar con indicadores de tan baja calidad.

Como ya hemos señalado, los ítems fueron recodificados en cuatro categorías. Ello se hizo empleando puntos que corte que permitieran lograr distribuciones discontinuas con diferentes grados de asimetría, para así evaluar el desempeño de cada procedimiento en diferentes distribuciones de respuestas. De esta forma, como se muestra en la Figura 1, se crearon tres tipos de distribuciones: los ítems tipo I representan distribuciones simétricas, los ítems tipo II representan asimetría media ($g_1 = 1.1$), mientras los ítems tipo III representan respuestas de alta asimetría ($g_1 = 1.7$). No se simuló mayores niveles de asimetría porque ellos implicarían situaciones en que las personas emplearían menos alternativas de respuesta que las puestas a su disposición, transformando los ítems en virtualmente dicotómicos.

Finalmente, los tamaños de las muestras fueron escogidos para representar situaciones habitualmente utilizadas en la investigación aplicada, concretamente: 100, 200, 500, 1000 y 2000 sujetos.

Según los criterios de Harwell, Stone, Hsu, y Kirisci (1996), se crearon 500 réplicas para aquellas condiciones en que se esperaba mayores varianzas en los resultados (condiciones de 100 y 200 sujetos, o condiciones con 500 sujetos en una estructura tri-dimensional con ítems altamente asimétricos) y 250 replicas para el resto de las condiciones.

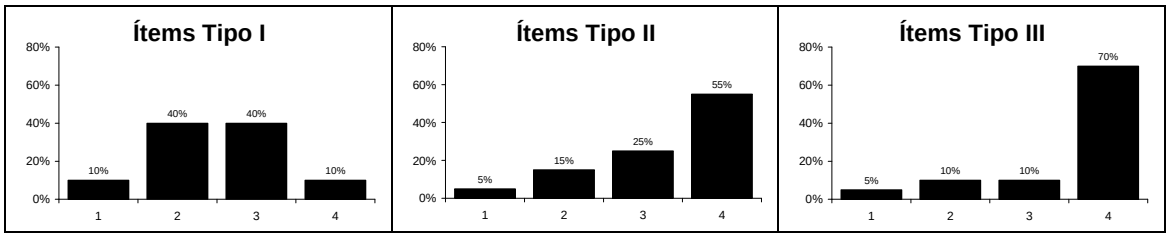

Figura 1. Tipos de ítems simulados.

En síntesis, se generaron 210 condiciones: 180 estructuras multidimensionales (2 y 3 factores x 3 niveles de correlación entre ellos x 2 tamaños de parámetros lambda x 3 niveles de asimetría x 5 tamaños de muestra) y 30 estructuras unidimensionales (2 tamaños de parámetros lambda x 3 niveles de asimetría x 5 tamaños de muestra).

Análisis de la Efectividad de los Procedimientos de Estimación

Para evaluar el desempeño de cada procedimiento de estimación ($DWLS_{PO}$, ULS_{PO}, WLS_{PE}, y ML_{PE}) al usar ítems tipo Likert con cuatro alternativas de respuesta, se implementó un análisis factorial confirmatorio (CFA, por sus siglas en inglés) usando el programa LISREL 8.8 (Jöreskog y Sörbom, 2006).

Cada procedimiento fue evaluado en su capacidad de producir estimaciones de parámetros no sesgados y estables. Por ello, consideramos: (a) La tasa de convergencia y soluciones admisibles obtenidas por cada procedimiento. Las soluciones no convergentes fueron aquellas para las cuales el procedimiento de estimación no alcanzó una solución después de 250 iteraciones, mientras que soluciones no admisibles fueron aquellas que produjeron valores fuera de rango o casos *Heywood* (i.e., varianzas negativas, parámetros lambda estandarizados mayores que 1). Por simplicidad, de aquí en adelante la tasa de convergencia y las soluciones admisibles serán referidas simplemente como tasa de convergencia (TC). Como ha sido sugerido por algunas investigaciones previas (Flora y Curran, 2004), las soluciones no convergentes y no admisibles no fueron consideradas en el resto de los análisis; (b) El sesgo relativo de estimación de los parámetros lambda (SRL), que

se definió como el porcentaje de sub o sobreestimación de los parámetros lambda simulados, promediados para todas las réplicas de cada condición; (c) la desviación estándar de estimaciones de los parámetros lambda (DSL) dentro de cada condición; (d) el sesgo absoluto de estimación del parámetro de correlación (SAC), el cual es la magnitud de sobre o subestimación de la correlación entre factores en valores absolutos, promediados a para todas las réplicas de de cada condición (el sesgo relativo de correlación entre los factores no fue considerado porque en algunas condiciones el valor ρ simulado fue nulo, con lo que el valor del SAC no está definido para esas situaciones) y; (e) la desviación estándar de las estimaciones de correlaciones (DSC), que consiste en la desviación estándar de la estimación de la correlación entre factores, promediados a través de todas las réplicas de cada condición.

El análisis de datos combina el empleo de pruebas ANOVA multivariadas, la estimación del tamaño del efecto usando el estadístico eta-cuadrado parcial (η^2_p) y el análisis descriptivo de resultados. Para los análisis descriptivos, se consideró como tamaños de efecto grandes aquellos valores que exceden .25 (Ferguson, 2009), se consideró inaceptables aquellas TC menores a 80% (Forero y Maydeu-Olivares, 2009), y se valoró como relevante cualquier sesgo mayor a 5% y *SD* mayores que 0.1 (Hoogland y Boomsma, 1998).

RESULTADOS

Análisis preliminares mostraron que ni la complejidad del modelo factorial (número de factores simulados), ni la presencia y magnitud de correlación entre factores, tenían un efecto estadísticamente importante en la explicación de las diferencias entre procedimientos de estimación; por lo tanto, esos resultados son omitidos del capítulo.

Tasa de Convergencia

La TC es altamente relevante para investigación aplicada porque refleja la probabilidad de lograr una solución aceptable al seleccionar un procedimiento estadístico.

La Tabla 1 muestra que los procedimientos de estimación considerados en este estudio no tuvieron un efecto importante en la capacidad de lograr soluciones válidas. Este resultado es muy interesante, pues en este estudio incluimos dos procedimientos del AF clásico que corrientemente no son recomendados en la literatura para tratar datos ordinales y cuyos sus resultados en TC fueron similares a los procedimientos AFI.

Tabla 1. *Análisis de varianza de las tasas de convergencia*

Variable	*F* (df [a])	η^2_p
PE	1.67 (3)	.01
Magnitud de los lambda	554.62 (1) **	.41
Asimetría de los items	10.37 (2) **	.03
Tamaño de la muestra	168.92 (4) **	.46
PE x lambda	1.50 (3)	.01
PE x asimetría	0.01 (6)	.00
PE x tamaño de la muestra	0.25 (12)	.00

Nota: PE = procedimiento de estimación. *F*(df) = *F* de Fischer-Snedecor y grados de libertad. η^2_p = Eta al cuadrado parcial. [a.] Grados de libertad del error = 808. * $p < .05$; ** $p < .01$.

Consecuentemente con la tabla 1, la Figura 2 muestra que los procedimientos tuvieron desempeños similares en sus tasas de convergencia a través de las 210 condiciones. Sin embargo, es interesante observar que ML_{PE} tiende a obtener una proporción ligeramente más baja de réplicas convergentes cuando se lo compara con los otros procedimientos y que WLS_{PE} evidenció mejores resultados que ML_{PE}. También debe notarse que no se encontró efectos de interacción significativos entre los procedimientos de estimación y el tamaño de muestra (ver Tabla 1), lo que implica que la TC de WLS_{PE} no fue afectada por la presencia de tamaños de muestras pequeñas, resultado que parece contradecir la evidencia reportada por aquellos estudios que han usado WLS con matrices de correlación tetra o policóricas - WLS_{PO}- (DiStefano, 2002; Flora y Curran, 2004). Para confirmar que este resultado

inesperado es correcto, decidimos probar WLS_{PO} en nuestros datos, y, como era de esperar, éste si obtuvo tasas de convergencia más bajas que los otros procedimientos en muestras menores de 500 sujetos, a diferencia de lo encontrado para $WLS_{PE.}$

Por otra parte, las variables que si mostraron un significativo efecto en la TC fueron: (a) la magnitud de los parámetros lambda simulados, donde en presencia de ítems de baja calidad ($\lambda = .3$) se obtuvo una tasa de TC inaceptable (69.7%), la cual mejoró notoriamente (hasta llegar a una TC casi perfecta) cuando la calidad de los ítems fue alta ($\lambda = .6$) y; (b) el tamaño de las muestras, donde se obtuvo un TC inaceptable con muestras de 100 sujetos (57.8%), el cual mejoró hasta un nivel satisfactorio (95.6%) para muestras de 500 casos y a nivel óptimo (99.2%) en muestras de 1000 sujetos. En consecuencia, independiente del procedimiento de estimación, se puede obtener aceptables TC con tamaños de muestras iguales o mayores a 500 sujetos si la calidad de los ítems es baja, mientras que sólo 100 sujetos son bastantes para estimar un modelo, cuando la calidad de los ítems es alta ($\lambda = .6$).

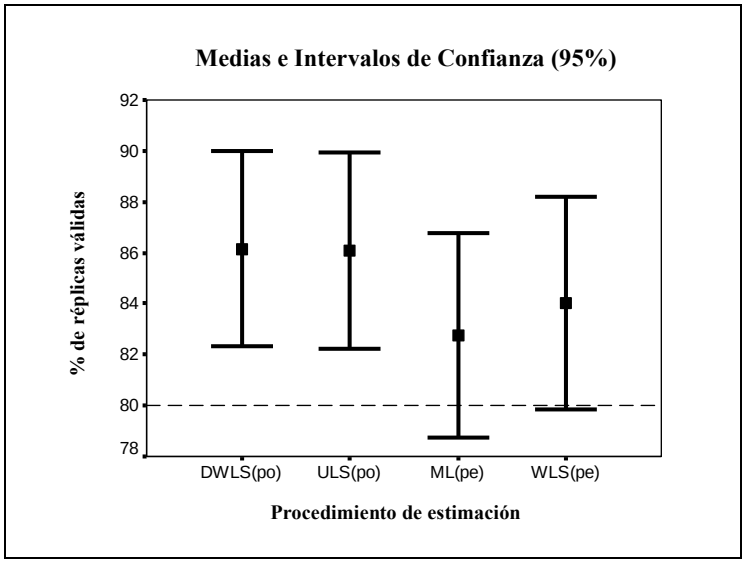

Figura 2. Medias e intervalos de confianza de las tasas de convergencia.

Sesgos Relativos de la Estimación de los parámetros lambda

Los parámetros lambda constituyen una información clave en el proceso de construcción de Escalas Likert pues sólo si se dispone de las cargas factoriales correctas entre los ítems y sus factores, se asegura una la adecuada eliminación de los ítems menos informativos al construir una escala uni o multidimensional.

Tabla 2. *Análisis de varianza del sesgo relativo de los parámetros lambda*

Variable	*F* (df [a])	η^2_p
PE	385.92 (3) **	.59
Magnitud de los lambda	174.10 (1) **	.18
Asimetría de los items	54.49 (2) **	.12
Tamaño de la muestra	257.76 (4) **	.56
PE x lambda	3.70 (3) *	.01
PE x asimetría	34.35 (6) **	.20
PE x tamaño de la muestra	33.04 (12) **	.33

Nota: PE = procedimiento de estimación. *F*(df) = *F* de Fischer-Snedecor y grados de libertad. η^2_p = Eta al cuadrado parcial. [a.] Grados de libertad del error = 808. * $p < .05$; ** $p < .01$.

Como se muestra en la Tabla 2, los procedimientos de estimación tuvieron un efecto estadísticamente significativo y grande en el SRL. Para examinar este efecto en detalle, la Figura 3 muestra el desempeño de cada procedimiento. Aquí podemos apreciar que $DWLS_{PO}$ y ULS_{PO} lograron resultados relativamente precisos (aunque levemente mejores para ULS_{PO}) con solo una ligera sobreestimación del parámetro simulado. Sorpresivamente, WLS_{PE} se desempeñó razonablemente bien, evidenciando un bajo sesgo de subestimación (menos de 5%) el cual fue solo ligeramente mayor que el sesgo mostrado por los procedimientos AFI. Por consiguiente, a diferencia del procedimiento ML_{PE} que manifestó un sesgo importante al estimar los parámetros lambda, WLS_{PE} podría ser considerado un procedimiento alternativo para obtener dichos parámetros en forma relativamente insesgada en ítems tipo Likert de 4 opciones de respuesta. Sin embargo, la magnitud de los efectos de interacción entre los

procedimientos de estimación y los tamaños de las muestras, tanto como con la asimetría de los ítems (ver Tabla 2), indican que la situación es más compleja.

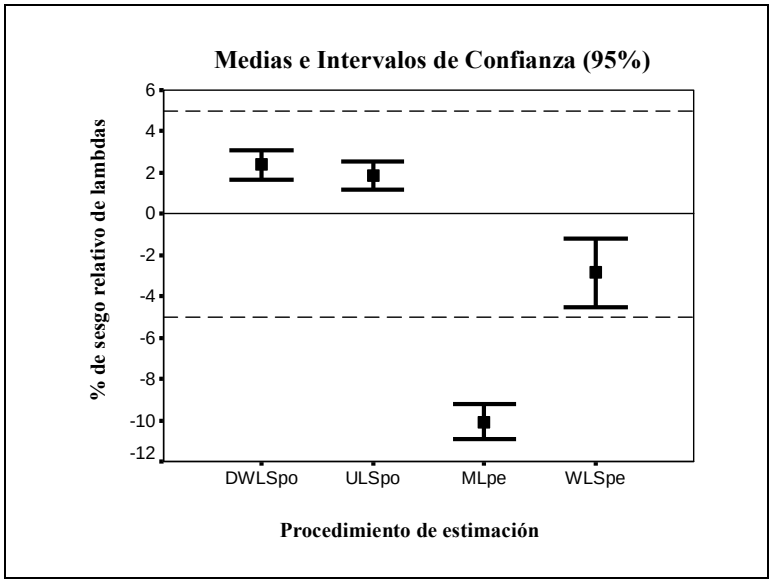

Figura 3. Medias e intervalos de confianza del sesgo relativo de los parámetros lambda.

De hecho, tal como se muestra en la Figura 4, WLS_{PE} sólo logró resultados equivalentes a ULS_{PO} y WLS_{PO} en presencia de ítems simétricos y muestras de 200 sujetos. Muestras más pequeñas tienden a presentar sobreestimaciones inaceptables, mientras que muestras iguales o mayores a 500 sujetos evidencian parámetros inaceptablemente subestimados. Mas aún, analizando detalladamente los resultados obtenidos por WLS_{PE} es posible determinar que su sesgo cercano a cero en muestras de 200 sujetos es el resultado de un desempeño inestable donde se obtienen grandes sesgos de signos opuestos que se compensan al obtener un promedio. Así, para muestras de 200 sujetos, WLS_{PE} sobreestima los parámetros lambda cuando la calidad de los ítems es baja ($\lambda = .3$), y este sesgo tiende a decrecer cuando la asimetría de los ítems aumenta, mientras que para ítems de alta calidad ($\lambda = .6$) se sobreestima el parámetro simulado y este sesgo tiende a aumentar cuando la asimetría

del ítem crece. Por lo tanto, WLS_{PE} no es un procedimiento confiable para calcular cargas factoriales en ítems tipo Likert.

En suma, observando la Figura 4, podemos concluir que procedimientos ULS_{PO} y $DWLS_{PO}$ mostraron desempeños similares (aunque ULS_{PO} parece ligeramente mejor), ambos son relativamente robustos a asimetría de ítems y muestras de 200 sujetos parece ser suficiente para alcanzar resultados aceptables –aunque 500 sujetos son requeridos para obtener precisión óptima-.

En contraste, ML_{PE} tiende a subestimar los parámetros lambda en todas las condiciones, especialmente cuando los ítems no son simétricos y, sorprendentemente, aumentar el tamaño de la muestra permite solamente la estabilización del sesgo de subestimación alrededor del 10%, sin resolver el problema.

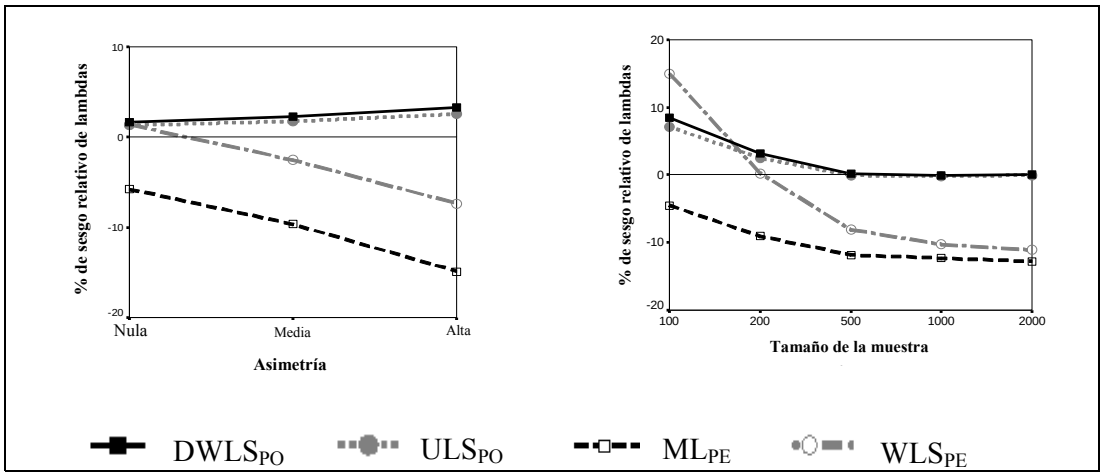

Figura 4. Sesgo relativo de estimación por asimetría y tamaño de muestra.

Desviación Estándar de la Estimación de los Parámetros Lambda

La DSL es un indicador relevante de la estabilidad de los parámetros estimados por un procedimiento estadístico. Valores grandes de *SD* muestran que un procedimiento de estimación produce estimaciones muy diferentes frente a datos equivalentes (parte de una misma condición) y, por tanto, evidencia que sus estimaciones no son exactas; en contraste, aquellos procedimientos que obtienen un bajo *SD* serán más exactos al estimar un parámetro.

Como se muestra en la Tabla 3, los procedimientos de estimación tuvieron un efecto estadísticamente significativo sobre la estabilidad de las estimaciones del parámetro; sin embargo, el tamaño de este efecto es casi irrelevante. Por lo tanto, se puede afirmar que los procedimientos de estimación no son demasiado diferentes en sus niveles de inestabilidad al estimar el parámetro. Además, el análisis descriptivo de resultados mostró que todos los procedimientos presentaron resultados dentro del rango aceptable.

Tabla 3. *Análisis de varianza de la desviación estándar de estimación de los parámetros lambda*

Variable	*F* (df [a])	η^2_p
PE	4.35 (3) **	.02
Magnitud de los lambda	3204.52 (1) **	.80
Asimetría de los items	162.94 (2) **	.29
Tamaño de la muestra	2431.55 (4) **	.92
PE x lambda	1.37 (3)	.01
PE x asimetría	0.43 (6)	.03
PE x tamaño de la muestra	2.27 (12) **	.03

Nota: EP = procedimiento de estimación. *F*(df) = *F* de Fischer-Snedecor y grados de libertad. η^2_p = eta al cuadrado parcial. [a.] Grados de libertad del error = 808. * $p < .05$; ** $p < .01$.

Las variables que tienen al menos un efecto moderado sobre la inestabilidad de las estimaciones del parámetro son la asimetría de ítems, la magnitud de los parámetros lambda y los tamaños de la muestra. Sin embargo, las diferencias relacionadas con la asimetría de los ítem son poco relevantes (por ejemplo, para ítems altamente asimétricos *SD* = 0.09, mientras que para ítems simétricos *SD* = 0.07). Con respecto a la magnitud de los parámetros lambda, cuando la calidad de los ítems fue baja ($\lambda = .3$) los parámetros fueron estimados justo en el límite superior de la inestabilidad aceptable (*DE* = 0.11), mientras que para ítems de más alta calidad ($\lambda = .6$) las estimaciones del parámetro fueron estables (*DE* = 0.06). Finalmente, para muestras iguales o menores que 100 sujetos, se observa una gran inestabilidad en las estimaciones (*DE* = 0.15), las que tienden a alcanzar valores completamente aceptables para muestras de 500 sujetos o más (*DE* = 0.07).

Sesgo Absoluto de la Estimación de las Correlaciones

Una estimación inapropiada de la correlación entre factores puede llevar a una errónea representación de la estructura dimensional del constructo que se desea medir. Por ello, los procedimientos de estimación deben ser analizados respecto de su capacidad relativa de reproducir la correlación simulada.

La Tabla 4 evidencia que se encontró una relación estadísticamente significativa entre los procedimientos de estimación y el SAC. Sin embargo, el tamaño de efecto de esta relación es moderado, ya que el sesgo empírico absoluto estuvo dentro del rango de -0.02 y 0.02, con lo que sólo ligeras diferencias fueron encontradas producto de que ML_{PE} obtuvo valores negativos de sesgo mientras que WLS_{PE} y los procedimientos AFI ($DWLS_{PO}$ y ULS_{PO}) presentaron asimetrías positivas.

Tabla 4. *Análisis de varianza del sesgo de estimación de las correlaciones entre factores*

Variable	F (df [a])	η^2_p
PE	27.04 (3) **	.11
Magnitud de los lambda	4.24 (1) *	.01
Asimetría de los items	6.89 (2) **	.02
Tamaño de la muestra	2.96 (4) *	.02
PE x lambda	8.42 (3) **	.04
PE x asimetría	1.47 (6)	.01
PE x tamaño de la muestra	5.75 (12) **	.09

Nota: PE = procedimiento de estimación. F(df) = F de Fischer-Snedecor y grados de libertad. η^2_p= eta al cuadrado parcial. [a.] Grados de libertad del error = 808. * $p < .05$; ** $p < .01$.

Como se observa en la Tabla 4, se encontró efectos significativos para varias variables independientes, pero el único relevante es una interacción entre los procedimientos de estimación y el tamaño de la muestra. La Figura 5 ilustra que este efecto consistente en una cierta tendencia de algunos procedimientos a manifestar mayores sesgos en presencia de tamaños de muestra pequeños. En este escenario, ML_{PE} tiende a subestimar la correlación y

WLS_{PE} tiende a sobreestimarla, mientras que $DWLS_{PO}$ y ULS_{PO} se muestran más robustos en esas condiciones.

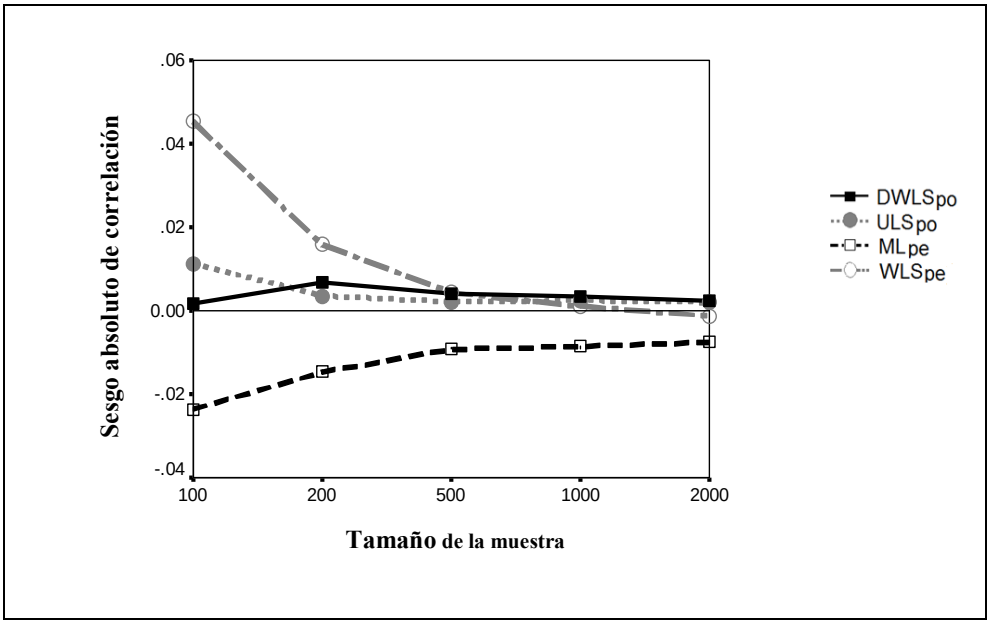

Figura 5. Sesgo absoluto de estimación de la correlación según tamaño de la muestra por procedimiento de estimación.

Desviación Estándar de la Estimación de las Correlaciones

Observando la Tabla 5 podemos afirmar que no hay diferencias estadísticamente significativas entre los procedimientos de estimación como efecto principal. De hecho, todos los procedimientos de estimación tienden a calcular la correlación entre factores con el mismo grado de inestabilidad, el cual estuvo levemente sobre el nivel aceptable ($DE > 0.1$).

Además, la Tabla 5 muestra que no existe ningún efecto de interacción entre los procedimientos de estimación y las otras variables independientes, lo cual indica que ningún procedimiento supera a los otros en cualquier situación.

Solamente dos efectos estadísticamente significativos y fuertes fueron encontrados para DSC: la magnitud de los parámetros lambda y el tamaño de la muestra. A semejanza de los análisis previos, se obtuvo mejores resultados con ítems de buena calidad y más pobres para ítems de baja calidad (cuando $\lambda = .3$ DSC = 0.18 y para $\lambda = .6$ DSC = 0.08), mientras

que la heterogeneidad de las estimaciones fue mayor en muestras más pequeñas (cuando *n* = 100 DSC = 0.23 y para *n* = 2000 DSC = 0.06).

Tabla 5. *Análisis de varianza de la desviación estándar de la estimación de correlaciones entre factores*

Variable	*F* (df [a])	η^2_p
PE	0.38 (3)	.00
Magnitud de los lambda	1669.83 (1) **	.71
Asimetría de los items	30.46 (2) **	.08
Tamaño de la muestra	614.02 (4) **	.78
PE x lambda	1.19 (3)	.01
PE x asimetría	0.18 (6)	.00
PE x tamaño de la muestra	0.58 (12)	.01

Nota: PE = procedimiento de estimación. *F*(df) = *F* de Fischer-Snedecor y grados de libertad. η^2_p = eta al cuadrado parcial. [a.] Grados de libertad del error = 808. * $p < .05$; ** $p < .01$.

En general, los resultados muestran que, para alcanzar un nivel aceptable de heterogeneidad (DSC < 0.1), se requiere muestras de 2000 sujetos cuando la calidad de los ítems es baja ($\lambda = .3$), mientras que una muestra de 500 sujetos podría ser suficiente si la calidad de los ítems es mediana ($\lambda = .6$).

CONCLUSIONES

Este estudio tenía como objetivo determinar el mejor procedimiento para estimar modelos factoriales aplicados a escalas compuestas por ítems tipo Likert de 4 alternativas de respuesta en escenarios uni y multidimensionales. Por ello, se comparó procedimientos AFI con procedimientos AF clásicos y, en conjunto, encontramos que los procedimientos AFI fueron netamente superiores a la perspectiva clásica.

De acuerdo a nuestros hallazgos, aunque todos los procedimientos mostraron una similar capacidad para producir soluciones válidas y lograr parámetros lambda y de correlación estables, ULS_{PO} y $DWLS_{PO}$ evidenciaron sesgos notablemente menores al estimar

ambos tipos de parámetros y fueron robustos en los escenarios más difíciles: distribuciones de ítems asimétricos, baja calidad de los ítems ($\lambda = .3$) y tamaños de muestra pequeños.

Ha sido claramente confirmado, entonces, que emplear procedimientos de estimación clásica en datos ordinales con cuatro alternativas de respuestas es inapropiado. Esto es consistente con la investigación previa que revela la sobreestimación de parámetros claves del modelo cuando se emplean procedimientos AF clásicos (Beauducel y Herzberg, 2006; DiStefano, 2002; Dolan, 1994; Holgado–Tello et al., 2010; Rhemtulla et al., 2012).

Sin embargo, en este tema dos puntos deben ser destacados: (a) primero, que usar AF con estimaciones WLS nunca es una opción viable para datos ordinales, dado lo obtenido en nuestra investigación empleando matrices de correlación de Pearson y considerando sus pobres resultados con matrices de correlación tetra y policóricas reportadas en la investigación previa (Flora y Curran, 2004) y; (b) segundo, que el pobre desempeño de ML_{PE} podría ser debido al empleo de correlaciones producto-momento de Pearson, en vez de poder ser atribuidas al procedimiento de estimación ML en si mismo, pues varios estudios han mostrado que usar estimación ML con matrices de correlación tetra o policóricas logra resultados bastante similares a $DWLS_{PO}$ y ULS_{PO}, especialmente en presencia de muestras grandes (Dolan, 1994; Rigdon y Ferguson, 1991; Yang-Wallentin et al., 2010).

De acuerdo a nuestros hallazgos, AFI debiera ser considerado el procedimiento estándar para analizar ítems ordinales de cuatro alternativas, porque su menor sesgo de estimación de lambdas y correlaciones, garantiza una selección más precisa de ítems para la escala final, y así, la generación de instrumentos más confiables.

Por otro lado, al comparar la calidad relativa de los procedimientos AFI ($DWLS_{PO}$ y ULS_{PO}), encontramos que apenas hay diferencias. De hecho, aunque ULS_{PO} parece mejor situado que $DWLS_{PO}$, esta ventaja es demasiado pequeña para hacer algunas recomendaciones con vistas a la investigación aplicada. Estos hallazgos son consistentes con

aquellos reportados por Rigdon y Ferguson (1991) y Yang-Wallentin et al. (2010) y de alguna manera divergen de aquellos reportados por Forero et al. (2009), pues la ventaja en favor de ULS_{PO} que ellos reportaron podría ser producto de trabajar con ítems dicotómicos, diluyéndose la diferencia en presencia de ítems con más alternativas de respuesta. Por lo tanto, nuestra recomendación es que los investigadores aplicados seleccionen a voluntad ULS_{PO} o $DWLS_{PO}$ al analizar escalas Likert multidimensionales compuestas de ítems politómicos.

Nuestro principal consejo para investigación aplicada se facilita dado que los procedimientos AFI están ampliamente implementados en varios programas bien conocidos y amigables, tales como: Factor (Lorenzo-Seva y Ferrando, 2006) el que permite estimar modelos AFI exploratorios, LISREL (Jöreskog y Sörbom, 2006), que puede ser usado para estimar modelos AFI confirmatorios, y MPlus (Muthén y Muthén, 2011), el cual permite desarrollar análisis AFI exploratorios y confirmatorios.

Además de la pregunta central de nuestra investigación, nuestro estudio también intentaba determinar los requerimientos mínimos para utilizar procedimientos AFI en ítems tipo Likert con cuatro alternativas de respuesta. A este respecto, nuestra investigación nos permite afirmar que si un investigador supone que la calidad de los ítems de su escala será baja ($\lambda = .3$), debería disponer de una muestra mínima de 500 sujetos para asegurar obtener resultados logrados admisibles (una solución convergente y sin casos Heywood) y una estimación de los parámetros clave del modelo relativamente insesgada y estable. Evidentemente, si se sospecha que los ítems reflejan el constructo latente en una mejor manera ($\lambda = .6$), se puede alcanzar estimaciones precisas y estables con muestras más pequeñas (200 o incluso 100 sujetos) si las distribuciones de los ítem son simétricas o sólo levemente asimétricas.

Para resumir, los hallazgos de la presente investigación revelan que el clásico AF no es robusto a disponer de datos ordinales en el caso de escalas tipo Likert de cuatro alternativas de respuesta. En consecuencia su empleo debe ser fuertemente desalentado en este escenario particular, aunque aún podría ser utilizado en otros escenarios con un mayor número de alternativas de respuestas –idealmente 7 o más- (Beauducel y Herzberg, 2006; Dolan, 1994; Rhemtulla et al., 2012).

Aunque estos hallazgos y directrices son muy interesantes y prometedores para la investigación aplicada, se deben señalar al menos tres importantes limitaciones del estudio para impedir que se hagan inferencias más allá de los límites de la evidencia disponible.

Primero, esta investigación consideró solamente modelos AFI confirmatorios. En consecuencia, se necesita más investigación para evaluar si estos hallazgos podrían ser extendidos a modelos factoriales exploratorios.

Segundo, nosotros consideramos solamente ítems tipo Likert de cuatro alternativas de respuesta, por lo que nuestros resultados no pueden ser completamente extrapolados a mayores o menores números de categorías. Dado que, a medida que el número de categorías de respuesta aumenta, los diferentes procedimientos tienden a mostrar mejores y más similares desempeños (Beauducel y Herzberg, 2006; Dolan, 1994; Savalei y Rhemtulla, 2013), se debería estudiar más detalladamente el caso de las escalas compuestas por ítems de tres alternativas, dado que la situación dicotómica ya ha sido ampliamente investigada.

Finalmente, esta investigación consideró solamente situaciones ideales (por ejemplo, una calidad homogénea de los ítems, sin la presencia de cargas cruzadas y sin datos perdidos). Por lo tanto, tendría mérito realizar un examen de los procedimientos de estimación AFI en situaciones más complejas y más cercanas a la investigación aplicada, como por ejemplo: incluir ítems de calidad heterogénea, considerar la presencia de factores débiles y fuertes, y simular número desigual de ítems por factor, entre otras posibilidades.

CAPITULO 2

COMPARANDO PROCEDIMIENTOS PARA LA EVALUACIÓN DE LA CALIDAD DE LOS ÍTEMS EN ESCALAS TIPO LIKERT: MODELOS POLITÓMICOS DE TRI VERSUS ANÁLISIS FACTORIAL DE ÍTEMS

RESUMEN

Con el objetivo de aportar a la discusión acerca de los mejores métodos para evaluar la calidad de ítems politómicos pertenecientes a escalas tipo Likert multidimensionales, se realizó un estudio Monte Carlo en que se comparó el desempeño de procedimientos de estimación derivados de la teoría de respuesta al ítem (TRI), con procedimientos asociados al análisis factorial de ítems (AFI). Los datos fueron generados según la versión factorial del modelo multidimensional de ojiva normal de respuesta graduada y las variables manipuladas fueron el tamaño de las muestras, el número y correlación entre los factores, y la asimetría y magnitud de las cargas factoriales de los ítems. Se evaluó la capacidad de recuperación del parámetro de discriminación de los ítems a través de dos procedimientos factoriales (ULS y DWLS) y uno de TRI (MLR). Contrariamente a lo que se esperaba, los tres procedimientos de estimación estudiados obtuvieron resultados relativamente similares y fueron afectados casi de igual forma por las mismas variables independientes. Además, se encontró que en condiciones subóptimas (tamaños de muestra pequeños y bajas cargas factoriales) los procedimientos de estimación por información limitada asociados al AFI alcanzaron resultados algo mejores. Para finalizar, se discute la importancia de las cargas factoriales y el tamaño de la muestra para alcanzar estimaciones óptimas, con independencia de los procedimientos de estimación utilizados.

INTRODUCCIÓN

Evaluar las propiedades de los ítems en escalas tipo Likert es una preocupación fundamental de los científicos sociales cuantitativos, pues de los resultados obtenidos dependerá la configuración final y la calidad de los instrumentos de medida usados para evaluar variables latentes complejas, como son las representaciones sociales, las opiniones, la personalidad o las actitudes de las personas (DeVellis, 1991; Spector, 1992).

Desde la década de 1980, una parte importante de la comunidad de expertos en psicometría han sostenido que las herramientas provistas por la teoría clásica de los tests (TCT) para evaluar la calidad y propiedades de tests e ítems (i.e., el cálculo de la discriminación de los ítems o el α de Cronbach, entre otros) tienen importantes limitaciones y no permiten un adecuado diagnóstico de las propiedades de los instrumentos de medida (Embretson y Reise, 2000; Sijtsma, 2009). Esas críticas han tenido una fuerte repercusión en algunos campos de la investigación social (i.e., la medición educacional y psicológica), en los cuales el empleo de modelos de la teoría de respuesta al ítem (TRI) ha sido alentado a fin de aprovechar las ventajas de esta teoría de los tests (De Ayala, 2009; Hambleton, Swaminathan, y Rogers, 1991). Pese a esta recomendación, los modelos TRI siguen siendo empleados principalmente para evaluar y diseñar tests unidimensionales compuestos por ítems dicotómicos, por lo que encontrar aplicaciones de este teoría de los tests fuera del campo de la medición educativa (donde priman ese tipo de instrumentos) es aún poco frecuente. Lo anterior posiblemente es consecuencia, entre otros factores, de la complejidad matemática de los procedimientos de estimación basados en la TRI, lo que aumenta muy rápidamente la carga computacional cuando se incrementa el número de factores del instrumento y el número de categorías de respuesta de los ítems (Forero y Maydeu-Olivares, 2009),

dificultando encontrar aplicaciones de esta teoría de los tests en el campo del diseño de escalas tipo Likert multidimensionales (Morren, Gelissen, y Vermunt, 2011).

Por otra parte, analizar escalas compuestas por ítems categóricos con el análisis factorial (AF) tradicional también presenta importantes problemas (DiStefano, 2002; Holgado–Tello et al., 2010), lo que ha impulsado el desarrollo del análisis factorial de ítems (AFI; Wirth y Edwards, 2007), el que es un tipo de AF especialmente adecuado para evaluar ítems categóricos y ordinales (es decir, dicotómicos y politómicos). El AFI estima los parámetros de los ítems tomando en cuenta la naturaleza discreta de los datos observados, lo que le permite superar el problema de la aparición de factores de dificultad habituales cuando se aplica el AF a ese tipo de datos (Bernstein y Teng, 1989). Además, su utilización en instrumentos multidimensionales y politómicos se ve facilitada por sus procedimientos de estimación más simples.

Dado que se ha demostrado que el AFI es matemáticamente equivalente a algunos de los modelos de TRI más populares (McDonald, 1997), se podría pensar que este procedimiento constituye una alternativa razonable para la estimación de los parámetros de ítems pertenecientes a escalas multidimensionales tipo Likert. Desafortunadamente, existe evidencia de que las estimaciones realizadas utilizando procedimientos derivados del AFI son menos precisas que las efectuadas desde la TRI (DeMars, 2012; Finch, 2010; Gosz y Walker, 2002; Tate, 2003). Sin embargo, esta diferencia se he demostrado con claridad sólo para datos dicotómicos. La única investigación actualmente disponible en la literatura que compara las estimaciones AFI y TRI en ítems politómicos (Forero y Maydeu-Olivares, 2009), si bien encontró algunas diferencias en la calidad de la recuperación de los parámetros de los ítems a favor de la TRI, no permite determinar en forma precisa si esas discrepancias se mantienen por igual en ítems dicotómicos que en politómicos.

Por lo tanto, la superioridad de los procedimientos de estimación asociados a la TRI todavía no ha sido establecida para datos politómicos, con lo que no se puede descartar que los procedimientos de estimación derivados del AFI pudieran ser una alternativa aceptable para analizar ese tipo de ítems. Este vacío de información puede ser de interés para los investigadores aplicados, pues dichos ítems, especialmente en formatos tipo Likert, son ampliamente utilizados en las ciencias sociales (Göb, McCollin, y Ramalhoto, 2007). En consecuencia, esta investigación tiene como objetivo evaluar y comparar la precisión de las estimaciones de parámetros realizadas con procedimientos vinculados a la TRI y al AFI en ítems politómicos y condiciones multidimensionales, con el fin de aportar algunas recomendaciones para la construcción de escalas tipo Likert en situaciones aplicadas.

PROCEDIMIENTOS MODERNOS PARA EVALUAR ÍTEMS EN ESCALAS LIKERT

La Teoría de Respuesta al Ítem

La TRI es una teoría de los tests compleja y diversa, constituida por una amplia variedad de modelos de medida para ítems dicotómicos y politómicos, así como para estructuras de datos unidimensionales y multidimensionales.

Los modelos de la TRI permiten la evaluación de las propiedades de sujetos e ítems asumiendo que las probabilidades de acertar o de contestar en una determinada categoría de respuesta de un ítem depende de uno o más atributos latentes continuos (por ejemplo, el nivel de conocimientos, las habilidades o las actitudes de las personas), de las características de los ítems (i..e, discriminación, dificultad, entre otras), y de una función link matemática que los vincula (de ojiva normal, logística o de punto ideal, entre las más habituales).

Generalmente, los procedimientos de estimación utilizados en la TRI emplean los patrones completos de respuestas de los sujetos, es decir, toda la información contenida en la base de datos, para calibrar ítems y sujetos. Por ello, se los denomina procedimientos por información completa (FI, por sus siglas en inglés). La estimación se realiza a través de algoritmos iterativos, que minimizan la diferencia entre las respuestas observadas y predichas a los ítems (Hambleton et al., 1991) empleando procedimientos por máxima verosimilitud o Bayesianos, los cuales logran estimaciones asintóticamente consistentes (Hambleton y Swaminathan, 1985).

La literatura psicométrica reconoce que el empleo de modelos de TRI tiene ventajas importantes sobre la TCT como herramienta para la construcción y evaluación de tests. Por ejemplo, la TRI posibilita la evaluación de sujetos e ítems en la misma escala, facilita conocer la precisión de la estimación obtenida (i.e., el error estándar) para cada nivel del rasgo y permitiría alcanzar estimaciones invariantes de los parámetros de ítems y sujetos, entre otras posibilidades (De Ayala, 2009; Embretson y Reise, 2000; Hambleton et al., 1991).

A pesar de estas ventajas teóricas, surgen ciertas dificultades prácticas cuando se ajustan modelos de TRI. En primer lugar, la TRI requiere grandes volúmenes de información (es decir, un importante número de ítems y una amplia muestra de sujetos) para obtener estimaciones de parámetros precisas, por lo que las propiedades asintóticas óptimas de los procedimientos por FI pueden no mantenerse en situaciones no óptimas (Forero y Maydeu-Olivares, 2009). En segundo lugar, como consecuencia del uso de procedimientos de estimación por FI, la carga computacional de los análisis crece rápidamente al aumentar la complejidad del modelo. De este modo, tomando en cuenta que el número de parámetros a estimar aumenta en la medida que el número de sujetos, de dimensiones y de categorías de respuesta de los ítems también lo hacen, la estimación de modelos de TRI a través de procedimientos por FI puede ser virtualmente imposible en presencia de instrumentos

politómicos con un gran número de ítems empleados para medir muchas dimensiones, situación habitual en ciencias sociales al emplear escalas tipo Likert para medir constructos complejos (Morren et al., 2011).

El Análisis Factorial de Ítems

El análisis factorial clásico asume que las variables que se va a analizar son continuas, por lo que emplearlo para estudiar datos ordinales o dicotómicos como los que producen habitualmente las preguntas de los tests que se utilizan en ciencias sociales, puede generar factores de dificultad espurios o una subestimación de los parámetros de discriminación de los ítems (Bernstein y Teng, 1989). Estos problemas motivaron el desarrollo del análisis factorial para datos ordinales (Christofferson, 1975, 1977; McDonald, 1982; Muthén, 1978, 1984), el cual permite trabajar con variables categóricas (dicotómicas o politómicas, siempre que estas últimas sean ordinales), lo que coincide con la mayor parte de las preguntas empleadas en los tests psicológicos y sociológicos, y con los ítems de las escalas tipo Likert. Ello ha llevado a algunos autores a rebautizar este análisis como análisis factorial de ítems (AFI; Wirth y Edwards, 2007).

A diferencia de la TRI, los modelos de análisis factorial tienden a estimar los parámetros del modelo empleando solo la información contenida en la matriz de varianza-covarianza de relaciones entre las variables observadas, es decir, emplean procedimientos de estimación por información limitada (LI, por sus siglas en inglés), el que se tiende a implementar en los modelos AFI a través de una metodología en tres pasos.

En primer lugar, las variables observadas se asumen como indicadores recodificados de las variables subyacentes continuas y con distribución normal que habrían constituido las respuestas de los sujetos, de no haber sido constreñidos por las alternativas de respuesta de los ítems. En consecuencia, el primer paso de la calibración consiste en estimar los *umbrales*

(τ) subyacentes a esa recodificación, los que son cercanos al valor z asociado a la proporción de respuestas acumuladas en cada categoría.

En un segundo paso, utilizando los umbrales anteriormente definidos y la distribución bivariada entre las variables observadas, se estima la asociación entre cada par de ellas empleando correlaciones tetracóricas o policóricas, pues estos estadísticos pueden entenderse como la estimación máximo verosímil de la correlación producto-momento que existiría entre los indicadores continuos con distribución normal que subyacen a los ítems categóricos observados.

Finalmente, se estiman los parámetros del modelo a través de procedimientos que minimizan la diferencia entre la matriz de correlaciones tetracóricas o policóricas estimada y reproducida por el modelo. Entre estos procedimientos, dos parecen lograr los mejores resultados en ítems politómicos y en este tipo de correlaciones (Forero, Maydeu–Olivares, y Gallardo-Pujol, 2009): mínimos cuadrados no ponderados (en adelante ULS por sus siglas en inglés; Muthén, 1993) y mínimos cuadrados diagonalmente ponderados (en adelante DWLS por sus siglas en inglés; Muthén, du Toit, y Spisic, 1997).

Las principales ventajas del AFI son las siguientes: (a) la facilidad de computación, que le permite estimar los parámetros de instrumentos más complejos que los actualmente alcanzables con procedimientos por FI en ordenadores de capacidad media; y (b) la equivalencia matemática entre el AFI y algunos modelos de la TRI (Christofferson, 1975; McDonald, 1982; Muthén, 1978; Takane y de Leeuw, 1987), implica que los parámetros estimables por ambos tipos de modelos son también equivalentes, aunque se encuentren en una métrica distinta. De esta forma, los parámetros λ corresponden a la discriminación (i.e., al parámetro *a*) de los ítems de algunos modelos de TRI, mientras que los τ son equivalentes a los parámetros de dificultad (i.e., parámetros *b*) de ítems dicotómicos o a los interceptos de los ítems de algunos modelos politómicos de TRI. Es más, con el AFI también es posible

estimar una puntuación factorial para cada sujeto, con lo que se obtiene un parámetro similar a la habilidad del sujeto en la TRI (θ).

Por lo tanto, el AFI permite la estimación de parámetros que son similares o equivalentes a aquellos obtenidos en los modelos de la TRI más ampliamente utilizados (i.e., modelos uni o multidimensionales, dicotómicos de uno o dos parámetros o politómicos de respuesta graduada), por lo que puede entenderse al AFI como una teoría de los tests alternativa, que aunque se encuentra restringida a un limitado conjunto de modelos, tiene el potencial de permitir trabajar con instrumentos factorialmente más complejos que los posibles de estimar actualmente por medio de la TRI en ordenadores normales, con la posible limitación de lograr una menor precisión en las calibraciones de los parámetros, como consecuencia del empleo de procedimientos de estimación que solo usan una porción de información contenida en las respuestas de los sujetos.

Equivalencia entre Modelos de TRI y AFI para Ítems Politómicos

En el marco de la TRI, las escalas tipo Likert pueden ser analizadas empleando modelos politómicos. Aun cuando existen muchos modelos de TRI para ese tipo de datos, el modelo de respuesta graduada (GRM, por sus siglas en inglés; Samejima, 1969) es actualmente uno de los más exitosos y frecuentemente utilizados (Maydeu-Olivares, Hernández, y McDonald, 2006), por lo que sus extensiones a situaciones multidimensionales son particularmente relevantes para la investigación aplicada.

Dos de esas extensiones son el modelo multidimensional de respuesta graduada (MGRM, por sus siglas en inglés), el cual emplea una función link logística para relacionar las probabilidades de respuesta con el rasgo latente, y el modelo multidimensional de respuesta graduada de ojiva normal (NOMGRM, por sus siglas en inglés), que utiliza una función link normal. El NOMGRM es de particular interés para este estudio porque, como señalan Forero y Maydeu-Olivares (2009), dicho modelo tiene un equivalente en el AFI.

El NOMGRM puede ser formalizado como (Muraki y Carlson, 1995):

$$P(u_{ij} = k \mid \boldsymbol{\theta}_i) = \frac{1}{\sqrt{2\pi}} \int_{\mathbf{a}_j'\boldsymbol{\theta}_j + d_{j.k+1}}^{\mathbf{a}_j'\boldsymbol{\theta}_i + d_{jk}} e^{\frac{-t^2}{2}} dt \,, \tag{2.1}$$

donde u_{ij} es la respuesta del sujeto *i* al ítem *j*, *k* son las alternativas de respuesta de los ítems, $\boldsymbol{\theta}_i$ es el vector de habilidades del sujeto, $\mathbf{a}_j$ es el vector de los parámetros de discriminación de los ítems, y d_{jk} son los interceptos de los ítems que indican la puntuación de theta donde es más probable responder *k*+1 que *k*.

El modelo AFI equivalente al NOMGRM es:

$$x_i^* = \lambda_{jl}\xi_l + \delta_j \,, \tag{2.2}$$

donde x_j^* son los *j* indicadores continuos subyacentes, los cuales son ordinalizados en *k* categorías por *k*-1 umbrales, produciendo los ítems categóricos observados x_j; λ_{jl} son las cargas factoriales de los indicadores *j* en los factores *l*; ξ_l son los factores latentes subyacentes *l,* y δ_j son los errores aleatorios de medida de los indicadores *j*.

La equivalencia entre los parámetros del CFA y el NOMGRM es obtenida por:

$$\boldsymbol{\lambda}_j = \frac{\mathbf{a}_j}{\sqrt{1 + \mathbf{a}_j' \boldsymbol{\Psi} \mathbf{a}_j}} \quad , \tag{2.3}$$

$$\tau_{jk} = \frac{-\mathbf{a}_j b_{jk}}{\sqrt{1 + \mathbf{a}_j' \boldsymbol{\Psi} \mathbf{a}_j}} \quad , \tag{2.4}$$

donde, $\boldsymbol{\Psi}$ es la matriz de correlación entre factores latentes; $\boldsymbol{\lambda}_j$ es el vector de parámetros lambda del AFI; τ_{jk} son los umbrales de los ítems en el modelo AFI; $\mathbf{a}_j$ es el vector de parámetros de discriminación y b_{jk} son los interceptos de los ítems en el modelo NOMGRM.

Dada esta equivalencia, ambos modelos podrían ser entendidos como diferentes parametrizaciones del mismo modelo subyacente (McDonald, 1997), cuya única diferencia es

la fuente de información empleada para la estimación de sus parámetros, es decir, habitualmente información limitada en al AFI e información completa en su contraparte TRI.

Precisión de las Estimaciones AFI y TRI

Como se mencionó anteriormente, los procedimientos de estimación asociados a la TRI teóricamente deberían lograr mejores calibraciones (i.e., menos sesgadas y con menor varianza) que aquellas alcanzadas por el AFI, pues los primeros emplean mucha mayor proporción de la información contenida en las respuestas de los sujetos.

Algunos estudios Monte Carlo han intentado evaluar esta superioridad, encontrando que ambos tipos de procedimientos tienen desempeños similares en un amplio rango de situaciones cuando los datos son simulados desde un modelo poblacional que tiene equivalente en el AFI, pero que ello no ocurre en caso contrario. Por ejemplo, Tate (2003) y Finch (2010) reportan que las estimaciones del AFI contienen mucho mayor sesgo y varianza que las estimaciones de la TRI cuando los datos se generan a partir de un modelo logístico de tres parámetros (lo que es esperable puesto que el AFI carece de un parámetro de seudo azar), pero esta diferencia disminuye notoriamente cuando los datos se simulan desde un modelo de dos parámetros, el cual si tiene equivalente en el AFI.

Sin embargo, también existe evidencia de que las estimaciones por FI asociadas con la TRI son más precisas que las por LI del AFI incluso para modelos equivalentes. Por ejemplo, Boulet (1996) y De Mars (2012) encontraron que cuando los factores simulados no tienen distribución normal, los procedimientos TRI logran mejores estimaciones; Gosz y Walker (2002) confirmaron dicho hallazgo para pruebas con un gran número de ítems con cargas cruzadas. Reiser y VanderBerg (1994) encontraron que los procedimientos AFI son menos precisos que las estimaciones asociadas a la TRI cuando las muestras están constituidas por menos de 500 sujetos, y Muraki y Engelhard (1985) reportaron que los procedimientos por FI obtienen mejores resultados que los por LI cuando los ítems son más difíciles.

Lamentablemente, todos los estudios citados en el párrafo anterior han sido realizados con datos dicotómicos. El único trabajo en el que se emplean ítems politómicos (Forero y Maydeu-Olivares, 2009) también reporta diferencias que favorecen a las estimaciones basadas en la TRI en las situaciones más difíciles (ítems altamente asimétricos, muestras iguales o menores a 500 sujetos, y tres ítems por factor), pero debe notarse que sus resultados no diferencian la evidencia obtenida a partir de datos dicotómicos de la producida por datos politómicos y las condiciones más difíciles incluyeron casi exclusivamente ítems dicotómicos (cf., Forero y Maydeu-Olivares, 2009, pp. 287-290). De ahí que las conclusiones sacadas de estos estudios previos no debieran ser extendidas a situaciones politómicas sin mayor evidencia.

En consecuencia, una pregunta que continúa abierta y que es de gran interés desde un punto de vista metodológico y práctico, especialmente en los campos de las ciencias sociales que más regularmente emplean escalas tipo Likert para medir constructos latentes, es saber si la superioridad teórica de las estimaciones TRI sobre las del AFI se mantienen en escenarios politómicos y multidimensionales, especialmente en condiciones subóptimas. El presente estudio pretende ser un aporte a responder esa pregunta, con el objetivo de proveer a los investigadores de recomendaciones prácticas para el desarrollo de escalas tipo Likert.

MÉTODO

Se llevó a cabo un estudio Monte Carlo para evaluar y comparar la precisión de la recuperación de parámetros por parte de procedimientos por FI e LI en situaciones politómicas y multidimensionales. Los datos fueron generados con el software PRELIS 2 (Jöreskog y Sörbom, 2002) de acuerdo a la Ecuación (2.2). Los ξ_l fueron simulados en función de una distribución normal estándar y en condiciones tanto de independencia, como

de asociación lineal entre sí. En tanto, las δ_j fueron generadas a partir de una distribución Normal (0, σ), mientras que la desviación estándar del término de error fue fijada de manera que todo indicador observado tuviera una varianza igual a 1. Finalmente, se generó una estructura simple de relación entre los ítems y los factores, sin cargas cruzadas ($\lambda_{jl} = 0$ para la relación entre cada ítem y los factores que no eran el propio). Después de la generación de datos, x_j^* fue recodificada en cuatro categorías para representar ítems tipo Likert de cuatro alternativas de respuesta (x_{ij}), el cual es un formato ampliamente utilizado en la investigación aplicada cuando se busca forzar a los sujetos a tomar una posición con respecto a un constructo en situaciones en que la alternativa intermedia puede presentar problemas (Garland, 1991; Raaijmakers et al., 2000).

Las variables manipuladas fueron:

- La magnitud de los parámetros lambda, simulados en dos categorías (λ = .3 y λ = .6, equivalente a parámetros a = .56 y a = 1.27, respectivamente, en métrica logística), para representar ítems considerados de adecuada y mínima calidad (i.e., capacidad de reflejar la variable latente), en la investigación aplicada. Si bien desde el análisis factorial es posible encontrar autores (e.g., Stevens, 1992) que señalan λ ≥ .4 como el umbral para considerar que un ítem es un indicador adecuado de un factor, esa no es una recomendación aceptada universalmente, por lo que es posible encontrar un importante grupo de estudios que emplean como punto de corte valores de hasta .3 (Henson y Roberts, 2006; Peterson, 2000). La intensión de la presente investigación es poner a prueba las consecuencias de trabajar con esos indicadores de calidad mínima;
- El grado de asimetría de los ítems. Esto fue realizado ajustando la frecuencia relativa de cada categoría de respuesta en tres tipos de distribuciones que representan ítems simétricos (asimetría = 0), con asimetría media (asimetría = 1.1), y altamente asimétricos (asimetría = 1.7), tal como se representa en la Figura 6. Se descartó simular mayores

niveles de asimetría, porque habría implicado producir distribuciones de frecuencia de respuesta que hubieran dejado virtualmente vacías algunas alternativas, alterando el carácter politómico de los ítems.

- El número de factores, considerando escenarios de una, dos o tres dimensiones.
- Para condiciones multidimensionales, la correlación entre los factores fue ajustada en tres niveles: nula ($\rho = 0$), baja ($\rho = .3$), y alta ($\rho = .6$).
- El tamaño de muestras, según cinco niveles: 100, 200, 500, 1000, y 2000 sujetos, para representar muestras comúnmente usadas en la investigación aplicada y para evaluar los procedimientos ante cantidades de sujetos inferiores a los simulados en investigaciones anteriores (cf., Flora y Curran, 2004).

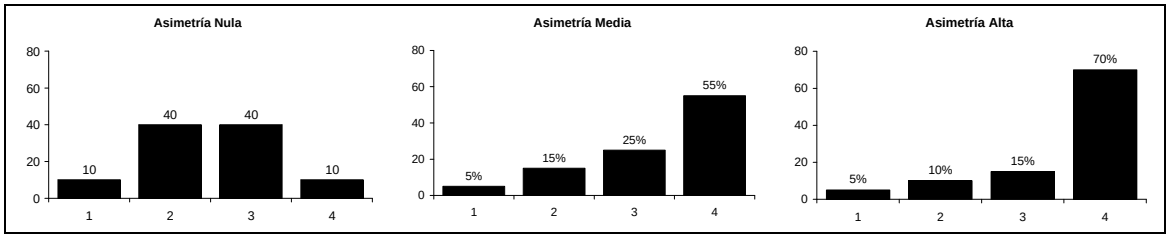

Figura 6. Distribución de los ítems según asimetría.

El número de ítems por dimensión fue fijado en seis siguiendo el consejo de Fabrigar et al. (1999), con respecto al número adecuado de ítems para lograr una buena especificación de los factores, por lo que se crearon 6, 12 y 18 ítems para las condiciones unidimensionales, bidimensionales y tridimensionales, respectivamente. Por lo tanto, se simularon 30 condiciones para el caso unidimensional (dos tamaños lambda x cinco tamaños de muestra x tres niveles de asimetría) y 180 para el caso multidimensional (las anteriores 30 x dos tipos de dimensiones x tres grados de correlación entre los factores).

Para definir el número de réplicas por condición, se siguió el consejo de Harwell y colaboradores (1996) relativo a realizar más réplicas en aquellas condiciones en que se espera una mayor varianza en los resultados. Por ello, se generaron 500 réplicas en las muestras de

100 y 200 sujetos, y en las de 500 sujetos cuando ello coincidió con simular estructuras tridimensionales y asimetrías altas. En las condiciones restantes, se generaron 250 réplicas.

Cada conjunto de datos fue calibrado con tres procedimientos de estimación (PE). Se seleccionó dos procedimientos por LI: ULS y DWLS; y un procedimiento por FI: la estimación por máxima verosimilitud con errores estándar robustos (MLR, por sus siglas en inglés).

Contrariamente a la práctica más habitual en los estudios de simulación, que tienden a emplear siempre que sea posible el mismo software para comparar dos o más procedimientos -buscando así controlar el efecto del programa informático sobre los resultados-, dado el enfoque aplicado de esta tesis, se ha optado por realizar los análisis de datos con los programas más utilizados por los investigadores en ciencias sociales en cada procedimiento: el Lisrel 8.8 (Jöreskog y Sörbom, 2006) para las estimaciones por LI y el MPlus 6.11 (Muthén y Muthén, 2011) para las estimaciones por FI. Si bien esta opción implica el riesgo de introducir al software como variable interviniente en los productos del estudio, ello también aumenta la capacidad de generalizar los hallazgos a lo que encontrarán los investigadores aplicados. En cualquier caso, para descartar que el programa informático ejerza una influencia relevante sobre los resultados, en algunas condiciones aleatoriamente seleccionadas se empleó MPlus 6.11 para obtener las estimaciones por ULS y DWLS, confirmando la presencia de diferencias insignificantes respecto a las alcanzadas con Lisrel 8.8. Adicionalmente, otras condiciones unidimensionales fueron recalibradas con el software Multilog 7.03 (Thissen, 2003) evidenciándose mínimas diferencias respecto de las obtenidas con MPlus 6.11.

Considerando que la estimación de los umbrales del modelo AFI en el programa Lisrel 8.8 sólo implica encontrar el valor z de la proporción acumulada de respuestas en cada categoría, y no depende del procedimiento de minimización empleado (i.e., ULS o DWLS),

se descartó evaluar la precisión de la recuperación de este parámetro. Por tanto, la investigación se enfocó en la precisión de la recuperación de los parámetros lambda o de discriminación, los que representan el grado con que cada ítem refleja el constructo subyacente y, por tanto, constituyen un aspecto central de la calidad de un instrumento.

Con el fin de realizar la comparación entre las estimaciones, las obtenidas con el procedimiento MLR fueron transformadas de acuerdo con las ecuaciones (2.3) y (2.4) para así expresar todos los parámetros en la métrica estandarizada propia del AFI, por lo que de aquí en adelante se usará la denominación *lambda* para hacer referencia a los parámetros lambda y/o de discriminación.

Se ha evaluado la capacidad de cada procedimiento de lograr estimaciones viables, insesgadas y estables, en función de las siguientes variables:

- Tasa de Convergencia (TC): se registró el porcentaje de réplicas no convergentes o impropias (parámetros fuera de rango) por cada condición. Se consideró que obtener una TC menor al 80% era inaceptable (cf., Forero y Maydeu-Olivares, 2009). Las soluciones no convergentes e impropias fueron eliminadas de posteriores análisis (cf., Flora y Curran, 2004).
- Sesgo relativo (SR): representa la proporción de infra o sobreestimación de los parámetros estimados en cada condición y se define como el parámetro estimado menos el parámetro poblacional dividido por el parámetro poblacional, multiplicado por 100. Un SR $\geq$ 5% fue considerado inaceptable (cf., Hoogland y Boomsma, 1998).
- Raíz de la media cuadrática de los errores de estimación (en adelante RMSE por sus siglas en inglés): representa la desviación respecto de lo simulado en cada réplica, por lo que constituye un índice no compensatorio del sesgo de estimación. Se lo define como la raíz del promedio de la diferencia cuadrática entre los parámetros simulados y estimados en cada réplica, donde valores altos indican un mayor sesgo. Como no se

encontró una sugerencia de punto de corte para interpretar este estadístico, hemos considerado como criterio ad-hoc que valores mayores o iguales a .1 son inaceptables.

- Desviación estándar de la estimación (DEE): representa la eficiencia del procedimiento de estimación, donde valores más altos indican una mayor inestabilidad de las estimaciones obtenidas. Se consideró inaceptable obtener valores mayores o iguales a .1 (cf., Hoogland y Boomsma, 1998).

Se llevo a cabo una de serie de pruebas ANOVA para cada variable dependiente (i.e., TC, SR, RMSE y DEE) con el fin de analizar, tanto los efectos principales de los PE y las variables manipuladas (variables independientes), como los efectos de interacción entre ellas. Evaluamos la magnitud de los efectos encontrados usando Eta cuadrado parcial (η^2_p), considerándose relevantes los efectos moderados a grandes ($\eta^2_p \geq .25$ cf., Ferguson, 2009). Con estos análisis se simplificó las tablas descriptivas de resultados.

RESULTADOS

Tasa de Convergencia

De acuerdo los resultados de la prueba de ANOVA, el emplear distintos PE no influyo en la TC. Sólo se encontraron efectos relevantes y estadísticamente significativos en relación con el tamaño muestral ($F(4, 507) = 113.12$; $p < .001$; $\eta^2_p = .47$) y la magnitud de los parámetros lambda poblacionales ($F(2, 507) = 324.38$; $p < .001$; $\eta^2_p = .39$). El grado de asimetría de los ítems y el número de dimensiones simuladas también dieron lugar a un efecto significativo en la TC, pero el tamaño de esos efectos fue muy pequeño (i.e., .03 y .01, respectivamente).

Como se muestra en la Tabla 6, en las condiciones en que se simuló lambdas de magnitud media ($\lambda = .6$), siempre se observaron tasas de convergencia óptimas, y solo hubo una leve tendencia a obtener TC más bajos para el procedimiento DWLS en muestras

pequeñas (n = 100), ítems altamente asimétricos y estructuras tridimensionales. En contraste, para las condiciones en que se generaron lambdas de tamaño bajo (λ = .3), frecuentemente se obtuvo tasas de convergencia inferiores al 80% cuando los tamaños de muestra eran iguales o menores a 200 sujetos (con independencia del PE utilizado), y sólo tendieron a alcanzar niveles aceptables u óptimos cuando el tamaño muestral igualó o excedió los 500 sujetos. Nótese que pese a no alcanzar niveles significativos, la estimación MLR tuvo una pequeña ventaja por sobre las dos versiones de estimación por LI, cuando la asimetría de los ítems fue alta, las muestras de tamaño medio, y cuando se simularon estructuras tridimensionales.

Sesgo Relativo

La prueba de ANOVA evidenció un efecto significativo, pero de pequeña magnitud (F(2, 508) = 4.4; p = .013; η^2_p = .02), entre los PE y el SR. Con relación al SR y las otras variables independientes, se encontraron efectos principales estadísticamente significativos y relevantes para el tamaño de los parámetros lambda (F(2, 507) = 166.04; p < .001; η^2_p = .25) y el tamaño muestral (F(4, 507) = 98.23; p < .001; η^2_p = .44). Por su parte, la correlación entre los factores, la asimetría de los ítems y el número de dimensiones simuladas, no mostraron efectos principales o de interacción significativos con el SR.

Tabla 6. *Tasas de convergencia*

		λ = .3									λ = .6								
Modelo		AFI						TRI			AFI						TRI		
PE		DWLS			ULS			MLR			DWLS			ULS			MLR		
Tipo Asim.		I	II	III	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III
Factores	*N*																		
1	100	**59.2**	**50.0**	**32.6**	**59.4**	**48.4**	**31.0**	**62.0**	**53.6**	**35.8**	100	100	99.6	100	100	99.6	100	100	99.8
	200	85.0	**75.0**	**62.0**	85.8	**74.2**	**60.2**	86.0	**78.6**	**63.6**	100	100	100	100	100	100	100	100	100
	500	99.6	98.0	93.2	99.6	97.6	93.2	99.2	98.0	95.2	100	100	100	100	100	100	100	100	100
	1000	100	100	99.6	100	100	99.6	100	100	99.2	100	100	100	100	100	100	100	100	100
	2000	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
2	100	**30.6**	**18.7**	**9.3**	**30.7**	**18.1**	**9.2**	**35.5**	**20.9**	**14.5**	100	100	99.1	100	100	99.3	100	100	99.3
	200	**70.7**	**57.0**	**36.6**	**71.7**	**56.8**	**34.3**	**72.3**	**55.6**	**42.4**	100	100	100	100	100	100	100	100	100
	500	98.8	97.6	89.2	98.7	97.1	88.3	98.8	95.9	90.4	100	100	100	100	100	100	100	100	100
	1000	100	100	99.7	99.9	100	99.6	100	100	99.7	100	100	100	100	100	100	100	100	100
	2000	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
3	100	**15.1**	**7.8**	**2.5**	**15.9**	**8.6**	**2.7**	**17.0**	**11.7**	**3.1**	100	99.9	97.7	100	99.9	98.1	100	99.9	99.5
	200	**58.2**	**37.6**	**17.2**	**58.2**	**38.7**	**15.8**	**61.1**	**28.9**	**14.3**	100	100	100	100	100	100	100	100	100
	500	97.7	93.9	**79.8**	97.6	94.0	**78.7**	97.6	95.3	83.7	100	100	100	100	100	100	100	100	100
	1000	100	99.6	98.8	100	99.7	98.7	100	99.7	98.7	100	100	100	100	100	100	100	100	100
	2000	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

Nota: AFI = análisis factorial de ítems. TRI = teoría de respuesta al ítem. DWLS = estimación por mínimos cuadrados diagonalmente ponderados. ULS = estimación por mínimos cuadrados no ponderados. MLR = estimación por máxima verosimilitud con errores estándar robustos. PE = procedimiento de estimación. I = ítems con distribución simétrica. II = ítems con distribución medianamente asimétrica. III = ítems con distribución altamente asimétrica. λ = parámetros lambda. Factores = número de dimensiones simuladas. *N* = tamaño de la muestra. En negrilla = resultado inaceptable, menos del 80% de las réplicas válidas.

Los resultados descriptivos de las variables que mostraron efectos significativos en el SR se presentan en la Tabla 7. En ella se puede apreciar que la presencia de lambdas de magnitud media (λ = .6) dio lugar a niveles óptimos de SR, independientemente de las otras variables, aunque la estimación MLR alcanzó un sesgo ligeramente mayor al de ULS y DWLS, con independencia del tamaño de las muestras. En contraste, cuando la magnitud del parámetro lambda poblacional fue baja (λ = .3), el tamaño muestral se volvió una variable crítica para el SR. De esta forma, en presencia de muestras pequeñas (n = 100), el SR fue inaceptable -independientemente del PE utilizado- y una situación similar se encontró con muestras de 200 sujetos, combinadas con asimetría de los ítems media o alta. En consecuencia, la asimetría pareció influenciar el SR solo en presencia de pequeñas cantidades de sujetos, volviéndose una variable poco relevante ante muestras de 500 o más.

Con respecto a las diferencias entre los PE, ULS y DWLS produjeron un sesgo ligeramente menor que MLR, lo cual podría explicar el efecto significativo encontrado en las pruebas ANOVA. Por otro lado, cuando se comparan los procedimientos de estimación por LI, ULS dio lugar a un SR menor que DWLS en las condiciones más rigurosas, logrando resultados insesgados incluso cuando las cargas factoriales fueron bajas, los tamaños de muestras iguales a 200 y la asimetría de los ítems no era alta.

RMSE

Para el RMSE, el análisis de ANOVA mostró efectos principales significativos (en orden decreciente) para el tamaño muestral (F(4, 507) = 1491.45; p < .001; η^2_p = .92), el tamaño de los parámetros lambda poblacionales (F(2, 507) = 2243.82; p < .001; η^2_p = .82), la asimetría de los ítems (F(2, 507) = 116,94; p < .001; η^2_p = .32), la correlación entre factores (F(2, 507) = 21.36; p < .001; η^2_p = .08), y el procedimiento de estimación (F(2, 507) = 9.52; p < .001; η^2_p = .04). Además, se encontró un efecto de interacción

significativo pero pequeño entre el PE y el tamaño de los parámetros lambda (F(2, 507) = 7.54; p < .001; η^2_p = .03). Los análisis descriptivos respecto a esta variable se muestran en la Tabla 8.

Tabla 7. *Porcentaje de sesgo relativo de los parámetros lambda*

Modelo		AFI						TRI		
PE		DWLS			ULS			MLR		
Tipo asim.		I	II	III	I	II	III	I	II	III
λ	*N*									
.3	100	**12.0**	**15.3**	**21.7**	**10.9**	**13.9**	**19.9**	**12.2**	**13.2**	**20.6**
	200	3.4	**5.6**	**9.2**	2.9	4.8	**8.3**	4.6	4.2	**8.2**
	500	0.1	0.4	0.7	-0.1	0.1	0.1	0.5	-0.7	1.3
	1000	-0.1	-0.1	-0.1	-0.3	-0.3	-0.4	0.4	-0.3	0.6
	2000	0.1	0.1	0.2	0	0	0	0.4	-0.3	0.8
.6	100	0.7	0.5	0.4	-0.1	-0.6	-1.3	0.9	0.8	0.9
	200	0.4	0.3	0.1	-0.1	-0.2	-0.7	0.4	0.6	0.8
	500	0.1	0.1	0.2	0	-0.1	-0.1	0.3	0.6	0.6
	1000	0.1	0.1	0	0	0	-0.2	0.2	0.3	0.4
	2000	0	0.1	0	0	0	-0.1	0.2	0.2	0.3

Nota: AFI = análisis factorial de ítems. TRI = teoría de respuesta al ítem. DWLS = estimación por mínimos cuadrados diagonalmente ponderados. ULS = estimación por mínimos cuadrados no ponderados. MLR = estimación por máxima verosimilitud con errores estándar robustos. PE = procedimiento de estimación. I = ítems con distribución simétrica. II = ítems con distribución medianamente asimétrica. III = ítems con distribución altamente asimétrica. λ = parámetros lambda. *N* = tamaño de la muestra. En negrilla = resultado inaceptable, sesgo mayor que 5%.

En dicha tabla se puede observar que cuando las cargas factoriales simuladas fueron de magnitud media (λ = .6) casi todas las condiciones dieron lugar a resultados aceptables, excepto cuando las muestras eran muy pequeñas (n = 100). Por lo tanto, muestras de 200 o más sujetos aseguraron estimaciones adecuadas. En estas condiciones, la influencia de otras variables fue insignificante, excepto que la presencia de una distribución simétrica parece mejorar ligeramente la precisión de la estimación. En contraste, cuando las cargas factoriales simuladas fueron pequeñas (λ = .3), el

tamaño muestral, la asimetría de la distribución de los ítems y la correlación entre factores tuvieron un importante impacto en la precisión de las estimaciones. De hecho, cuando el tamaño muestral fue igual o menor a 200 sujetos, el RMSE fue alto e inaceptable en la mayor parte de las condiciones, mientras que con muestras de 500 sujetos, se encontraron resultados aceptables cuando la distribución de los ítems era simétrica o los factores estaban altamente correlacionados (ρ = .6). Por su parte, disponer de muestras iguales o mayores a 1000 sujetos implicó obtener resultados aceptables en todas las condiciones.

Finalmente, la influencia de los PE en el RMSE fue bajo, especialmente cuando las cargas factoriales de los ítems fueron medias (λ = .6). Para cargas factoriales más bajas, las dos versiones de estimación por LI lograron resultados un poco mejores que MLR. Por su parte, cuando se compara los procedimientos de estimación propios del AFI, se puede observar que DWLS y ULS obtuvieron casi los mismos resultados; sin embargo, el primero logró mejores resultados que el segundo, cuando los ítems eran más asimétricos y se simuló una correlación nula entre los factores.

Tabla 8. *RMSE de estimación de los parámetros lambda*

		λ = .3									λ = .6								
Modelo		AFI						TRI			AFI						TRI		
PE		DWLS			ULS			MLR			DWLS			ULS			MLR		
Tipo Asim.		I	II	III	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III
ρ	*N*																		
0	100	**0.18**	**0.19**	**0.14**	**0.18**	**0.19**	**0.21**	**0.21**	**0.22**	**0.23**	**0.10**	**0.11**	**0.14**	**0.10**	**0.11**	**0.14**	**0.10**	**0.12**	**0.14**
	200	**0.15**	**0.16**	0.09	**0.14**	**0.16**	**0.17**	**0.16**	**0.18**	**0.19**	0.07	0.08	0.09	0.07	0.08	0.09	0.07	0.08	0.09
	500	0.09	**0.11**	0.06	0.09	**0.11**	**0.13**	**0.10**	**0.11**	**0.14**	0.04	0.05	0.06	0.04	0.05	0.06	0.05	0.05	0.06
	1000	0.06	0.07	0.04	0.06	0.07	0.09	0.07	0.08	0.09	0.03	0.03	0.04	0.03	0.03	0.04	0.03	0.04	0.04
	2000	0.04	0.05	0.03	0.04	0.05	0.06	0.05	0.05	0.06	0.02	0.02	0.03	0.02	0.02	0.03	0.03	0.03	0.03
.3	100	**0.17**	**0.19**	**0.21**	**0.17**	**0.19**	**0.21**	**0.20**	**0.22**	**0.23**	**0.10**	**0.11**	**0.13**	**0.10**	**0.12**	**0.14**	**0.10**	**0.12**	**0.13**
	200	**0.14**	**0.15**	**0.17**	**0.14**	**0.15**	**0.17**	**0.15**	**0.16**	**0.19**	0.07	0.08	0.09	0.07	0.08	0.09	0.07	0.08	0.09
	500	0.09	**0.10**	**0.12**	0.09	**0.10**	**0.12**	0.09	**0.10**	**0.12**	0.04	0.05	0.06	0.04	0.05	0.06	0.05	0.05	0.06
	1000	0.06	0.07	0.08	0.06	0.07	0.08	0.06	0.07	0.09	0.03	0.03	0.04	0.03	0.04	0.04	0.03	0.04	0.04
	2000	0.04	0.05	0.06	0.04	0.05	0.06	0.05	0.05	0.06	0.02	0.02	0.03	0.02	0.03	0.03	0.02	0.03	0.03
.6	100	**0.16**	**0.18**	**0.20**	**0.16**	**0.18**	**0.20**	**0.18**	**0.20**	**0.22**	**0.10**	**0.11**	**0.13**	**0.10**	**0.11**	**0.13**	**0.10**	**0.11**	**0.13**
	200	**0.12**	**0.14**	**0.15**	**0.12**	**0.14**	**0.15**	**0.13**	**0.15**	**0.17**	0.07	0.08	0.09	0.07	0.08	0.09	0.07	0.08	0.09
	500	0.08	0.09	**0.10**	0.08	0.09	**0.10**	0.08	0.09	**0.11**	0.04	0.05	0.06	0.04	0.05	0.06	0.04	0.05	0.06
	1000	0.05	0.06	0.07	0.05	0.06	0.07	0.06	0.06	0.08	0.03	0.03	0.04	0.03	0.03	0.04	0.03	0.03	0.04
	2000	0.04	0.04	0.05	0.04	0.04	0.05	0.04	0.04	0.05	0.02	0.02	0.03	0.02	0.02	0.03	0.02	0.02	0.03

Nota: AFI = análisis factorial de ítems. TRI = teoría de respuesta al ítem. DWLS = estimación por mínimos cuadrados diagonalmente ponderados. ULS = estimación por mínimos cuadrados no ponderados. MLR = estimación por máxima verosimilitud con errores estándar robustos. PE = procedimiento de estimación. I = ítems con distribución simétrica. II = ítems con distribución medianamente asimétrica. III = ítems con distribución altamente asimétrica. λ = parámetros lambda. ρ = correlación entre los factores. *N* = tamaño de la muestra. En negrilla = resultado inaceptable, RMSE mayor o igual que 0.1.

Desviación Estándar de la Estimación de los Parámetros Lambda

Respecto a esta variable, la prueba ANOVA detectó efectos principales relevantes y significativos con las siguientes variables independientes: el tamaño del parámetro lambda poblacional ($F(2, 570) = 2452.54$; $p < .001$; $\eta^2_p = .83$), la asimetría de la distribución de los ítems ($F(2, 570) = 121.97$; $p < .001$; $\eta^2_p = .32$), y los tamaños de las muestras ($F(4, 570) = 1654.45$; $p < .001$; $\eta^2_p = .93$). Además, el grado de correlación entre los factores produjo un efecto significativo pero pequeño en la DEE ($F(2, 570) = 22.99$; $p < .001$; $\eta^2_p = .08$) al igual que los PE ($F(2, 570) = 9.51$; $p < .001$; $\eta^2_p = .04$). Esta última variable también tuvo efectos de interacción pequeños pero significativos con la magnitud del parámetro lambda poblacional ($F(2, 570) = 11.32$; $p < .001$; $\eta^2_p = .04$) y con el tamaño de las muestras ($F(8, 570) = 10.93$; $p < .001$; $\eta^2_p = .03$). La Tabla 9 presenta estos resultados desde una perspectiva descriptiva.

De manera similar a los resultados anteriores, cuando las escalas estuvieron compuestas de ítems de buena calidad ($\lambda = .6$) las estimaciones de lambda fueron notablemente estables, excepto cuando el tamaño muestral era igual a 100 sujetos, condiciones donde la DEE obtuvo valores inaceptables. En este escenario de lambdas medios, los tres PE se desempeñaron muy bien, obteniendo resultados casi iguales: estimaciones razonablemente estables en muestras de 200 sujetos y claramente precisas con 500 o más casos.

Una situación diferente se encontró cuando los lambdas poblacionales fueron bajos ($\lambda = .3$). En este caso, se detectó estimaciones altamente inestables aun con muestras de 500 sujetos, y la influencia de otras variables independientes se volvió más evidente: al aumentar la asimetría de los ítems y disminuir la correlación entre los factores, la heterogeneidad de las estimaciones tendió a incrementarse, aunque estos

efectos fueron menores a los producidos por la magnitud del parámetro lambda y el tamaño muestral.

Aun cuando los PE tuvieron desempeños similares, los resultados sugieren una ligera tendencia en favor de aquellos por LI. De esta forma, el MLR tendió a generar resultados que fueron levemente más inestables que las dos versiones asociadas al AFI, especialmente frente a ítems de baja calidad y tamaños de muestra pequeños. Por su parte, no se encontraron diferencias destacables entre los procedimientos DWLS y ULS.

DISCUSIÓN Y CONCLUSIONES

El objetivo de esta investigación fue evaluar el desempeño de los procedimientos de estimación asociados al AFI y la TRI en escenarios politómicos y multidimensionales, con el objeto de poner a prueba las ventajas teóricas de los procedimientos por FI frente a los por LI.

De este estudio se pueden extraer dos conclusiones principales: (a) los procedimientos de estimación de ambos tipos de modelos tienden a producir resultados altamente similares, por lo que no se pudo demostrar las ventajas de los procedimientos por FI; (b) las variables más críticas para lograr estimaciones precisas y estables, fueron la calidad de los ítems (i.e., la magnitud de los parámetros lambda) y el tamaño muestral, y no los procedimientos de estimación utilizados.

Tabla 9. *Desviación estándar de la estimación de los parámetros lambda*

		λ = .3									λ = .6								
Modelo		AFI						TRI			AFI						TRI		
PE		DWLS			ULS			MLR			DWLS			ULS			MLR		
Tipo asim		I	II	III	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III
ρ	N																		
0	100	**0.18**	**0.18**	**0.20**	**0.17**	**0.18**	**0.19**	**0.21**	**0.21**	**0.22**	**0.10**	**0.11**	**0.13**	**0.10**	**0.11**	**0.14**	**0.10**	**0.12**	**0.14**
	200	**0.14**	**0.16**	**0.17**	**0.14**	**0.16**	**0.17**	**0.16**	**0.18**	**0.19**	0.07	0.08	0.09	0.07	0.08	0.09	0.07	0.08	0.09
	500	0.09	**0.11**	**0.13**	0.09	**0.11**	**0.13**	**0.10**	**0.11**	**0.14**	0.04	0.05	0.06	0.04	0.05	0.06	0.04	0.05	0.06
	1000	0.06	0.07	0.09	0.06	0.07	0.09	0.07	0.07	0.09	0.03	0.03	0.04	0.03	0.03	0.04	0.03	0.03	0.04
	2000	0.04	0.05	0.06	0.04	0.05	0.06	0.05	0.05	0.06	0.02	0.02	0.03	0.02	0.02	0.03	0.02	0.02	0.03
.3	100	**0.17**	**0.18**	**0.19**	**0.17**	**0.19**	**0.20**	**0.19**	**0.22**	**0.22**	**0.10**	**0.11**	**0.13**	**0.10**	**0.12**	**0.14**	**0.10**	**0.12**	**0.13**
	200	**0.14**	**0.15**	**0.17**	**0.14**	**0.15**	**0.17**	**0.15**	**0.16**	**0.18**	0.07	0.08	0.09	0.07	0.08	0.09	0.07	0.08	0.09
	500	0.09	**0.10**	**0.12**	0.09	**0.11**	**0.12**	0.09	**0.10**	**0.12**	0.04	0.05	0.06	0.04	0.05	0.06	0.04	0.05	0.06
	1000	0.06	0.07	0.08	0.06	0.07	0.08	0.06	0.07	0.09	0.03	0.03	0.04	0.03	0.04	0.04	0.03	0.03	0.04
	2000	0.04	0.05	0.06	0.04	0.05	0.06	0.04	0.05	0.06	0.02	0.02	0.03	0.02	0.02	0.03	0.02	0.02	0.03
.6	100	**0.16**	**0.17**	**0.18**	**0.16**	**0.17**	**0.19**	**0.18**	**0.19**	**0.20**	**0.12**	**0.11**	**0.13**	**0.10**	**0.11**	**0.13**	**0.10**	**0.11**	**0.13**
	200	**0.12**	**0.14**	**0.15**	**0.12**	**0.13**	**0.15**	**0.13**	**0.15**	**0.17**	0.07	0.08	0.09	0.07	0.08	0.09	0.07	0.08	0.09
	500	0.08	0.09	**0.10**	0.08	0.09	**0.10**	0.08	0.09	**0.11**	0.04	0.05	0.06	0.04	0.05	0.06	0.04	0.05	0.05
	1000	0.05	0.06	0.07	0.05	0.06	0.07	0.06	0.06	0.08	0.03	0.03	0.04	0.03	0.03	0.04	0.03	0.03	0.04
	2000	0.04	0.04	0.05	0.04	0.04	0.05	0.04	0.04	0.05	0.02	0.02	0.03	0.02	0.02	0.03	0.02	0.02	0.03

Nota: AFI = análisis factorial de ítems. TRI = teoría de respuesta al ítem. DWLS = estimación por mínimos cuadrados diagonalmente ponderados. ULS = estimación por mínimos cuadrados no ponderados. MLR = estimación por máxima verosimilitud con errores estándar robustos. PE = procedimiento de estimación. I = ítems con distribución simétrica. II = ítems con distribución medianamente asimétrica. III = ítems con distribución altamente asimétrica. λ = parámetros lambda. ρ = correlación entre los factores. N = tamaño de la muestra. En negrilla = resultado inaceptable, SDE mayor o igual que 0.1.

En consideración al primer hallazgo, los resultados muestran que en la mayoría de las condiciones examinadas, los PE por FI e LI obtienen estimaciones muy similares, especialmente en los escenarios más favorables (i.e., cuando las muestras son grandes, los ítems son simétricos, y sus cargas factoriales son medias). En todo caso, se observó una ligera ventaja de las estimaciones por LI sobre MLR en las condiciones más rigurosas (i.e., muestras iguales o menores a 200 sujetos, ítems con distribuciones altamente asimétricas y que reflejan de manera más limitada el factor latente). Por otra parte, cuando se contrastaron los procedimientos de estimación asociados al AFI (i.e., ULS y DWLS), no se encontraron diferencias sustantivas entre ellos. Por lo tanto, las ventajas teóricas de los procedimientos de estimación asociados a la TRI sobre aquellos relacionados con el AFI no fueron confirmados para datos politómicos en las condiciones subóptimas estudiadas, aunque se debe señalar que el procedimiento MLR si mostró una cierta tendencia a lograr mayores tasas de soluciones convergentes.

En este contexto, se puede argumentar que la precisión relativa y la mayor facilidad de cómputo de los procedimientos por LI asociados al AFI, los convierten en una excelente alternativa para estimar los parámetros de escalas tipo Likert, sobre todo si se dispone de tests con muchas dimensionales y de gran longitud, pues emplear modelos multidimensionales de TRI (MTRI; Rekease 2009) puede requerir capacidades de procesamiento y tiempo superiores a los disponibles por los investigadores aplicados.

Estos resultados difieren parcialmente de aquellos reportados en estudios previos que comparan los procedimientos de estimación asociados a la TRI y al AFI en situaciones dicotómicas (e.g., Boulet, 1996; DeMars, 2012; Finch, 2010; Gosz y Walker, 2002; Muraki y Engelhard, 1985; Tate, 2003; VanderBerg, 1994). Algunas de estas diferencias son explicables por el diseño de investigación implementado, pues en este estudio se trabajó

con condiciones favorables al AFI, ya que: (a) se simularon con distribución normal, tanto los factores, como las variables subyacentes correspondientes a cada ítem observado; (b) se estimó un modelo politómico en que existe equivalencia entre la TRI y el AFI.

Sin embargo, otra parte de la discrepancia con la literatura previa podría explicarse por las diferentes exigencias que enfrentan los procedimientos de estimación cuando trabajan sobre datos dicotómicos o politómicos. Por ejemplo, cuando se dispone de datos politómicos, usualmente el número de parámetros a estimar es mucho mayor, lo que puede poner en tensión a procedimientos que utilizan información completa, requiriéndose una mayor muestra para lograr estimaciones estables. Por el contrario, tal como puede ser inferido del estudio de Forero y Maydeu-Olivares (2009), una variable crítica para la calidad de las estimaciones derivadas del AFI, es la asimetría de las distribuciones de respuesta de los ítems, habiéndose demostrado que las correlaciones tetracóricas son mucho más vulnerables a altos niveles de asimetría que las correlaciones policóricas (Timmerman y Lorenzo-Seva, 2011), a lo que se agrega que si los ítems realmente funcionan como politómicos (es decir, existe un mínimo de personas que marca cada alternativa de respuesta), este tipo de variables difícilmente alcanzarán los niveles de asimetría posibles en ítems dicotómicos.

Con respecto al segundo hallazgo de este estudio, los resultados muestran la relevancia de la magnitud de los parámetros lambda (i.e., la capacidad de los ítems de reflejar el constructo latente) y del tamaño de las muestras, para lograr estimaciones precisas de los parámetros de los ítems en modelos politómicos multidimensionales. De hecho, la magnitud del parámetro lambda fue la variable independiente con mayor efecto en la calidad de las estimaciones, de manera que en las condiciones en que se simuló ítems de relativa buena calidad ($\lambda = .6$), las estimaciones de parámetros fueron relativamente

insesgadas y precisas incluso con muestras de 200 sujetos, con virtual independencia del PE empleado y de las otras variables independientes. En contraste, cuando los ítems simulados reflejaban pobremente la variable latente ($\lambda = .3$), el tamaño muestral se volvió relevante para conseguir estimaciones precisas y estables. En esta situación, usualmente se requirió un tamaño muestral de 500 sujetos para lograr estimaciones de calidad razonable y de 1000 casos para evitar con total seguridad caer en errores de estimación inaceptables.

También vale la pena notar que las otras variables independientes, como la complejidad del modelo (i.e., el número de factores y su grado de correlación lineal) y la asimetría de los ítems, tuvieron poca capacidad para explicar la precisión de las estimaciones logradas y sólo adquirieron relevancia en algunas condiciones específicas. Para el caso de la asimetría de los ítems, ello posiblemente se explica porque no simulamos magnitudes muy altas de esta variable, dado el carácter politómico de los datos con que se trabajó.

Por otro lado, aunque existe evidencia previa de que la baja calidad de los ítems puede ser compensada utilizando un número superior de ellos -variable no manipulada en este estudio- (cf., Marsh et al., 1998), esto no debe ocultar el hecho de que la magnitud de los lambdas tuvo un gran impacto en los resultados, y que las consecuencias perjudiciales de disponer de ítems de baja calidad se hicieron más evidentes cuando se los combinó con pequeños tamaños de muestra, y/o altos niveles de asimetría de los ítems, situaciones posiblemente frecuentes en la práctica.

En base a los hallazgos presentados, se proponen tres recomendaciones para la investigación aplicada: Primero, los procedimientos de estimación por LI basados en el AFI parecen constituir una muy buena alternativa a la estimación de modelos MTRI politómicos, para la evaluación de la calidad de los ítems y el diseño de escalas tipo Likert

multidimensionales, sobre todo en presencia de tests largos y con muchos factores, por lo que se sugiere su empleo en esas condiciones.

Segundo, si se dispone de un tamaño muestral pequeño (pero no menor a 200 sujetos), y el investigador sospecha que los ítems de su escala no reflejan muy adecuadamente el constructo latente (i.e., sus lambdas posiblemente serán bajos), parece más prudente emplear el AFI en vez de la TRI para estimar el modelo porque, aunque en este escenario no está garantizado obtener un nivel óptimo de precisión en la estimación de los parámetros, aquellas obtenidas por LI posiblemente serán más estables y menos sesgadas que aquellas estimadas por FI. Por otra parte, si se dispone de un tamaño muestral menor a 200 sujetos, no recomendamos estimar un modelo politómico multidimensional, pues existe un elevado riesgo obtener estimaciones no convergentes o altos niveles de sesgo e inestabilidad en las estimaciones.

Por último, en términos más generales, se recomienda a los investigadores aplicados prestar especial atención, tanto a la calidad media de los ítems que componen las escalas que empleen, como a los tamaños de muestra que utilicen en sus investigaciones, pues esas dos variables parecen ser cruciales para obtener estimaciones precisas. Por lo mismo, el desarrollo de instrumentos de medida compuestos por ítems de calidad debiera ser una prioridad de la investigación social cuantitativa.

Aun cuando confiamos que estas recomendaciones serán útiles para los investigadores aplicados, debemos aclarar que, en rigor, los resultados de los que ellas se desprenden están restringidos a las condiciones examinadas en este estudio, por lo que aún se necesita más investigación para extenderlas a situaciones diferentes a las simuladas aquí. Por ejemplo, no es posible asegurar que las tendencias mostradas en esta investigación se mantengan sin cambios ante modelos multidimensionales más complejos, factores

distribuidos no normalmente, ítems con cargas cruzadas, mayor número de ítems por factor, diferentes niveles de asimetría en cada ítem o lambdas no homogéneos, por nombrar algunas posibilidades.

Por último, nuestra recomendación respecto de emplear procedimientos de estimación asociados al AFI, en vez de aquellos relacionados con la TRI, se limita a la estimación de modelos para los cuales exista equivalencia matemática entre ambos, lo cual incluye a los modelos de TRI más conocidos y utilizados en la investigación aplicada, dentro de los que están algunos de los que parecen más apropiados para analizar escalas tipo Likert, como son las distintas versiones del modelo de respuesta graduada (i.e., unidimensional, multidimensional, logístico y de ojiva normal).

CAPITULO 3

REVISITANDO LOS PROCEDIMIENTOS DE ESCALAMIENTO DE SUJETOS: PUNTUACIONES BRUTAS VERSUS ESTIMACIONES DE THETA

RESUMEN

La presente investigación compara la validez de los escalamientos de sujetos generados por los modelos clásicos de la teoría de respuesta al ítem (TRI) (i.e., modelos logísticos dicotómicos de uno, dos, y tres parámetros y el modelo politómico de respuesta graduada), con respecto a la suma no ponderada de puntuaciones o puntuaciones brutas (PB) de la teoría clásica de los test (TCT), por medio de dos estudios Monte Carlo. El primer estudio evaluó el grado de linealidad de la relación entre las puntuaciones verdaderas (PV) de la TCT y el rasgo latente, manipulando las discriminaciones y dificultades de los ítems y el nivel de inadecuación de los test. El segundo estudio evaluó la correlación entre las PB y las estimaciones theta de la TRI, así como su capacidad para reproducir el rasgo latente en un conjunto seleccionado de condiciones. Si bien la relación entre las PV y el rasgo latente se confirmó como no lineal, los resultados revelaron que el grado de no linealidad es muy variable, llegando a ser despreciable en algunas situaciones. Consecuentemente, se encontraron grandes correlaciones lineales entre las puntuaciones brutas (PB) y las estimaciones theta de la TRI, las que también evidenciaron una capacidad similar de reproducir el rasgo latente en muchas condiciones. Por el contrario, en situaciones óptimas, se observó una pequeña pero significativa ventaja de las estimaciones de la TRI por sobre las PB. Se discuten estos resultados intentando aclarar la contradictoria evidencia previa disponible en literatura psicométrica, con el fin de orientar a los investigadores aplicados interesados en escalar a sus sujetos de estudio.

INTRODUCCIÓN

Las ventajas de la teoría de respuesta al ítem (TRI) por sobre la teoría clásica de los test (TCT) para el diseño y desarrollo de instrumentos de medición han sido ampliamente difundidas en la literatura psicométrica (e.g., De Ayala, 2009; DeMars, 2010; Embretson y Reise, 2000; Hambleton, Swaminathan, y Rogers, 1991); sin embargo, su potencial superioridad con respecto al escalamiento de los sujetos aún no han sido suficientemente establecida (Ferrando y Chico, 2007; Xu y Stone, 2012).

De acuerdo a la teoría psicométrica, las puntuaciones de los sujetos estimadas con la TRI (en lo sucesivo $\hat{\theta}$) estarían linealmente relacionadas con el verdadero rasgo latente subyacente (i.e., θ), mientras que las PB tendrían una relación no lineal con θ (Embretson y Reise, 2000; Harwell y Gatti, 2001; Lord, 1980; Reise y Haviland, 2005). En consecuencia, las PB constituirían una representación distorsionada del rasgo latente, por lo que emplearlas podría tener importantes consecuencias negativas para la validez de los análisis que se pueda realizar con ellas (cf., Embretson, 1996; Kang y Waller, 2004; Morse, Johanson, y Griffeth, 2012). Sin embargo, la evidencia empírica en general revela la existencia de correlaciones lineales muy fuertes ($r \geq .85$) entre las $\hat{\theta}$ y las PB, tanto en datos simulados (MacDonald y Paunonen, 2002), como empíricos (Breithaupt, 2000; Fan, 1998; Ndalichako y Rogers, 1997; Progar, Socan, y Slovejija, 2008; Sukirno y Siengthai, 2010; Tomkowitcs y Rogers, 2005), así como también una capacidad similar de ambos métodos para correlacionar con otras variables (Ferrando y Chico, 2007; Xu y Stone, 2012).

A pesar de estas similitudes, algunos estudios han intentado determinar que variables permiten que se exprese en los escalamientos de sujetos la superior validez teórica

de $\hat{\theta}$ por sobre las PB (Embretson, 1996; Ferrando y Chico, 2007; Xu y Stone, 2012); sin embargo, hasta el momento, dichas variables no han sido claramente identificadas.

En este contexto de evidencia poco clara, la presente investigación evalúa las situaciones en que $\hat{\theta}$ podría ser un estimador más válido de las habilidades de los sujetos que las PB y aquellas donde ambos escalamientos debieran producir resultados similares. Para ello, se comparó las PB con las theta estimadas por los modelos más comunes de TRI -modelos logísticos dicotómicos de uno, dos, y tres parámetros y el modelo politómico de respuesta graduada (GRM)- por medio de dos estudios Monte Carlo. En el primer estudio, se evaluó la relación entre ambos procedimientos en situaciones donde el error de estimación está ausente, es decir, se estudió la relación entre las puntuaciones verdaderas (PV) con θ. En el segundo estudio, incluimos el error de estimación, comparando las PB con $\hat{\theta}$.

Definir las situaciones donde $\hat{\theta}$ constituye una mejor estimación del rasgo latente puede ser de gran interés para los investigadores aplicados, pues les permitiría conocer las ventajas que podrían o no obtener de escalar a los sujetos empleando modelos de TRI, comparado con opciones más simples como el cálculo de las PB.

LAS TEORÍAS DE LOS TEST Y EL ESCALAMIENTO

Los principales objetivos que se tiene al estimar las puntuaciones de los sujetos en las investigaciones aplicadas son los siguientes: (a) organizar o jerarquizar a los sujetos de acuerdo a su nivel en el rasgo latente medido, (b) tomar decisiones prácticas basadas en las puntuaciones de los sujetos (e.g., diagnosticar a las personas) y, (c) determinar la relación entre el rasgo latente y otras variables latentes u observadas. En consecuencia, es posible evaluar la calidad de los procedimientos de escalamiento examinando la validez de los

órdenes producidos (i.e., la capacidad de las estimaciones para reproducir el rasgo latente subyacente) y su relación con otras variables.

Dos procedimientos de escalamiento se han sugerido habitualmente desde las teorías de los test: el cálculo de las PB desde la TCT y la estimación de theta desde la TRI.

La TCT y las Puntuaciones Brutas

Habitualmente se piensa que la TCT es una teoría de los tests unificada, sin embargo, hoy en día es posible afirmar la existencia de dos versiones distintas de ella: la versión tradicional que se describe en los textos clásicos (Gulliksen, 1950/1987; Lord y Novick, 1968) y una versión más moderna que aproxima la TCT al análisis factorial para datos ordinales (Joreskög, 1971; Kohli, Koran, y Henn, 2014; Raykov y Marcoulides, 2015).

La TCT tradicional no constituye realmente un modelo –pues no es posible de falsar- (Raykov y Marcoulides, 2011), ni tiene la pretensión de modelar las respuestas de los sujetos a los ítems, focalizándose en la manera de faccionar la puntuación total obtenida por los sujetos en una aplicación única de un test multicomponente (llamada puntuación observada o X), en una parte sistemática (denominada puntuación verdadera o PV) y una parte aleatoria propia de cada aplicación del test (denominada error o E). Para ello, se define X como una combinación linear de las puntuaciones verdaderas más un error aleatorio con esperanza igual a cero (i.e., X = PV + E) y realizando deducciones a partir de esa formula, se establecen métodos para estimar la PV y determinar la fiabilidad de X. Adicionalmente, esta teoría de los tests incluye varios procedimientos que permiten estimar las propiedades de los ítems, como su discriminación o su contribución a la validez de X (Gulliksen, 1950/1987; Lord y Novick, 1968).

Con respecto a la forma de obtener X a partir de las respuestas de los sujetos, dentro de la TCT se sugieren varios procedimientos (Thissen y Wainer, 2001), algunos de los cuales ponderan el peso de los ítems por su contribución a la fiabilidad o a la validez de X, pero lo más habitual es emplear la simple suma de aciertos en el caso de ítems dicotómicos o la suma no ponderada de puntos en ítems politómicos cuyas respuestas han sido numeradas con enteros sucesivos (i.e., 1, 2, 3, 4, etc.) (Gulliksen, 1950/1987), con lo que, si bien se simplifica el cálculo, no se consideran las propiedades de los ítems para obtener X. En este capítulo emplearemos esta versión no ponderada de X, por lo que en adelante la denominaremos puntuación bruta o PB.

Finalmente, es interesante notar que la TCT tradicional no especifica las relaciones entre el rasgo latente y las PB, y que las PV no son iguales a dicho rasgo latente, pues ellas solo constituyen el resultado esperado de múltiples aplicaciones del test, pudiendo incluir distorsiones sistemáticas respecto de θ. De hecho, es posible demostrar que existe una relación no lineal entre esta definición de la PV y el rasgo latente (Lord, 1953b; Raykov y Marcoulides, 2011).

Por su parte, la nueva versión de la TCT deriva, tanto del denominado modelo congenérico (Jöreskog, 1971) que propone una forma de modelar las respuestas de los sujetos a los ítems agregando supuestos adicionales a la TCT, como del análisis factorial para variables discontinuas (Christoffersson, 1975), que asimila la PV al factor latente del modelo de factor común (Jöreskog, 1971; McDonald, 1999), con lo que la TCT se hace virtualmente indistinguible del análisis factorial de ítems (i.e., Raykov y Marcoulides, 2015).

Un estudio reciente (Kohli, Koran, y Henn, 2014) compara los escalamientos TRI y TCT en esta nueva versión factorial, mostrando la amplia equivalencia entre ambos

procedimientos, resultado que es compatible con la equivalencia matemática entre algunos modelos de TRI y el modelo de factor común para datos ordinales (Takane y Leeuw, 1987) y con la literatura que muestra la similitud entre los parámetros de los ítems estimados a partir de uno u otro tipo de modelo (e.g., De Mars, 2012; Forero y Maydeu-Olivares, 2009).

Debe notarse que, producto de la equivalencia entre el análisis factorial de ítems y la TRI, en esta nueva versión de la TCT las PV deberían tener una relación lineal con el rasgo latente, a diferencia de su formulación en la TCT tradicional.

Pese al interés de esta reciente versión de la teoría clásica de los tests, en este capítulo hemos optado por trabajar con su versión tradicional, pues: (a) aunque constituye una teoría de los tests simple y antigua, aún muchos tests se evalúan y escalan empleándola, lo que aumenta la relevancia práctica de nuestro estudio; (b) la versión tradicional de la TCT es la teoría de los tests que más habitualmente se ha empleado como patrón de comparación del escalamiento producido por la TRI, pese a lo cual, no se ha logrado establecer con claridad las variables que inciden en la equivalencia o diferencia entre ambos escalamientos; (c) no existe equivalencia entre los modelos de TRI y TCT tradicional, con lo que lo que asegura que estamos comparando los productos de dos modelos distintos y no solamente los procedimientos es estimación de dos modelos matemáticamente equivalentes, como ocurriría al comparar la nueva versión de la TCT con los modelos clásicos de TRI.

La TRI y la Estimación de Theta

La TRI constituye una teoría de los tests compleja, cuyo objetivo es superar las carencias de la TCT (Embretson y Reise, 2000; Hambleton et al., 1991) y que comprende diversos tipos de modelos. En este estudio, hemos escogido trabajar sólo con los modelos TRI más

populares (considerados clásicos debido a su temprana aparición), es decir, con los modelos logísticos dicotómicos unidimensionales de uno, dos, y tres parámetros (1P, Rasch, 1960/1980; 2P y 3P, Birnbaum, 1968) y el GRM (Samejima, 1969), que constituye una extensión politómica del modelo 2P.

Los modelos de TRI mencionados en el párrafo anterior asumen que todos los ítem del test reflejan un solo constructo latente unidimensional, el que se relaciona con la probabilidad de acertar los ítems dicotómicos (o con la probabilidad de responder en o sobre una determinada categoría en un ítem politómico) por una función sigmoidea. La función no linear que relaciona el rasgo latente y las probabilidades de endosar una categoría de respuesta es la consecuencia de la discontinuidad de respuestas, puesto que existe una relación esencialmente lineal entre θ y la respuesta continua subyacente (Ω) que habría sido obtenida si el sujeto tuviera la oportunidad de responder a ítems continuos (De Ayala, 2009).

Varios procedimientos basados en la teoría estadística frecuentista o Bayesiana permiten obtener $\hat{\theta}$ en el marco de la TRI. Todos ellos intentan obtener las $\hat{\theta}$ más probables dados los patrones de respuesta y los parámetros de los ítems (DeMars, 2010). Entonces, si el mecanismo subyacente que explica las respuestas a los ítems es equivalente al modelo TRI usado para los análisis, la esperanza de $\hat{\theta}$ será igual a θ (Hambleton y Swaminathan, 1985; Lord, 1980) y la relación entre $\hat{\theta}$ y θ será lineal.

Considerando que las PB y las PV están linealmente relacionadas y que lo mismo es cierto para la relación entre θ y $\hat{\theta}$, evaluar la relación entre las PB y θ es equivalente a estudiar la relación entre las PB y $\hat{\theta}$ en escenarios donde no existe error de estimación. Entonces, si definimos θ_j como el nivel de rasgo del sujeto *j* y PV_j como la puntuación

verdadera del sujeto *j*, la PV_j puede ser entendida como el valor esperado (*E*) de la suma de probabilidades de acierto para cada ítem *i*, condicional al nivel del rasgo de cada sujeto *j*. Por lo tanto, si las probabilidades de acierto de un ítem *i* están en función de un modelo de TRI, como podría ser el modelo 2P, entonces la PV_j puede ser definida como:

$$PV_j = E\left\{\sum_{i=1}^{n} P_i\left(x_i = 1 \mid \theta_j\right)\right\} = \sum_{i=1}^{n} \frac{1}{1 + e^{-Da_i(\theta_j - b_i)}} \qquad (3.1)$$

De la Ecuación 3.1 se desprende que toda vez que el modelo generador de las respuestas de los sujetos sea un modelo de TRI, las PV se podrán definir como la suma de las curvas características de los ítem del respectivo modelo de TRI, lo que también es conocido como la curva característica del test (CCT) (Lord, 1980).

La naturaleza no lineal de la CCT le permitió a Lord (1953b) argumentar que "the relation of true score to ability is in general curvilinear; consequently true score and ability are not identical concepts" [la relación de la puntuación verdadera con la habilidad es en general curvilínea; consecuentemente la puntuación verdadera y la habilidad no son conceptos idénticos] (p. 530). El grado de no linealidad entre ambos conceptos variará según la función CCT de cada test, por lo que en algunas condiciones "would be virtually a straight line except at its extremes" [sería virtualmente una línea recta excepto en sus extremos] (Lord, 1980, p. 49), mientras que en otras será más curvilínea. Sin embargo, "it may frequently happen that the curvilinearity encountered in actual practice is wholly negligible" [frecuentemente podría suceder que la curvilinealidad encontrada en la práctica real sea completamente insignificante] (Lord, 1953b, p. 529) cuando la dificultad de los ítems iguala la habilidad de los sujetos o cuando las correlaciones entre los ítem que componen el test no sean muy altas (Lord, 1953b)

Aún cuando la relación no lineal entre las PV y θ ha sido documentada en la literatura reciente (cf., Embretson y Reise, 2000; Reise y Haviland, 2005), las condiciones en que dicha no linealidad aumenta o disminuye aún están insuficientemente descritas, por lo que, excluyendo las breves referencias ya citadas de Lord, se desconoce en qué situaciones las PV constituyen una aproximación relativamente adecuada o distorsionada de θ. En el primer estudio, nuestro objetivo es llenar ese vacío.

ESTUDIO 1: FORMA Y RELACIÓN ENTRE LAS PUNTUACIONES VERDADERAS Y THETA

Este estudio evaluará el grado de linealidad entre las PV y θ de acuerdo a ciertas características de los ítems (i.e., discriminación y dificultad) y la longitud de los tests, con la meta de establecer la relación entre los procedimientos de escalamiento de la TRI y la TCT en ausencia de errores de estimación.

Método

Los datos fueron simulados en dos pasos utilizando el software R 2.15.2 (R Development Core Team, 2012). En primer lugar, se generó una secuencia de valores de θ desde -3 a 3 a intervalos de 0.001. Luego, la Ecuación (3.1), y su generalización trivial a otros modelos TRI, fueron usadas para calcular las PV para cada valor de θ de acuerdo a los modelos 1P, 2P, 3P, y GRM.

Para el modelo 1P, el parámetro de discriminación común fue ajustado a cuatro niveles (i.e., 1.0, 1.5, 2.0 y 3.0). Para los modelos de 2P, 3P y el GRM, los parámetros de

discriminación de cada ítem (i.e., a_j) fueron simulados a partir de distribuciones uniformes con distintos límites inferior y superior para representar ítems de calidad relativamente baja, intermedia, alta y excelente (i.e., Uniforme (0.5, 1.5), Uniforme (0.5, 2.5), Uniforme (1.0, 3.0) y Uniforme (2.5, 3.5), respectivamente). En consecuencia, la media de los parámetros a_j en las condiciones de los modelos de 2P, 3P y GRM fueron iguales al parámetro de discriminación común en el modelo de 1P equivalente.

Los parámetros de dificultad (i.e., b_i) fueron generados a partir de una distribución Normal (μ_b; σ_b) en donde μ_b fue simulado según cuatro valores diferentes para representar *test apropiados* ($\mu_b = 0$) y tres tipos de *tests inadecuados* (i.e., $\mu_b = 0$, 1.0, y 1.5), lo cual, de acuerdo a Embretson (1996), ocurre cuando la dificultad media de los ítems de un test es diferente de la habilidad promedio de los sujetos que lo responden (que en el caso de la presente simulación es cero). Adicionalmente, simulamos tres niveles de σ_b para crear test con parámetros de dificultad homogéneos y heterogéneos (i.e., σ_b = 0.1, 0.5, y 1, respectivamente). En los escenarios politómicos, los parámetros umbral (τ) fueron generados a partir de una distribución Normal (μ_t; σ_t) en donde μ_t toma los mismos valores que μ_b en los escenarios dicotómicos, mientras que σ_t fue ajustado dentro de tres niveles (i.e., σ_t = 1, 1.5, y 2) para producir umbrales más heterogéneos que los parámetros *b* en los escenarios dicotómicos, con el fin de evitar que umbrales muy cercanos entre ellos transformen los ítems politómicos en virtualmente dicotómicos.

Los parámetros de seudo azar (i.e., c_i) fueron simulados a partir de una distribución Uniforme (0, 0.25) para el modelo 3P, mientras que este parámetro fue definido como 0 para los modelos de 1P, 2P y GRM. Adicionalmente, la longitud de los tests fue ajustada a cuatro condiciones (i.e., 10, 25, 50, y 100 ítems), y 500 réplicas fueron generadas para cada condición.

Se empleó un gráfico de dispersión para representar la relación entre θ y las PV. El grado de linealidad de la relación entre ellas fue evaluado, tanto por medio del cálculo de la correlación de Pearson, como por la estimación de la distorsión de la linealidad (DL), entre ambas variables. La DL fue calculada a partir de la diagonal del gráfico de dispersión entre las PV y θ, y la CCT obtenida en cada réplica. Ya que la diagonal que divide el gráfico en dos partes iguales simultáneamente refleja una perfecta relación lineal entre θ y las PV, el área entre la CCT y la diagonal representa la DL. Por lo tanto, calculamos la DL como el porcentaje de área bajo o sobre la diagonal respecto de la mitad del área total del gráfico, a fin de garantizar que la DL se mantenga en valores entre cero, cuando la DL sea máxima, y uno, cuando la relación entre θ y las PV sea completamente lineal.

Resultados

La correlación de Pearson promedio entre θ y las PV fue alta en todos los modelos y todas las condiciones ($r \geq .92$). Sin embargo, se encontró que estas altas correlaciones enmascaraban grados importantes de no linealidad en algunas situaciones, pues la proporción de la DL alcanzó valores entre .1 y .4. En las Tablas 10 y 11 se presentan los resultados de las nueve condiciones que más claramente reflejan las variables que afectan la linealidad de la relación entre θ y las PV para los modelos dicotómicos y politómicos, respectivamente.

Tabla 10. *Linealidad entre Theta y las Puntuaciones Verdaderas para Modelos Dicotómicos*

Condiciones				Modelo	Modelo logístico de un parámetro				Modelo logístico de dos parámetros				Modelo logístico de tres parámetros			
	μ_a	μ_b	σ_b	*Ni*	10	25	50	100	10	25	50	100	10	25	50	100
				Índices												
A	**1.0**	0	0.5	*r*(θ,PV)	.993	.993	.993	.993	.992	.992	.992	.992	.992	.992	.992	.992
				DL	.119	.119	.119	.119	.122	.123	.122	.123	.123	.122	.122	.122
B	**1.5**	0	0.5	*r*(θ,PV)	.982	.983	.983	.983	.982	.983	.983	.983	.982	.983	.983	.983
				DL	.188	.189	.189	.189	.184	.183	.184	.183	.183	.183	.183	.183
C	**2.0**	0	0.5	*r*(θ,PV)	.972	.973	.973	.973	.974	.974	.975	.975	.973	.974	.974	.975
				DL	.237	.239	.239	.239	.227	.230	.229	.229	.229	.229	.230	.229
D	1.5	**0.5**	0.5	*r*(θ,PV)	.975	.976	.976	.977	.976	.977	.977	.977	.976	.977	.977	.977
				DL	.216	.217	.216	.215	.209	.208	.207	.207	.207	.208	.208	.207
E	1.5	**1.0**	0.5	*r*(θ,PV)	.955	.956	.957	.957	.958	.959	.959	.959	.958	.959	.959	.959
				DL	.296	.294	.293	.294	.280	.279	.279	.279	.278	.277	.279	.278
F	1.5	**1.5**	0.5	*r*(θ,PV)	.926	.925	.926	.927	.930	.931	.932	.932	.930	.931	.932	.932
				DL	.404	.409	.406	.406	.385	.384	.383	.384	.382	.383	.385	.383
G	3.0	0	**0.1**	*r*(θ,PV)	.943	.943	.943	.943	.943	.943	.943	.943	.943	.943	.943	.943
				DL	.344	.344	.344	.344	.343	.342	.343	.343	.342	.343	.342	.342
H	3.0	0	**0.5**	*r*(θ,PV)	.957	.958	.959	.959	.957	.958	.959	.960	.957	.959	.959	.959
				DL	.296	.295	.296	.295	.294	.294	.294	.294	.295	.294	.295	.294
I	3.0	0	**1.0**	*r*(θ,PV)	.975	.979	.980	.981	.975	.979	.981	.981	.975	.979	.980	.981
				DL	.201	.199	.199	.199	.200	.199	.197	.199	.201	.199	.200	.198

Nota: θ = rasgo latente. PV = puntuación verdadera. *ni* = numero de ítems. *r* = correlación de Pearson. DL = proporción de distorsión de la linealidad en la relación entre θ y las PV. μ_a = media de los parámetros *a*. μ_b = media de los parámetros *b*. σ_b = desviación estándar del parámetro *b*. Negrilla = variable modificada en cada una de las tres condiciones.

Tabla 11. *Linealidad entre Theta y las Puntuaciones Verdaderas para el modelo GRM*

Condiciones				Modelo	Modelo de Respuesta Graduada			
	μ_a	μ_τ	σ_τ	*Ni*	10	25	50	100
				Índices				
A	**1.0**	0	1.5	*r*(θ,PV)	.991	.991	.991	.991
				DL	.070	.072	.071	.073
B	**1.5**	0	1.5	*r*(θ,PV)	.983	.983	.983	.984
				DL	.093	.087	.089	.091
C	**2.0**	0	1.5	*r*(θ,PV)	.976	.976	.977	.976
				DL	.101	.103	.107	.107
D	1.5	**0.5**	1.5	*r*(θ,PV)	.980	.979	.980	.981
				DL	.116	.118	.122	.115
E	1.5	**1.0**	1.5	*r*(θ,PV)	.971	.971	.972	.972
				DL	.184	.184	.186	.184
F	1.5	**1.5**	1.5	*r*(θ,PV)	.957	.959	.960	.955
				DL	.275	.267	.258	.281
G	3.0	0	**1.0**	*r*(θ,PV)	.963	.964	.966	.964
				DL	.199	.198	.193	.201
H	3.0	0	**1.5**	*r*(θ,PV)	.965	.966	.964	.964
				DL	.118	.123	.124	.126
I	3.0	0	**2.0**	*r*(θ,PV)	.957	.958	.958	.958
				DL	.072	.073	.089	.079

Nota: θ = rasgo latente. PV = puntuación verdadera. *ni* = numero de ítems. *r* = correlación de Pearson. DL = proporción de distorsión de la linealidad en la relación entre θ y las PV. μ_a = media de los parámetros *a*. μ_b = media de los parámetros *b*. σ_b = desviación estándar del parámetro *b*. Negrilla = variable modificada en cada una de las tres condiciones.

En las citadas tablas podemos observar que las variables con mayor impacto en el grado de linealidad entre θ y las PV son: (a) el tamaño de las discriminaciones de los ítems (i.e., a mayor discriminación, aumenta la no linealidad); (b) la desviación estándar de las dificultades de los ítems (i.e., a mayor heterogeneidad, aumenta la no linealidad); y (c) el grado de inadecuación del test (i.e., cuando la dificultad promedio de los ítems se aproxima a la media de habilidad de los sujetos, disminuye la no linealidad). La longitud del test y el

modelo poblacional a partir del cual se generaron los datos, produjeron diferencias despreciables en el grado de linealidad; sin embargo, para el GRM se encontraron correlaciones lineales más fuertes y DL más pequeñas, como consecuencia de la heterogeneidad de los umbrales simulados en este modelo, comparados con la variabilidad de los parámetros *b* generados en los modelos dicotómicos.

Para entender la distorsión generada por las PV con respecto a θ, se presenta la Figura 7, la que representa la relación promedio observada en 500 réplicas entre las PV y el θ para un modelo 2P compuesto por 100 ítems. Se puede observar que la CCT (línea en negrita) difiere de la linealidad en diferentes grados. La relación entre las PV y θ tiende a acercarse a una línea recta cuando: las discriminaciones de los ítems son pequeñas (compare la Figura 7A, con las Figuras 7B y 7C), la media de la dificultad de los ítems es cercana a cero (compare la Figura 7D, con las Figuras 7E y 7F), y la heterogeneidad de los parámetros de dificultad se incrementa (compare la Figura 7I, con las Figuras 7G y 7H). Se puede notar también que, en todas las situaciones, la no linealidad se observa en las colas del rasgo latente (i.e., valores altos o bajos de θ) como una consecuencia de los efectos de techo y/o suelo de la métrica de las PV, en donde, bajo ciertas condiciones (como en las Figuras 7C y 7G), los sujetos con valores θ bajo -1 o mayores a 1, obtienen PV casi idénticas.

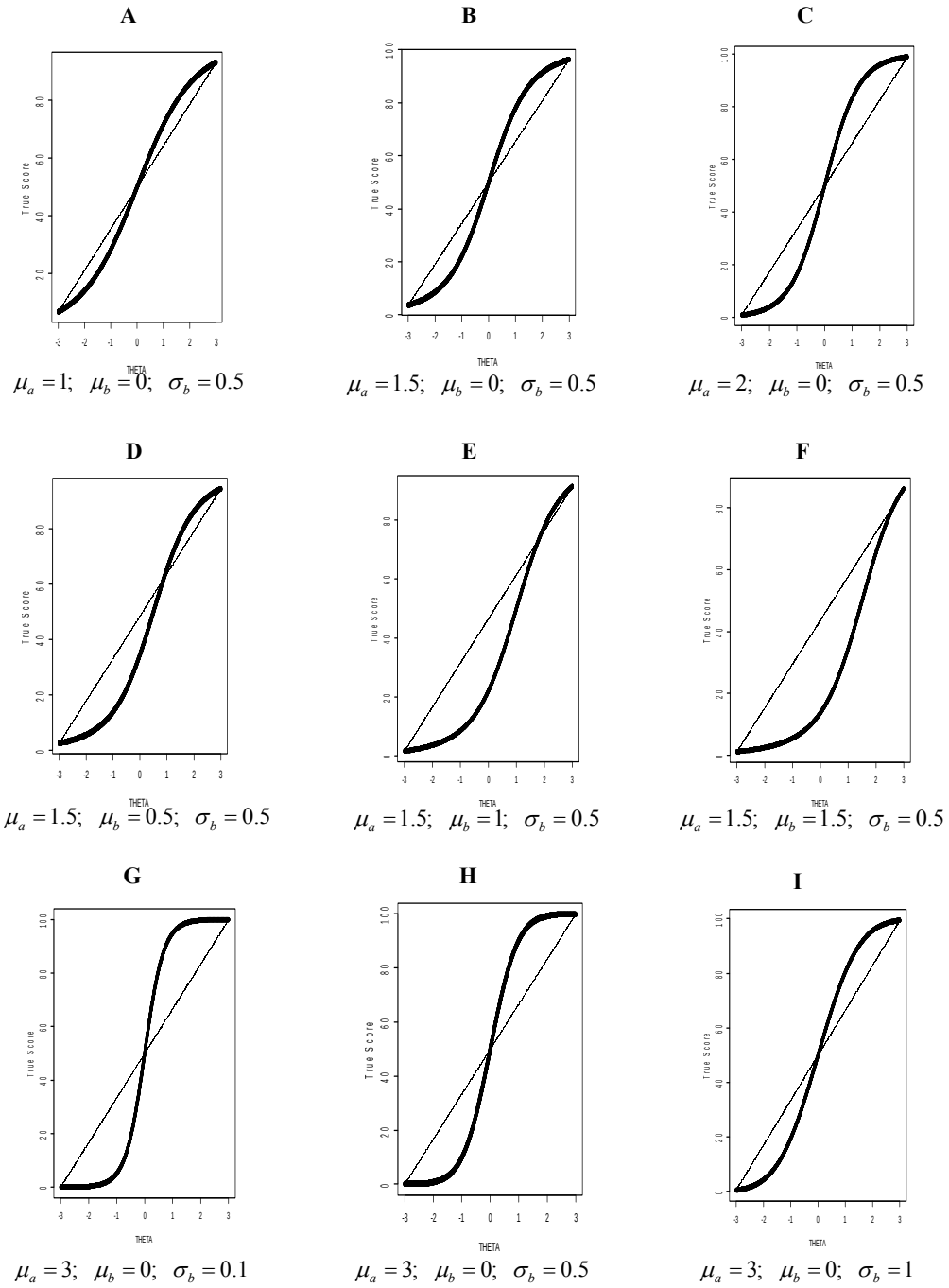

Figura 7. Curva característica del test para el modelo logístico de dos parámetros. Línea en negrilla = relación empírica entre la puntuación verdadera y el rasgo latente. Línea normal = relación lineal perfecta esperada si la puntuación verdadera fuese una perfecta representación del rasgo latente.

Por lo tanto, si la dificultad del test iguala las habilidades de los sujetos (i.e., la dificultad media de los ítems es igual a la media del rasgo latente) las PV serán una representación distorsionada del rasgo latente para los sujetos ubicados en las colas de la distribución de la habilidad. Si existe un desajuste entre la dificultad del test y las habilidades de los sujetos (i.e., los ítems son muy fáciles o muy difíciles para la población evaluada o, equivalentemente, la media de los parámetros *b* está muy por bajo o por sobre la media de θ), las PV no discriminarán entre sujetos con niveles moderadamente bajos o altos en la distribución de la habilidad. En contraste, las PV serán una aproximación razonable de θ para sujetos con un nivel de habilidad más cercano a la media de del rasgo, porque la relación entre las PV y θ es más lineal en el segmento intermedio del continuo de habilidad.

Aún cuando algunas de las condiciones presentadas aquí son improbables de encontrar en la investigación aplicada (e.g., test compuestos por ítems con discriminaciones muy altas y homogéneas), estas condiciones fueron seleccionadas para lograr claridad respecto de las variables que afectan la forma de la relación entre las PV y θ. En consecuencia, dada la evidencia presentada, se espera que la relación empírica entre ambas sea más cercana a la linealidad en la mayoría de las situaciones aplicadas. Para ilustrar lo anterior, la Figura 8 representa la relación esperada promedio (en 500 réplicas) entre las PV y θ en test dicotómicos compuestos por 40 ítems con características más frecuentemente encontradas en investigación aplicada. La Figura 8A representa tests cuyos parámetros han sido definidos a partir de distribuciones empleadas para representar *test habituales* en estudios simulados (cf., Tay y Drasgow, 2012), en donde los parámetros de discriminación fueron generados a partir de una distribución Log-Normal (0, 0.5) y los parámetros de dificultad fueron simulados a partir de una distribución normal estándar. La Figura 8B

representa tests con discriminaciones de los ítems simulados a partir de una distribución Log-Normal (0, 0.34) y dificultades generadas a partir de una distribución Normal (0.4, 0.65), lo que describe las características de los ítems presentados en el reporte técnico del TIMMS 2007 (Olson, Martin, y Mullis, 2008) basados en más de 1200 calibraciones de ítems con el modelo 2P.

Como se observa en la Figura 8, la relación entre las PV y θ es muy cercana a la linealidad en ambas simulaciones, y esto se confirma por una muy alta correlación de Pearson promedio (r = .99 en ambas simulaciones) y una DL baja (DL = .11 en la Figura 8A y .14 en la Figura 8B).

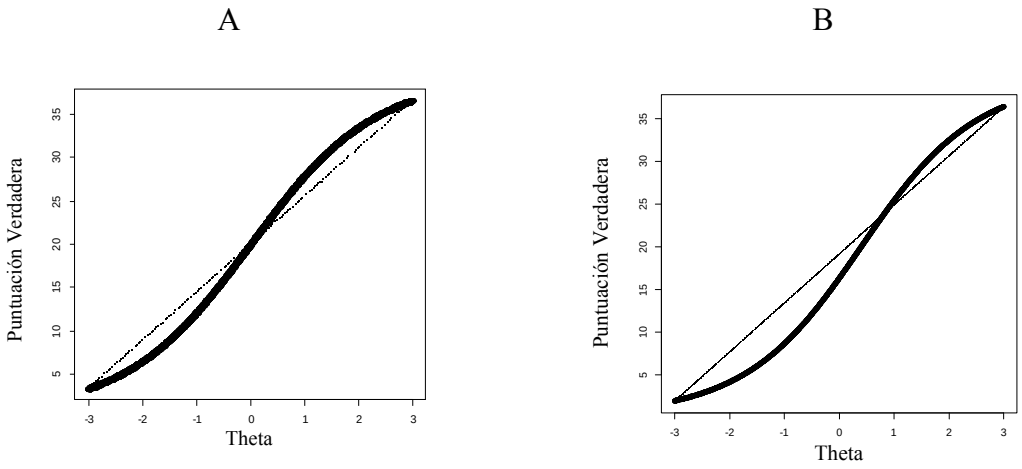

Figura 8. Curva característica del test para el modelo logístico de dos parámetros en test compuestos por 40 ítems dicotómicos con características generalmente encontradas en la investigación aplicada.
A = tests con discriminaciones de los ítems simuladas desde una distribución Log-Normal (0, 0.5) y dificultades de los ítems generadas desde una distribución normal estándar. B = tests con discriminaciones de los ítems simuladas desde una distribución Log-Normal (0, 0.34) y dificultades de los ítems generadas desde una distribución Normal (0.4, 0.65). Línea en negrilla = relación empírica entre la puntuación verdadera y el rasgo latente. Línea normal = relación lineal perfecta esperada si la puntuación verdadera fuese una perfecta representación del rasgo latente.

Conclusiones

A pesar de que la correlación lineal entre las PV y θ fue muy alta en la mayoría de los casos, el primer estudio reveló que esta correlación podría enmascarar niveles significativos de DL en las colas de la distribución del rasgo. Las variables que afectan la linealidad de la relación entre las PV y θ son la magnitud de las discriminaciones de los ítems, y la heterogeneidad y el grado de centralidad de los parámetros de dificultad de los ítems. En consecuencia, la no linealidad entre las PV y θ será mayor para los test compuestos por ítems con discriminaciones altas, valores homogéneos de parámetros *b*, y para los test que son difíciles o fáciles con respecto a la población medida.

ESTUDIO 2: RELACIÓN ENTRE LAS PUNTUACIONES BRUTAS Y THETA ESTIMADO

En el primer estudio, examinamos la relación entre las PV y θ para ilustrar la relación entre las puntuaciones brutas y las estimaciones de theta sin las distorsiones resultantes del error de estimación. Sin embargo, ya que la investigación empírica está enfocada en la estimación de las PB o $\hat{\theta}$, es importante indagar en la relación entre éstas últimas, así como en su validez relativa (i.e., su capacidad para reproducir en forma precisa el rasgo latente θ).

La mayoría de las investigaciones que han profundizado en la relación entre las PB y $\hat{\theta}$ han sido realizadas sobre datos reales. La evidencia encontrada revela habitualmente fuertes correlaciones lineales (i.e., $r \geq .9$) entre ambas estimaciones (Breithaupt, 2000; Fan, 1998; Ndalichako y Rogers, 1997; Progar et al., 2008; Sukirno y Siengthai, 2010;

Tomkowitcz y Rogers, 2005), y sólo unas pocas investigaciones (e.g., Dumenci y Achenbach, 2008) reportan una correlación más baja.

Por su parte, pese a que se le ha dado menos atención a la evaluación de la relación entre las PB y $\hat{\theta}$ en datos simulados, es posible encontrar algunas investigaciones en este campo. Por ejemplo, en concordancia con la mayor parte de los estudios empíricos, MacDonald y Paunonen (2005) realizan un estudio Monte Carlo en que reportan altas correlaciones lineales entre las PB y $\hat{\theta}$, y muestran que dichas correlaciones aumentan cuando la heterogeneidad de los parámetros de dificultad también lo hace. Por otro lado, a pesar de las expectativas relacionadas con la superioridad de las estimaciones provenientes de la TRI, los procedimientos de escalamiento de la TCT y la TRI mostraron aceptables y similares niveles de validez, aunque esta última fue levemente superior para $\hat{\theta}$ que para las PB. Por otra parte, Xu y Stone (2012) informaron que cuando los datos se generan con el modelo GRM, en la mayoría de los casos $\hat{\theta}$ y las PB presentan una capacidad similar para reproducir un resultado, aunque en ciertas condiciones (i.e., test cortos y muestras de tamaño pequeño) las PB parecen ser más adecuadas que $\hat{\theta}$. Además, se ha informado que cuando existe inadecuación del test, las PB incrementan en mucho mayor medida que $\hat{\theta}$ el error Tipo I en la detección de efectos de interacción en pruebas de ANOVA y en análisis de regresión múltiple moderados (cf., Embretson, 1996; Kang y Waller, 2005; Morse et al., 2012), sin embargo, otros autores han informado de que en esos mismos escenarios las PB son estimadores más fiables que $\hat{\theta}$ (Culppeper, 2013).

Hipotetizamos que estos resultados contradictorios podrían ser la consecuencia de las diferentes capacidades de los procedimientos de escalamiento para reproducir (o distorsionar) la distribución del rasgo latente, dependiendo en las condiciones estudiadas.

En consecuencia, se ha realizado este segundo estudio a fin de distinguir en que situaciones las PB y $\hat{\theta}$ constituyen representaciones aceptables y equivalentes de θ, y en cuales podríamos esperar encontrar diferencias significativas entre ellas. Esto permitirá a los investigadores aplicados evaluar en cada escenario concreto las consecuencias de utilizar un procedimiento simple (como las PB) en vez de uno más complejo (como $\hat{\theta}$), para escalar a los sujetos y las ganancias o pérdidas potenciales que obtendrían en uno u otro caso.

Método

En consonancia con la investigación de MacDonald y Paunonen (2002), en este estudio se evaluó la relación entre las PB y $\hat{\theta}$, así como su capacidad de reproducir el rasgo latente simulado. Como variables de manipulación se seleccionó aquellas que mostraron mayor impacto en la relación entre ambos escalamientos en el primer estudio, aunque los valores específicos generados en esta oportunidad fueron modificados para hacerlos más similares a los encontrados en las investigaciones aplicadas. Adicionalmente, dada la poca influencia que tuvo en el primer estudio el modelo específicamente simulado, en esta oportunidad solo se ha trabajado con dos modelos dicotómicos (i.e., 1P y 2P).

Para el modelo 1P, la discriminación común fue ajustada a tres niveles (i.e., 0.625, 1.5 y 2.25). Para el modelo 2P, los parámetros de discriminación fueron generados a partir de distribuciones pensadas para representar tests compuestos por ítems con discriminaciones bajas, heterogéneas y altas (i.e., Uniforme (0.5, 0.75), Uniforme (0.5, 2.5) y Uniforme (2.0, 2.5), respectivamente). Los parámetros de dificultad fueron simulados de acuerdo a una distribución Normal (μ_b, σ_b) con dos medias poblacionales diferentes ($\mu_b = 0$ y $\mu_b = 1$) para representar test apropiados e inapropiados, respectivamente (dado que los

valores de θ fueron generados a partir de una distribución normal estándar), y dos grados de heterogeneidad ($\sigma_b = 0.5$ y $\sigma_b = 1$).

Las bases de datos fueron generadas para muestras de 1000 sujetos y 30 ítems, así como para muestras de 300 sujetos y 15 ítems, con el objetivo de representar condiciones *óptimas* y *mínimas* para estimar modelos TRI. Se simularon quinientas réplicas en cada condición utilizando el software R 2.15.2 (R Development Core Team, 2012). Las estimaciones fueron realizadas con el paquete LTM (Rizopoulus, 2006). En conformidad con el estudio de Kang y Waller (2005) empleamos la estimación por máxima probabilidad marginal (MML, por sus siglas en inglés) para obtener los parámetros de los ítems y la estimación esperada a posteriori (EAP, por sus siglas en inglés) para los parámetros de los sujetos. Las réplicas no-convergentes o inapropiadas (i.e., réplicas convergentes que producen valores de parámetros diez veces mayores al valor poblacional) fueron descartadas y vueltas a simular hasta alcanzar las 500 réplicas válidas por condición. Es importante notar que la proporción de réplicas no-convergentes e inapropiadas fueron menos que el 10% de las originales en todas las condiciones, lo que es considerado aceptable (Flora y Curran, 2004).

Se empleó la correlación de Pearson promedio por condición para evaluar, tanto la relación entre las PB y $\hat{\theta}$, así como la relación de ambas con el θ simulado.

Resultados

La correlación entre las PB y $\hat{\theta}$ fue siempre mayor que las correlaciones entre ambas estimaciones y el rasgo latente subyacente (ver Tabla 12), revelando que las similitudes entre los procedimientos de escalamiento examinados, son mayores que su capacidad de reproducir θ.

Tabla 12. *Correlación Lineal entre Theta estimado de la TRI, la Puntuación Bruta y el Rasgo Latente*

Modelo			Modelo logístico de un parámetro						Modelo logístico de dos parámetros					
Ni / *ns*			15 / 300			30 / 1000			15 / 300			30 / 1000		
μ_b	σ_b	*a* / *r*	0.625	1.5	2.25	0.625	1.5	2.25	U(0.5, 0.75)	U(0.5, 2.5)	U(2.0, 2.5)	U(0.5, 0.75)	U(0.5, 2.5)	U(2.0, 2.5)
0	0.5	$r(\hat{\theta}, PB)$	.999	.995	.988	.999	.991	.983	.978	.984	.987	.995	.986	.983
		$r(\theta, \hat{\theta})$	.753	.914	.936	.850	.951	.961	.738	.909	.935	.849	.949	.961
		$r(\theta, PB)$	.753	.909	.926	.850	.944	.948	.751	.898	.926	.850	.938	.948
		ΔVE(%)	0.0	0.9	1.9	0.0	1.3	2.5	-1.9	2.0	1.7	-0.2	2.1	2.5
0	1.0	$r(\hat{\theta}, PB)$	.999	.997	.992	.999	.995	.989	.975	.982	.991	.995	.987	.989
		$r(\theta, \hat{\theta})$	.747	.906	.936	.845	.945	.964	.728	.903	.935	.843	.948	.964
		$r(\theta, PB)$	.746	.903	.930	.845	.942	.957	.742	.891	.930	.843	.938	.957
		ΔVE (%)	0.1	0.5	1.1	0.0	0.6	1.3	-2.1	2.2	0.9	0.0	1.9	1.3
1.0	0.5	$r(\hat{\theta}, PB)$	.999	.984	.965	.998	.974	.951	.975	.973	.966	.993	.971	.955
		$r(\theta, \hat{\theta})$	.741	.882	.895	.840	.931	.932	.724	.874	.895	.838	.926	.932
		$r(\theta, PB)$	.740	.868	.867	.838	.907	.891	.736	.854	.867	.838	.900	.890
		ΔVE (%)	0.1	2.5	4.9	0.3	4.4	7.5	-1.8	3.5	4.9	0.0	4.7	7.7
1.0	1.0	$r(\hat{\theta}, PB)$	.999	.989	.979	.999	.984	.972	.970	.975	.979	.994	.978	.971
		$r(\theta, \hat{\theta})$	.737	.887	.913	.838	.936	.949	.716	.880	.913	.834	.933	.950
		$r(\theta, PB)$	.736	.878	.895	.837	.922	.924	.733	.863	.897	.835	.914	.926
		ΔVE (%)	0.1	1.6	3.3	0.2	2.6	4.7	-2.5	3.0	2.9	-0.2	3.5	4.5

Nota: *ni* = número de ítems. *ns* = tamaño de la muestra. u_b = media de la dificultad de los ítems. σ_b= desviación estándar de la dificultad de los ítems. *a* = parámetro de discriminación. *r* = Correlación de Pearson. θ = rasgo latente. PB = puntuación bruta. $\hat{\theta}$ = theta estimado por el modelo de teoría de respuesta al ítem. ΔVE (%) = porcentaje de incremento en la varianza explicada por θ cuando se emplea $\hat{\theta}$ en vez de las PB.

Se encontró un correlación promedio entre las PB y $\hat{\theta}$ muy alta ($r \geq .951$) en todas las condiciones, y esta relación se incrementó cuando las discriminaciones de los ítems fueron bajas, la heterogeneidad de las dificultades de los ítems fueron altas y cuando la dificultad promedio de los ítems igualó la media de las habilidades de los sujetos. En contraste, la correlación promedio entre cada estimador y el rasgo latente simulado varió notoriamente a través de las condiciones, con lo que la superioridad de cada procedimiento de escalamiento dependió de las situaciones examinadas. Las diferencias observadas en la validez de las PB y $\hat{\theta}$ tendió a disminuir cuando la discriminación de los ítems, el tamaño de la muestra o el número de los ítems decrecía, y cuando los tests estaban compuestos por ítems heterogéneos en sus dificultades o cuya dificultad promedio igualaba la habilidad media de la población.

En conjunto, $\hat{\theta}$ produjo resultados más válidos que las PB, evidenciando una leve pero no irrelevante mayor capacidad para reproducir θ. En condiciones óptimas para la utilización de modelos de TRI (i.e., test compuestos por 30 ítems y muestras de 1000 sujetos), $\hat{\theta}$ mejoró la estimación de los parámetros de sujetos entre 4.5% y 7.7% respecto de las PB cuando la discriminación de los ítems era alta (a = 2.25) y la media de los parámetros b era diferente de la media de θ. En contraste, las PB superaron a las estimaciones de la TRI sólo en algunas de las condiciones mínimas para la utilización de modelos TRI (i.e., test de diez ítems y muestras de 300 sujetos) y en que, además, se calibran modelos de 2P y se dispone de tests compuestos por ítems con bajas discriminaciones. Sin embargo, incluso en ese tipo de escenario, la superioridad de las PB frente a $\hat{\theta}$ fue pequeña (de 1.8% a 2.5%) y, en esas condiciones ambos estimadores tendieron a obtener pobres aproximaciones a θ, como se puede observar de sus bajas correlaciones con el rasgo simulado.

Para ejemplificar estos resultados, la Figura 9 representa la relación entre los procedimientos de escalamiento y el rasgo latente simulado a través de 500 réplicas para un subconjunto de condiciones contrastantes en donde la inadecuación del test está presente y se observaron las mayores diferencias en validez entre las PB y $\hat{\theta}$. El Panel I muestra que $\hat{\theta}$ y las PB están linealmente relacionadas con el rasgo latente cuando los parámetros de discriminación son bajos, la desviación estándar de los parámetros de dificultad es alta, y el número de ítems y el tamaño de la muestra es pequeña. Sin embargo, en este escenario ni $\hat{\theta}$ ni las PB son una representación precisa de θ y podrían no ser útiles para la investigación aplicada pues producen gran dispersión alrededor del verdadero rasgo latente.

En contraste, el Panel II muestra que cuando las discriminaciones de los ítems son altas, la desviación estándar de las dificultades de los ítems es pequeña, y el número de ítems y tamaño de la muestra son grandes, $\hat{\theta}$ es una mejor representación de θ porque su relación con el rasgo latente es más lineal comparado con las PB. Sin embargo, debido a que en el caso representado las dificultades de los ítems no coinciden con el nivel de rasgo promedio de la población, también se observa un efecto piso para $\hat{\theta}$, revelando una estimación menos precisa del rasgo para los sujetos con menores niveles de habilidad.

Por su parte, la relación no lineal entre las PB y θ mostrada en el Panel II es la consecuencia de efectos piso y techo, en que los sujetos con niveles de habilidad más bajos obtienen PB iguales o muy cercanas a cero, lo que resulta en una muy pequeña dispersión de puntuaciones. Una situación similar pero inversa se observa en los niveles de habilidad más altos. Esta precisión espuria de las PB en las colas de θ es mayor cuando la inadecuación del test se incrementa. En otras palabras, la no linealidad de las PB como consecuencia del efecto piso es más alta en cuanto el test se vuelve más difícil

para la población evaluada, y la no linealidad de las PB como consecuencia del efecto de techo es más alta cuando el test se vuelve más fácil para la población evaluada.

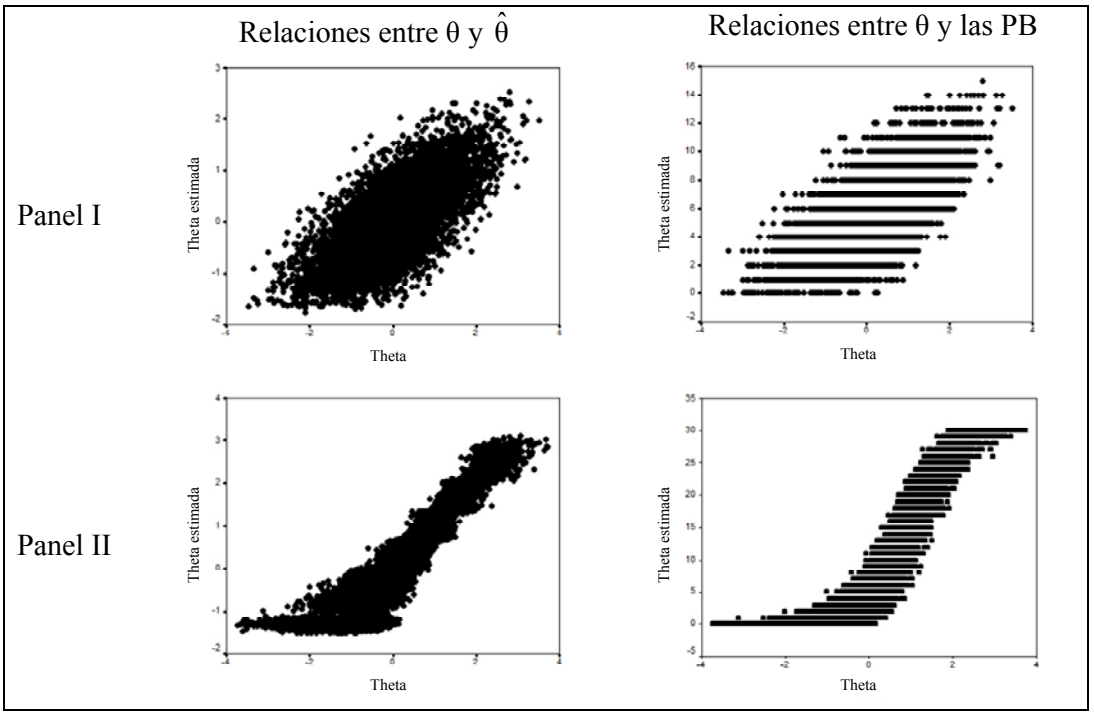

Figura 9. Gráficos de Dispersión para la relación entre el Rasgo Latente, las estimaciones de los sujetos de la TRI y las Puntuaciones Brutas de la TCT para algunas Condiciones Contrastantes a lo largo de 500 réplicas.
Panel I = tests compuestos por 15 ítems y 300 sujetos, con parámetros *a* simulados a partir de una distribución Uniforme (0.5, 0.75) y parámetros *b* simulados desde una distribución Normal (1, 1). Panel II = tests compuestos por 30 ítems y 1000 sujetos, con parámetros *a* simulados a partir de una distribución Uniforme (2, 2.5) y parámetros *b* simulados desde una distribución Normal (1, 0.5).

Conclusiones

En concordancia con el primer estudio, los resultados del segundo evidenciaron que $\hat{\theta}$ constituye una representación más precisa del verdadero rasgo latente que las PB, aunque su mayor validez es generalmente pequeña y no se observa en todas las situaciones. La superioridad de las estimaciones de la TRI por sobre las de la TCT son más prominentes cuando interactúan cinco elementos: muestras grandes, test de mayor

longitud, ítems con altas discriminaciones, tests más inadecuados e ítems de dificultad homogénea.

En contraste, cuando las discriminaciones de los ítems son más bajas y/o cuando la dificultad media de los ítems está bien equiparada al nivel promedio de habilidad de la población, las diferencias entre $\hat{\theta}$ y las PB tienden a desvanecerse e incluso pueden revertirse cuando las discriminaciones de los ítems son bajas y se emplean test cortos con muestras pequeñas. La mayor validez relativa de las PB en esas situaciones puede ser consecuencia, tanto de una relación más lineal entre θ y las PB en presencia de ítems con bajas discriminaciones (como fue mostrado en el primer estudio), como de la menor calidad de las estimaciones de theta en condiciones no óptimas para la utilización de modelos de TRI.

DISCUSIÓN Y CONCLUSIONES GENERALES

En esta investigación, se examinó la relación entre los procedimientos de la TRI y la TCT para estimar las puntuaciones de los sujetos, por medio de dos estudios Monte Carlo.

El primer estudio demostró que aún cuando las PV y θ están no linealmente relacionadas, el grado de alejamiento de la linealidad varía dramáticamente como consecuencia de las características de los ítems y del test. En suma, la relación entre las PV y θ fue más cercana a la linealidad cuando las discriminaciones de los ítems simulados fueron pequeñas, y cuando las dificultades medias de los ítems eran heterogéneas y se equiparaban a la habilidad promedio de la población objetivo. Por el contrario, el grado de no linealidad de la relación entre ambas variables fue mayor (especialmente para los sujetos con un nivel de habilidad más cercano a las colas de la

distribución del rasgo), cuando las discriminaciones de los ítems eran mayores y sus dificultades eran homogéneas o en promedio se diferenciaban de la habilidad media de la población.

El segundo estudio reveló que cuando el error de estimación está presente, la no linealidad entre ambos procedimientos de escalamiento tiende a desaparecer, especialmente para los test compuestos por ítems con parámetros de discriminación bajos o intermedios, resultando en estimaciones con una capacidad similar para reproducir el rasgo latente. En contraste, cuando se simuló ítems altamente discriminadores, las estimaciones de la TRI superaron a las PB de la TCT en la reproducción del rasgo latente, especialmente cuando se combinaban con disponer de muestras grandes, ítems con dificultades homogéneas y test largos con importantes grados de inadecuación. La ventaja de las estimaciones de la TRI fue demostrada por la mayor proporción de varianza del rasgo explicada por $\hat{\theta}$, aún cuando se debe notar que, incluso en el mejor caso, la ventaja de las estimaciones de la TRI alcanzó sólo un 7.7% más que la varianza explicada por las PV.

Estos resultados permiten interpretar la evidencia contradictoria disponible en la literatura psicométrica donde, por ejemplo, la superioridad teórica del escalamiento de la TRI contrasta con la similitud que ésta muestra con las PB según lo reportado en la investigación aplicada (e.g., Breithaupt, 2000; Fan, 1998; Progar et al., 2008). De acuerdo a los resultados presentados en este capítulo, dicha evidencia no es realmente contradictoria. Las altas correlaciones entre los procedimientos de escalamiento de la TCT y la TRI observadas en estudios aplicados podrían ser interpretadas como resultado de haber hecho la comparación en tests apropiados para la población evaluada (i.e., test que no son muy fáciles o muy difíciles para los sujetos evaluados), compuestos por ítems con discriminaciones relativamente bajas o medias. Por su parte, las

correlaciones relativamente menores reportadas por otros autores (e.g., Dumenci y Achenbach, 2008) podrían ser el resultado de tests inadecuados y compuestos por ítems más discriminadores.

En conjunto, los resultados presentados en este capítulo confirman la superioridad del procedimiento de escalamiento de la TRI en la mayor parte de las condiciones: efectivamente dicha estimación permite una representación más precisa del rasgo latente suavizando los efectos techo y el piso en las colas de la distribución del rasgo de las PB, especialmente cuando las discriminaciones de los ítems son altas. Ese suavizado fue más prominente cuando existía inadecuación del test y las dificultades de los ítems eran homogéneas.

Sin embargo, debe ser notado que la mayor validez del escalamiento de la TRI usualmente fue pequeña, llegando incluso a revertirse en condiciones sub-óptimas (i.e., bajas discriminaciones de los ítems, test cortos y muestras pequeñas). Este resultado adquiere más relevancia si consideramos que las condiciones que hemos denominado mínimas tienden a ser bastante habituales en la investigación psicométrica aplicada (e.g., Henson y Roberts, 2006), por lo que, dadas las prácticas actuales, emplear el escalamiento de la TRI no implicaría mejoramientos sustantivos en la validez de las puntuaciones asignadas a los sujetos en la mayoría de los estudios contemporáneos. Además, es posible suponer que la mayor parte de quienes construyen y aplican tests, intentan producir o emplear instrumentos con dificultades heterogéneas y apropiadas para sus poblaciones objetivo, condiciones en que la ganancia obtenida por el escalamiento TRI es menor.

Además, como consecuencia de las altas correlaciones entre ambos procedimientos de puntuación, la ganancia de emplear el escalamiento de la TRI podría ser insignificante para el análisis de las relaciones lineales entre el constructo medido y

otras variables, o para la estimación de las diferencias promedio entre grupos evaluados en el rasgo latente. Por el contrario, las ventajas del escalamiento TRI podrían ser más prominentes para los análisis no lineales (e.g., la evaluación de efectos de interacción), los que son altamente sensibles a las transformaciones no lineales de las variables (Davison y Sharma, 1990). En efecto, la relación no lineal observada en el estudio actual entre ambos escalamientos podría explicar el mayor error Tipo I que obtienen las PB frente a $\hat{\theta}$ en varios estudios que emplean ANOVA y regresión múltiple moderada para detectar efectos de interacción (e.g., Embretson, 1996; Kang y Waller, 2004; Morse et al., 2012).

Finalmente, es posible señalar que el procedimiento de escalamiento de la TRI: (a) maximiza los beneficios de disponer de tests de buena calidad (i.e., test compuestos por ítems con altas discriminaciones) y condiciones de estimación óptimas (i.e., test largos y muestras grandes); (b) permite disponer de un procedimiento de estimación más robusto a aplicar tests compuestos de ítems con dificultades homogéneas y no ajustadas a la aptitud media de los sujetos. Sin embargo, emplear la TRI no compensará las consecuencias negativas que emergen de tener instrumentos de baja calidad, test cortos o muestras pequeñas. En escenarios normales o sub-óptimos, el procedimiento de escalamiento de la TRI producirá resultados equivalentes (o incluso inferiores) a las PB de la TCT, como ha sido reportado en estudios previos (Ferrando y Chico, 2007; Xu y Stone, 2012). Esto implica que en esas condiciones, las PB pueden ser consideradas una aproximación un poco tosca, pero razonablemente adecuada, al rasgo latente, especialmente para escalar a sujetos de habilidades intermedias. Por otro lado, para obtener todo el potencial del escalamiento TRI se debe mejorar la calidad general de la medición (i.e., generando ítems más discriminadores, empleando muestras de mayor

tamaño o aplicando tests de mayor longitud), teniendo en cuenta que los modelos TRI no son herramientas milagrosas para el escalamiento de los sujetos.

Las conclusiones reseñadas en este documento tienen al menos cuatro limitaciones que deben ser consideradas. Primero, los resultados presentados respecto de la TRI se limitan a evaluar el desempeño de los procedimientos de estimación MML y EAP para los parámetros de los ítems y de los sujetos, respectivamente. Por lo tanto, aún cuando las estimaciones MML y EAP sean extensamente recomendadas y utilizadas en la práctica -y deberían generar productos resultados relativamente equivalentes a otros procedimientos-, los resultados presentados aquí no deberían de ser generalizados a cualquier procedimiento de estimación sin evidencia adicional.

Segundo, las conclusiones derivadas del presente estudio se limitan a la utilización de las PB según las hemos calculado aquí, es decir, basadas en una suma no ponderada de las respuestas a los ítems. Aún cuando este método de cálculo es el más común en la investigación aplicada, debe de ser notado que, en ciertas circunstancias, el grado de no linealidad de la relación entre las PB y el rasgo latente puede ser reducido mediante la transformación de las PB al valor *Z* correspondiente a la proporción de respuestas correctas implicadas en cada PB (cf., Fan, 1998). Por ello, las conclusiones de este capítulo no deberían de ser generalizadas a estas transformaciones sin adicionales análisis.

Tercero, en este capítulo hemos empleado la versión tradicional de la TCT, por lo que nuestros resultados no son extrapolables a quienes utilicen la versión más actual de esta teoría de los tests (Kohli, Koran, y Henn, 2014; Raykov y Marcoulides, 2015), línea de investigación mucho más emparentada con los estudios respecto de las relaciones entre la TRI y el análisis factorial para datos ordinales, que con nuestra

perspectiva y con la mayor parte de los estudios que comparan los escalamientos TCT y TRI.

Por último, la presente investigación ha sido realizada con un tipo de modelos de TRI (i.e., modelos de *dominancia*) que asumen que las funciones de repuesta de los ítems aumentan monótonamente cuando el rasgo latente se incrementa (Drasgow, Chernyshenko, y Stark, 2010). Por lo tanto, las conclusiones de este estudio no deberían de ser generalizadas a otros tipos de modelos de TRI, como los modelos de *punto ideal* (cf., Drasgow et al., 2010; Roberts, Donoghue, y Laughlin, 2000), que no suponen una relación monotónica entre el rasgo latente y la probabilidad de acierto y, por tanto, donde los procedimientos de escalamiento de la TCT y la TRI pueden producir resultados mucho más divergentes.

A pesar de las limitaciones descritas arriba, se espera que la presente investigación haya contribuido a establecer las ganancias específicas que se puede obtener al emplear los procedimientos de escalamiento de los modelos clásicos de TRI y la utilidad de las PB como una aproximación razonable al rasgo latente bajo algunas condiciones.

CAPITULO 4

EL CANTO DE LAS SIRENAS EN PSICOMETRÍA: UNA REVISIÓN CRÍTICA DE LA PROPIEDAD DE INVARIANZA DE LOS MODELOS DE TRI

RESUMEN

La idea de que los modelos de teoría de respuesta al ítem (TRI) producen estimaciones de parámetros invariantes es ampliamente aceptada entre los investigadores de ciencias sociales y del comportamiento y se considera un paso adelante para lograr mediciones realmente científicas. A partir de la definición conceptual y matemática de invarianza métrica y empleando datos simulados, en este trabajo se presenta un análisis crítico de los estudios empíricos y teóricos que respaldan la propiedad de invarianza en la TRI. Como resultado del estudio, y para aclarar el sentido y límites de la invarianza en dicha teoría de los tests, se proponen los conceptos de *invarianza interna* e *invarianza externa*, planteando que el ámbito de la invarianza como propiedad de la TRI se circunscribe estrictamente al primero. Finalmente, se discuten las consecuencias de ser seducido por los *cantos de sirenas* que proclaman haber conseguido mediciones totalmente invariantes en las ciencias sociales y del comportamiento.

INTRODUCCIÓN

En su conocido libro Contra el Método, Feyerabend (1974) desarrolla una tesis respecto de la fuerza de atracción que las creencias y posturas teóricas previas ejercen en los científicos y argumenta sobre como esa fuerza disminuye la capacidad de éstos para percibir los límites que tienen sus pruebas empíricas. Feyerabend ejemplifica su teoría describiendo la forma en que Galileo defendió el sistema heliocéntrico utilizando argumentos poco apegados a los datos astronómicos que realmente poseía, mostrando como la presencia de creencias previas muy definidas puede hacer que los científicos sean propensos a tratar de convencer a los demás (y a ellos mismos) utilizando estrategias discursivas, en lugar de pruebas concretas.

El presente documento argumenta que algo similar ha ocurrido en el campo de psicometría, donde una parte importante de la comunidad psicométrica, especialmente aquella que se ha dedicado a la difusión de la teoría de respuesta al ítem (TRI), ha sido impulsada a promover la idea de que los modelos TRI poseen la propiedad de lograr mediciones invariantes, a pesar de que dicha propiedad está débilmente definida, limitadamente demostrada y su significado real para la investigación aplicada parece bastante más restringido que lo dado a entender en una parte de la literatura.

Para apoyar esta tesis, proporcionaremos una definición conceptual y matemática de la propiedad de invarianza y examinamos críticamente la evidencia teórica y empírica que fundamenta el logro de estimaciones invariantes en la TRI. Adicionalmente, introduciremos la distinción entre invarianza interna e invarianza externa para aclarar el significado de esta propiedad en la TRI. Finalmente, discutiremos algunas de las consecuencias negativas de ser seducido por el *canto de las sirenas* de

quienes proclaman haber conseguido mediciones invariantes en las ciencias sociales y del comportamiento.

LA INVARIANZA EN LA PSICOMETRÍA: IMPERATIVO Y DEFINICIÓN

En el campo de las matemáticas y de la física, se entiende a la invarianza como una propiedad de algunos sistemas reales o formales consistente en que las relaciones entre los elementos del sistema no se modifican frente a determinadas transformaciones. Entonces, si entendemos como un sistema al conjunto de mediciones realizadas con un único instrumento sobre varios objetos, se podría definir como invarianza de la medición, o invarianza métrica, a la obtención las mismas relaciones entre las medidas cuando se utiliza un segundo instrumento para evaluar al mismo conjunto de objetos. En algunos campos de la ciencia, la obtención de invarianza en las mediciones puede ser trivial pues medir un objeto utilizando instrumentos igualmente válidos y confiables (e.g., termómetros basados en diferentes principios) produce usualmente resultados equivalentes. Por desgracia, la invarianza métrica no está garantizada en las ciencias sociales y del comportamiento, donde con frecuencia se encuentran diferencias en las mediciones obtenidas cuando se utilizan dos instrumentos válidos y confiables para medir el mismo constructo, o el mismo instrumento presenta propiedades notoriamente diferentes cuando es aplicado en distintas muestras.

La anterior dificultad no ha impedido que la invarianza métrica haya sido considerada en psicometría como un atributo esencial de una verdadera medición científica (Jones, 1960), por lo que su logro es estimado como una “matter of life and death to the science of mental measurement” [materia de vida o muerte para la ciencia de la medición mental] (Write, 1968, p. 1). En consecuencia, es natural que exista un

cierto consenso en torno a considerar la falta de invarianza como una situación muy insatisfactoria para una ciencia que pretende ser científica (De Ayala, 2009; Embretson, 1999; Hambleton, Swaminathan, y Rogers, 1991; Wright, 1968), y que se haya destinado significativos esfuerzos a definir las situaciones en que es posible suponer la existencia de invarianza métrica en las ciencias sociales y del comportamiento.

En el campo de psicometría, la invarianza métrica ha sido definida generalmente –siguiendo a Meredith (1993) y Millsap (2008)- como la equivalencia entre la probabilidad de acertar un ítem X_i por un sujeto *j* perteneciente a una población Q_k dado su nivel de una variable latente θ_j, y la probabilidad de acertar el mismo ítem dado únicamente θ_j, lo que se expresa formalmente en la siguiente igualdad:

$$P(X_i \mid \theta_j, Q_k) = P(X_i \mid \theta_j) \qquad (4.1)$$

Por lo tanto, si se obtienen iguales estimaciones de parámetros de los ítems (dentro de los niveles de error de estimación) al aplicar el mismo test a distintos grupos de sujetos, independientemente de la población Q_k a la que ellos pertenezcan, se dirá que el test y los ítems que lo componen obtienen mediciones invariantes.

Dado que los sujetos pueden pertenecer a un numeroso conjunto de poblaciones Q_k, las que pueden tener distintas características distintivas C_l, la condición necesaria y suficiente para lograr mediciones invariantes consiste en que ninguna de las C_l de las distintas poblaciones estudiadas estén asociadas con la probabilidad de acertar los ítems dado θ (McDonald, 1982). Lo anterior es otra forma de decir que la probabilidad de acertar los ítems de un test debe ser exclusivamente función del rasgo latente y que, si todos los ítems del test son invariables, todos los sujetos con el mismo nivel de habilidad (i.e., con el mismo valor en el rasgo latente), tendrán la misma estimación de resultados ($\hat{\theta}$) -dentro de los niveles del error de estimación- independientemente de la población a la que pertenezcan. Por lo tanto, la invarianza es una propiedad condicional,

que sólo es relevante en el contexto de la comparación entre múltiples poblaciones (Rupp y Zumbo, 2006), o al menos cuando se pueda suponer la existencia de dos poblaciones con diferentes características C_l que podrían interactuar con las probabilidades de acertar los ítem del test dado θ.

Teniendo esto en cuenta, es posible preguntarse si es posible lograr mediciones invariantes en el contexto de las ciencias sociales y de la conducta y, en caso afirmativo, indagar que tipo de herramientas metodológicas y estadísticas serían necesarias para garantizar esos resultados invariables.

LA INVARIANZA EN LA TRI: PROPIEDAD INTRÍNSECA

Desde los primeros desarrollos de la psicometría, varios expertos invirtieron importantes esfuerzos en elaborar procedimientos metodológicos que generaran mediciones invariantes. Por ejemplo, Throndike (1922), propuso emplear las "transmuting scores" para comparar las puntuaciones entre distintos grupos, mientras Thurstone (1927) desarrolló el "absolute scaling" y Goodman (1950) el "scalogram analysis" (cf., Engelhard, 1984, 2008). A mediados del siglo XX, la búsqueda de invarianza en psicometría se trasladó a la TRI. Por ejemplo, Lord (1952) argumentó que era posible lograr invarianza métrica en modelos de variables latentes al escribir: "it is nevertheless possible under certain conditions to define a metric for the ability such that the frequency distribution of ability in the group tested will remain the same even though the composition of the test is changed" [es sin embargo posible definir, bajo ciertas condiciones, una métrica para la habilidad tal que la distribución de frecuencia de la habilidad en el grupo testeado seguirá siendo la misma, aunque la composición del test cambie] (pp. 1-2).

En los modelos TRI, la invarianza métrica se produce cuando los ítems presentan las mismas curvas características del ítem (CCI) en distintos grupos de sujetos o, lo que es equivalente, los ítems obtienen las mismas estimaciones de sus parámetros en los diversos grupos en que se evalúan (Embretson y Reise, 2000). En consecuencia, la invarianza métrica debería un tema de investigación empírica, pues es una situación que podría o no ocurrir entre dos poblaciones, y por tanto, tendría que ser abordada a través de estudios del funcionamiento diferencial de los ítems (DIF), un tipo de investigaciones que cuentan con tradición y herramientas específicas dentro de la TRI (cf., Camilli y Shepard, 1994; Holland y Wainer, 1993). Sin embargo, un grupo influyente de difusores de esta teoría de los tests han propagado la creencia de que la invarianza métrica es una propiedad intrínseca de la TRI y, como tal, es posible garantizarla utilizando este tipo de modelos, dadas algunas condiciones.

Uno de los primeros expertos en psicometría que argumentó que la invarianza es una propiedad de los modelos TRI fue Rasch (1960/1980). Él propuso el modelo que actualmente lleva su nombre y que es reconocido como una propuesta que permite obtener mediciones invariantes debido a que tiene la propiedad de *objetividad específica*. Dicha propiedad permitiría comparar sujetos independientemente del conjunto específico de ítems o instrumentos utilizados en el proceso de medición. De esta forma, si las respuestas a los ítems ajustan al modelo de Rasch, dicha propiedad garantiza que: (a) la diferencia en los logaritmos de las probabilidades de acertar un ítem que exista entre cualquier par de sujetos será igual, independiente del ítem que sea empleado para la comparación; (b) la diferencia en los parámetros de dificultad entre cualquier par de ítems será igual, independiente del grupo de sujetos empleados para estimar aquellos parámetros. En consecuencia, esta propiedad permitiría el logro de mediciones invariantes de los parámetros de ítems y sujetos.

Dado que objetividad específica es una característica exclusiva de los modelos de Rasch, algunos autores sostienen que la propiedad de invarianza métrica sólo es intrínseca a esos modelos (Fisher y Molenaar, 1995; Wright, 1999). Pese a ello, otros autores (i.e., De Ayala, 2009; Embretson y Reise, 2000; Hambleton, Swaminathan, y Rogers, 1991; Reise y Haviland, 2005) generalizan dicha propiedad a todos los modelos TRI, argumentando, por ejemplo, que ellos: (a) estiman los parámetros de los sujetos tomando en cuenta las propiedades de los ítems, y viceversa (De Ayala, 2009; Embretson, 1996); (b) aseguran que las probabilidades de acierto a los ítems dependen únicamente de las CCI y no de las habilidades promedio del grupo de sujetos empleados para la calibración (Hambleton y Swaminathan, 1985); y (c) adoptan la forma de una regresión (aunque no-lineal), y estimar una regresión es invariante porque no depende de la distribución de las habilidades del grupo evaluado (Lord, 1980).

La consecuencia de esa generalización es que en la mayoría de los textos y artículos donde se promueve la TRI, la propiedad de invarianza es destacada como una de las principales ventajas de los modelos TRI en comparación con lo que se puede lograr trabajando con la teoría clásica del test (TCT). Por ejemplo, Hambleton, Swaminathan, y Rogers (1991) sostienen que la invarianza métrica "is the cornerstone of IRT and its major distinction from classical test theory" [es la piedra angular de TRI y su principal distinción respecto de la teoría clásica del test] (p. 19). Por su parte, Reise, Ainsworth y Haviland (2005) la consideran una de las principales características de los modelos TRI porque sin invarianza sería "virtually impossible to administer a common measure to different groups, compute raw scores, and make meaningful comparisons" [prácticamente imposible administrar una medida común a diferentes grupos, calcular puntuaciones brutas, y hacer comparaciones significativas] (p. 97),

mientras que Embretson y Reise (2000) indican que se trata de una de las *nuevas reglas de la medición* introducidas por el uso de la mencionada teoría de los tests.

En consecuencia, para muchos promotores de la TRI, la invarianza métrica constituye un objetivo logrado, puesto que ella es una propiedad intrínseca de dichos modelos, generando lo que en palabras de Wright es una revolución en el campo de la psicometría: "a new measurement in psychology has emerged from a confluence of scientific and social science methodology" [una nueva medición ha emergido en psicología de la confluencia de la metodología científica y de las ciencias sociales] (1999, p. 65).

Aunque los autores que consideran a la invarianza como una propiedad intrínseca de la TRI tienden a no dar una definición formal de su significado y alcances, de sus declaraciones es posible deducir que una parte importante considera que ella permitiría la completa independencia de las estimaciones de parámetros respecto de las muestras de sujetos e ítems utilizados en la calibración. Por ejemplo, Embretson (1999) afirma que las estimaciones de TRI son "population-free" y "test-free" (p. 8), mientras que De Ayala (2009) explica que "with IRT it is possible to have invariant of both person and item characterizations" [con la TRI es posible tener invarianza tanto de las personas como de los ítems] (p. 7) y agrega que "this property is desirable and useful because it frees the practitioner from the specific characteristics of the instrument and samples used" [esta propiedad es deseable y útil, ya que libera al practicante de las características específicas del instrumento y las muestras utilizadas] (p. 409). Por su parte, Hambleton y Russell (1993), aclaran que en la TRI "person parameters or abilities are estimated independently of the particular test items" [los parámetros de las personas o habilidades son estimadas independientemente de los ítems particulares] (p. 257), mientras que Reise, Ainsworth, y Haviland (2005) afirman que las estimaciones de la

TRI “do not depend on the characteristics of a particular population. Also, the scale of the trait does not depend on any particular item set, but exists independently” [no dependen de las características de una población particular. Además, la escala del rasgo no depende de ningún grupo de ítems en particular, sino que existe independientemente] (p. 96).

Más aún, algunos autores también sugieren que la propiedad de invarianza de la TRI permite la calibración de tests con muestras sesgadas de la población objetivo. Por ejemplo, Embretson y Reise (2000) sostienen que “unbiased estimates of item properties may be obtained from unrepresentative samples” [se puede obtener estimaciones insesgadas de las propiedades de los ítems a partir de muestras no representativas] (p. 23), mientras que Hambleton y Russell (1993) subrayan que “the property of sample invariance inherent within IRT means that test developers do not need a representative sample of the examinee population to calibrate test items” [la propiedad de invarianza muestral inherente a los modelos TRI, significa que los desarrolladores de tests no necesitan una muestra representativa de la población objetivo para calibrar los ítems de la prueba] (p. 260).

Por lo tanto, es posible inferir de los argumentos citados en el párrafo anterior, que la TRI permitiría la calibración valida de un test con una muestra sesgada, la que podría sobre-representar o subrepresentar a cualquier grupo o subgrupo (i.e., una etnia, un género, un nivel socioeconómico), o incluir sólo sujetos con altas o bajas habilidades respecto de la media de la población objetivo.

Sin embargo, cabe señalar que lo anterior no implica que se sostenga que la propiedad de invarianza de la TRI signifique que se obtendrá exactamente la misma estimación de parámetros cuando se ajuste el modelo en diferentes muestras (i.e., sujetos con alta y baja habilidad) porque, debido a la indeterminación de la estimación

(es decir, la asignación de valores arbitrarios para la media y la escala del rasgo latente), las estimaciones de parámetros sólo estarán linealmente relacionadas (DeMars, 2010; Rupp y Zumbo, 2004). Como resultado de ello, "IRT properties are invariant only within a linear transformation" [las propiedades de la TRI son invariantes solo dentro de una transformación lineal] (Reise, Ainsworth, y Haviland, 2005, pág. 96), por lo que únicamente después de equiparar los parámetros a la misma métrica, ambos conjuntos de datos serán equivalentes.

Además, para los divulgadores de TRI, la propiedad de invarianza es una potencialidad que se plasmará en un conjunto de datos concreto sólo cuando el modelo ajuste a los datos. Por ejemplo, Reise y Haviland (2005) afirman que "any advantages that IRT modeling may have relative to CTT can only be realized in practice when data are judged appropriate for IRT models and the estimated IRT model parameters fit the observed data" [cualquier ventaja que los modelos TRI puedan tener con respecto a la TCT sólo puede concretarse en la práctica cuando los datos se consideren apropiados para los modelos TRI y los parámetros estimados para un modelo de TRI ajusten a los datos observados] (p. 230). Por su parte, De Ayala (2009) afirma que "theoretically, IRT item parameters are invariant... However, whether invariance is realized in practice (i.e., with parameter estimates) is contingent on the degree of model-data fit" [en teoría, los parámetros de los ítems estimados por la TRI son invariables... Sin embargo, si la invarianza se realiza en la práctica (i.e., con estimaciones de parámetros) depende del grado de ajuste de los datos al modelo] (p. 61).

En otras palabras, la invarianza de los modelos de TRI es sólo una potencialidad derivada de su función matemática y permanecerá en ese estado hasta que se demuestre el ajuste entre el modelo y los datos. En este escenario, ¿cuáles son los límites (si los

hay) para la invarianza lograda cuando las respuestas de un grupo de sujetos a un conjunto de ítems ajustan al modelo? En las páginas siguientes, abordaremos este punto.

LA INVARIANZA EN LA TRI: CUESTIONANDO LA PRETENSIÓN

Como se mencionó anteriormente, los modelos TRI se consideran una herramienta metodológica que produce estimaciones invariantes de parámetros cuando existe evidencia del ajuste del modelo a los datos. Por tanto, se afirma que si un modelo TRI ajusta a una población determinada de ítems y respuestas, cualquier submuestra de ítems obtenidos de ella producirá la misma estimación de parámetros de los sujetos (tras igualar la métrica de las calibraciones), y consiguientemente, se obtendrán parámetros de ítems equivalentes cuando se calibre la prueba con cualquier submuestra de sujetos extraídos de la población para la cual el modelo ajusta.

Sin embargo, no es posible suponer que la invarianza se mantendrá si el mismo conjunto de ítems se aplica a una muestra extraída de una población de individuos distinta o a una subpoblación que se comporta de manera diferente con respecto al rasgo latente, ya que en ese escenario, el modelo TRI puede no ajustar a los datos o, lo que es más importante, puede ajustar al mismo tipo de modelo de TRI, pero con diferentes parámetros, lo que no es suficiente para la obtención de invarianza, por lo que en ese caso no será cierto que sea irrelevante estimar el modelo con sujetos poseedores de niveles de habilidad más altos o más bajos.

Para ilustrar este último punto, se realizó una simulación Monte Carlo similar a las que se emplean para estudiar el DIF, generándose datos de acuerdo al modelo Rasch (aunque la generalización de los resultados a otros modelos TRI es trivial) para test unidimensionales dicotómicos de 40 ítems aplicados a muestras de 2000 sujetos donde

los valores de θ fueron extraídos de una distribución Normal (0, 1). Se crearon 5 condiciones mediante la manipulación de los parámetros de dificultad de los sujetos con altos niveles de habilidad, por lo que los datos fueron producidos en base a modelos de Rasch ligeramente diferentes en cada condición.

La condición I se generó como línea base. En ella, se simulan sujetos que -independiente de su nivel de habilidad- habrían respondido a los ítems de acuerdo a exactamente el mismo modelo Rasch (es decir, un modelo con los mismos parámetros), cuyos su parámetros de dificultad fueron extraídos de una distribución Normal (0, 1). Por su parte, para las condiciones II a V, primero se generaron los parámetros de dificultad para todos los ítems de acuerdo a la misma distribución Normal (0, 1), pero a continuación los parámetros de dificultad fueron modificados aleatoriamente por números aleatorios, extraídos de una distribución Uniforme (-1.5, 1.5) (para las condiciones II y III) y una distribución $\chi 2(1)$ (para las condiciones IV y V), exclusivamente para los sujetos e ítems con niveles de habilidad y dificultad superior a la media (es decir, $\theta > 0$ y $b > 0$). En las condiciones II y IV, los números aleatorios obtenidos fueron sumados a los parámetros de dificultad original, mientras que en las condiciones III y V, dichos valores fueron multiplicados por el parámetro original. Esta simulación implicó que en todas las condiciones se simulan sujetos cuyas respuestas ajustan al modelo de Rasch, pero en las condiciones II a la V, la subpoblación de encuestados con mayor nivel de habilidad respondió los ítems difíciles de acuerdo a parámetros de dificultad ligeramente diferentes a los de la población de encuestados con niveles más bajos de habilidad.

Se realizó un total de 500 réplicas por condición utilizando el software R 2.15.2 (R Development Core Team, 2012). Después de la generación de los datos, cada muestra fue dividida en dos submuestras de tamaños iguales, de acuerdo a la mediana

de θ (que llamaremos muestra A, para los sujetos con baja habilidad y muestra B en el caso de los sujetos con alta habilidad). Por último, cada submuestra fue calibrada independientemente utilizando el modelo de Rasch y el Paquete LTM (Rizopoulus, 2006) y se calculó la correlación entre las estimaciones de parámetros de las submuestras A y B en cada réplica de cada condición.

La correlación de Pearson promedio de las estimaciones de los parámetros de dificultad en las submuestras A y B a lo largo de las 500 réplicas fue igual a .994 en la condición I y .845, .598, .773 y .738 en las condiciones de la II a la V, respectivamente. Para ilustrar estos resultados, en la Figura 10 se muestra la relación observada entre las estimaciones de parámetros en las submuestras A y B para una replica seleccionada al azar.

La fila superior de la Figura 10 muestra la fuerte relación lineal entre las estimaciones de parámetros en las dos submuestras en la condición I, mientras que la segunda y tercera fila muestran la baja relación entre la estimación de los parámetros en ambas submuestras en las condiciones II, III, IV y V. Nótese que las bajas relaciones y la gran variabilidad observada en las condiciones II, III, IV y V son consecuencia de verdaderas diferencias poblacionales en los parámetros de dificultad entre los sujetos con niveles altos y bajos de la habilidad y no son resultado de errores de estimación.

La correlación casi perfecta que se obtuvo para la condición I muestra que cuando todos los datos son generados desde exactamente el mismo modelo de Rasch, las estimaciones de los parámetros son invariantes incluso si la estimación es realizada en muestras con niveles de habilidad sesgados (por ejemplo, alta o baja habilidad). Por el contrario, el nivel de correlación más bajo obtenido en las condiciones II, III, IV y V muestran que la invarianza no es alcanzada cuando el modelo es estimado en muestras de sujetos para los cuales los ítems presenten diferentes niveles de habilidad, aún si sus

respuestas son generadas siempre de acuerdo a un modelo de Rasch. Esto significa que, siempre que entre dos o más poblaciones Q_k -en este caso, dos poblaciones definidas por su nivel de θ- al menos una tenga una característica C_l que se correlacione con la probabilidad de acertar algunos de los ítems dado θ, no se sostendrá la propiedad de invarianza, a pesar de que el modelo ajuste a los datos dentro de cada subpoblación.

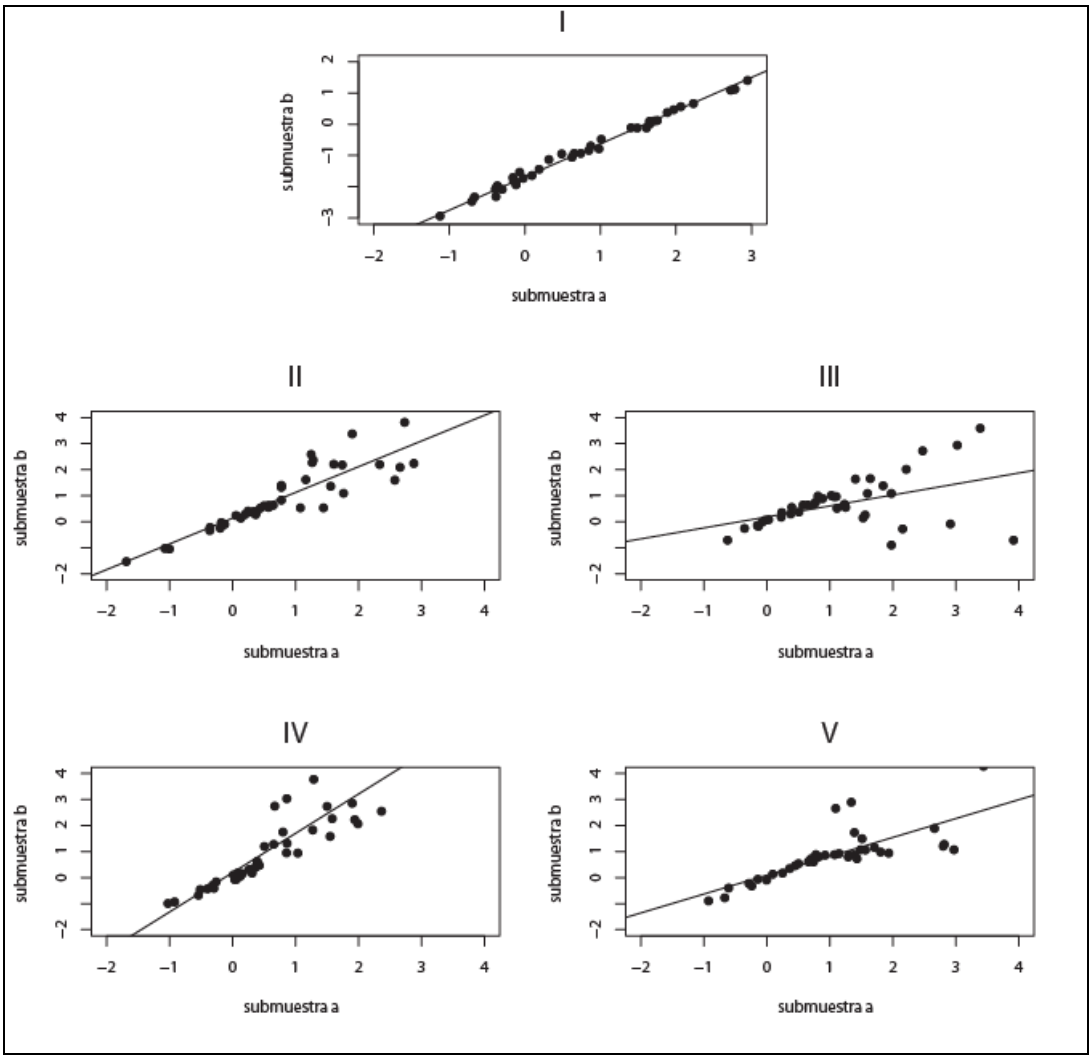

Figura 10. Relaciones entre los parámetros de dificultad estimados para submuestras de bajos y altos niveles de habilidad sobre una réplica seleccionada aleatoriamente en las condiciones I a V.

Este ejemplo muestra una importante limitación a la posibilidad de generalizar la invarianza de los modelos TRI, es decir, la diferencia entre *invarianza interna* e

invarianza externa. Los modelos TRI poseen invarianza interna porque producen las mismas estimaciones de parámetros *dentro* de una muestra de ítems y sujetos para la cual exista evidencia de ajuste exactamente al mismo modelo de TRI (es decir, con iguales parámetros estimados de los ítems para todos los subgrupos de sujetos en que sea posible dividir la muestra). Sin embargo, esta propiedad no se mantiene para ninguna otra muestra si ella tiene al menos una característica que interfiere con las respuestas condicionales de los sujetos a los ítems. En consecuencia, no es correcto afirmar que si hay ajuste de un modelo de TRI a los datos, es irrelevante que nuestra muestra esté compuesta de personas con un nivel de habilidad más alto o más bajo que la población objetivo, pues nada autoriza a pensar que en sujetos de habilidad diferente ajustará exactamente el mismo modelo de TRI. Por el contrario, dado que usualmente las subpoblaciones humanas con distintos niveles de habilidad difieren también en múltiples otras variables, es probable que alguna de ellas interfiera con las probabilidades condicionales de respuesta a los ítems.

En otras palabras, en contra de la interpretación tradicional de la propiedad de invarianza en la TRI, sostenemos que el ajuste de modelos TRI, sólo permite invocar la existencia de invarianza interna, mientras que no faculta a sostener la hipótesis de invarianza externa de los resultados (es decir, la invariancia *entre* poblaciones y muestras).

La distinción entre invarianza interna y externa no parece haber sido reportada previamente en la literatura psicométrica, tanto así que las demostraciones dicha propiedad posibles de leer en papers y manuales dedicados a la difusión de la TRI, solo demuestran la invarianza interna, aunque parecen dar a entender que esta teoría de los tests poseería además invarianza externa. Para ilustrar ello, a continuación

presentaremos y discutiremos tres de estas demostraciones extraídas de los más citados manuales de TRI.

En primer lugar, presentamos el trabajo de De Ayala (2009), quien afirma que la invarianza "is not present in the application of CTT, but it is exhibited in IRT" [no está presente en la aplicación de la TCT, pero si es exhibida en la TRI] (p. 409). Para demostrar ello, De Ayala condujo un pequeño experimento Monte Carlo simulando datos de acuerdo al modelo logístico de un parámetro en test dicotómicos de 20 ítems con una muestra de 1000 sujetos. Luego, el test fue dividido en dos subgrupos de 10 ítems fáciles y difíciles, para finalmente calcular las puntuaciones brutas obtenidas por cada sujeto en cada subtest. Los resultados revelaron que la correlación existente entre ambas series de puntuaciones fue igual a .71, lo que (en opinión del autor) demostraría que la TCT no obtiene parámetros invariantes. Un poco sorprendentemente, De Ayala continúa señalando que las $\hat{\theta}$ en los dos subconjuntos de ítems presentaron una correlación igual a .75, lo que esta bastante cerca del valor obtenido por la TCT, y por tanto, no parece servir para demostrar la presencia de invarianza en la TRI. Sin embargo, en este punto De Ayala afirma que en el escenario simulado fue el error de estimación y no la falta de invarianza lo que explica los resultados, ya que theta se estima mal con tests de sólo 10 ítems. Por ello, el autor repitió el ejercicio usando 2 subgrupos de 50 ítems cada uno, encontrando una correlación igual a .93 entre ambas series de $\hat{\theta}$, con lo cual consideró que existía evidencia de la invarianza en los modelos TRI. Curiosamente, De Ayala no abordó el hecho de que la TCT también podría ser afectada por disponer de tests cortos, por lo tanto, las puntuaciones brutas también deberían presentar una mayor correlación al comparar dos subtest de 50 ítems.

Para evaluar esta hipótesis, replicamos el trabajo de De Ayala en un estudio Monte Carlo con 500 réplicas y confirmamos que, para subtests de 10 ítems, la media

de correlación entre las estimaciones de theta era baja (i.e., .773) al igual que la correlación entre las puntuaciones brutas (i.e., .717). Por otro lado, cuando se ocuparon subgrupos de 50 ítems, la correlación entre ambas $\hat{\theta}$ llegó a .92, mientras que la correlación entre las puntuaciones brutas alcanzó un valor de .82, lo que parece sustentar la idea de una mayor invarianza en la TRI. Sin embargo, si las puntuaciones brutas se sustituyen por el valor *Z* de la proporción de respuestas correctas obtenidas por cada sujeto, como Fan (1998) ha sugerido para evitar que se generen efectos techo y/o piso como consecuencia de la métrica censorizada de las puntuaciones brutas, la correlación promedio entre los escalamientos de los dos subtest aumentó a .91. Por lo tanto, a condición que se realice una pequeña transformación de puntuaciones, la TCT alcanza el mismo nivel de invarianza que la TRI.

Un segundo ejemplo de la forma como se demuestra la propiedad de invarianza de la TRI en la literatura, se encuentra en el trabajo de Embretson y Reise (2000). Dichos autores generan las respuestas de 3000 sujetos a un test de 30 ítems, a partir de un modelo de Rasch con los parámetros theta y de dificultad producidos en función de una distribución normal estándar. Luego de simulados los datos, dividieron la muestra en dos de acuerdo a la mediana de θ para producir un grupo con alta y otro con baja habilidad, los que fueron calibrados independientemente. Basados en la fuerte correlación (r = .997), que se observó entre los parámetros b de ambas muestras, los autores llegaron a la conclusión que las estimaciones de TRI son invariantes. Al mismo tiempo, descartaron la existencia de invarianza en la TCT al encontrar que la correlación entre la proporción de respuestas correctas para cada ítem (es decir, el parámetro *p*) en ambas muestras fue bastante menor (r = .8) e incluso mostró una relación no lineal (aunque monótona creciente). Lo interesante es que la falta de linealidad de la relación entre los parámetros *p* nuevamente pareció ser consecuencia de efectos techo y piso en

la métrica de *p*, más que de la falta de invarianza. Por ello, hipotetizamos que si los parámetros *p* se hubieran reemplazado por el valor *Z* correspondiente a cada *p,* la falta de linealidad tendería a desaparecer y la TCT también mostraría importantes niveles de invarianza. Para poner a prueba esta hipótesis, replicamos el ejercicio de Embretson y Reise generando 500 réplicas, encontrando una correlación promedio entre los parámetros de dificultad b de la TRI igual a .962, mientras que la correlación media de los parámetros *p* fue .814, lo que confirma los resultados reportados por esos autores. Ahora bien, cuando se transformó la *p* de cada ítem por Probit(*p*), la correlación media entre los parámetros de dificultad de la TCT fue .996, evidenciando que la aparente falta de invarianza era casi completamente explicada por la métrica de los *p*.

Por último, Hambleton, Swaminathan, y Rogers (1991) pusieron a prueba la propiedad de invarianza en la TRI trabajando con datos reales. Ellos sugirieron que un acercamiento razonable para evaluar la invarianza es dividir la muestra en dos grupos según dos procedimientos: primero, asignando aleatoriamente los sujetos a ambos grupos, y segundo, dividiendo la muestra de acuerdo a la mediana de $\hat{\theta}$. Como resultado, los autores encontraron una correlación igual a .86 entre $\hat{\theta}$ en los grupos creados aleatoriamente y una correlación igual a .8, en los grupos que se crearon en función de la mediana de $\hat{\theta}$. De acuerdo con sus conclusiones, estos resultados son tan similares que evidenciarían la existencia de invarianza, pues dividir la muestra por un criterio relevante como es la media de theta es igual que dividirla al azar, con lo que se confirma la propiedad de invarianza de la TRI en esos datos.

A pesar de que los tres ejemplos descritos anteriormente efectivamente demuestran la presencia de invarianza en estimaciones de parámetros realizados con la TRI, ellos solo han puesto a prueba lo que hemos denominado invarianza interna, pues todos se circunscribieron a una muestra específica que mostró ajuste a un modelo

específico de TRI. Por lo tanto, las conclusiones que podamos extraer de estas demostraciones no deberían ser generalizadas a más allá del subconjunto de ítems disponible y de la muestra de sujetos utilizados, pues otros ítems que midan el mismo constructo y las respuestas a ellos de sujetos pertenecientes a otras poblaciones podrían ajustar a un modelo distinto de TRI. Por otra parte, aunque en todos los ejemplos la muestra fue dividida en dos grupos muy diferentes (con predilección por los niveles altos y bajos de habilidad de los sujetos o dificultad de los ítems), esta demostración de invarianza es bastante limitada debido a lo siguiente: (a) en los primeros dos ejemplos, los datos fueron creados de acuerdo al mismo modelo TRI, por lo que evidentemente no existían características C_l que pudieran interferir con las probabilidades de acierto a los ítems dado θ, con lo que la invarianza sólo es una tautología generada por el procedimiento de simulación. Ello no asegura que las diferencias de habilidad sean irrelevantes en situaciones reales; y (b) el tercer ejemplo solo demuestra la presencia de cierto grado de invarianza con respecto a la magnitud de θ, lo que no garantiza que diferentes divisiones de la muestra de acuerdo a otras variables (i.e., edad, sexo, etc.) producirían también resultados invariantes. En consecuencia, es engañoso dar por demostrada globalmente la invarianza en esa muestra.

En consecuencia, creemos que la invarianza externa no es demostrada ni considerada en los ejemplos que se citan frecuentemente para apoyar el empleo de modelos TRI como protección contra la influencia de características específicas de las distintas poblaciones estudiadas (i.e., como protección contra la falta de invarianza externa).

LA INVARIANZA EN LA TRI: CONSIDERANDO EL AJUSTE AL MODELO Y EL TAMAÑO DE LAS SUBMUESTRAS

Tal como se mencionó anteriormente, la posibilidad de atribuir la propiedad de invarianza a cualquier conjunto particular de respuestas a un test esta inevitablemente relacionado con la existencia de ajuste de esos datos a un modelo del TRI; de hecho, para algunos autores "invariance and model-data fit are equivalent concepts" [invarianza y ajuste del modelo son conceptos equivalentes] (Hambleton, 1994, p. 540). En consecuencia, dado que la invarianza es propiedad de una muestra específica, extrapolarla a un universo poblacional mayor depende de las características del procedimiento de muestreo utilizado para producir la muestra. Por lo tanto, examinar los límites de la propiedad de invarianza en la TRI, requiere incluir una discusión de los problemas relacionados con el ajuste de los datos al modelo y el diseño muestral.

En cuanto a la invarianza y el ajuste del modelo a los datos, emergen dos problemas. Por un lado, se ha reconocido que “invariance is a property of the model parameters not the estimates” [la invarianza es una propiedad de los parámetros del modelo, no de las estimaciones] (Hambleton, 1994, p. 538), y que la invarianza “only holds when the fit of the model to the data is exact in the population” [sólo se sostiene cuando el ajuste de los datos al modelo es exacto en la población] (Hambleton, Swaminathan, y Rogers, 1991, p. 23). Sin embargo, debido a la naturaleza ideal de los modelos TRI y la complejidad de la realidad, siempre será posible encontrar grados variables de desajuste y falta de invarianza en los datos reales, por lo que será muy improbable encontrar parámetros completamente invariantes (DeMars, 2010). Por otra parte, incluso sin considerar lo anterior, los investigadores aplicados deben basarse en pruebas estadísticas para evaluar la concordancia del modelo con los datos, bajo el

supuesto que un ajuste relativo es suficiente para obtener un grado aceptable de invarianza. Sin embargo, determinar cuando un desajuste de los datos con el modelo es suficientemente significativo como para rechazar un modelo TRI es altamente complejo en la actualidad. Es conocido que algunos estadísticos de ajuste tienden a rechazar modelos esencialmente correctos en presencia de magnitudes insignificantes desajuste cuando el tamaño de la muestra es grande (Embretson y Reise, 2000; Hambleton, Swaminathan, y Rogers, 1991) y/o muestran una potencia inaceptable o altas tasas de error Tipo I (Liu y Maydeu-Olivares, 2013; Orlando y Thissen, 2000). Por lo tanto, es posible pensar que actualmente las herramientas estadísticas para evaluar el ajuste del modelo con los datos no están lo suficientemente desarrolladas como para permitir un fácil y eficaz proceso de toma de decisiones orientada a evaluar la invarianza de los resultados en la investigación aplicada.

Por otro lado, en relación al diseño muestral, los manuales de TRI más citados (i.e., Embretson y Reise, 2000; Hambleton, Swaminathan, y Rogers, 1991; van der Linden y Hambleton, 1997) generalmente profundizan en los métodos y complejidad estadística de la estimación de los parámetros de los modelos de TRI, pero usualmente no discuten sus requerimientos en términos muestrales. De hecho, cuando se mencionan brevemente cuestiones relativas a la muestra, los autores sólo señalan que para estimar modelos de TRI se requieren muestras grandes y heterogéneas (Hambleton y Russell, 1993), agregando que no son necesarias representativas (Embretson y Reise, 2000; Hambleton y Russell, 1993), pues "because of the invariance property, the sample *theoretically* does not need to be a random sample from the population of interest" [debido a la propiedad de invarianza, la muestra *teóricamente* no necesita ser una muestra aleatoria de la población de interés] (DeMars, 2010, p. 32, cursiva en el original).

Sin embargo, la moderna teoría del muestreo demuestra que las inferencias a la población basadas en una muestra solo son sustentables cuando la muestra es representativa de esa población, en el sentido de que refleja las características de interés presentes en la población en proporciones relativamente equivalentes a las de esta última (Lohr, 2009). Por lo tanto, pese a que el tamaño de la muestra es importante para obtener estimaciones de los parámetros, no se debe olvidar que "large unrepresentative samples can perform as badly as small unrepresentative samples. A large unrepresentative sample may do more damage than a small one because many people think that large samples are always better than small ones" [muestras grandes no representativas pueden tener desempeños tan malos como las muestras pequeñas no representativas. Una muestra grande no representativa puede provocar más daño que una pequeña, porque mucha gente cree que las muestras grandes son siempre mejores que las pequeñas] (Lohr, 2009, pp. 8-9). Como consecuencia de ello, resaltar el tamaño de la muestra como la única característica clave para realizar inferencias en los modelos TRI, descuida la relevancia que tiene el diseño muestral y la representatividad para la investigación de invarianza.

Debido a la naturaleza probabilística de los modelos TRI, las inferencias que podrían sustentar la invarianza externa de los resultados al universo poblacional del cual se ha extraido la muestra, solo se sostendrán si la evidencia respecto de la invarianza interna es obtenida a partir de muestras representativas. En otras palabras, si se quiere afirmar que el ajuste a los datos de una muestra es propiedad de un universo mayor de sujetos del cual esa muestra es sólo una selección, se debe disponer de una muestra representativa. Como ya hemos señalado, asumir que las estimaciones de parámetros obtenidos de una muestra sesgada son apropiadas para propósitos inferenciales al universo poblacional, implica subestimar el riesgo de que exista alguna característica

particular de los grupos que están sobre o sub representados en la muestra que interfiera con las probabilidades condicionales de acertar los ítems.

Además, aunque un ajuste aceptable a un modelo de TRI puede ser obtenido de una muestra representativa de una población general, dado que la mayoría las poblaciones humanas son heterogéneas y pueden ser entendidas como la sumatoria de muchas subpoblaciones, las que pueden tener características C_l que se correlacionen con las probabilidades condicionales de acertar los ítems de un test, es difícil determinar el grupo o subgrupo de subpoblaciones para las cuales se sostiene la invarianza. Esto es porque incluso si el tamaño total de la muestra es grande, el tamaño de la muestra de alguna de las subpoblaciones puede no ser lo suficientemente grande como para generar un desajuste de la muestra global, aunque las respuestas de ese subgrupo ajusten a un modelo de TRI diferente o no ajusten a ninguno.

Para ilustrar esta situación, se llevó a cabo una simulación Monte Carlo con 3 condiciones. Se creó un total de 500 réplicas para cada condición, considerando muestras de 1000 sujetos y respuestas a test de 50 ítems, de acuerdo a un modelo de 2 parámetros donde las habilidades de los sujetos y los parámetros de dificultad fueron generados a partir de una distribución normal estándar, mientras que los parámetros de discriminación fueron producidos en función de una distribución Uniforme (5, 2.5). A modo de control, en la condición I, todas las respuestas de los sujetos fueron creadas según un mismo modelo TRI. En las condiciones II y III, los sujetos fueron asignados aleatoriamente a las subpoblaciones A y B, que fueron representadas en diferentes proporciones. En la condición II, el 95% de los sujetos fueron asignados a la subpoblación A y el 5% a la B, mientras que en la condición III, el 55% de los sujetos fueron asignados a la subpoblación A y el 45% a la B. En las condiciones II y III, los parámetros de discriminación fueron los mismos para todos los ítems y sujetos de las

subpoblaciones A y B, pero los parámetros de dificultad fueron distintos en cada subpoblación (aunque fueron extraídos del mismo tipo de distribución). Con ello, hemos querido representar poblaciones heterogéneas compuestas de subgrupos de diferente tamaño, que responden al test en función de diversos parámetros del mismo modelo genérico de TRI.

La Tabla 13 muestra el promedio de los estadísticos de bondad de ajuste para las 500 réplicas en cada condición. Los resultados muestran un buen ajuste de las respuestas de los sujetos en la condición I y un ajuste aceptable en la condición II, mientras que una significativa proporción de los ítems presentan un desajuste en la condición III, especialmente en el análisis de pares y tripletes de ítems (lo que podría ser mal interpretado como un problema de dependencia local). Estos resultados confirman que cuando las poblaciones son heterogéneas y están compuestas de subpoblaciones que responden los ítems en función de diferentes parámetros de un modelo de TRI, la obtención de un buen ajuste relativo en la muestra total sólo demuestra el ajuste de las subpoblaciones numerosas dentro de la muestra, pudiendo quedar ocultos desajustes en algunas submuestras, si es que ellas son lo suficientemente pequeñas en la población o se encuentran subrepresentadas en la muestra analizada.

Tabla 13. *Estadísticos de Bondad de Ajuste para cada condición*

Estadísticos de bondad de ajuste	Condición		
	I	II	III
Media de χ^2 por ítem	9.31	9.86	10.89
% de ítems con desajuste de acuerdo a χ^2	6.9%	9.6%	14.0%
Media de χ^2 por pares de ítems	1.03	2.02	9.96
% de ítems con desajuste de acuerdo al χ^2 de pares de ítems	0.2%	4,0%	25.6%
Media de χ^2 por tripletes de ítems	3.64	7.78	32.26
% de ítems con desajuste de acuerdo al χ^2 de tripletes de ítems	0.4%	9.7%	47.5%
Media de Lz por sujeto	0.23	0.23	0.24
% de sujetos con desajuste de acuerdo a Lz	2.3%	5.1%	2.0%

Nota: Lz = Versión estandarizada del estadístico L_0 (Levine y Rubin, 1979).

En definitiva, evaluar el ajuste a los datos de un modelo de TRI en poblaciones heterogéneas es una tarea compleja que requiere una investigación en sí misma (Hambleton, Swaminathan, y Rogers, 1991), así como muestras grandes y representativas para cada subpoblación relevante, que permita la evaluación empírica de la invarianza métrica en cada subpoblación y realizar inferencias poblacionales, como en cualquier otra investigación en ciencias sociales y del comportamiento.

CONSECUENCIAS DE LA EXCESIVA CONFIANZA EN LA INVARIANZA DE LA TRI

La evidencia presentada hasta ahora desincentiva pensar que las estimaciones de parámetros obtenidos aplicando modelos de TRI son independientes de las muestras de sujetos o ítems empleados. Pese a lo generalizado de la creencia contraria, es posible encontrar algunos expertos en psicometría que comparten algunos de los argumentos acá señalados.

Por ejemplo, McDonald (1999) considera que la invarianza es más una tautología matemática que una propiedad de los modelos TRI, porque "if the item parameters from two groups cannot be rescaled so as to coincide... we can always use population membership as a 'latent trait' and make a model whose parameters are tautologically invariant" [si los parámetros de los ítems de dos grupos no pueden ser re-escalados para que coincidan… siempre se podrá utilizar la pertenencia a una población como 'rasgo latente' y diseñar un modelo cuyos parámetros sean tautológicamente invariantes] (p. 326). Sin embargo, en modelos donde no se define la población como una variable latente, la invarianza no está garantizada y debe ser empíricamente evaluada para cada subpoblación (McDonald, 1986), tal y como se hace en los estudios

DIF, en los que no se asume la invarianza externa como una propiedad garantizada (Camilli y Shepard, 1994; Holland y Wainer, 1993), y como también es habitual en la tradición del análisis factorial (Maydeu-Olivares, Morera, y D'Zurilla, 1999).

Por su parte, Rupp y Zumbo (2004, 2006) sostienen que la invarianza es una propiedad relacional de múltiples poblaciones, por lo que tiene poco sentido evaluarla cuando sólo se dispone de una población, como ocurre en la mayor parte de las investigaciones en las que se demuestra la propiedad de invarianza. Desde otro punto de vista, Millsap (2008) considera que aunque la invarianza fuese una propiedad teórica de los modelos TRI "in truth, these theoretical properties have little role to play in any actual investigation of invariance" [en verdad, esas propiedades teóricas tienen poco rol que jugar en cualquier investigación real de invarianza] (p. 196), porque "invariance is an empirical property of items that may or may not hold, but is not mandated by the structure of a particular latent variable model" [la invarianza es una propiedad empírica de los ítems que podría sostenerse o no, pero no es mandatada por la estructura de un modelo particular de variable latente] (p. 197).

Por otra parte, la declaración de Hambleton, Swaminathan, y Rogers (1991) sobre la equivalencia entre la invarianza y el ajuste del modelo, puede ser interpretada como un reconocimiento implícito de que la invarianza está restringida a muestras y poblaciones para las cuales existe evidencia de ajuste y no puede ser generalizada más allá. Finalmente, Muñiz y Hambleton (1992) también sugieren una interpretación interna de la propiedad de invarianza en la TRI cuando afirman que,

> *Cuando se habla de invarianza de las mediciones respecto de los tests utilizados se está refiriendo a test compuestos por ítems pertenecientes a un banco y calibrados en la misma escala, de lo contrario no hay tal. . . . Sin la existencia*

de un banco de ítems la TRI no reporta ninguna ventaja fundamental frente a la TCT (p. 53).

Estas precauciones contrastan con la creencia generalizada en que la propiedad de invarianza de la TRI permite a las mediciones ser independientes de los tests y las muestras. También contrastan con la idea de que la invarianza (es decir, invarianza interna y externa como se ha definido anteriormente) esta asegurada cuando hay evidencia del ajuste de los datos al modelo. De hecho, la falta de claridad sobre el significado y límites de la propiedad de invarianza en la TRI ha generado al menos 3 importantes consecuencias negativas:

1. Se han oscurecido las diferencias efectivas entre la TCT y la TRI, las que están más vinculadas a diferencias en los procedimientos y métricas usadas para estimar las propiedades de los ítems y sujetos (es decir, censorizada en la TCT y no censorizada en la TRI), o a la capacidad distintiva que tiene la TRI de modelar las relaciones entre un rasgo latente y las respuestas de los sujetos a los ítems, en lugar de la propiedad de invarianza.
2. Ha generado confusión entre los científicos sociales y del comportamiento, quienes han intentado encontrar pruebas empíricas para demostrar la ventaja de la TRI en el campo de la invarianza externa, lo que les hace propensos a forzar sus datos para confirmar sus expectativas. Por ejemplo, Adedoyin, Nenty, y Chilisa (2008) intentaron demostrar la invarianza de las estimaciones TRI a través de la comparación de la media de los parámetros de los sujetos estimados en un conjunto de pares de muestras reales, considerando como prueba de ello que esas medias eran iguales y cercanas a 0 para todos los pares de muestras calibradas con la TRI, sin notar que los software de TRI fijan arbitrariamente la media del rasgo latente en

torno a cero, por lo que sus resultados no demuestran ninguna propiedad del modelo.

3. Ha llevado a algunos investigadores a creer que la invarianza de la TRI libera al investigador del empleo de muestras representativas cuando está calibrando tests. Por ejemplo, Breithaupt y Zumbo (2002) sostienen que los modelos TRI, "are not theoretically sensitive to examinee characteristics unrelated to ability (such as gender, or average group performances)" [no son teóricamente sensibles a características de los examinados no relacionadas con la habilidad (tales como el género, o el promedio de desempeño de un grupo)] (p. 391); mientras que Chernyshenko, Stark, Drasgow, y Roberts (2007) desarrollaron una escala destinada para la población en general y la calibraron con una muestra de sólo estudiantes, afirmando que "whereas it is true that college samples likely show higher means on order than does the general U.S. population, it is important to note that IRT item parameters are subpopulation invariant" [si bien es cierto que las muestras de estudiantes tienden a mostrar promedios mayores que la población estadounidense en general, es importante notar que los parámetros de los ítems de la TRI son subpoblacionalmente invariantes] (p. 95). Sin embargo, como hemos mostrado en las páginas anteriores, las características de una población (que en esta caso son los parámetros que asumen los ítems al aplicarse a esa población) pueden ser estimadas a partir de una muestra con un grado de precisión apropiado, solo cuando las muestras son representativas (Lohr, 2009).

En el presente documento hemos argumentado que las características de la población, tales como la edad, el sexo, el idioma, la nacionalidad, la raza, o incluso las diferencias en habilidades que los sujetos puedan tener, pueden estar relacionados con la probabilidad de acertar uno o más ítems dado θ, y que el utilizar modelos TRI (o

cualquier otro modelo estadístico) no ofrece ninguna protección contra esa influencia. Por lo tanto, para asegurarse de que una característica de una población (por ejemplo, el sexo) no interfiere con la medición de un constructo, se debe obtener una muestra lo suficientemente grande de cada una de las categorías de esa variable como para permitir no sólo demostrar el ajuste del modelo a los datos en el conjunto de la muestra, sino también la presencia de ajuste del modelo en cada submuestra y además debe realizarse un análisis de funcionamiento diferencial del ítem o del test (DIF y DTF respectivamente, por sus siglas en inglés) que asegure que ambas muestras ajustan exactamente al mismo modelo de TRI.

Pasar por alto las restricciones a la inferencia de los parámetros desde una muestra hacia una población, y la necesidad de usar una muestra grande y representativa de cada subpoblación relevante para las cuales el instrumento está destinado, son las principales consecuencias negativas de la ambigua definición de la propiedad de invarianza en la TRI. Estas omisiones podrían tener consecuencias negativas para sujetos evaluados con tests calibrados con una muestra sesgada o en una población diferente a la que pertenecen.

COMENTARIOS FINALES: EL SEDUCTOR CANTO DE LAS SIRENAS

Este documento ha analizado el concepto de invarianza en la TRI, así como la evidencia teórica y empírica que sostiene el considerarla su propiedad intrínseca. A partir de la crítica de dicha evidencia, se han argumentado los límites empíricos de esa pretensión y se han discutido las consecuencias de una delimitación poco clara de dicho concepto para la investigación aplicada en las ciencias del comportamiento.

A pesar de que un grupo importante de difusores de la TRI ha promovido la idea de que esta teoría de los tests es capaz de lograr mediciones invariantes, el presente documento ha proporcionado evidencia para sustentar la idea de que la TRI es sólo internamente invariante. Esto significa que las inferencias que se pueden hacer de esa propiedad están limitadas a las poblaciones de ítems y sujetos que están adecuadamente representadas en la muestra utilizada en la calibración (siempre que el modelo ajuste a los datos) y no son generalizables a otras poblaciones o muestras sin mayor evidencia.

Ejemplos como los que se presentaron en este documento y los extendidos fenómenos de DIF y DTF, evidencian esta limitación y el riesgo de asumir invarianza externa sin evidencia empírica sólida. Por lo tanto, para afirmar que los resultados de las mediciones obtenidas con la TRI son invariables a través de diferentes poblaciones y/o instrumentos (es decir, apelar a la existencia de invarianza externa), se debe realizar una amplia investigación de comparación entre poblaciones. En consecuencia, se aconseja a los investigadores aplicados a actuar con prudencia, evitando creer que el uso de modelos de TRI los protege contra la influencia de variables que distorsionen la medición.

Por otro lado, a pesar de que la idea de invarianza interna en la TRI puede parecer menos atractiva que la de invarianza en un sentido amplio, sus consecuencias empíricas siguen siendo relevantes para la investigación empírica. La invarianza interna de la TRI permite, por ejemplo, el desarrollo de test adaptativos informatizados (CATs, por sus siglas en inglés) garantizando que cualquier subconjunto de ítems procedentes de un banco de ítems que ajuste a los datos producirá resultados equivalentes a cualquier otro subconjunto de ítems procedentes del mismo banco. Naturalmente, el empleo válido de los CATs siempre estará limitado a la población de sujetos a la que los parámetros estimados pueden ser extrapolados, ya que no hay ninguna razón para creer

que diferentes poblaciones producirán las mismas estimaciones. Ello implica que al calibrar los CATs también se debe dar la mayor importancia a la obtención de muestras representativas de las poblaciones objetivo.

Alcanzar mediciones verdaderamente científicas en las ciencias sociales y del comportamiento es un objetivo muy deseable, y esta deseabilidad ha llevado a los divulgadores de la TRI a mal interpretar el alcance de la propiedad de invarianza en esta teoría de los tests, o por lo menos, a ser ambiguos al describirla, permitiendo interpretaciones erróneas. Si bien es cierto que esta confusión le dio mayor legitimidad a la TRI (en comparación con la TCT), pues la legitimó como procedimiento ante los investigadores aplicados, nos parece que a largo plazo esta situación genera más dificultades que beneficios, confundiendo las diferencias efectivas entre la TCT y la TRI y llevando a descuidar la necesidad de seleccionar muestras representativas cuando se desea calibrar tests. En definitiva, se ha pasado por alto el hecho de que los modelos de TRI son herramientas estadísticas que no remplazan la necesidad de disponer de muestras insesgadas cuando se estiman las propiedades de los instrumentos, especialmente si luego se piensa suponer que esas estimaciones son propiedad de la población.

La historia de la ciencia muestra que los avances en la construcción del conocimiento se producen cuando hay claridad respecto de las limitaciones de las herramientas de medición y no como consecuencia de una sobrevaloración de sus posibilidades. Por lo tanto, creer que los modelos TRI son herramientas que producen invarianza interna y externa implica y rendirse a un seductor canto de sirenas frente al cual es mejor resistir por el bien de nuestras disciplinas.

DISCUSIÓN Y CONCLUSIONES GENERALES DE LA TESIS

Antes de comenzar la reflexión respecto de los resultados de las investigaciones que componen esta tesis, es conveniente recordar la pregunta general que la ha guiado: ¿es posible obtener estimaciones de parámetros de los ítems y de los sujetos de calidad y con propiedades similares a las de la TRI empleando procedimientos alternativos en condiciones cercanas a la investigación aplicada?

Este foco en el investigador aplicado implicó centrar la atención en un subconjunto de modelos de TRI (i.e., modelos uni y multidimensionales de ojiva normal o logísticos, dicotómicos de uno a tres parámetros y politómico de respuesta graduada), que constituyen los modelos más cercanos a la experiencia de ese tipo de investigador, por lo que potencialmente serán aquellos que más frecuentemente estén dispuestos a emplear. Dado que la TRI tiene ventajas evidentes y no discutibles en algunos campos cercanos al interés del investigador aplicado (i.e., la evaluación del error de medida condicionado a la habilidad de los sujetos o la estimación de la probabilidad de acierto al azar de los ítems de aptitud), esta tesis se ha enfocado en tres temas críticos en los que puede existir un cierto grado de debate: (a) la obtención de estimaciones más precisas de los parámetros de los ítems; (b) la estimación más precisa de puntuaciones de los sujetos; (c) la obtención de parámetros invariantes de ítems y sujetos.

¿Qué evidencian los estudios presentados respecto de esas tres preocupaciones? Las investigaciones descritas en los capítulos uno y dos muestran que los procedimientos ULS y DWLS asociados al AFI que operan por LI, permiten obtener parámetros de muy superior precisión y estabilidad a aquellos procedimientos

relacionados al análisis factorial clásico y casi completamente equivalentes a aquellos posibles de alcanzar con procedimientos de estimación por información completa asociables a la TRI, al menos en la estimación del parámetro de discriminación en ítems politómicos. Incluso es posible afirmar que en algunas de las condiciones más complejas (i.e., menores tamaños de muestra e ítems con parámetros de discriminación más bajos), las estimaciones AFI pudieran ser levemente de mejor calidad que las estimables con procedimientos TRI.

Por su parte, el capítulo tres muestra que, si bien los escalamientos de sujetos que produce la TRI en los modelos estudiados (i.e., dicotómicos de uno, dos y tres parámetros y modelo politómico de Samejima) son más precisos que los posibles de obtener con la simple suma de puntuaciones derivada de la versión tradicional de la TCT, la diferencia entre ambos escalamientos es usualmente baja y se amplia moderadamente en algunas condiciones poco habituales en la investigación real: tests con alta inadecuación, compuestos por ítems muy discriminativos y con dificultades homogéneas. Además, los escalamientos que produce la TRI serían más precisos en presencia de buenas condiciones para la estimación (i.e., tests largos y muestras grandes) y cuando se dispone de instrumentos compuestos por ítems que reflejan en mayor medida el constructo latente (i.e., ítems con más altas discriminaciones), mientras que la calidad de su estimación de las puntuaciones de los sujetos se deteriora notablemente (incluso en mayor grado que las puntuaciones brutas) cuando se dispone de tests compuestos por ítems de baja discriminación y de condiciones subóptimas para realizar la estimación (e.g., bajo número de ítems y sujetos), puesto las propiedades que aseguran calidad (i.e., ausencia de sesgo, varianza mínima) a estos estimadores, son asintóticas, es decir, convergen a su valor verdadero cuando la muestra y el número de ítems tiende a infinito.

Finalmente, el capítulo cuatro evidencia que si bien la propiedad de invarianza de la TRI es un hecho concreto y verificable empíricamente, tiene un alcance a nivel aplicado bastante menor que el afirmado por algunos difusores de esta teoría de los tests. De esta forma, el ajuste de un conjunto de datos a un determinado modelo de TRI asegura la presencia de invarianza interna a los ítems y sujetos que componen esa muestra, pero ella no es extrapolable necesariamente a otras poblaciones de sujetos o ítems. En otras palabras, emplear procedimientos TRI no asegura la existencia de lo que en esta tesis se ha denominado como invarianza externa. Pese a que la ausencia de invarianza externa puede resultar menos atractiva para los investigadores aplicados, pues el solo uso de modelos de TRI no garantizaría mayor generalización de las estimaciones obtenidas, la presencia de invarianza interna si puede ser de mucho interés para los investigadores especializados, ya que facilita realizar procesos psicométricos más complejos, como por ejemplo: la equiparación de puntuaciones, la construcción de bancos de ítems o el diseño de tests adaptativos informatizados.

En consecuencia, los resultados obtenidos permitirían afirmar que es posible obtener estimaciones de parámetros de los ítems y de los sujetos de calidad y propiedades similares a los estimables a partir de la TRI empleando procedimientos alternativos, al menos para los modelos considerados en esta tesis.

¿Cómo se explican los resultados obtenidos? A la base de la similitud en la calidad de la estimación de parámetros obtenidos empleando modelos de TRI y AFI se encuentra la equivalencia matemática entre el AFI y algunos modelos de TRI. No obstante, puede sorprender la posibilidad de obtener parámetros de discriminación de calidad igual o incluso mejor -en algunas condiciones puntuales-, empleando procedimientos de información limitada. Si bien en principio un procedimiento que emplea toda la información contenida en el patrón de respuestas de los sujetos debería

obtener mejores resultados que otro que sólo utiliza información relativa a la relación bivariada entre los ítems, ello no necesariamente tendría que ocurrir en algunas de las situaciones simuladas, pues:

a. El carácter politómico de los ítems que se simularon impuso limitaciones a la magnitud de asimetría razonable de generar sin transformarlos en ítems virtualmente dicotómicos, producto de la ausencia de respuestas en algunas alternativas. Esto redujo sustancialmente el impacto de la asimetría en las estimaciones, siendo que ella se ha reconocido como uno de los principales factores que deteriora las estimaciones AFI (Forero y Maydeu-Olivares, 2009).
b. Como afirma Roderick McDonald (1995), la distribución bivariada de respuestas a los ítems contiene la mayor parte de la información de un conjunto de datos, por lo que la ventaja de emplear los patrones de respuesta completos muy probablemente constituye una ventaja marginal para la investigación aplicada, y esto podría explicar por ejemplo, la similitud entre las estimaciones AFI y TRI;
c. Algunas de las condiciones simuladas en esta tesis constituyeron contextos notoriamente desfavorables a los complejos procesos de estimación de la TRI, pues consideraban muestras bastante pequeñas, tests cortos, e ítems de baja discriminación. En estas situaciones la frecuencia de los patrones de respuestas puede resultar afectada por mínimas variaciones aleatorias, lo cual puede haber introducido mucho error en las estimaciones basadas en ella. En tanto, al sostenerse sólo en la información bivariada, el AFI sería menos sensible a este tipo de variaciones, lo que podría explicar su buen desempeño en condiciones difíciles.

En relación a la relativa similitud de los resultados de los escalamientos TCT y TRI, existirían al menos dos motivos que podrían explicar los hallazgos encontrados.

En primer lugar, tanto la literatura (Lord, 1953b, 1980) como la evidencia en esta tesis muestran que, para los modelos que se han estudiado, existe una relación perfecta, aunque no lineal, entre los escalamientos TRI y TCT en ausencia de error de estimación. Dado que el grado de no linealidad entre ambos escalamientos es bastante moderado en la mayor parte de las situaciones realistas, ambos escalamientos debieran producir resultados similares con mucha frecuencia.

En segundo lugar, al incluir el error de estimación, se puede constatar que también aparecen equivalencias entre los modelos de TRI estudiados y la suma de puntuaciones. Así, por ejemplo, la propiedad de suficiencia estadística del modelo de Rasch (Fischer, 1922; Rasch, 1960/1980) implica que para cada puntuación bruta obtenida por los sujetos, sólo existirá una estimación TRI, por lo que toda la falta de correlación lineal entre ambas medidas se deberá al limitado grado de no linealidad de la relación entre ambas series de escalamientos, por lo que es perfectamente esperable que exista una alta coincidencia entre ellos. Por otro lado, para el modelo de dos parámetros y el de Samejima, es posible esperar correlaciones lineales un poco menores, pues los escalamientos que producen esos modelos implican tomar en cuenta la discriminación de los ítems al estimar el rasgo, cosa que la simple suma de puntos evidentemente no realiza. Sin embargo, es esperable que esa diferencia sea de poca importancia si los ítems tienen discriminaciones relativamente homogéneas (e.g., similarmente altas o bajas) o incluso que discriminaciones heterogéneas se tiendan a compensar al menos parcialmente en tests largos, como fue observado en algunas condiciones. En consecuencia, se podría esperar una importante relación lineal basal entre ambos tipos de escalamientos, la que deberían tender a ser más fuerte en el caso del modelo de un parámetro y más bajo en el modelo de dos parámetros, sobre todo en

condiciones de discriminaciones heterogéneas, lo que es consistente con los hallazgos de esta tesis.

Por otro lado, hipotetizamos que la explicación a que la simple suma de puntuaciones obtenga resultados incluso levemente más precisos en las condiciones más difíciles (i.e., muestras pequeñas y tests cortos) que las estimaciones de la TRI, se debe a la misma razón que ya hemos señalado anteriormente: en tests cortos, muestras pequeñas e ítems de baja discriminación, la frecuencia de patrones de respuestas puede ser inestable a fluctuaciones aleatorias que no inciden tanto en la suma de puntuaciones.

En consecuencia, los primeros estudios presentados en esta tesis llevan a pensar en los modelos y procedimientos de estimación de la TRI como herramientas capaces de sacar todo el partido posible a datos de buena calidad e importantes volúmenes de información, pero que al mismo tiempo realizan una contribución sólo limitada al ser aplicados a conjuntos de datos de regular calidad, o incluso pueden llegar a ser contraproducentes ante situaciones claramente adversas.

Finalmente, respecto del estudio que sugiere una reformulación más estrecha de la propiedad de invarianza dentro de la TRI, los resultados que alcanzamos podrían explicarse producto que las propiedades de un modelo estadístico abstracto sólo se pueden atribuir a datos que manifiesten un amplio grado de ajuste a dicho modelo, condición que hace notoriamente restrictivas las aplicaciones prácticas de esas propiedades, pues en la investigación empírica con datos reales no debería ser frecuente encontrar esos niveles de ajuste y nada faculta para suponer que, aún cuando éste sea obtenido en un conjunto de datos concreto, ello se repetirá en cualquier otro conjunto de datos diferente al primero. McDonald (1986) resume esa idea al señalar que la propiedad de invarianza es en el mejor de los casos un fuerte supuesto falseable, que debe ser testeado en aplicaciones concretas siempre que sea posible.

Hipotetizamos que una posible explicación de la confusión respecto del sentido de la invarianza en la TRI, proviene de considerar que un aspecto que guió el desarrollo de esta teoría de los tests fue la solución de un problema de las estimaciones de parámetros de los ítems en la TCT, cual es su dependencia respecto del nivel promedio de habilidad del grupo en que se los estima (Gulliksen, 1950/1987). La TRI efectivamente logra una solución a ese problema, lamentablemente, esa invarianza respecto del nivel de habilidad de los sujetos, parece haberse transformado con el tiempo y la presión por legitimar la TRI, en una declaración abstracta de mucho más amplio e impreciso alcance (De Ayala, 2009; Embretson, 1999; Hambleton y Russell, 1993; Reise, Ainsworth, y Haviland, 2005) que proponemos aclarar y delimitar con los conceptos de invarianza interna y externa.

Evidentemente toda investigación tiene limitaciones y vacíos que ponen restricciones a la generalización de sus resultados. Más allá de las reservas específicas que se señalan en cada capítulo, a continuación se discutirán algunas limitaciones y temas no desarrollados en los estudios presentados en esta tesis.

Respecto a los capítulos uno y dos, si bien se mostró que es posible obtener parámetros equivalentes e incluso en algunas ocasiones más precisos que los obtenidos con TRI utilizando procedimientos AFI, ello se limita al contexto de dos situaciones particulares de las simulaciones que se realizaron: (a) haber generado variables latentes con distribuciones normales, situación favorable al AFI. Algunos estudios que han mostrado que ante otros tipos de distribuciones subyacentes de los rasgos, las estimaciones AFI se deterioran sustancialmente en comparación con las estimaciones TRI (DeMars, 2012). Por lo tanto, los antecedentes presentados en esta tesis no debiesen ser generalizados fuera de este marco. Pese a ello, es importante notar que no es fácil evaluar la relevancia práctica de esta limitación, pues por definición la

distribución de la variable latente es siempre desconocida, con lo que no es posible determinar si el supuesto de distribución normal del constructo subyacente es legítimo o constituye una situación poco habitual; (b) haber simulado ítems politómicos donde se emplean todas las alternativas de respuestas en proporciones al menos mínimas (e.g., un mínimo del 5% de los casos en cada alternativa). Esto implica que la asimetría producida no fue tan alta y debe tenerse en cuenta que la magnitud de asimetría es una variable que tiende a deteriorar las estimaciones AFI (Forero y Maydeu-Olivares, 2009).

Por otro lado, si bien el estudio que compara los escalamientos TCT y TRI establece con claridad el impacto de la inadecuación del test, de la magnitud del parámetro de discriminación y del grado de homogeneidad de las dificultades de sus ítems, sobre el grado de relación entre ambos escalamientos; esta investigación no aborda en profundidad el impacto de la heterogeneidad y distribución de la discriminación de los ítems sobre la precisión relativa de ambos escalamientos, tema potencialmente relevante pues es posible pensar que en algunas condiciones que impliquen manipular esas variables se podría obtener alguna ventaja un poco mayor para el modelo de 2 parámetros de la TRI, dada su capacidad para considerar la discriminación de los ítems en las estimaciones del rasgo.

Otro aspecto en que habría sido interesante profundizar en materia de escalar a los sujetos, es el comparar la estimación del rasgo basado en modelos TRI con la estimación de las puntuaciones factoriales con procedimientos AFI. Si bien ya se ha realizado en un estudio reciente sobre este tema (cf., Kohli, Koran, y Henn, 2014), no se ha evaluado el impacto relativo de diferentes procedimientos de estimación de las puntuaciones factoriales (DiStefano, Zhu, y Mindrila, 2009). Un estudio que tratase ese

tema habría servido para extender nuestras conclusiones respecto de la similitud de los escalamientos TCT y TRI con el AFI.

Respecto al estudio de invarianza, hemos afirmado que cuando los datos ajustan a un determinado modelo de TRI se obtiene invarianza interna a esos ítems y sujetos. Un posible vacío de este estudio ha sido el no evaluar si bajo dicho ajuste de los datos es posible detectar también la presencia de invarianza interna para las estimaciones de parámetros del AFI y de la TCT. En principio, dado que la equivalencia matemática entre algunos modelos de TRI y el AFI permite concebirlos como reparametrizaciones del modelo contrario, se hipotetiza que para el caso del AFI la respuesta a esta pregunta debería ser afirmativa, mientras que la no linealidad de la relación entre los parámetros de la TRI y la TCT impulsaría a creer que si bien para esta última teoría de los tests podría existir cierto nivel de invarianza (situación que ha sido reportada en algunos estudios previos, cf., Fan, 1998; MacDonald y Paunonen, 2002), ella sería sólo aproximada y variable de acuerdo al grado de no linealidad de la relación TRI - TCT.

Pese a estas limitaciones, creemos que la presente tesis ha demostrado que, bajo algunas condiciones y modelos, es posible obtener parámetros equivalentes a los de la TRI con procedimientos alternativos mucho más simples, con lo que esta teoría de los tests podría perder algún grado de atractivo para el investigador aplicado pues en esas condiciones: i) no mejoraría significativamente el escalamiento de los sujetos; b) no aportaría mucha mejor información respecto de la calidad de los ítems; c) las estimaciones de parámetros obtenidas no serían más generalizables que las producidas por otros procedimientos. ¿Implica ello que se debería recomendar a los investigadores aplicados no hacer uso nunca de la TRI? Nuestra respuesta a esta pregunta es claramente negativa por los siguientes motivos.

En primer lugar, si bien en algunas áreas relevantes para el investigador aplicado la TRI no presenta ventajas de gran magnitud, se debe recordar que, tanto en la estimación de la fiabilidad e información de ítems y tests, como en la estimación de la probabilidad del acierto al azar de los ítems de aptitud, los modelos de TRI clásicos tienen ventajas evidentes que por lo mismo no hemos puesto a prueba. De esta forma, salvo que explícitamente no se esté trabajando con tests de rendimiento o no se requiera más que una evaluación global de la precisión del instrumento, la TRI debería ser considerada una herramienta ventajosa para calibrar las respuestas a un test.

En segundo lugar, pese a que se ha mostrado la existencia de una relativa equivalencia entre los parámetros estimados con procedimientos alternativos y los obtenidos con la TRI, esta última teoría de los tests parece la mejor diseñada para sacar un óptimo partido de la existencia de buenas condiciones para la estimación modelos de medida (i.e., muestras grandes, ítems con altos niveles de discriminación y tests de mayor longitud). Por ello, los procedimientos de TRI deberían ser recomendados ampliamente en esas favorables condiciones.

En tercer lugar, pese a que en esta tesis se puso el acento en aquellas funciones de la psicometría que pueden ser de mayor interés para el investigador aplicado no experto en psicometría, no debemos olvidar que, en ocasiones, ellos pueden requerir modelos más sofisticados y sin duda la TRI tiene inmensas ventajas en ese terreno pues dispone de modelos más flexibles, que por ejemplo pueden proponer relaciones no sigmoides entre la variable latente y la probabilidad de respuesta o permitirían modelar los procesos cognitivos de respuesta de los sujetos, entre otras muchas posibilidades.

Pese a estas ventajas de la TRI por sobre teorías de los tests más simples o más cercanas a la práctica de los investigadores no psicómetras, los resultados de la presente investigación resultan relevantes, pues muestran que existen alternativas a la TRI al

menos para aquellas condiciones y operaciones en que nos hemos focalizado. ¿Implica ello que se debería recomendar el uso de la TCT para los investigadores aplicados? Esto no parece del todo adecuado pues la TCT no constituye realmente un modelo psicométrico (Raykov y Marcoulides, 2011), no modela las respuestas de los sujetos a los ítems, no dispone de criterios de ajuste a los datos, los parámetros que calcula están sujetos a efectos suelo y techo que los hacen vulnerables a la inadecuación del test y generan relaciones no lineales con los parámetros estimados por la TRI, entre otras dificultades. En función de la evidencia de esta tesis, sólo sería posible recomendar utilizar la suma de puntuaciones de los sujetos como un primer tamizaje de resultados para sujetos que no hayan obtenido puntuaciones extremas, en condiciones en que se pueda suponer que no existe una alta inadecuación del test, o como una aproximación del rasgo cuando se dispone de condiciones en las que no resulte recomendable el empleo de procedimientos de estimación TRI, como puede ocurrir en presencia de ítems de bajo poder discriminador, tests cortos o pequeños tamaños de muestra.

Por otro lado, los resultados obtenidos en esta investigación y la evidencia previa contenida en la literatura permiten dar una respuesta mucho más positiva a la posibilidad de recomendar el AFI como alternativa a la TRI para la estimación de parámetros en condiciones similares a las que se ha estudiado. De esta forma, la equivalencia entre ambos modelos asegura que los parámetros que se obtendrán con ambos procedimientos tendrán una relación esencialmente lineal (Kohli, Koran, y Henn, 2014) y que, salvo la presencia de niveles muy altos de asimetría en ítems dicotómicos (Forero y Maydeu-Olivares, 2009), los parámetros obtenidos serán similares en precisión, ausencia de sesgo y estabilidad, situación que será más clara en presencia de ítems politómicos. Adicionalmente, la estimación por información limitada del AFI disminuye la complejidad y los tiempos del procesamiento (aspecto que pese al

mejoramiento continúo de los ordenadores aún es relevante hoy en día en condiciones multidimensionales), con lo que es posible obtener rápidamente estimaciones de modelos con mayor número de dimensiones latentes e ítems que los procedimientos que emplean los patrones de respuesta completos.

Desde un punto de vista aplicado, producto de su habitual uso para determinar la dimensionalidad de un conjunto de datos, es probable que el AFI y sus especificaciones matriciales sean mucho más familiares a los investigadores aplicados que los procedimientos y modelos de la TRI, facilitando así su difusión y uso. Por ello, también es posible que dichos investigadores tengan mayor cercanía con los estadísticos de ajuste del AFI que con los de la TRI, con lo que también se hace más probable que los empleen al evaluar sus instrumentos de medida. Finalmente, dada la mayor cercanía entre los procedimientos AFI y los modelos de ecuaciones estructurales, difundir el uso del AFI tiene la potencialidad de mejorar la conexión que los investigadores puedan hacer entre el diseño de instrumentos y la construcción y evaluación de modelos de relaciones estructurales, contribuyendo potencialmente a que los mejoramientos en los modelos de medida se traduzcan en otros tipos de descubrimientos.

Más aún, si se acepta la delimitación de la TRI que se ha propuesto en la introducción de esta tesis, incluso se podría proponer al AFI como tercera teoría de los tests pues, a semejanza de los dos existentes, se le han incorporado procedimientos de escalamiento de los sujetos (DiStefano, Zhu, y Mindrila, 2009), formas de determinar el error de medida de la estimación de los sujetos (i.e., estadístico Ω del error de estimación de las puntuaciones factoriales), estrategias de calibración de los parámetros de los ítems (i.e., parámetros λ y τ) y procedimientos de evaluación de la invarianza y equiparación de puntuaciones.

En consecuencia, se propone pasar de considerar al análisis factorial como herramienta para la determinación de la dimensionalidad de un conjunto de datos, y al AFI en particular como aproximación por información limitada a la TRI, a entender -en línea con la propuesta de McDonald (1999)-, al AFI como una teoría de los tests alternativa para operaciones de diseño de instrumentos en los cuales se pueda suponer una relación lineal entre el factor o rasgo y las respuestas continuas subyacentes a los ítems categóricos observados.

Evidentemente esta propuesta tiene al menos cuatro limitaciones: (a) el AFI es fuertemente dependiente del supuesto de distribución normal de la variable latente a diferencia de una TRI mucho más versátil; (b) El AFI está restringido a modelar modelos uni y multidimensionales de uno y dos parámetros logísticos y de ojiva normal para ítems dicotómicos y politómicos, lo que contrasta con la versatilidad de la TRI que puede modelar una infinidad de otros modelos; afortunadamente, los modelos posibles de trabajar con AFI constituyen los más utilizados por los investigadores aplicados (Maydeu-Olivares, Hernández, y McDonald, 2006), quizá con la sola excepción del modelo de tres parámetros para datos dicotómicos de la TRI que no tiene equivalente en AFI pero que en cualquier caso es aplicable principalmente en tests de rendimiento; (c) La TRI permite obtener funciones de información y estimaciones del error de medida para cada patrón de respuestas, lo que no es posible en el contexto de AFI; (d) El AFI es relativamente vulnerable a datos dicotómicos con altos niveles de asimetría, pero ello podría ser subsanado utilizando otros procedimientos factoriales, como por ejemplo NOHARM (Fraser y McDonald, 1988) que trabaja con información limitada, que ha demostrado ser robusto a la asimetría de las respuestas y que permite estimar el modelo de tres parámetros (aunque hoy en día de forma no totalmente satisfactoria).

Esta propuesta de delimitación, con el AFI como teoría de los tests focalizado en el investigador aplicado y la TRI orientada hacia el experto en psicometría, tiene la potencial ventaja de facilitar la difusión de procedimientos de diagnóstico y diseño de tests más avanzados entre los investigadores aplicados de ciencias sociales, rompiendo con el predominio que aún tiene la TCT en muchos campos de investigación y que para el caso de los tests psicológicos más empleados en España ha confirmado recientemente por Elosua (2012). Asimismo, esta distinción y división de tareas entre la TRI y el AFI facilitaría que, con el objetivo de elevar los niveles de manejo psicométrico de los investigadores aplicados, los textos de enseñanza y difusión de la psicometría moderna pudieran destinar mayores esfuerzos en orientar a los lectores hacia una mejor comprensión del AFI como estación intermedia, permitiendo al mismo tiempo una presentación de la TRI más equilibrada, que dejara claras tanto sus potencialidades como sus limitaciones, las que a nuestro entender están poco desarrolladas en parte de la literatura. Es interesante señalar que algunos textos recientes parecen estar dando pasos en el sentido señalado, al dar mucha más importancia al análisis factorial clásico y de ítems en sus presentaciones (i.e., Abad, Olea, Ponsoda, y García, 2011; McDonald, 1999; Raykov y Marcoulides, 2011).

Por otro lado, entre las limitaciones de la TRI que esta opción facilitaría explicar estarían, por ejemplo, los siguientes temas derivados de los hallazgos de esta tesis: (a) el sentido limitado de la propiedad de invarianza de la TRI y la necesidad de contar con muestras representativas para realizar procesos de inferenciales respecto del comportamiento del universo poblacional como es propio de cualquier análisis estadístico; (b) los riesgos de emplear los procedimientos de estimación de parámetros de la TRI en presencia de malas condiciones de estimación, pues si bien es cierto en algunos textos se señalan los tamaños muestrales mínimos para la aplicación de

modelos de TRI (i.e., Muñiz, 1997), no es habitual que se explique la interacción de ese criterio con la discriminación de los ítems y longitud del test, o se detallen las consecuencias de disponer de pobres condiciones para realizar la calibración. Además se podría presentar en esos casos al AFI o el cálculo de las puntuaciones brutas como alternativas útiles bajo algunas condiciones; (c) la mayor potencialidad de la TRI ante ítems con altas discriminaciones y el poco rendimiento adicional que ofrece ante ítems poco discriminadores, dejando en claro que un procedimiento más preciso y sofisticado no suple la necesidad de disponer de ítems e instrumentos de buena calidad.

Finalmente, creemos que si bien es indispensable que los investigadores aplicados mejoren sus conocimientos de medición y evaluación de instrumentos de medida, pues incluso ante una revisión informal de investigaciones empíricas que miden constructos latentes es posible encontrar problemas y carencias relevantes, es posible que resulte demasiado ambicioso esperar que la resolución de esas carencias venga por la vía de difundir la TRI en esa amplia y heterogénea comunidad de investigadores sin escalas intermedias. Creemos que el AFI constituye una herramienta que se encuentra muy bien situada entre los investigadores aplicados para cumplir este rol intermedio que, al tiempo que mejora las prácticas de diseño de instrumentos de los investigadores aplicados, prepara el terreno para que aquellos más interesados comiencen a interiorizarse en los modelos y procedimientos de estimación de la TRI.

En consecuencia, hipotetizamos que la indispensable mejoría en el dominio psicométrico de las futuras generaciones de científicos sociales podría pasar por difundir el AFI como herramienta de calidad para la medición de variables latentes y por introducir a la TRI como una teoría de los tests más sofisticada y más versátil para los investigadores más avanzados.

REFERENCIAS

Abad, F. J., Olea, J. D., Ponsoda, V. G., y García, C. (2011). *Medición en ciencias sociales y de la salud*. Madrid: Síntesis.

Adedoyin, O. O., Nenty, H. J., y Chilisa, B. (2008). Investigating the invariance of item difficulty parameter estimates based on CTT and IRT. *Educational Research and Reviews*, *3*(2), 83-93.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561-573.

Arbuckle, J. L. (2010). *Amos (Version 19.0)* [Computer Program]. Chicago: IBM Company.

Asún, R., y Zúñiga, C. (2008). Ventajas de los modelos politómicos de teoría de respuesta al ítem en la medición de actitudes sociales: El análisis de un caso. *Psykhe (Santiago)*, *17*(2), 103-115.

Beauducel, A., y Herzberg P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling:* a *Multidisciplinary Journal, 13*(2), 186-203.

Bernstein, I. H., y Teng, G. (1989). Factoring items and factoring scales are different: Spurious evidence for multidimensionality due to item categorization. *Psychological Bulletin, 105*(3), 467-477.

Birnbaum, A. (1968). Some latent trait models and their use in inferring a examinee's ability. En F. M. Lord y M. R. Novick, *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Boote, A. S. (1981). Reliability testing of psychographic scales: Five-point or seven-point? Anchored or labeled? *Journal of Advertising Research, 21*, 53-60.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*(3), 425-440.

Boulet, J. R. (1996). *The effect of nonnormal ability distributions on IRT parameter estimation using full-information and limited-information methods* (Doctoral Thesis, University of Ottawa). Retrieved from http://hdl.handle.net/10393/9725.

Breithaupt, K. J., y Zumbo, B. D. (2002). Sample invariance of the structural equation model and the item response model: a case study. *Structural Equation Modeling*, *9*(3), 390-412.

Breithaupt, K. J. (2000). *A comparison of the sample invariance of item statistics from the classical test model, item response model, and structural equation model: A case study of real response data* (Doctoral Thesis, University of Ottawa). Retrieved from http://www.ruor.uottawa.ca/en/handle/10393/9132.

Brown, G., Widing, R. E., y Coulter, R. L. (1991). Customer evaluation of retail salespeople using the SOCO scale: A replication, extension, and application. *Journal of the Academy of Marketing Science, 9*, 347-351.

Browne, M. W. (1984). Asymptotic distribution free methods in the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37*, 127–141.

Camilli, G., y Shepard, L. A. (1994). *Methods for identifying biased test items.* Thousand Oaks, CA: Sage.

Carifio, J., y Perla, R. J. (2007). Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes. *Journal of Social Sciences 3*(3), 106-116.

Chang, L. (1994). A psychometric evaluation of 4-point and 6-point Likert-type scales in relation to reliability and validity. *Applied Psychological Measurement, 18*, 205-215.

Chernyshenko, O. S., Stark, S., Drasgow, F., y Roberts, B. W. (2007). Constructing personality scales under the assumptions of an ideal point response process: Toward increasing the flexibility of personality measures. *Psychological Assessment*, *19*(1), 88.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika, 40*(1), 5–32.

Christoffersson, A. (1977). Two-step weighted least squares factor analysis of dichotomized variables. *Psychometrika, 42*(3), 433–438.

Cox III, E. P. (1980). The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research, 17*, 407-422.

Crocker, L., y Algina, J. (1986). *Introduction to classical and modern test theory*. Harcourt Brace Jovanovich College Publishers: Philadelphia.

Culppeper, S. A. (2013). The reliability and precision of total scores and IRT estimates as a function of polytomous irt parameters and latent trait distribution. *Applied Psychological Measurement, 37*(3) 201–225.

Davison, M. L., y Sharma, A. R. (1990). Parametric statistics and levels of measurement: Factorial designs and multiple regression. *Psychological Bulletin, 107*(3) 394-400.

De Ayala, R. J. (2009). *Theory and practice of item response theory*. New York: Guilford Publications.

DeMars, C. E. (2010). *Item response theory*. Oxford University Press.

DeMars, C. E. (2012). A comparison of limited-information and full-information methods in Mplus for estimating item response theory parameters for nonnormal populations. *Structural Equation Modeling: A Multidisciplinary Journal, 19*(4), 610-632.

DeVellis, R. F. (1991). *Scale development, theory and applications.* Newbury Park: Sage.

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling:* a *Multidisciplinary Journal, 9*, 327-346.

DiStefano, C., Zhu, M., y Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation*, *14*(20), 1-11.

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology, 47*, 309–326.

Drasgow, F., Chernyshenko, O. S., y Stark, S. (2010). 75 years after Likert: Thurstone was right! *Industrial and Organizational Psychology*, *3*(4), 465-476.

Dumenci, L., y Achenbach, T. M. (2008). Effects of estimation methods on making trait-level inferences from ordered categorical items for assessing psychopathology. *Psychological Assessment*, *20*(1), 55-62.

Elosua, P., y Zumbo, B. D. (2008). Coeficientes de fiabilidad para escalas de respuesta categórica ordenada. *Psicothema*, *20*(4), 896-901.

Embretson, S. E. (1996). Item response theory models and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement*, *20*(3), 201-212.

Embretson, S. E. (1999). Issues in the measurement of cognitive abilities. En S. Embretson y S. Hershberger (Eds.). *The new rules of measurement: What every psychologist and educator should know* (pp. 1-15). Mahwah, N.J.: Psychology Press.

Embretson, S. E., y Reise, S. P. (2000). *Item response theory for psychologists*. New Jersey: Lawrence Erlbaum Associates.

Engelhard, G. (1984). Thorndike, Thurstone, and Rasch: A comparison of their methods of scaling psychological and educational tests. *Applied Psychological Measurement*, *8*(1), 21-38.

Engelhard, G. (2008). Historical perspectives on invariant measurement: Guttman, Rasch, and Mokken. *Measurement*, *6*(3), 155-189.

Fabrigar, L. R., Wegener D. T., MacCallum R. C., y Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272-299.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item / person statistics. *Educational and Psychological Measurement*, *58*(3), 357-381.

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice, 40*(5), 532-538.

Ferrando, P. J., y Chico, E. (2007). The external validity of scores based on the two-parameter logistic model: Some comparisons between IRT and CTT. *Psicologica: International Journal of Methodology and Experimental Psychology, 28*(2), 237-257.

Feyerabend, P. (1975). *Against method: Outline of an anarchistic theory of knowledge*. New Jersey: Atlantic Highlands.

Finch, H. (2010). Item parameter estimation for the MIRT model: Bias and precision of confirmatory factor-analysis based models. *Applied Psychological Measurement, 34*(1), 10-26.

Fischer, G. H., y Molenaar, I. W. (Eds.). (1995). *Rasch models: Foundations, recent developments, and applications*. New York: Springer.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 309-368.

Flora, D. B., y Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*(4), 466-491.

Forero, C. G., y Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: limited versus full information methods. *Psychological Methods*, *14*(3), 275-299.

Forero, C. G., Maydeu-Olivares, A., y Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling:* A *Multidisciplinary Journal, 16,* 625–641.

Fraser, C., y McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research*, *23*(2), 267-269.

Gaito, J. (1980). Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin, 87*, 564-567.

Galbraith, J. I., Moustaki, I., Bartholomew, D. J., y Steele, F. (2002). *The analysis and interpretation of multivariate data for social scientists*. London: CRC Press.

Garland, R. (1991). The mid-point on a rating scale: Is it desirable? *Marketing Bulletin, 2*(1), 66-70.

Garner, W. R. (1960). Rating scales, discriminability and information transmission. *Psychological Review, 67*, 343-352.

Göb, R., McCollin, C, y Ramalhoto, M. F., (2007). Ordinal methodology in the analysis of Likert scales. *Quality and Quantity, 41*(5), 601-626.

González-Romá, V., y Espejo, B. (2003). Testing the middle response categories "not sure", "in between" and "?" in polytomous items. *Psicothema, 15*(2), 278-284.

Gorin, J. S., y Embretson, S. E. (2006). Item diffficulty modeling of paragraph comprehension items. *Applied Psychological Measurement*, *30*(5), 394-411.

Gosz, J. K., y Walker, C. M. (2002, April). *An empirical comparison of multidimensional item response data using TESTFACT and NOHARM*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Gulliksen, H. (1950/1987). *Theory of mental tests*. New Jersey: Lawrence Erlbaum Associates.

Guttman, L. (1950). The basis for scalogram analysis. En S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, y J. A. Clausen (Eds.), *Measurement and prediction* (Vol. IV, pp. 60–90). Princeton, NJ: Princeton University Press.

Hair, J. F. Jr., Anderson, R. E., Tatham, R. L., y Black, W. C. (1998). *Multivariate Data Analysis* (5ª edición). Upper Saddle River, NJ: Prentice Hall.

Hambleton, R. K. (1994). Item response theory: a broad psychometric framework for measurement advances. *Psicothema*, *6*(3), 535-556.

Hambleton, R. K., y Russell J. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38-47.

Hambleton, R. K., y Swaminathan, H. (1985). *Item response theory: Principles and applications*. New York: Springer.

Hambleton, R. K., Swaminathan, H., y Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park: Sage.

Hancock, G. R., y Klockars A. J. (1991). The effect of scale manipulations on validity: targeting frequency rating scales for anticipated performance levels. *Applied Ergonomics, 22*, 147-154.

Haraway, D. (1995). *Ciencia, cyborgs y mujeres. La reinvención de la naturaleza*. Cátedra: Valencia.

Harwell, M. R., y Gatti, G. G. (2001). Rescaling ordinal data to interval data in educational research. *Review of Educational Research*, *71*(1), 105-131.

Harwell, M., Stone, C. A., Hsu, T., y Kirisci, L. (1996). Montecarlo Studies in item response theory. *Applied Psychological Measurement, 20*(2), 101-125.

Hau, K., y Marsh, H. W. (2004). The use of items parcels in structural equation modelling: Non-normal data and small sample sizes. *British Journal of Mathematical Statistical Psychology, 57*, 327-351.

Henson, R. K., y Roberts, J. K. (2006). Use of exploratory factor analysis in published research common errors and some comment on improved practice. *Educational and Psychological measurement*, *66*(3), 393-416.

Holgado–Tello, F. P., Chacón–Moscoso, S., Barbero–García, I., y Vila–Abad, E. (2010). Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality & Quantity, 44*(1), 153-166.

Holland, P. W., y Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hoogland, J. J., y Boomsma, A. (1998). Robustness studies in covariance structural modeling: An overview and a meta-analysis. *Sociological Methods & Research, 26*(3), 329–367.

Jamieson, S. (2004). Likert scales: How to (ab)use them. *Medical Education, 38*, 1212-1218.

Jones, L. V. (1960). Some invariant finding under the method of successive intervals. En H. Gulliksen y S. Messick (Eds.), *Psychological scaling: Theory and applications* (pp. 7–20). New York: John Wiley & Sons, Inc.

Jöreskog, K. G., y Sörbom, D. (2002). *PRELIS 2: User's reference guide.* Lincolnwood: Scientific Software International, Inc.

Jöreskog, K. G., y Sörbom, D. (2006). *LISREL 8.8: User's reference guide.* Lincolnwood: Scientific Software International, Inc.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, *36*(2), 109-133.

Kang, S. M., y Waller, N. G. (2005). Moderated multiple regression, spurious interaction effects, and IRT. *Applied Psychological Measurement*, *29*(2), 87-105.

Kohli, N., Koran, J., y Henn, L. (2014). Relationships among classical test theory and item response theory frameworks via factor analytic models. *Educational and Psychological Measurement*. Advance online publication. doi:10.1177/0013164414559071.

Kolen, M. J., y Brennan, R. L. (2004). *Test equating, scaling, and linking*. New York: Springer.

Kulas, J. T., Stachowski, A. A., y Haynes, B. A. (2008). Middle response functioning in Likert-responses to personality items. *Journal of Business and Psychology, 22*(3), 251-259.

Levine, M. V., y Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational and Behavioral Statistics*, *4*(4), 269-290.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22*(140), 44-55.

Likert, R., Roslow, S., y Murphy, G. (1934). A simple and reliable method of scoring Thurstone attitudes scales. *The Journal of Social Psychology, 5*(2), 228-238.

Liu, Y., y Maydeu-Olivares, A. (2013). Local dependence diagnostics in IRT modeling of binary data. *Educational and Psychological Measurement*, *73*(2), 254-274.

Lohr, S. (2009). *Sampling: design and analysis*. Cengage Learning.

Loken, B., Pirie, P., Virnig K. A., Hinkle, R. L., y Salmon, C. T. (1987). The use of 0-10 scales in telephone surveys. *Journal of the Market Research Society, 29*(3), 353-362.

Lord, F. M., y Novick, F. M. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lord, F. M. (1952). A theory of test scores. *Psychometric monographs, nº7*. Richmond, V.A.: Psychometric Corporation.

Lord, F. M. (1953a). On the statistical treatment of football numbers. *American Psychologist, 8*, 750-751.

Lord, F. M. (1953b). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13*(4), 517-549.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. New Jersey: Lawrence Erlbaum Associates.

Lorenzo-Seva, U., y Ferrando, P. J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavioral Research Methods, Instruments and Computers, 38*(1), 88-91.

MacDonald, P., y Paunonen, S. V. (2002). A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement*, *62*(6), 921-943.

Maij-de Meij, A. M., Kelderman, H., y van der Flier, H. (2008). Fitting a mixture item response theory model to personality questionnaire data: Characterizing latent classes and investigating possibilities for improving prediction. *Applied Psychological Measurement*, *32*(8), 611-631.

Marsh, H. W., Hau, K-T., Balla, J. R., y Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33*(2), 181-220.

Martínez-Arias, M. R., Hernández-Lloreda, M. J., y Hernández-Lloreda, M. V. (2006). Psicometría. *Madrid: Alianza Editorial*.

Matell, M. S., y Jacoby, J. (1971). Is there an optimal number of alternatives for Likert scale items? Study 1: Reliability and validity. *Educational and Psychological Measurement, 31*, 657-674.

Matsumoto, M., y Nishimura, T. (1998). Mersenne Twister: A 623-dimensionally equidistributed uniform pseudo-random number generator, *ACM Transactions on Modeling and Computer Simulation*, *8*(1), 3–30.

Maydeu-Olivares, A., Hernández, A., y McDonald, R. P. (2006). A Multidimensional ideal point item response theory model for binary data. *Multivariate Behavioral Research, 41*(4), 445–472.

Maydeu-Olivares, A., Morera, O., y D'Zurilla, T. J. (1999). Using graphical methods in assessing measurement invariance in inventory data. *Multivariate Behavioral Research*, *34*(3), 397-420.

McDonald, R. P., y Mok, M. M. C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research*, *30*(1), 23-40.

McDonald, R. P. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement*, *6*(4), 379-396.

McDonald, R. P. (1986). Describing the elephant: Structure and function in multivariate data. *Psychometrika*, *51*(4), 513-534.

McDonald, R. P. (1997). Normal ogive multidimensional model. En W. van der Linden y R. Hambleton (Eds). *Handbook of modern item response theory* (pp. 257-269). New York: Springer.

McDonald, R. P. (1999). *Test theory: A unified approach*. Mahwah, NJ: Lawrence Erlbaum.

Mellenbergh, G. J. (1994a). Generalized linear item response theory. *Psychological Bulletin*, *115*(2), 300-307.

Mellenbergh, G. J. (1994b). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research, 29*, 223–236.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, *58*, 525-543.

Merenda, P. F. (1997). A guide to the proper use of factor analysis in the conduct and reporting of research: Pitfalls to avoid. *Measurement & Evaluation in Counseling & Development, 30*(3), 156-164.

Michell, J. (2008). Is psychometrics pathological science? *Measurement*, *6*(1-2), 7-24.

Michell, J. (2009). The psychometricians' fallacy: Too clever by half? *British Journal of Mathematical Statistical Psychology, 62*, 41-55.

Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review 63*(2), 81-97.

Millsap, R. E. (2008). Model-implied invariance in psychometrics: Be skeptical when theory suggests data are not needed. *Measurement: Interdisciplinary Research and Perspectives, 6*(3), 195-197.

Mokken, R. J. (1997). Nonparametric models for dichotomous responses. En W. van der Linden y R. Hambleton (Eds). *Handbook of modern item response theory* (pp. 351-367). New York: Springer.

Morren, M., Gelissen, J. P., y Vermunt, J. K. (2011). Dealing with extreme response style in cross-cultural research: A restricted latent class factor analysis approach. *Sociological Methodology, 41*(1), 13-47.

Morse, B. J., Johanson, G. A., y Griffeth, R. W. (2012). Using the graded response model to control spurious interactions in moderated multiple regression. *Applied Psychological Measurement*, *36*(2), 122-146.

Muñiz, J., y Hambleton, R. K. (1992). Medio siglo de teoría de respuesta a los ítems. *Anuario de psicología*, (52), 41-66.

Muñiz, J. (1997). *Introducción a la teoría de respuesta a los ítems*. Madrid: Pirámide.

Muñiz, J. (2001). *Teoría clásica de los tests*. Madrid: Pirámide.

Muñiz, J. (2010). Las teorías de los tests: teoría clásica y teoría de respuesta a los ítems. *Papeles del psicólogo*, *31*(1), 57-66.

Muraki, E., y Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement, 19*(1), 73-90.

Muraki, E., y Engelhard, G. H. (1985). Full-information item factor analysis: Applications of EAP scores. *Applied Psychological Measurement, 9*(4), 417-430.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika, 43*(4), 551–560.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variables indicators. *Psychometrika, 49*(1), 115-132.

Muthén, B. (1989). Dichotomous factor analysis of symptom data. *Sociological Methods & Research, 18*(1), 19-65.

Muthén, B. (1993). Goodness of fit with categorical and other non-normal variables. En K. A. Bollen y J. S. Long (Eds.), *Testing Structural Equation Models* (pp. 205-243). Newbury Park, CA: Sage.

Muthén, B., du Toit, S. H. C., y Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*. Retrieved June 11, 2013 (http://pages.gseis.ucla.edu/faculty/muthen/articles/Article_075.pdf).

Muthén, L. K., y Muthén, B. (2011). *Mplus version 6.11*. Los Angeles: Muthén & Muthén.

Ndalichako, J. L., y Rogers, W. T. (1997). Comparison of finite state score theory, classical test theory, and item response theory in scoring multiple-choice items. *Educational and Psychological Measurement*, *57*(4), 580-589.

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education, 15*(5), 625-632.

Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.

Olson, J. F., Martin, M. O., y Mullis, I. V. (Eds.). (2008). *TIMSS 2007 technical report*. IEA TIMSS & PIRLS.

Orlando, M., y Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*(1), 50-64.

Peterson, R. A. (2000). A meta-analysis of variance accounted for and factor loadings in exploratory factor analysis. *Marketing Letters, 11*(3), 261-275.

Preston, C. C., y Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica, 104*, 1-15.

Progar, S., Socan, G., y Slovejija, M. P. (2008). An empirical comparison of item response theory and classical test theory. *Horizons of Psychology*, *17*(3), 5-24.

R Development Core Team. (2012). *R: A language and environment for statistical computing version 2.1.5.2.* [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/

Raaijmakers, Q. A. W., van Hoof, J. T. C., Verbogt, T. F. M. A., y Vollebergh W. A. M. (2000). Adolescents' midpoint response on Likert-type scale items: Neutral or missing values? *International Journal of Public Opinion Research, 12*(2), 208-216.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research. (Expanded ed., 1980, Chicago: University of Chicago Press).

Raykov, T., y Marcoulides, G. A. (2011). *Introduction to psychometric theory*. New York: Routledge.

Raykov, T., y Marcoulides, G. A. (2015). On the relationship between classical test theory and item response theory from one to the other and back. *Educational and Psychological Measurement*. Advance online publication, doi: 10.1177/0013164415576958

Reckase, M. (2009). *Multidimensional item response theory*. New York: Springer.

Reise, S. P., y Haviland, M. G. (2005). Item response theory and the measurement of clinical change. *Journal of personality assessment*, *84*(3), 228-238.

Reise, S. P., y Henson, J. M. (2003). A discusión of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment, 81*, 93-103.

Reise, S. P., Ainsworth, A. T., y Haviland, M. G. (2005). Item response theory fundamentals, applications, and promise in psychological research. *Current Directions in Psychological Science*, *14*(2), 95-101.

Reiser, M., y VandenBerg, M. (1994). Validity of the chi-square test in dichotomous variable factor analysis when expected frequencies are small. *British Journal of Mathematical and Statistical Psychology, 47*(1), 85-107.

Rhemtulla, M., Brosseau-Liard, P. É., y Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*(3), 354-373.

Rigdon, E. E., y Ferguson Jr., C. E. (1991). The performance of the polychoric correlation coefficient and selected fitting functions in confirmatory factor analysis with ordinal data. *Journal of Marketing Research, 28*, 491-497.

Rittle-Johnson, B., Matthews, P. G., Taylor, R. S., y McEldoon, K. L. (2011). Assessing knowledge of mathematical equivalence: A construct-modeling approach. *Journal of Educational Psychology*, *103*(1), 85-105.

Rizopoulos, D., y Moustaki, I. (2008). Generalized latent variable models with non-linear effects. *British Journal of Mathematical and Statistical Psychology*, *61*(2), 415-438.

Rizopoulos, D. (2006). LTM: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, *17*(5), 1-25.

Roberts, J. S., Donoghue, J. R., y Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement*, *24*(1), 3-32.

Rupp, A. A., y Zumbo, B. D. (2004). A note on how to quantify and report whether IRT parameter invariance holds: When Pearson correlations are not enough. *Educational and Psychological Measurement*, *64*(4), 588-599.

Rupp, A. A., y Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement*, *66*(1), 63-84.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., y Flannery, B. P. (1992). *Numerical recipes in FORTRAN. The art of scientific computing*. Cambridge, England: Cambridge University Press.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph,* Nº17.

Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika, 38*(2), 203-219.

Santisteban, C. (2009). *Principios de psicometría*. Madrid: Síntesis.

Savalei, V., y Rhemtulla, M. (2013). The performance of robust test statistics with categorical data. *British Journal of Mathematical and Statistical Psychology, 66*(2), 201-223.

Schrage, L. (1979). A more portable FORTRAN random number generator. *ACM Transactions on Mathematical Software, 5*(2), 132-138.

Sijtsma, K., y Molenaar I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, California: Sage.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107-120.

Spearman, C. (1904). "General Intelligence" objectively determined and measured. *American Journal of Psychology, 5*, 201-293.

Spector, P. E. (1992). *Summating rating scale construction: An introduction.* Newbury Park: Sage.

Stevens, J. P. (1992). *Applied Multivariate Statistics for the Social Sciences* (2ª edición). Hillsdale, NJ: Erlbaum.

Steyer, R. (1989). Models of classical psychometric test theory as stochastic measurement models: Representation, uniqueness, meaningfulness, identifiability, and testability. *Methodika*, *3*, 25-60.

Sukirno, y Siengthai, S. (2010). The comparison of graded response model and classical test theory in human resource research: A model fitness test. *Research and Practice in Human Resource Management, 18*(2), 77-90.

Takane, Y., y De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*, 393–408.

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement, 27*(3), 159–203.

Tay, L., y Drasgow, F. (2012). Adjusting the adjusted χ2/df ratio statistic for dichotomous item response theory analyses: Does the model fit? *Educational and Psychological Measurement, 72*(3), 510-528.

Thissen D., y Wainer, H. (2001). *Test scoring*. Mahwah: Lawrence Erlbaum Associates.

Thissen, D. (2003). *MULTILOG version 7.03*. Lincolnwood, IL: Scientific Software International.

Thorndike, E. L. (1922). On finding equivalent scores in tests of intelligence. *Journal of Applied Psychology, 6*, 29-33.

Thurstone, L. L. (1927). The unit of measurement in educational scales. *Journal of Educational Psychology, 18*, 505-524.

Timmerman, M. E., y Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods, 16*(2), 209-220.

Tomkowicz, J., y Rogers, W. T. (2005). The use of one-, two-, and three-parameter and nominal item response scoring in place of number-right scoring in the presence of test-wiseness. *Alberta Journal of Educational Research*, *51*(3), 200-215.

van der Linden, W. J., y Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York: Springer.

Velleman, P. F., y Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *American Statistician, 47,* 65-72.

Wirth, R. J., y Edwards, M. C. (2007). Item factor analysis: current approaches and future directions. *Psychological Methods, 12*(1), 58-79.

Wright, B. D. (1968). Sample-free test calibration and person measurement. En *Proceedings of the 1967 Invitational Conference on Testing Problems* (pp. 85-101). Princeton, NJ: Educational Testing Service.

Wright, B. D. (1999). Fundamental measurement for psychology. En S. Embretson y S. Hershberger (Eds.). *The new rules of measurement: What every psychologist and educator should know* (pp. 65-104). Mahwah, N.J.: Psychology Press.

Xu, T., y Stone, C. A. (2012). Using IRT trait estimates versus summated scores in predicting outcomes. *Educational and Psychological Measurement, 72*(3), 453-468.

Yang-Wallentin, F., Jöreskog, K. G., y Luo, H. (2010). Confirmatory factor analysis of ordinal variables with misspecified models. *Structural Equation Modeling:* A *Multidisciplinary Journal, 17*, 392–423.