LA NECESARIA COMPLEMENTARIEDAD ENTRE TEORÍA CLÁSICA DE LA MEDICIÓN (TCM) Y TEORÍA DE RESPUESTA AL ÍTEM (IRT):
ASPECTOS CONCEPTUALES Y APLICACIONES*

Jorge Manzi y Ernesto San Martín

Este artículo tiene por objetivo mostrar la complementariedad conceptual y práctica que existe entre la Teoría Clásica de Medición (TCM) y la Teoría de Respuesta al Ítem (IRT). En la primera parte se presentan las similitudes estructurales entre ambos modelos, esto es, que ambas teorías se construyen a partir de dos hipótesis fundamentales: (1) que los instrumentos miden un rasgo unidimensional no observable, y (2) el axioma de independencia local. Luego se presentan los modelos IRT como una extensión natural de la TCM, la cual permite modelar el puntaje verdadero de un individuo no sólo en función de su rasgo latente propio, sino también por medio de las características de un ítem, como son su dificultad y su grado de discriminación. En la segunda parte del artículo se destacan las ventajas que se obtienen al analizar datos educacionales con un modelo estructuralmente más rico (es decir, los modelos IRT) en

JORGE MANZI. Psicólogo, P. Universidad Católica de Chile; Ph.D. en Psicología, Universidad de California, Los Angeles. Profesor Adjunto de la Escuela de Psicología de la Pontificia Universidad Católica de Chile.

ERNESTO SAN MARTÍN. Ingeniero Civil Matemático, Universidad de Chile; Ph.D. en Estadística, Universidad Católica de Lovaina, Bélgica. Profesor Auxiliar del Departamento de Estadística de la Pontificia Universidad Católica de Chile.

* Los autores han realizado consultorías para el SIMCE relativas a la comparabilidad de los puntajes de las pruebas y agradecen el acceso otorgado por el SIMCE a la información que ha sido empleada para ilustrar los temas abordados en este artículo.

comparación con la TCM. Dicha comparación se ilustra usando resultados de una prueba SIMCE aplicada a estudiantes de segundo año de enseñanza media. Los autores del artículo concluyen señalando que el enfoque IRT debe formar parte de la agenda de trabajo de todo esfuerzo serio en el ámbito de la medición educacional en gran escala.

1. Introducción

Las teorías de la medición son el fundamento conceptual que hace posible asociar números a los objetos que nos interesa medir. Por ello, han acompañado desde sus orígenes a las disciplinas científicas que fundan su quehacer en la medición. Esto es particularmente relevante en disciplinas como la psicología y la educación, que se interesan por medir aspectos que no son directamente observables, y que, por tanto, deben inferir atributos subyacentes a partir de los comportamientos observables de los examinados.

Las teorías de la medición tienen su origen hacia fines del siglo XIX. La formalización de la hoy llamada Teoría Clásica de la Medición (TCM) se remonta a comienzos del siglo XX. El desarrollo del análisis factorial fue una contribución complementaria que potenció dicha teoría y facilitó su aplicación para resolver problemas de gran interés científico, como ha sido la medición de inteligencia y personalidad. Luego de su gran dinamismo en la primera mitad del siglo XX, la TCM se constituyó en un cuerpo bien establecido de principios y procedimientos que han servido de base para el desarrollo de la mayoría de los más importantes instrumentos de medición que hoy conocemos. El texto de Gulliksen publicado en 1950 representa la consolidación del conocimiento acumulado por la TCM.

El dinamismo en torno a la medición se volcó en la segunda mitad del siglo pasado al desarrollo de modelos complementarios a la TCM, intentando superar los límites que se conocía tenía ese cuerpo de conocimientos. Dos fueron los desarrollos más relevantes: la teoría de la generalizabilidad (Cronbach, Gleser y Rajaratnam, 1963) y la Teoría de Respuesta al Ítem (IRT), también llamada teoría de rasgo latente. De ellas, esta última es la que ha concentrado marcadamente la atención académica a nivel internacional. Los trabajos pioneros de Rasch a fines de los años 50, y el impulso de Birnbaum y Lord en los años 60, sirvieron de base para el desarrollo de una teoría que hoy da cuenta de la mayor parte de la producción científica en este campo, y que comienza a ser considerada como el marco de referen-

cia básico para el diseño de nuevos instrumentos de medición. Es importante constatar que en este caso ha existido un importante rezago temporal entre el desarrollo de las ideas y sus aplicaciones prácticas, el que se explica fundamentalmente por las complejidades computacionales que supone el enfoque IRT. Sin embargo, a medida que se han desarrollado algoritmos y programas computacionales para resolver los problemas de estimación de parámetros y de habilidad de este modelo, su difusión ha comenzado a ampliarse. De hecho, una rápida revisión de revistas científicas no especializadas en materias psicométricas revela que el modelo IRT está siendo usado en múltiples contextos, tales como la medición de psicopatología (Rouse, Finger y Butcher, 1999; Waller, Thompson, y Wenk, 2000), personalidad (Fraley, Waller y Brennan, 2000; Gray-Little, Williams y Hancock, 1997; Panter, Swygert, Dahlstroim, Grant y Tanaka, 1997), inteligencia (Ellis, 1989; Godber, Anderson y Bell, 2000; Maller, 2000), etcétera.

Las primeras aplicaciones de este enfoque, que siguen siendo las más relevantes, se han concentrado en el campo educacional. Grandes programas de medición educacional a nivel internacional, tales como el TIMSS (matemática y ciencias), la medición de competencias lectoras (SIALS) o la evaluación de conocimientos y actitudes cívicas, se han apoyado en el enfoque IRT para desarrollar sus instrumentos y para reportar sus resultados. La medición del progreso educacional es también un área común, como lo muestra el uso de IRT en EE.UU. (NAEP) o en Chile (SIMCE). Más aun, en años recientes, se ha optado por este mismo enfoque para medidas educacionales individuales, como son la prueba norteamericana para seleccionar alumnos a programas de postgrado (el Graduate Record Examination, GRE) o la prueba que mide competencias lingüísticas en inglés (TOEFL).

Por ello, es relevante clarificar el rol y aporte que este enfoque representa. Como se describe en este trabajo, el enfoque IRT surge desde la TCM y representa una extensión de ella. No se trata, como algún observador ingenuo pudiera haber supuesto, de teorías alternativas o competitivas para explicar un mismo fenómeno. Por el contrario, se trata de teorías complementarias, que en la práctica deben ser usadas conjuntamente para resolver los problemas básicos que toda medición representa.

Como se describe en este trabajo, el enfoque IRT busca resolver algunas limitaciones importantes de la TCM. En ciertos casos, esto ha llevado a desarrollos que no tienen paralelo en la Teoría Clásica (como la medición adaptativa) y, en otros, ha significado la ampliación de los recursos métricos para evaluar la calidad de un instrumento.

En este trabajo se representan las aplicaciones más establecidas y consolidadas del enfoque IRT, especialmente las que dicen relación con el análisis de instrumentos que emplean ítemes puntuados en forma dicotómica (que es lo usual en el contexto educacional). Para este tipo de aplicaciones existe, por lo demás, una variada oferta de programas computacionales. No consideraremos, en cambio, desarrollos y extensiones más recientes de este enfoque, tales como los modelos IRT multidimensionales, los modelos para ítemes puntuados en varias categorías de respuesta y los modelos no paramétricos (Van der Linden y Hambleton, 1997).

2. ¿Qué significa medir?

En nuestro país existe una amplia tradición de medición educacional. Las pruebas nacionales SIMCE y las pruebas de ingreso a la universidad son los principales ejemplos de ello. Las bases de datos generadas por dichas mediciones son utilizadas tanto para análisis rutinarios de medición, como para nuevos desarrollos teóricos y aplicados en el campo de la medición. Tanta actividad profesional y académica nos debe motivar a volver, cada cierto tiempo, sobre *cuestiones fundamentales*. Éstas, junto a la experiencia aplicada acumulada, nos permitirá tener una mejor comprensión de lo que significa medir y de los desafíos que ello implica.

¿Qué es medir? Parece una pregunta trivial. Sin embargo, si tratamos de explicar qué hacemos cuando medimos el largo de un lápiz, veremos que no es asunto fácil. Tanto más, si nos reducimos al fascinante campo de la medición educacional. En lo que sigue queremos revisar algunos elementos básicos de la *teoría de la medición*. Cuando medimos algún atributo de una clase de objetos o eventos (por ejemplo, medir el largo de un lápiz), *asociamos números (u otra entidad matemática familiar) con los objetos de una manera tal que las propiedades del atributo son completamente representadas por propiedades numéricas*. Supongamos que tenemos dos trozos rígidos de varillas, llamados *a* y *b*, cuyos largos queremos medir. Si ponemos una varilla al lado de la otra, entonces hay sólo tres posibilidades que pueden ocurrir: el extremo de la varilla *a* puede estar más allá del extremo de la varilla *b*, o el de *b* más allá del de *a*, o ambos extremos coinciden. En el primer caso decimos que *a* es más largo que *b*, en el segundo que *b* es más largo que *a*, y en el tercero que *a* y *b* tienen largos equivalentes. Por brevedad, escribimos $a>b$, $b>a$ o $a=b$, respectivamente.

Un primer elemento subyacente a toda medición es, por tanto, una *comparación cualitativa*; toda situación de medición involucra dicha com-

paración. Ejemplos más complejos son el de preferencias de canastas, grados de creencia frente a juegos de azar, mediciones astronómicas, etc. Un segundo aspecto fundamental subyacente a toda medición es la *evaluación cuantitativa de la comparación cualitativa*. El problema consiste, por tanto, en asignar números $\phi(a)$, $\phi(b)$, etc. a objetos *a, b,* etc. tal que el orden establecido cualitativamente se mantenga. Las evaluaciones numéricas se llaman *escala de medición*.

Un procedimiento natural para asignar números es el siguiente: asignamos a la primera varilla seleccionada un número *cualquiera.* Si el largo de la segunda varilla seleccionada excede el de la primera, le asignamos un número mayor; en caso contrario, uno menor. Hacemos lo mismo con una tercera varilla, excepto si su largo está entre las dos anteriores, en cuyo caso le asignamos un número que esté entre los dos anteriores. Este procedimiento puede ser realizado de manera indefinida. La simplicidad del ejemplo muestra que la asignación numérica (escala) es *arbitraria*, excepto que debe reflejar el orden cualitativo establecido u observado entre los objetos en cuestión.

Medir significa, por tanto, *asignar números tales que* $\phi(a)>\phi(b)$ *si y sólo si* $a>b$. No se trata sólo de definir una escala cualquiera tal que si $a>b$, entonces $\phi(a)>\phi(b)$, sino también que la escala sea tal que cuando $\phi(c)>\phi(d)$ entonces se pueda afirmar que $c>d$. Así, los problemas planteados por toda medición son los siguientes: (1) ¿qué propiedades debe satisfacer la relación de orden cualitativa que se observa o establece entre los objetos o atributos de una clase? Por ejemplo, puede requerirse un ordenamiento *total* de todos los elementos de la clase, de la misma manera en que están ordenadas las letras en el alfabeto; o tal vez podría tenerse una situación donde es imposible que todos los elementos de la clase sean comparables, por lo que sólo se puede definir un orden *parcial;* (2) ¿cómo asegurar la existencia de una escala de medición que preserve el orden cualitativo? Los resultados que se desean establecer dicen relación tanto con la *existencia* de una escala de medición (la función ϕ) como sus propiedades, todo lo cual depende esencialmente de las propiedades que el orden cualitativo satisface. Por ejemplo, si *A* es un conjunto de atributos, cuya cantidad de elementos es igual a todos los números naturales, y < es un orden simple[1], entonces *existe* una escala ϕ sobre *A* tal que para todo *a* y *b* elementos de *A*, $a<b$ *si y sólo si* $\phi(a)>\phi(b)$. Digamos de paso que cuando se habla de escalas ordinales, escalas afines o escalas de intervalos, lo que se está requiriendo

[1] Esto es, que para todo *a* y *b* elementos de *A*, se tiene que $a<b$ o $b<a$; y además la relación < es transitiva.

son ciertas propiedades que la función ϕ debe satisfacer. La escala ϕ satisface más propiedades cuanto más rica sea la estructura del orden cualitativo.

Estas consideraciones parecen suficientes para entrever lo complejo y fascinante que resulta el aplicar la teoría de medición en el ámbito de la medición educacional. Por lo pronto, baste decir que, si se quiere "medir", el objetivo de un instrumento administrado a un grupo de estudiantes es ayudar a definir u observar una relación de orden cualitativa para *comparar alumnos*, estableciendo un ordenamiento entre ellos en base a un atributo o rasgo medido por un instrumento. El segundo problema, de no menor complejidad, consiste en encontrar (si es que existe) una escala que represente numéricamente dichas ordenaciones.

Para terminar esta sección, es importante mencionar que los primeros resultados de Teoría de Medición fueron establecidos a finales del siglo antepasado y principios del pasado; véase Cantor (1895), Helmholtz (1895) y Hölder (1901). A pesar de esto, la teoría de medición lleva desarrollándose de manera muy formal desde hace más de 60 años; sus principales resultados aún siguen apareciendo en el *Journal of Mathematical Psychology*. Para una exposición bastante completa y formal de la teoría, véase Pfanzagl (1968), Ellis (1968), Krantz *et al.* (1971) y Michell (1990). Para aspectos filosóficos y metodológicos, véase, entre otros, Swistak (1990), Hand (1996), Michell (1997a, 1997b).

3. ¿Cuáles son las hipótesis subyacentes a la TCM?

Cuando aplicamos un modelo estadístico a un conjunto de datos, un objetivo fundamental que se persigue es extraer *la información relevante contenida en ellos.* En la década de los 20, los métodos estadísticos enfatizaban la especificación y estimación de la distribución de probabilidades que generan los datos observados. Fisher (1922) escribió uno de los artículos seminales a partir del cual se desarrolló esta perspectiva. Sin embargo, las aplicaciones en econometría y psicometría motivaron una reformulación de los puntos de vista de Fisher. A principios de la década de los 50, en un artículo publicado en una de las mejores revistas de estadística matemática, *The Annals of Mathematical Statistics*, Koopmans y Reiersøl (1950) propusieron una reformulación del problema de la modelización tal y como lo concebía Fisher: "en muchos campos de aplicación, el objetivo de todo investigador no es sólo la población en el sentido de la distribución de probabilidades de las variables observables, sino también una estructura proyectada subyacente a esta distribución, por la cual esta última se asume

es generada" (p. 165). Estas ideas no sólo encontraron en Koopmans y Reiersøl (1950) su formulación formal, sino que motivaron todo un campo de investigación en psicometría y econometría; para detalles, véase Goldberger (1971, 1972), Novick (1980), Cox (1990), Manski (1995), McCullagh (2002) y Mouchart y San Martín (2003), San Martín (2003).

La Teoría Clásica de los Tests, así como la Teoría de Respuesta al Ítem, se han desarrollado teniendo como telón de fondo estas ideas. Los esfuerzos han estado concentrados en justificar, en la medida de lo posible, los modelos estadísticos con *elementos sustantivos*. Así, no sólo interesa la bondad de ajuste de los modelos estadísticos, sino también su significación sustantiva. El caso extremo sería un modelo que ajusta con un 99% de bondad, pero que su contenido sustantivo es tan pobre que no sirve para explicar los datos, ni mucho menos para aprender de ellos. Basta pensar, por ejemplo, en una regresión lineal con una cantidad bastante grande de variables explicativas.

En el campo de la medición cabe entonces preguntarse cómo se puede evaluar la significación sustantiva de un modelo estadístico, información que debe ser complementada con las más rutinarias como "buen ajuste estadístico". Ya los antiguos psicometras que desarrollaron la Teoría Clásica de la Medición sugirieron ciertos lineamientos, los cuales han sido ampliamente desarrollados desde entonces. En efecto, no se trata de presentar un modelo estadístico, en este caso la TCM, listando un conjunto de supuestos o hipótesis, sino mostrar tanto su significación como sus implicaciones lógicas. Es necesario decir que estas consideraciones están ausentes de los actuales manuales de TCM (como de IRT) y, posiblemente, del uso profesional que se hace de los mismos. Sin embargo, si consultamos los trabajos originales, uno verá con sorpresa cómo estas consideraciones eran de gran importancia para el desarrollo de la teoría misma. Una de las referencias más importantes es Lord y Novick (1968), texto que no sólo desarrolla la TCM, sino también, y de manera integrada, los modelos IRT y los modelos factoriales[2] . Otros trabajos que es importante mencionar aquí son Guttman (1945, 1953), Novick (1966) y Novick y Lewis (1967), entre otros. En esta sección queremos revisar dichos elementos, utilizando una notación más reciente a fin hacer explícita las relaciones estructurales que hay entre la TCM y los modelos IRT.

[2] Dicho libro fue escrito en colaboración con A. Birnbaum, uno de los estadísticos más conocidos por sus desarrollos en fundamentos de la Estadística. Birnbaum fue invitado a escribir los capítulos del libro que presentaron formalmente los fundamentos de la teoría de respuesta al ítem.

3.1. La estructura básica de la TCM

La TCM es, en apariencia, un modelo bastante simple. La variable dependiente corresponde al *puntaje total observado* de una persona particular. Las variables independientes son el *puntaje verdadero* y el *error de medición.* Se asume que estas variables están relacionadas *aditivamente.* Las hipótesis básicas del modelo son las siguientes: (a) el valor esperado del error de medición es cero; (b) el error de medición para una persona determinada no está relacionado con otras variables como el puntaje verdadero y el error de medición asociado a otros ítemes (véase Crocker y Algina, 1986; Gulliksen, 1950; Thissen y Wainer, 2001). Sin embargo, los "supuestos'" (a) y (b) pueden ser deducidos de hipótesis más fundamentales, muchas de las cuales subyacen también a la Teoría de Respuesta al Ítem (IRT). En efecto, la TCM distingue dos procesos de aleatorios de selección: (i) la selección de los individuos, los cuales denotaremos por *s*; (ii) la selección de los ítemes respondidos por un individuo *s* dado, los cuales son denotados por *i.* La primera hipótesis fundamental subyacente a cualquier teoría de medición es que las observaciones miden características individuales que no son directamente observables. Más precisamente,

> *Hipótesis 1*: La respuesta del individuo *s* al ítem *i* mide un rasgo latente unidimensional, el cual caracteriza sólo al individuo *s.* Dicha variable la denotamos por θs.

Para hacer explícita la relación entre la observación (medición) con el constructo latente θs denotamos por *Xis* la respuesta del individuo *s* al ítem *i*. Entonces *siempre* podemos escribir la siguiente relación aditiva:

$$X_{is} = E[X_{is} \mid \theta = \theta_s] + (X_{is} - E[X_{is} \mid \theta = \theta_s])$$

donde $E[Xis \mid \theta = \theta s]$ denota la esperanza condicional de *Xis* dado el constructo o variable latente θs. Esta esperanza condicional asume implícitamente que el individuo representado por θs ha sido escogido con respecto al proceso de selección (ii) mencionado arriba. En TCM, dicha esperanza condicional se llama *puntaje verdadero,* mientras que la diferencia $Xis - E[Xis \mid \theta = \theta s]$ se llama *error de medición.* Ambas cantidades *son funciones de la variable latente* θs y, en consecuencia, no son observables directamente. Si $\tau i(\theta s)$ denotando el puntaje verdadero, y $\varepsilon i(\theta s)$ denota el error de medición, la relación aditiva anterior se escribe equivalentemente como $Xis = \tau i(\theta s) + \varepsilon i(\theta s)$. En el caso dicotómico (i.e. cuando cada ítem

tiene una y sólo una alternativa correcta) el puntaje verdadero se reduce a la probabilidad condicional de responder correctamente un ítem *i* dada la variable latente θs, esto es, $\tau i(\theta s) = P[Xis=1 \mid \theta = \theta s]$.

Una pregunta relevante es la siguiente: ¿qué relación funcional hay que suponer para expresar el puntaje verdadero $\tau i(\theta s)$? Es evidente que esta pregunta no tiene una respuesta única. La TCM supone que esta es una relación lineal en θ asumiendo por tanto que τ es una función lineal y que todos los ítemes de una prueba miden un mismo rasgo latente; véase Novick (1966). Hipótesis adicionales pueden ser introducidas a fin de obtener otras relaciones funcionales que deban ser ajustadas a los datos. Sin embargo, en lo que resta de esta sección, seguiremos haciendo la discusión independiente de la forma funcional, pues los resultados no dependen de la misma.

Usando las propiedades básicas de la esperanza condicional y sus relaciones con la covarianza[3] es inmediato deducir que (i) $E[\varepsilon i(\theta s)] = E[E(\varepsilon i(\theta s) \mid \theta = \theta s] = 0$, y (ii) $\text{cov}(\tau i(\theta s), \varepsilon i(\theta s)) = 0$. Estas propiedades son habitualmente presentadas como "supuestos" del modelo, sin embargo dependen de la hipótesis 1 ya introducida. Para obtener la propiedad según la cual la correlación entre los errores de medición debidos a dos ítemes distintos es nula se obtiene a partir de la siguiente hipótesis, que no es otra cosa que el *axioma de independencia local*:

> *Hipótesis 2:* Para cada individuo *s*, las mediciones *X1s, ..., Xis, ..., Xms* a los *m* ítemes de una prueba son condicionalmente independientes dada la variable latente θs.

El axioma de independencia local afirma que las cantidades latentes (en el caso de la TCM, los puntajes verdaderos) son los únicos factores importantes, y que una vez que éstos han sido determinados, el comportamiento es aleatorio. Una excelente discusión heurística de este axioma puede encontrarse en Lazarsfeld (1959). Otras referencias importantes son Anderson (1959), Novick (1966), Holland y Rosenbaum (1980), Bartholomew (1987) y Sobel (1997). Así, usando otra propiedad fundamental de la covarianza[4], se deduce que la hipótesis 2 implica que $cor(\varepsilon i(\theta s), \varepsilon j(\theta s))=0$ para dos ítemes distintos *i* y *j*.

---

[3] En particular, las siguientes dos propiedades: (i) si E(X) existe, E(X)=E[E(X | Y)]; (ii) si E(X), E(Y) y E(XY) existen, entonces cov(X,Y)=cov[X,E(Y | X))].

[4] A saber, si X, Y y Z son variables aleatorias tales que sus esperanzas existen, entonces cov(X,Y)=E[cov(X,Y | Z)] + cov[E(X | Z),E(Y | Z)].

3.2. Algunas propiedades de los ítemes con respecto a la muestra

Cuando se analizan los ítemes de una prueba, interesa estudiar su comportamiento con respecto a una determinada muestra o *población de referencia*. Para la discusión que sigue, consideremos el caso dicotómico, siendo la extensión a otros casos inmediata. La dificultad media de un ítem *i* con respecto a una población de referencia de tamaño *n* se define por

$$X_i = \frac{1}{n}\sum_{s=1}^{n} X_{is}$$

Tanto el puntaje verdadero como el error de medición de un ítem dado pueden ser expresados en términos de una población de referencia. Para ello, es necesario considerar el proceso de selección aleatoria (ii) mencionado arriba. Así, el puntaje verdadero y el error de medición de un ítem *i* con respecto a una población de referencia están respectivamente dados por

$$\tau_i \equiv E[\tau_i(\theta)], \qquad \tau_i(\theta) = \frac{1}{n}\sum_{s} \tau_i(\theta_s) \qquad y \qquad \varepsilon_i(\theta) = \frac{1}{n}\sum_{s} \varepsilon_i(\theta_s)$$

Las propiedades discutidas en la sección anterior se heredan a nivel poblacional. Por lo tanto, las hipótesis 1 y 2 anteriormente introducidas implican que, con respecto a una población de referencia, la TCM satisface las siguientes propiedades:

(i) $Xi = \tau i(\theta) + \varepsilon i(\theta)$, para todo $i=1,...,m$.
(ii) $E[\varepsilon i(\theta),] = 0$, para todo $i=1,...,m$.
(iii) $cor(\tau i(\theta), \varepsilon i(\theta)) = 0$, $i = 1,...,m$.
(iv) $cor(\varepsilon i(\theta), \varepsilon j(\theta)) = 0$ para todo *i* distinto de *j*.

En particular, se deduce que $E[\tau i] = E[Xi]$, esto es, que el puntaje verdadero esperado del ítem *i* es igual a su puntaje observado esperado.

Un "buen ítem" será aquel que tiene una "alta'' correlación entre su puntaje observado *X*i y su puntaje verdadero τi. De las propiedades anteriores se deduce que

$$cor^2(X_i, \tau_i(\theta)) = \frac{Var(\tau_i(\theta))}{Var(X_i)}.$$

Esta relación representa, por tanto, la precisión con que diferencias en el puntaje verdadero entre personas son estimadas por la diferencia entre

el puntaje observado entre personas. Por ello, esta relación se llama *confiabilidad del ítem i.* Mencionemos, sin embargo, que dicha confiabilidad no es calculable directamente pues depende de una cantidad no observable, a saber, θ. Es posible expresar la confiabilidad de un ítem bajo hipótesis adicionales, las cuales dicen relación con el concepto de medidas paralelas. Para detalles, véase Lord y Novick (1968, capítulo 3).

En conclusión, la TCM depende de dos hipótesis fundamentales: primero, de que las observaciones miden, con error, un rasgo latente, el cual a su vez caracteriza a cada individuo; y, segundo, del axioma de independencia local.

4. Los supuestos básicos de los modelos IRT

En la sección anterior, hemos mencionado el problema de cómo especificar el puntaje verdadero. La Teoría de Respuesta al Ítem ofrece respuestas a este problema. En efecto, suponiendo las hipótesis 1 y 2 asumidas por la TCM, la teoría IRT introduce hipótesis adicionales que permite *derivar relaciones funcionales específicas*. En efecto, los llamados *modelos 1PL* o *modelo Rasch* asumen las siguientes hipótesis adicionales:

> *Hipótesis 3:* la función de distribución $P[Xis=1 \mid \theta = \theta s]$ como función de θ es continua y estrictamente.
>
> *Hipótesis 4:* El puntaje total de cada individuo obtenido en una prueba es un estadístico suficiente para θs. Es decir, el puntaje total contiene toda la información relevante provista por las observaciones *X1s,..., Xis,..., Xms*

Usando las hipótesis 1, 2, 3 y 4 se deduce que existen constantes $\beta 1,..., \beta i,..., \beta m$ tales que

$$P\left[X_{is}=1 \mid \theta=\theta_s\right]=\frac{\exp\left(a\left(\theta_s-\beta_i\right)\right)}{1+\exp\left(a\left(\theta_s-\beta_i\right)\right)}$$

Cuando $a=1$, este modelo se conoce como *modelo Rasch.* Para detalles, véase Fischer (1995a, 1995b) y Junker (2001). Como es sabido, los βi's representan la dificultad de cada ítem *i*. Así, el modelo 1PL puede ser visto como una extensión de la TCM en el sentido de introducir hipótesis adicionales a fin de especificar funcionalmente el puntaje verdadero de la TCM.

El modelo 1PL, y por tanto sus hipótesis subyacentes, implican una representación en una escala común del rasgo latente θs y de los parámetros de dificultad βi de los ítemes. Más aún, ambos parámetros están representados en la *escala del logito*. En efecto, si denotamos por $P(Xis)$ la relación funcional que define el modelo 1PL, entonces (para a=1, siendo los otros casos de misma interpretación)

$$\ln \frac{P(X_{is})}{1-P(X_{is})} = \theta_s - \beta_i .$$

El cuociente de probabilidades corresponde a la *razón de chance* que un individuo *s* tiene de responder correctamente el ítem *i* y de responder incorrectamente el mismo ítem. Dicha razón depende de dos características diferentes: la primera debida al individuo *s* (presente, por medio de la hipótesis 1, en la TCM), la segunda debida a los ítemes (presente en el modelo IRT gracias a las hipótesis 3 y 4). Cuando $\theta s > \beta i$, el individuo *s* tiene mayor *chance* de responder correctamente el ítem que de responderlo incorrectamente; la constante de proporcionalidad está dada por $exp(\theta s - \theta i)$. Similarmente, cuando $\theta s < \beta i$, el individuo *s* tiene menor *chance* de responder correctamente el ítem que de responderlo incorrectamente. Hay un punto de inflexión cuando $\theta s = \beta i$, lo cual significa que la probabilidad de responder correctamente el ítem es igual a la probabilidad de responderlo incorrectamente; de hecho, ambas son iguales a 0.5.

De la ecuación de razón de *chance* ya mencionada, se puede deducir de forma inmediata que la diferencia entre el logaritmo de la razón de *chance* del individuo *s* con respecto al ítem *i* y con respecto al ítem *j* es igual a la diferencia βi-βj. En otras palabras, la escala en que se ordenan los ítemes es equivalente a la escala del logito en que se ordenan las razones de *chances*.

Una observación importante es la siguiente: para una persona *s* dada (luego, condicionalmente a θs), su puntaje total $X1s + X2s + ... + Xms$ es utilizado en la TCM como una estimación de su puntaje verdadero en una prueba de largo infinito; véase Lord y Novick (1968, sección 5.4). En los cálculos que se hacen habitualmente en TCM, se reporta el puntaje total observado (o la proporción correspondiente). Siendo que el objetivo de cualquier teoría de medición es ordenar dos o más personas, la TCM ordena (compara) a los individuos de acuerdo a su puntaje total observado. Puesto que el puntaje total de la persona *s* en la prueba es un estadístico suficiente para el rasgo latente θs, el modelo 1PL *produce la misma ordenación entre los individuos que la producida por la TCM*. Desde este punto de vista, no hay diferencia alguna entre ambos modelos (o conjunto de hipótesis).

Todas estas propiedades estructurales del modelo son debidas a las hipótesis 3 y 4 añadidas a las hipótesis 1 y 2 usadas por la TCM. Como hemos ya ilustrado en los párrafos anteriores, esto permite obtener un modelo más rico en términos de interpretación que lo que podemos hacer con TCM.

Birnbaum desarrolló el llamado modelo 2Pl en el capítulo 17 de Lord y Novick (1968). La relación funcional que se asume es la siguiente:

$$P[X_{is} = 1 \mid \theta = \theta_s] = \frac{\exp(a_i(\theta_s - \beta_i))}{1 + \exp(a_i(\theta_s - \beta_i))}$$

Los parámetros αi's son interpretados como los parámetros de discriminación de cada ítem *i*. Este modelo también puede ser deducido a partir de hipótesis adicionales a las hipótesis 1, 2 y 3 mencionadas arriba. Referimos al lector interesado a Fisher (1987, 1994a). Un detalle importante que merece ser mencionado es que el puntaje ponderado es también un estadístico suficiente para el rasgo latente θs. Esto significa que todos los individuos teniendo el mismo puntaje ponderado deben tener también una misma estimación para θs. En particular, individuos con diferentes patrones de respuesta pueden tener el mismo puntaje ponderado. El modelo 2PL introduce entonces la importancia o "peso" que se les dan a los diferentes ítemes de una prueba. Para otros detalles de estimación y discusión de los submodelos, véase Hambleton, Swaminathan y Rogers (1991) y Embretson y Reise (2000).

$$X_i = \frac{1}{n} \sum_{s=1}^{n} a_i X_{is}$$

Al igual que en el modelo 1PL, se puede deducir que la razón de *chance* de un individuo *s* con respecto a un ítem *i* es igual a $ai\ (\theta s - \beta i)$. Así, el parámetro de discriminación *ai* indica cuán mayor o menor es la probabilidad del individuo para responder correcta e incorrectamente el ítem. Dicho sea de paso que esto permite tener una caracterización alternativa de lo que significa el parámetro de discriminación.

5. Ventajas de los modelos IRT en relación a la TCM

Como hemos visto en las secciones anteriores, el enfoque IRT representa una extensión natural de la TCM, compartiendo algunos de sus supuestos fundamentales. Pese a sus evidentes similitudes con respecto a sus supuestos básicos, es importante apreciar también sus diferencias, pues

tanto en el plano conceptual como en sus aplicaciones prácticas, estas teorías poseen diferencias que es necesario considerar antes de resolver cuál de ellas es más pertinente para resolver qué tipo de problemas de medición. Estas diferencias son producidas por las hipótesis adicionales que el enfoque IRT asume, hipótesis que ayudan a extraer mayor información no sólo de los individuos, ni sólo de los ítemes, sino también de sus interacciones.

5.1. La relación entre puntajes observados y rasgos latentes

Ambas teorías se fundan en afirmar una relación monotónicamente creciente entre los puntajes observados en un instrumento de medición y el atributo latente θ al mismo instrumento. Esta relación es conceptualmente razonable si se asume, como lo hacen ambas teorías, que el instrumento de medición evalúa fundamentalmente un atributo latente, esto es, se asume la unidimensionalidad del rasgo latente (véase hipótesis 1)[5]. Como fue discutido en las secciones anteriores, ambas teorías difieren en la forma funcional específica que postulan para la relación monótona creciente: la TCM plantea un supuesto más simple, sosteniendo que esta relación es lineal, mientras que el enfoque IRT plantea como un postulado fundamental que dicha relación es no lineal (véase hipótesis 3). En este último caso, un determinado instrumento de medición tiene la posibilidad de discriminar en ciertos rangos bajo y alto de la escala de θ. Por ello existe un comportamiento asintótico inferior y superior en los puntajes observados, si se considera un rango suficientemente grande de θ.

¿Cuál de los planteamientos parece más adecuado para modelar datos obtenidos en mediciones como las educacionales? Si bien es conceptualmente indiscutible que el planteamiento del enfoque IRT es más completo y consistente con lo que cabe esperar de la relación entre θ y puntajes observados[6], es evidente que se trata, al mismo tiempo, de un modelo más complejo. Por otro lado, el planteamiento de la TCM es evidentemente más simple y, por lo mismo, más parsimonioso. Esta competencia entre precisión y parsimonia explica en gran medida la preferencia que en términos prácticos llevan a emplear uno u otro modelo. Quienes privilegian la sim-

[5] En ambas teorías la evaluación de este supuesto se lleva a cabo fundamentalmente con el apoyo del análisis factorial exploratorio. Para probar que un instrumento es unidimensional usualmente se compara la magnitud del valor propio asociado al primer factor con el de los restantes valores propios.

[6] Es relevante mencionar aquí que las discusiones de Rasch en cuanto a la pertinencia de los modelos IRT se basaron no sólo en caracterizaciones matemáticas, sino también en experimentos sustantivos. Para un excelente resumen acerca de cómo surgió el modelo Rasch, véase Andersen y Olsen (2001).

plicidad, probablemente preferirán el supuesto lineal. Quienes, en cambio, prefieren un modelo que resuelva en forma conceptualmente más precisa el problema planteado al comienzo de esta sección, preferirán IRT.

¿Qué nos muestra la realidad? Es útil ilustrar el dilema planteado a partir de datos empíricos basados en aplicaciones prácticas de las teorías de la medición. Con este propósito, trabajaremos con datos obtenidos de la aplicación de la prueba SIMCE, puesto que, por su carácter censal, se dispone de información que potencialmente refleja la realidad de toda la población de interés. En términos concretos, la mayoría de los ejemplos que se discutirán en el presente trabajo se basan en datos provenientes de la última aplicación de la prueba SIMCE de matemáticas a estudiantes de 2º medio, utilizando una de las formas de dicha prueba. Para el problema que nos interesa, la relación entre atributo latente y puntaje observado, mostraremos en primer lugar la aproximación más básica a dicha relación que puede derivarse de la TCM: la relación entre puntaje total en la prueba y la probabilidad de responder correctamente cada uno de los ítemes que conforman dicha prueba. Como se indicó más arriba, la TCM emplea como estimador de θ el puntaje total en un conjunto de ítemes (la suma o promedio de las respuestas que los examinados producen en el conjunto de ítemes). Para ilustrar la relación, se agruparon todos los examinados en rangos de 4 puntos en el total de ítemes[7]; luego se calculó el promedio de respuesta correcta de cada grupo para cada uno de los 45 ítemes de la prueba. Tal como puede apreciarse en los gráficos de la Figura 1 del Anexo, todos los ítemes muestran un comportamiento monotónicamente creciente con respecto al puntaje total. Adicionalmente, se observa que la mayoría de ellos presenta una relación no lineal con el puntaje total: en algunos casos (como el ítem 6), con una evidente asíntota inferior, en otros (como el ítem 1), con una clara asíntota superior, y en otros casos con una forma cercana a una ojiva. Todos estos patrones son consistentes con los planteamientos del enfoque IRT.

5.2. Los supuestos acerca del error de medición

Como discutimos en la sección 3.1., toda teoría de la medición asume que las respuestas observadas miden, con error, los atributos. Los errores de medición se definen con respecto a una población (como lo desarrollado en la sección 3.2.) o con respecto a un instrumento o test. En

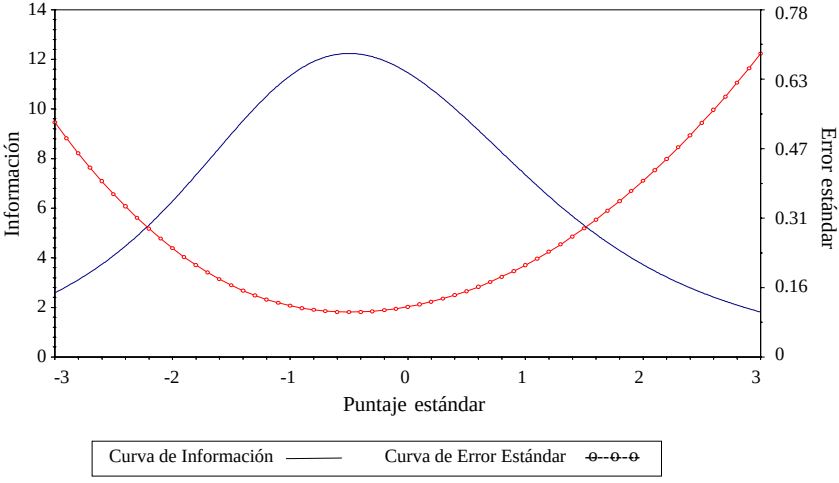
[7] Para estimar los puntajes totales de dicha prueba se emplearon 45 ítemes. Por tanto, los puntajes observados podían variar entre 0 y 45 puntos. Para agrupar los examinados, se crearon intervalos de 4 puntos en el puntaje total. La primera categoría, sin embargo, abarca de 0 a 9 puntos, para asegurar que un número mínimo de examinados quedara en cada categoría. Ninguna categoría incluye menos de 1.200 casos.

este último caso, los supuestos de la TCM implican que dicho error es constante para un mismo instrumento, transformándose así en un atributo que caracteriza la calidad de la medición y de las inferencias que pueden hacerse a partir del instrumento; véase Lord y Novick (1968, capítulo 3). Sin embargo, la intuición sugiere que dicho error difícilmente es constante: basta conocer cómo se construyen los instrumentos de medición para concluir que la precisión de la medición seguramente es desigual. Por ejemplo, si se está construyendo una prueba de carácter selectivo, seguramente se seleccionarán más preguntas difíciles que fáciles, con lo que la precisión debiera ser mayor en la región superior de los puntajes que en la inferior. Lamentablemente, la TCM emplea un supuesto simplificador, que en este caso puede resultar peligroso, puesto que lleva a muchos usuarios de instrumentos de medición a pensar que la calidad con que se está evaluando es constante, ignorando que el error de medición puede ser sensiblemente más alto en ciertas regiones de los atributos evaluados.

Los supuestos subyacentes al enfoque IRT permiten subsanar esta situación, incorporándose la noción de error no constante; en efecto, se deduce que dicho error varía de manera inversamente proporcional a la información que un instrumento de medición posee. Así, tal como puede observarse en la Figura 2, el error, siguiendo lo que nuestra intuición sugiere, es mayor en las regiones extremas de la escala de medición, y menor en la zona media. Esto refleja además el hecho que hay menos individuos con patrones de respuestas en las zonas extremas que en las zonas medias. Es importante tener presente que no necesariamente un instrumento debe mostrar el patrón que aparece en la Figura 2. De hecho, se puede construir un instrumento que resulte especialmente informativo en la región superior o inferior de un determinado atributo, si las aplicaciones prácticas del instrumento así lo aconsejen. Cómo resolver el problema de diseñar un instrumento de medición para satisfacer propósitos específicos, es una tarea para la cual IRT provee herramientas conceptuales y técnicas precisas. De hecho, aplicaciones de creciente importancia en la medición internacional, tales como la medición adaptativa o la medición basada en computadores (computer based testing), se fundan en las reglas del escalamiento óptimo que ofrece IRT[8]. Para este tipo de aplicaciones, la TCM no ofrece reglas ni principios que puedan resolver en forma satisfactoria y precisa el problema. De hecho, la única regla claramente deducible de la TCM cuando se desea incrementar la precisión de un instrumento, es incrementar su longitud, basándose en la profecía de Sperman-Brown (véase Lord y Novick, 1968;

[8] El escalamiento óptimo alude a la selección de ítemes que otorgan el máximo de información y, en consecuencia, el mínimo de error, en determinada región del atributo que interesa evaluar.

FIGURA 2: FUNCIÓN INFORMATIVA Y ERROR ESTÁNDAR DE MEDICIÓN DE LA PRUEBA SIMCE DE MATEMÁTICA



Crocker y Algina, 1986), lo que en muchos casos lleva a soluciones impracticables. En cambio, como lo hace ver Embretson y Reise (2000), IRT provee indicaciones directas para aumentar la precisión de un instrumento, sin que ello conlleve una extensión de su longitud.

5.3. Dependencia entre estimaciones de habilidad e ítemes de una prueba

Las teorías de la medición se basan en el supuesto que una mejor aproximación al puntaje verdadero de una persona se obtiene cuando se emplean varios indicadores o ítemes. En estas condiciones, lo deseable sería que, para un mismo número de ítemes, las estimaciones de habilidad que se obtengan sean relativamente independientes de los ítemes específicos que se empleen. Lamentablemente esta segunda condición no está incorporada en la TCM. En este enfoque, la estimación de habilidad está intrínsecamente ligada a los ítemes específicos que conforman una prueba. Si se alteran o cambian los ítemes, aunque los nuevos midan el mismo rasgo o atributo latente, se obtienen distintas estimaciones de habilidad. En términos prácticos, es fácil deducir que si de un conjunto de ítemes se extrae una muestra de un cierto tamaño de ítemes fáciles y otra de igual tamaño de ítemes difíciles, la mayor parte de los examinados obtendrá un

mayor puntaje (y por tanto, estimación más alta de su habilidad) si se les administra el primer conjunto de ítemes. En cambio, el enfoque IRT incorpora en la estimación de habilidad la calidad de los ítemes que han sido abordados y respondidos. Así, una persona que sólo responde correctamente un subconjunto de preguntas fáciles obtendrá en este enfoque un menor puntaje que una persona que responde correctamente un mismo número de preguntas, si estas últimas son más difíciles. Desde este punto de vista, la selección específica de los ítemes que conforman una prueba es una decisión que tiene consecuencias más críticas en la TCM que en el enfoque IRT.

Lo anterior tiene, nuevamente, importantes consecuencias en el diseño de pruebas. Para obtener una estimación adecuada de la habilidad de una persona, en el enfoque de la TCM se requiere que quienes son examinados sean expuestos a una muestra más o menos representativa del conjunto de ítemes que es posible concebir para el dominio que se está evaluando. En términos prácticos, esta condición es muy difícil de satisfacer, puesto que las pruebas sólo pueden contener un número relativamente limitado de ítemes o reactivos. Por lo demás, la TCM tampoco provee reglas precisas para establecer si se dispone de un banco de preguntas relativamente representativo. En cambio, empleando las reglas del enfoque IRT, es menos crítico cuáles ítemes específicos sean administrados a un examinado. Si los parámetros de los ítemes que se administran han sido calibrados en forma conjunta, una muestra relativamente pequeña de preguntas permite obtener estimaciones adecuadas de habilidad, y por lo demás comparables entre distintos examinados, aun si distintas personas son expuestas a diferentes preguntas.

5.4. Dependencia de los parámetros de las muestras en que son estimadas

Según lo visto hasta aquí, un aspecto fundamental de la construcción de instrumentos de medición es la calidad de los ítemes que conforman una prueba. Tanto la TCM como el enfoque IRT contemplan parámetros para evaluar la calidad de los ítemes disponibles. Los parámetros más importantes en la tradición de la TCM son los que aluden al grado de dificultad[9] de

[9] El término grado de dificultad es algo equívoco, puesto que es sólo pertinente para ítemes que miden habilidades cognitivas. Cuando se miden otros aspectos, como por ejemplo, actitudes, el promedio de respuesta de un ítem no puede interpretarse como grado de dificultad. Por otra parte, cuando la escala de respuesta no es dicotómica, el promedio de un ítem tampoco corresponde a una proporción de respuesta correcta, como es usualmente identificado el grado de dificultad. En consecuencia, el concepto más general sería la posición promedio del ítem en la escala de respuesta que se esté empleando.

un ítem y a su capacidad discriminativa. En este enfoque, un ítem es apropiado en la medida que su grado de dificultad se encuentra dentro de un rango aceptable (típicamente entre 0,2 y 0,8, o entre 0,1 y 0,9), y en la medida que su capacidad discriminativa sea adecuada (habitualmente esto se especifica con exigencias mínimas para la correlación entre el puntaje en el ítem y el puntaje en la prueba total[10]). En el enfoque IRT, las hipótesis adicionales subyacentes al modelo muestran que las características de un ítem son fundamentales para definir el puntaje verdadero de un individuo (o probabilidad de responder correctamente un ítem). En efecto, como vimos en la sección 4, el modelo de 1 parámetro (o modelo Rasch) incluye la posición de cada ítem (que tiene una interpretación análoga al grado de dificultad en la TCM). Más aún, el modelo de 1 parámetro también considera la capacidad discriminativa de los ítemes, la cual se asume constante para todos ellos. El modelo de 2 parámetros, por su parte, considera que tanto la posición como la capacidad discriminativa de los ítemes son atributos específicos para *cada* ítem. Mencionemos que existe un tercer modelo que involucra 3 parámetros; la relación funcional entre la variable observada y el atributo latente es tal que se acepta la posibilidad que la asíntota inferior de dicha función no se sitúe en 0. Típicamente esto último ocurre en ítemes de respuesta cerrada, donde existe la probabilidad de responder correctamente al azar.

Lo deseable sería que la estimación que se obtuviera de los parámetros de los ítemes pudiera ser relativamente independiente de las características de las muestras involucradas en su estimación. Lamentablemente ello no puede ser garantizado en la TCM, puesto que por definición el grado de dificultad de un ítem está definido como el promedio que obtienen en él quienes responden dicho ítem. Por tanto, si la muestra incluye a personas con un mayor nivel de habilidad, la proporción de respuestas correctas debiera ser, por definición, mayor que si el ítem es respondido por un grupo de examinados con menor nivel de habilidad. En consecuencia, la dificultad no es en este caso un atributo del ítem, sino que una interacción entre el ítem y los examinados que lo responden. Con respecto a la capacidad discriminativa, se puede asumir que será relativamente menos dependiente que el grado de dificultad de las características de la muestra, condicionado al grado de variabilidad que exista en la muestra. En contraste con lo anterior, el enfoque IRT separa la información debida a un individuo y la información debida a un ítem; esto permite decir (como fue discutido ya en la sección 4) que los parámetros que caracterizan los ítemes son relativa-

[10] Cuando los ítemes son puntuados dicotómicamente esta correlación adopta la forma de una correlación biserial o de una correlación biserial-puntual.

TABLA 1: PARÁMETROS TCM E IRT DE 45 ÍTEMES DE PRUEBA SIMCE DE MATEMÁTICAS

Ítem	PropCorrecta	Desv Est	$Corr_{it\text{-}test}$	$a_{j(todos)}$	ErrEst(a_j)	$b_{j(todos)}$	ErrEst(b_j)
m1	0.80	0.40	0.28	0.64	0.01	-1.55	0.02
m2	0.48	0.50	0.38	0.60	0.01	0.06	0.01
m3	0.59	0.49	0.32	0.52	0.01	-0.54	0.01
m4	0.58	0.49	0.37	0.63	0.01	-0.42	0.01
m6	0.21	0.40	0.31	0.49	0.01	1.83	0.02
m7	0.82	0.39	0.22	0.46	0.01	-2.12	0.04
m8	0.74	0.44	0.38	0.89	0.01	-0.97	0.01
m9	0.55	0.50	0.41	0.70	0.01	-0.29	0.01
m10	0.43	0.50	0.29	0.40	0.01	0.41	0.01
m11	0.57	0.49	0.37	0.62	0.01	-0.39	0.01
m12	0.48	0.50	0.30	0.45	0.01	0.10	0.01
m13	0.83	0.37	0.20	0.45	0.01	-2.31	0.04
m14	0.64	0.48	0.38	0.70	0.01	-0.66	0.01
m15	0.51	0.50	0.42	0.72	0.01	-0.11	0.01
m16	0.47	0.50	0.34	0.50	0.01	0.14	0.01
m17	0.49	0.50	0.33	0.50	0.01	-0.01	0.01
m18	0.42	0.49	0.38	0.59	0.01	0.31	0.01
m19	0.61	0.49	0.37	0.64	0.01	-0.56	0.01
m20	0.85	0.36	0.27	0.73	0.01	-1.70	0.02
m21	0.75	0.43	0.40	1.05	0.01	-0.95	0.01
m22	0.72	0.45	0.40	0.98	0.01	-0.87	0.01
m23	0.46	0.50	0.46	0.80	0.01	0.09	0.01
m24	0.48	0.50	0.37	0.56	0.01	0.06	0.01
m26	0.30	0.46	0.25	0.37	0.01	1.46	0.02
m27	0.36	0.48	0.29	0.42	0.01	0.83	0.02
m28	0.26	0.44	0.24	0.34	0.01	1.94	0.03
m29	0.58	0.49	0.42	0.78	0.01	-0.41	0.01
m30	0.63	0.48	0.44	0.91	0.01	-0.56	0.01
m31	0.51	0.50	0.37	0.59	0.01	-0.08	0.01
m32	0.64	0.48	0.39	0.76	0.01	-0.64	0.01
m33	0.50	0.50	0.39	0.60	0.01	-0.07	0.01
m34	0.61	0.49	0.43	0.83	0.01	-0.48	0.01
m35	0.34	0.47	0.40	0.63	0.01	0.72	0.01
m36	0.38	0.49	0.48	0.84	0.01	0.38	0.01
m37	0.37	0.48	0.34	0.52	0.01	0.67	0.01
m38	0.36	0.48	0.52	0.92	0.01	0.42	0.01
m39	0.38	0.49	0.36	0.56	0.01	0.56	0.01
m40	0.40	0.49	0.24	0.35	0.01	0.73	0.02
m42	0.53	0.50	0.38	0.41	0.01	2.22	0.03
m43	0.67	0.47	0.27	0.62	0.01	-0.21	0.01
m44	0.80	0.40	0.34	0.44	0.01	-1.09	0.02
m45	0.47	0.50	0.37	0.91	0.01	-1.21	0.01
m46	0.42	0.49	0.23	0.58	0.01	0.09	0.01
m47	0.31	0.46	0.16	0.32	0.01	0.60	0.02
m48	0.43	0.50	0.41	0.23	0.01	2.05	0.05
Promedio	0.53	0.47	0.35	0.61	0.01	-0.06	0.01

mente independientes del grupo de examinados que se emplee para estimarla.

Para ilustrar este aspecto, emplearemos nuevamente los datos del SIMCE como marco de referencia. La Tabla 1 contiene para cada uno de los 45 ítemes sus parámetros TCM e IRT estimados con todos los casos. Para evaluar la estabilidad de ellos en submuestras, se obtuvieron muestras al azar de tamaño algo superiores a los 2.000 estudiantes bajo 2 condiciones: muestras del conjunto de todos los examinados y muestras de grupos que poseen importantes diferencias de puntaje, como es el caso de quienes asisten a establecimientos de distinta dependencia[11] (municipalizados, particulares subvencionados y particulares pagados). En cada caso se estimaron los parámetros TCM e IRT y luego se compararon sus valores con los estimados en el conjunto de la población.

Los resultados, que se resumen en la Tabla 2, muestran con claridad que cuando se emplean muestras al azar, se obtienen estimaciones muy cercanas a los valores poblacionales tanto de los parámetros IRT como de

TABLA 2: CORRELACIONES ENTRE PARÁMETROS TCM E IRT DE LOS ÍTEMES ESTIMADOS EN DISTINTOS SUBGRUPOS
(Las correlaciones incluyen 45 ítemes de una de las formas de la prueba SIMCE de Matemática.)

	Grado de dificultad		Capacidad discriminativa	
	TCM (Pj)	IRT (bj)	TCM (rit-test)	IRT (aj)
Población - Aleatoria 1	1.00	1.00	.98	.97
Población - Aleatoria 2	1.00	1.00	.98	.98
Población - Particulares	.86	.96	.67	.63
Población - Subvencionados	1.00	.99	.95	.95
Población - Municipalizados	.99	.99	.92	.94
Particulares - Subvencionados	.87	.94	.54	.54
Particulares - Municipalizados	.82	.93	.45	.43
Subvencionados - Municipalizados	.98	.99	.95	.95

Población: estimaciones basadas en todos los estudiantes del país que rindieron la prueba.

Aleatoria 1 y Aleatoria 2 son dos muestras independientes obtenidas al azar.

Particulares, Subvencionados y Municipalizados corresponden a muestras obtenidas al azar de estudiantes de cada una de las dependencias.

Todas las muestras incluyen unos 2.200 estudiantes.

[11] Es importante considerar que estos 3 tipos de establecimientos poseen tamaños desiguales en el país, lo que explica que los valores poblacionales se asemejen más a los grupos de mayor tamaño (los municipalizados representan el 47,5% del total de estudiantes, los subvencionados el 42,9% y los particulares el 9,6%).

los parámetros de la TCM. Sin embargo, cuando se emplean grupos con diversa habilidad como base para estimar los parámetros, se obtienen resultados más divergentes. En concordancia con los supuestos teóricos, el grado de dificultad es menos consistente al estimarlo con TCM que al hacerlo con IRT. Por su parte, las estimaciones de la capacidad discriminativa de los ítemes muestran correlaciones relativamente semejantes, aunque algo inferiores a las obtenidas para los parámetros asociados al grado de dificultad.

5.5. Representación simultánea de habilidades y dificultades

Como hemos mencionado en la sección 4, los modelos IRT permiten una representación simultánea de habilidades y dificultades. Esto representa sin duda una ventaja adicional del modelo IRT sobre el de la TCM. En efecto, en la teoría clásica, las habilidades de los individuos y las dificultades de los ítemes son representadas en escalas diferentes: el grado de dificultad de un ítem se expresa (habitualmente) como una proporción de respuesta correcta, mientras que la habilidad de las personas se expresa como un puntaje total (en su forma bruta, como suma o promedio de respuestas correctas, o en su forma estandarizada, luego de transformar los puntajes a otra escala).

Esta representación simultánea habilidad-dificultad producida por los modelos IRT ofrece ventajas interpretativas adicionales. En efecto, gracias a ella es posible referir el rendimiento de un examinado al tipo de ítemes cuya localización está por debajo o por encima de él; esto, a su vez, facilita la posibilidad de construir estándares para interpretar los resultados de una prueba. Por otra parte, cuando la ubicación de los ítemes está en la misma escala que los examinados, es posible identificar con facilidad el tipo de ítemes que resultaría más informativo para evaluar a una determinada persona. Este es el principio que orienta la medición adaptativa, especialmente en su implementación secuencial, donde usualmente con el apoyo de un computador, se determina en cada etapa de una medición cuál sería el ítem que resultaría más informativo para estimar la habilidad de una persona, a partir del rendimiento que dicha persona haya obtenido en los ítemes que se le hayan administrado previamente. Esta propiedad, que no tiene paralelo en la TCM explica por qué el enfoque IRT es la única teoría de la medición que puede ser empleada como fundamento para la medición adaptativa. Este tipo de aplicaciones ya forma parte de mediciones educacionales en gran escala, como son la prueba para medir competencias lingüísticas

en inglés (TOEFL), o la prueba de selección para estudios de postgrado en EE.UU. (Graduate Record Examination).

5.6. Comparación de puntajes entre distintas pruebas (*equating*)

Finalmente, cabe mencionar que el enfoque IRT facilita otra importante aplicación de la teoría de la medición: la posibilidad de hacer comparables las puntuaciones de dos o más instrumentos de medición. Esta capacidad fue implícitamente aludida previamente al mencionar que en este enfoque las puntuaciones que obtienen las personas no dependen de los ítemes específicos que se le administran a una determinada persona. En concreto, si un conjunto de ítemes ha sido calibrado en forma conjunta, o sus parámetros han sido establecidos en una misma escala, se pueden construir diversas formas o pruebas a partir de dichos instrumentos, cuyos puntajes estarían expresados en una misma escala. Desde este punto de vista, la comparabilidad de puntuaciones (*equating*) es consustancial al modelo IRT. En cambio, en la TCM la posibilidad de realizar comparaciones entre puntajes de personas que han respondido diversas formas de una misma prueba, requiere el empleo de diseños y procedimientos especiales, puesto que, como se precisó antes, las puntuaciones en este enfoque no son comparables, salvo en el caso que se basen en formas estrictamente paralelas de un mismo instrumento.

6. Aplicaciones prácticas del enfoque IRT

Hasta aquí se ha aludido a las semejanzas y diferencias básicas entre los modelos TCM e IRT. En esta sección se ilustrará cómo puede aplicarse en forma práctica el enfoque IRT para resolver los problemas habituales de medición. Las aplicaciones prácticas de la teoría clásica son más conocidas y están por lo demás disponibles en software estadístico general, como SPSS, SAS o STATISTICA. IRT, en cambio, representa un modelo matemáticamente más complejo, cuya aplicación supone el empleo de técnicas y programas especializados. A continuación se ilustrará, recurriendo al ejemplo de la prueba SIMCE de matemática, algunas aplicaciones prácticas del modelo IRT.

6.1. Determinación de la dimensionalidad de los ítemes

El supuesto de unidimensionalidad, como fue mencionado previamente, constituye la base de la mayor parte de los modelos IRT. Aunque en

años recientes se han desarrollado modelos multidimensionales, e incluso se está desarrollando software para hacer posible su implementación, las aplicaciones prácticas de este enfoque suponen unidimensionalidad. Para evaluar este supuesto se puede recurrir a diversas técnicas (véase Hambleton, Swaminathan y Rogers, 1991; Embretson y Reise, 2000). La más relevante de ellas es el análisis factorial exploratorio. Con esta técnica se busca determinar el número de dimensiones que subyacen a un conjunto de ítemes, a partir del análisis de la matriz de intercorrelaciones entre ellos. Un problema práctico que aparece al aplicar esta técnica a ítemes puntuados dicotómicamente, es que la correlación regular de Pearson entre variables dicotómicas (usualmente denominada coeficiente Phi) no se acomoda al modelo factorial. Este tipo de correlaciones resulta atenuada con respecto a las que se obtendrían si las variables fueran continuas, con lo que se subestiman los pesos factoriales. Por otra parte, cuando las probabilidades de responder correctamente los ítemes difieren entre ellos, que es lo usual, el modelo factorial convencional sobreestima el número de factores. Finalmente, los ítemes dicotómicos no se relacionan en forma lineal con las dimensiones subyacentes continuas (factores). Por todas estas razones, la recomendación tradicional ha sido efectuar análisis factorial de ítemes a partir de correlaciones tetracóricas entre los ítemes (Mislevy, 1986; Woods, 2002). Por desgracia esta opción no está implementada en programas estadísticos convencionales, por lo que es necesario recurrir a programas ad hoc, como TESTFACT (Bock *et al.*, 2003). Una opción a dicho programa es Mplus (Muthén y Muthén, 1998), un programa especializado en sistemas de ecuaciones estructurales (SEM), que ha incorporado opciones para variables categóricas, lo que permite manejar ítemes dicotómicos.

Para el caso de la prueba SIMCE de matemática, se llevó a cabo el análisis factorial implementado en TESTFACT 4.0, calculando la matriz de correlaciones tetracóricas entre los ítemes. Los resultados demostraron, en primer lugar, que había razonable evidencia de unidimensionalidad en esta prueba, tal como lo atestigua el gráfico de los eigenvalues (Figura 3). En él se puede apreciar que el primer factor se destaca muy claramente de los restantes. Adicionalmente, los pesos factoriales de los ítemes en el primer factor demuestran que el primer factor se correlaciona en forma sustantiva con casi todos los ítemes. Sólo los ítemes 5 y 25 muestran bajas correlaciones, por lo que se decidió excluirlos de los análisis subsecuentes[12]. En suma, esta parte de los análisis indica que la estructura de esta prueba es consistente con los requerimientos del enfoque IRT.

[12] Estos dos ítemes, así como el 41 (por razones que se aclaran en la siguiente sección) fueron finalmente excluidos de la calibración de la prueba de matemática. Ello determinó que en definitiva se emplearan 45 de los 48 ítemes de esta forma de dicha prueba. Los análisis reportados en secciones anteriores de este trabajo habían excluido estos ítemes.

FIGURA 3: GRÁFICO DE EIGENVALUES BASADO EN EL ANÁLISIS FACTORIAL DE LA PRUEBA SIMCE DE MATEMÁTICA

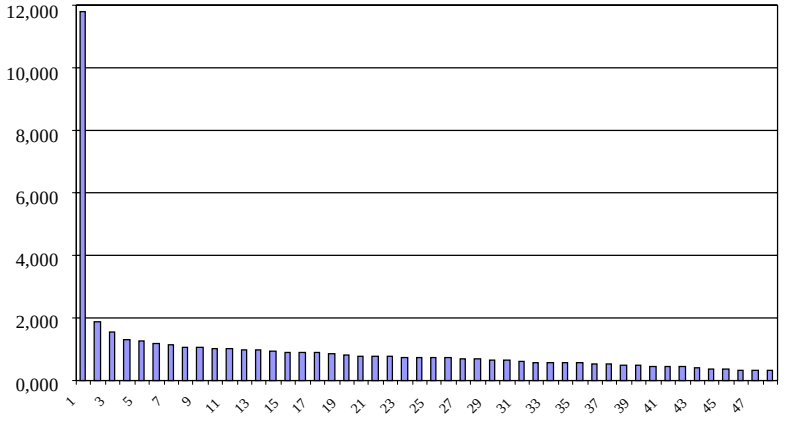


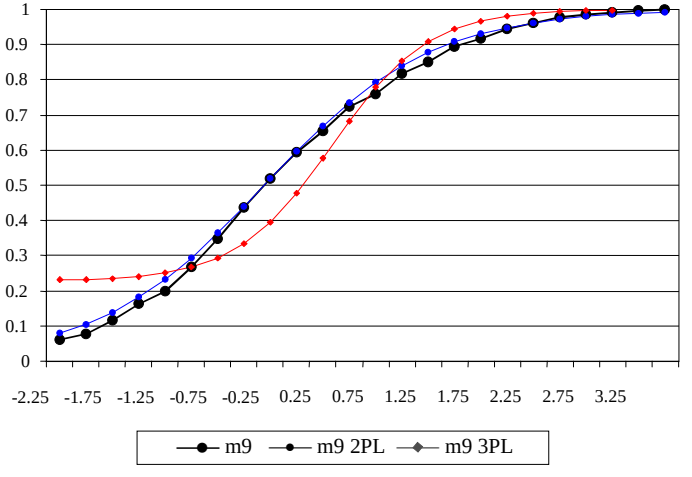
TABLA 3: PESOS FACTORIALES DE LOS 48 ÍTEMES DE LA PRUEBA SIMCE DE MATEMÁTICAS

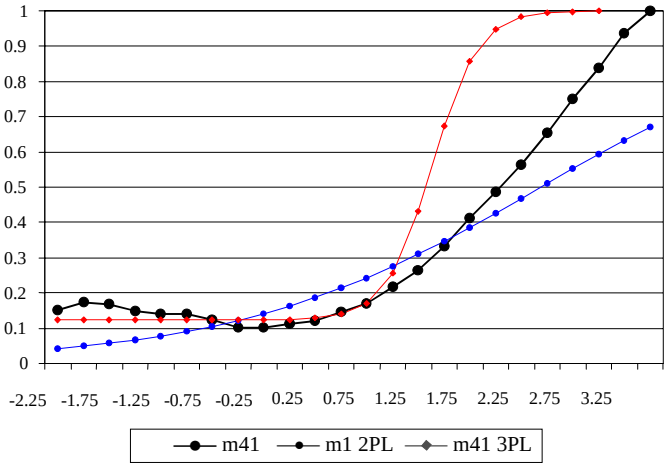
Ítem	Peso factorial	Ítem	Peso factorial
m01	0.44	m25	0.03
m02	0.50	m26	0.34
m03	0.40	m27	0.38
m04	0.52	m28	0.32
m05	0.16	m29	0.58
m06	0.50	m30	0.57
m07	0.32	m31	0.48
m08	0.54	m32	0.59
m09	0.55	m33	0.50
m10	0.39	m34	0.60
m11	0.51	m35	0.58
m12	0.38	m36	0.62
m13	0.31	m37	0.45
m14	0.54	m38	0.69
m15	0.54	m39	0.49
m16	0.44	m40	0.32
m17	0.47	m41	0.38
m18	0.53	m42	0.48
m19	0.52	m43	0.41
m20	0.43	m44	0.52
m21	0.59	m45	0.49
m22	0.56	m46	0.33
m23	0.66	m47	0.24
m24	0.51	m48	0.56

6.2. Comparación del ajuste de distintos modelos IRT

Tal como fue previamente mencionado, hay modelos IRT de 1, 2 y 3 parámetros. Una de las decisiones básicas en aplicaciones prácticas de este enfoque se refiere a la adopción de una de estas alternativas. Desde el punto de vista conceptual, el modelo de 3 parámetros sólo es pertinente cuando se está trabajando con preguntas de respuesta cerrada, donde existe la posibilidad de beneficiarse contestando al azar. Excepto por este criterio, que definitivamente excluye el modelo de 3 parámetros cuando no hay probabilidad de responder correctamente por azar, la opción entre los modelos debiera fundarse en un examen del grado en que cada uno de los modelos se ajusta a los datos. Aunque existen algunas aproximaciones estadísticas para apoyar esta decisión, en la práctica, siempre se recomienda el examen empírico del grado en que las curvas características resultantes de la aplicación de modelos de diverso número de parámetros, concuerdan con las curvas empíricas correspondientes. Las curvas empíricas se obtienen por un procedimiento análogo al empleado para construir las curvas descritas en la Figura 1, sólo que en este caso se usa como estimación de la habilidad de los examinados la que determina el modelo IRT (θ). En el caso de las pruebas SIMCE de matemática, se comparó el modelo de 2 parámetros con el de 3 parámetros. Dado que se trata de preguntas de selección múltiple, resultaba lógico contemplar la posibilidad que el modelo de 3 parámetros obtuviera un mejor ajuste que uno de 2 parámetros. Sin embargo, la mayoría de los ítemes de esta prueba mostró un mejor grado de ajuste del modelo de 2 parámetros. Los gráficos de la Figura 4 ilustran, con fines descriptivos,

FIGURA 4: COMPARACIÓN ENTRE CURVAS EMPÍRICAS Y CCI DE 2 Y 3 PARÁMETROS EN 2 ÍTEMES DE LA PRUEBA SIMCE DE MATEMÁTICA



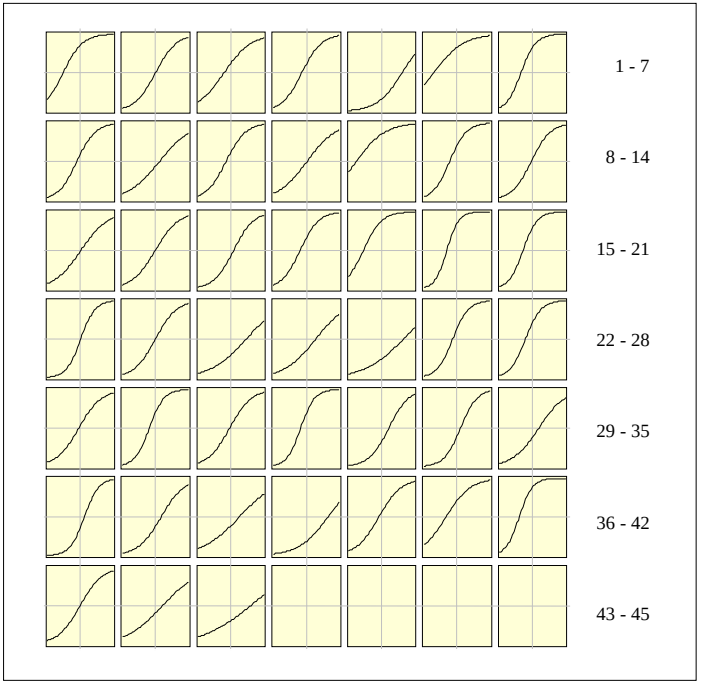


sólo dos ítemes: uno que muestra un buen ajuste al modelo de 2 parámetros, y otro con un pobre ajuste. Este último ítem corresponde al tercer ítem excluido de esta prueba.

6.3. Calibración de modelos IRT

La calibración de los ítemes al modelo IRT corresponde a la fase de estimación de los parámetros en el modelo correspondiente. Para llevar a cabo esta tarea se requiere el empleo de software especializado. Aunque se han desarrollado varios programas con este propósito, los más populares son los producidos por Darrell Bock y sus colaboradores: BILOG, PARSCALE y MULTILOG. En conjunto, estos programas permiten resolver una amplia variedad de problemas prácticos, incluyendo modelos de diverso número de parámetros, para ítemes puntuados dicotómica o policotómicamente. En este último caso, los programas ofrecen diversas opciones, como el modelo de respuesta graduada de Samejima, el de crédito parcial de Muraki o el de respuesta nominal de Bock, entre otros. Asimismo, ofrecen diversos algoritmos para la estimación de los parámetros o las habilidades de los examinados. Por último, cuentan con opciones para manejar múltiples grupos (lo que es conveniente para realizar análisis DIF), así como múltiples examinadores (cuando diversos jueces deben asignar puntajes a respuestas abiertas). Las versiones más recientes de estos programas incluyen una interfaz gráfica que permite visualizar diversos aspectos, tales como las curvas características de los ítemes, su función informativa, la función informativa del test, etc.

FIGURA 6: CURVAS CARACTERÍSTICAS DE LOS ÍTEMES DE LA PRUEBA SIMCE DE MATEMÁTICA



En el caso que nos interesa, la calibración final de la prueba se llevó a cabo con PARSCALE 4.1 (Muraki y Bock, 2003) empleando el modelo de 2 parámetros. Luego de 8 ciclos se obtuvo una solución satisfactoria: los parámetros resultantes se encuentran en un rango adecuado, con un valor medio de 0,61 para el parámetro de discriminación, y –0,06 para el parámetro de localización o dificultad. Además, tal como puede apreciarse en la Tabla 1, el error estándar asociado a todos los parámetros es homogéneamente bajo, lo que respalda el ajuste de los ítemes al modelo. La Figura 5, por último, muestra gráficamente las curvas características de los 45 ítemes de esta prueba.

6.4. Evaluación de la capacidad informativa de una prueba

La función informativa de la prueba es un importante indicador de la calidad de un instrumento. Esta función nos revela el grado de precisión de un instrumento para evaluar el atributo subyacente en distintas regiones del mismo. Como es sabido, en la Teoría Clásica de la Medición el concepto que revela la calidad de la medición, la confiabilidad, consiste en un indica-

dor único, que alude a la calidad global de un instrumento. Lamentablemente, es engañoso creer, como se mencionó en la sección 5.2., que un instrumento pueda ser homogéneamente preciso. El error de medición no es constante, por lo que es conveniente contar con una evaluación más específica y precisa de la calidad de la medición.

En el caso de la prueba que estamos analizando, su función informativa, que aparece en la Figura 1, revela que el instrumento es adecuadamente informativo en la región media de los puntajes. Dicha precisión decae en ambos extremos. Dado que el propósito de este instrumento es estimar la habilidad de grupos de estudiantes (pues el SIMCE se reporta a nivel de establecimientos), se puede constatar que para la gran mayoría de los puntajes que se estiman el instrumento produce una información adecuada.

6.5. Análisis del sesgo de preguntas (análisis DIF)

El sesgo de medición se ha transformado, en los últimos años, en un tema de gran importancia en la medición. En la medida que los instrumentos de medición han adquirido gran relevancia para tomar importantes decisiones, con claras consecuencias personales o sociales[13], la preocupación por la posibilidad que los instrumentos puedan inadvertidamente favorecer o perjudicar a determinados grupos, ha aumentado considerablemente. Ello ha repercutido directamente en la teoría de la medición, reflejándose en el desarrollo de técnicas y procedimientos para evaluar el sesgo en la medición, tanto a nivel de ítemes como de los puntajes totales de un instrumento. Los estándares vigentes para el desarrollo de instrumentos que tienen consecuencias sociales y personales relevantes establecen la necesidad de evaluar dicho sesgo.

El análisis del funcionamiento diferencial de los ítemes (conocido por su sigla en inglés DIF), se ha transformado en el enfoque dominante para estos propósitos. Con esta técnica se busca establecer si un determinado grupo (establecido a partir de diferencias de género, raciales, sociales u otras) puede verse beneficiado o perjudicado en sus puntajes (Camilli y Shepard, 1994). El enfoque IRT ofrece un marco de referencia muy claro para esta evaluación, puesto que, de acuerdo a sus fundamentos, la posibilidad de responder correctamente un ítem debiera depender sólo de la habilidad de una persona. En consecuencia, cuando la pertenencia a un grupo

[13] Para ilustrar esto basta pensar en el rol que hoy tienen en todo el mundo, incluido nuestro país, instrumentos de medición en decisiones como: acceso a la educación (primaria, secundaria y universitaria), acceso y evaluación laboral (selección laboral, evaluación del desempeño, etc.), peritajes judiciales, etc.

afecta los puntajes independientemente de su nivel de habilidad, se produce una violación de los supuestos fundamentales de la teoría. Se dice que un ítem tiene un efecto DIF cuando la pertenencia grupal de un examinado afecta su probabilidad de responder correctamente dicho ítem, más allá de su nivel de habilidad.

Para ilustrar este tipo de análisis emplearemos nuevamente información proveniente de las mediciones SIMCE. Sin embargo, deberemos recurrir a una medición distinta de la empleada en los ejemplos anteriores, puesto que para facilitar la interpretación del efecto DIF es conveniente presentar el contenido de los ítemes involucrados. Dado que los ítemes empleados en la medición SIMCE de 2º medio de 1998 fueron liberados al conocimiento público, podemos recurrir a ellos. En concreto, se presentan a continuación 2 ítemes de la prueba de lenguaje, uno con un efecto DIF favorable a las mujeres y otro con un efecto DIF favorable a los hombres. En ambos casos se trata de preguntas que se basan en un texto previo (omitido aquí, pues no resulta indispensable para interpretar los resultados).

Ejemplo 1: Ítem con efecto DIF favorable a las mujeres

¿Cuál es el principal sentimiento que expresa el relator hacia la mujer descrita?
a) Simpatía
b) Curiosidad
c) Amor
d) Crítica

Ejemplo 2: Ítem con efecto DIF favorable a los hombres

En el texto, ¿qué significa "Europa fue el teatro inicial del conflicto"?
a) Que la Segunda Guerra Mundial comenzó en Europa
b) Que al principio Europa fue una espectadora del conflicto
c) Que Europa comenzó la Segunda Guerra Mundial
d) Que al principio la Segunda Guerra Mundial fue una farsa

Para evaluar el efecto DIF se empleó el software PARSCALE, que ofrece la posibilidad de estimar modelos IRT con múltiples grupos. En concreto, se evaluó un modelo DIF comparando a hombres y mujeres. El análisis DIF se centró en el parámetro de localización o dificultad. Los resultados, que se representan gráficamente en la Figura 6a y 6b, muestran que en estos dos ítemes el parámetro b mostró diferencias estadísticamente

significativas. En el primer caso, la curva correspondiente a los hombres aparece desplazada hacia la derecha, lo que indica que en este ítem es necesario un mayor nivel de habilidad para responderlo correctamente. Lo contrario ocurre con el segundo ítem. La interpretación del efecto DIF en estos dos casos resulta relativamente simple, puesto que el contenido de ambas preguntas se vincula de manera más o menos directa con los aspectos tradicionalmente asociados a la socialización diferencial de hombres y mujeres: la primera pregunta alude a un tema usualmente considerado más femenino, la comprensión de emociones, mientras que la segunda se relaciona con un tema convencionalmente visto como masculino, la guerra.

Es importante dejar constancia que la detección de un efecto DIF no debe ser considerada como una razón para eliminar automáticamente un ítem. Más bien, esta evidencia debe ser incorporada como una señal de alerta, que debe ser analizada en conjunto con otros antecedentes para resolver la suerte de un ítem. En el caso que nos ocupa, el efecto, que resulta fácilmente interpretable, nos alerta acerca de los riesgos de incluir preguntas cuyo contenido o temática pueda resultar más familiar, cercana o significativa para un determinado grupo.

FIGURA 6a: CURVAS CARACTERÍSTICAS DE ÍTEM CON EFECTO DIF FAVORABLE A MUJERES

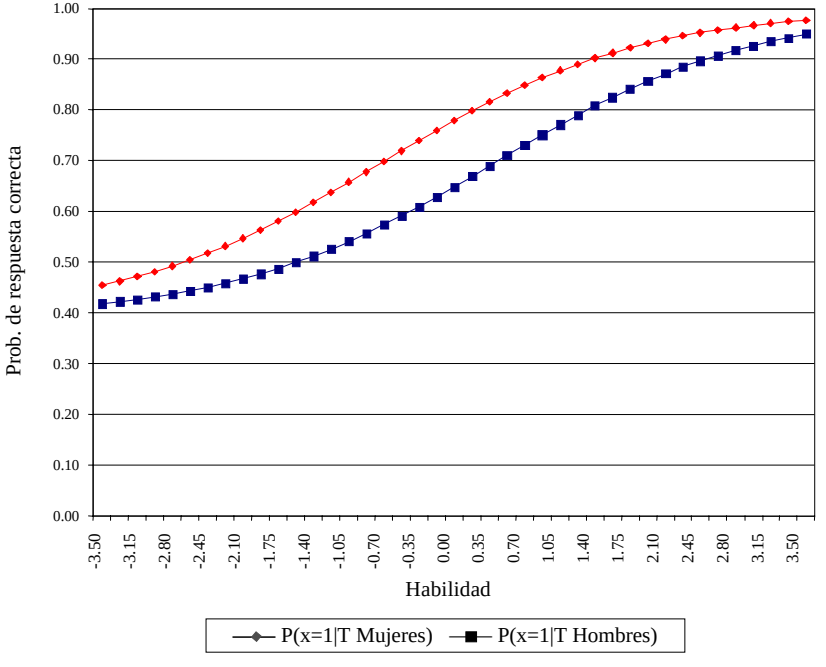
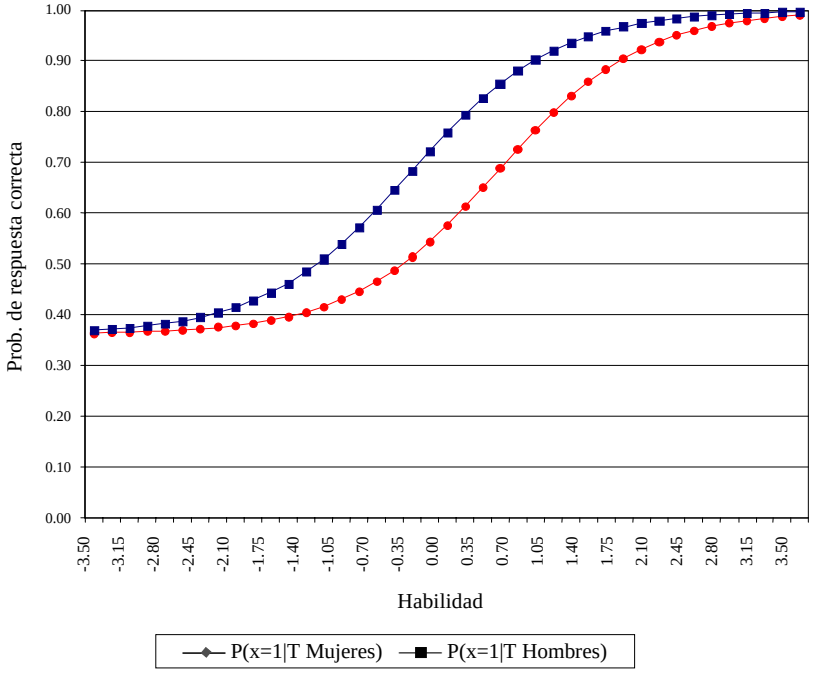


FIGURA 6b: CURVAS CARACTERÍSTICAS DE ÍTEM CON EFECTO DIF FAVORABLE A HOMBRES



Conclusiones

En este trabajo hemos revisado los fundamentos de las dos principales teorías de la medición desarrolladas durante el siglo pasado: la Teoría Clásica y la teoría moderna o Teoría de Respuesta al Ítem. Hemos hecho ver que ambas teorías fundan sus postulados básicos en supuestos semejantes, aunque difieren en el grado de complejidad con respecto a la forma en que modelan la relación entre los atributos subyacentes y las variables manifiestas (respuestas a las preguntas de un instrumento de medición). En gran medida el modelo IRT es una extensión de la TCM, por lo que estas teorías no pueden ser vistas como modelos competitivos o antagónicos.

Así como la TCM, los postulados básicos del enfoque IRT, al menos en lo referente al análisis de instrumentos de medición unidimensionales con preguntas puntuadas dicotómicamente, constituyen un cuerpo asentado

y bien establecido de principios, por lo que su empleo para resolver problemas prácticos de medición en el ámbito psicológico y educacional se encuentra sólidamente respaldado.

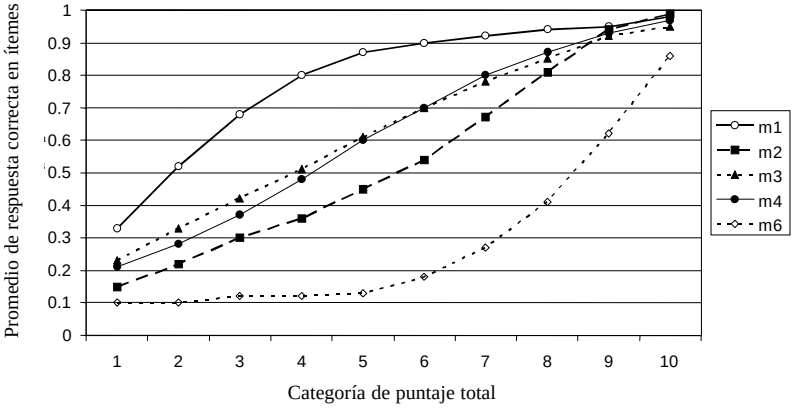
Es cierto que el modelo IRT ha demorado en ser incorporado masivamente en el desarrollo de instrumentos de medición. Excepto por las aplicaciones a la medición educacional en gran escala, su implementación en otros campos ha sido relativamente reciente, tal como se ilustró en la introducción de este trabajo. Esta demora, sin embargo, no refleja debilidades de la teoría, sino que barreras de orden práctico: la relativa escasez de programas computacionales capaces de manejar en forma eficiente y rápida los complejos algoritmos asociados a la estimación de los parámetros de los ítemes, las exigencias muestrales relativamente altas de esta teoría, relativas a las existentes en la TCM, y la mayor complejidad matemática de este enfoque en relación a su precedente.

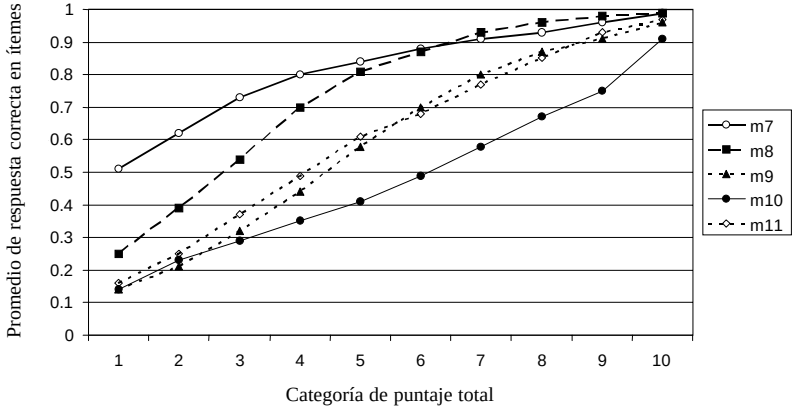
Es importante destacar que el interés académico en torno a la teoría de la medición se ha volcado casi enteramente en los últimos años a la Teoría de Respuesta al Ítem. Basta revisar el índice de las principales revistas científicas relacionadas con la psicometría (*Psychometrika, Applied Psychological Measurement* y *Psychological Methods*) para constatar que casi todos los artículos que se publican se refieren a este enfoque. De hecho, un examen de la revista más especializada en teoría de la medición, *Applied Psychological Measurement*, muestra que de los 25 artículos publicados en 2002, 24 tenían relación con IRT y 1 con la teoría de la generalizabilidad (ninguno se refería a la teoría clásica). Este dinamismo académico, sumado al creciente desarrollo de software especializado, permite anticipar que en los próximos años se verá un aumento sostenido del empleo de métodos basados en este enfoque para diseñar y evaluar instrumentos de medición tanto en educación como en psicología.

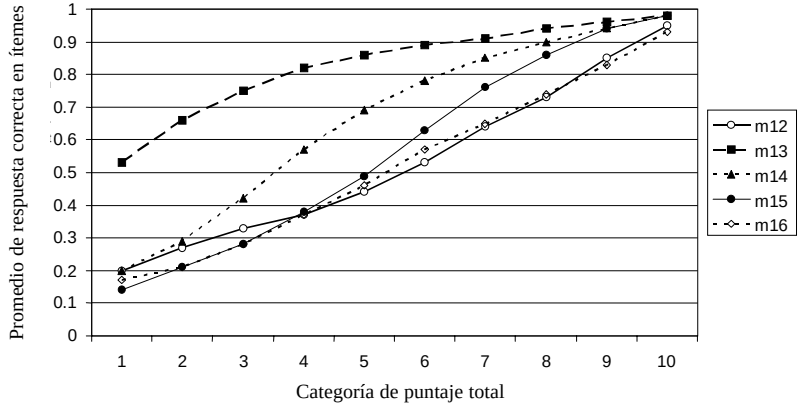
En consecuencia, lo único razonable es incorporar este enfoque en la agenda de trabajo de todo esfuerzo serio en materia de medición. No se trata, como esperamos haber clarificado en este artículo, de reemplazar a la Teoría Clásica, sino de complementarla con los procedimientos, técnicas y posibilidades que ofrece el enfoque IRT. Especialmente en los casos en que se dispone de grandes volúmenes de información, como es el caso de la medición educacional en gran escala, no hay razones científicamente fundadas para excluir el uso de este enfoque.

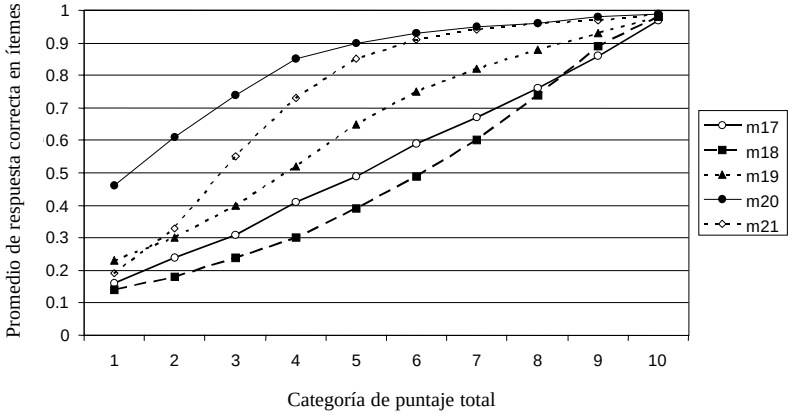
ANEXO

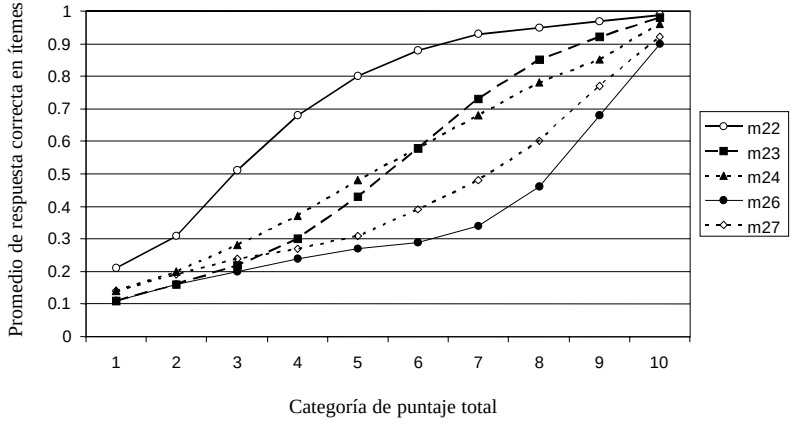
FIGURA 1: CURVAS EMPÍRICAS DE LOS ÍTEMES DE LA PRUEBA SIMCE DE MATEMÁTICA

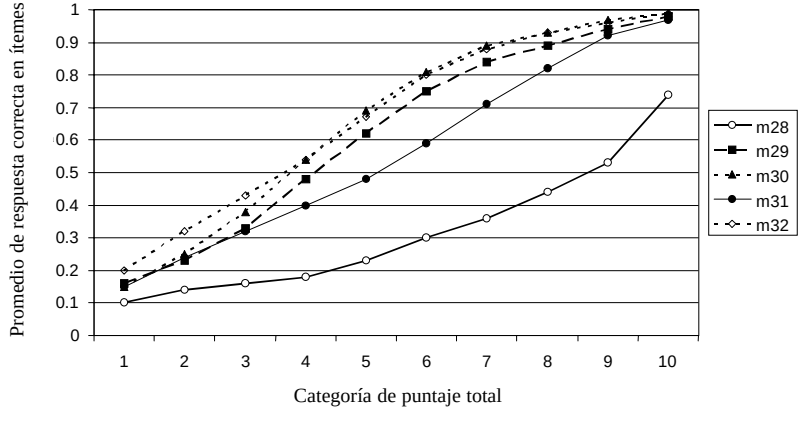


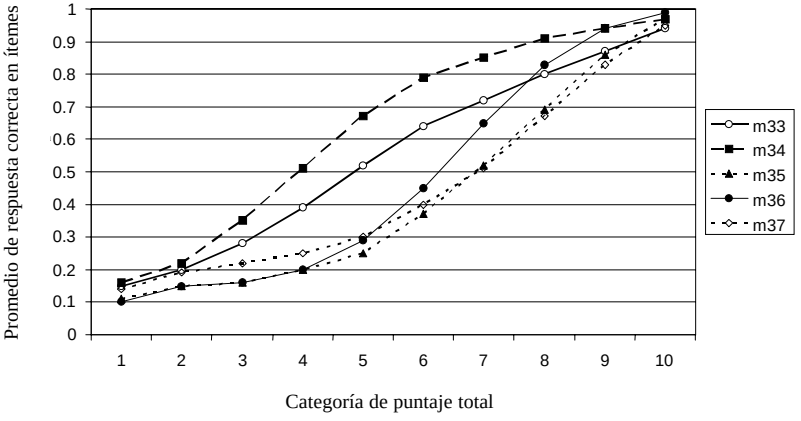


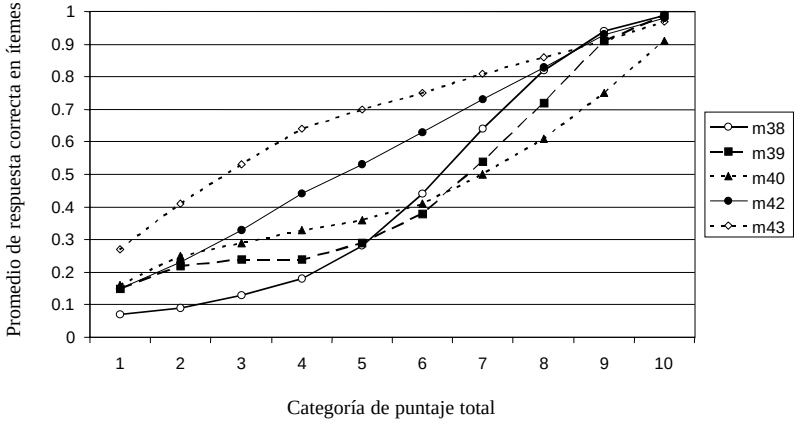


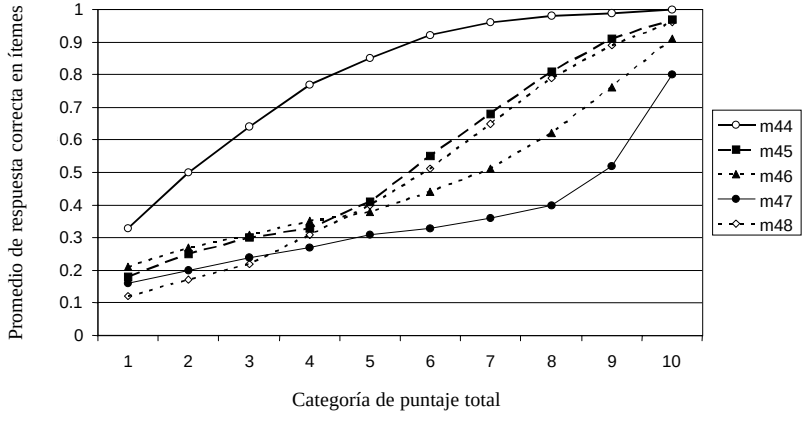

Promedio de respuesta correcta en ítemes
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1 2 3 4 5 6 7 8 9 10
Categoría de puntaje total
m17
m18
m19
m20
m21


Promedio de respuesta correcta en ítemes
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1 2 3 4 5 6 7 8 9 10
Categoría de puntaje total
m22
m23
m24
m26
m27


Promedio de respuesta correcta en ítemes
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1 2 3 4 5 6 7 8 9 10
Categoría de puntaje total
m28
m29
m30
m31
m32


Promedio de respuesta correcta en ítemes
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1 2 3 4 5 6 7 8 9 10
Categoría de puntaje total
m33
m34
m35
m36
m37


Promedio de respuesta correcta en ítemes
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1 2 3 4 5 6 7 8 9 10
Categoría de puntaje total
m38
m39
m40
m42
m43


Promedio de respuesta correcta en ítemes
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1 2 3 4 5 6 7 8 9 10
Categoría de puntaje total
m44
m45
m46
m47
m48

BIBLIOGRAFÍA

Andersen, E.B., y Olsen, L. W. "The Life of George Rasch as a Mathematician and as a Statistician". En A. Boomsma *et al.* (eds.), *Essays on Item Response Theory*. New York: Springer, 2001.

Anderson, T. W. "Some Scaling Models and Estimation Procedures in the Latent Class Model". En U. Grenander (ed.), *Probability and Statistics.* New York: Wiley, 1959, pp. 9-38.

Bartholomew, D. J. *Latent Variable Models and Factor Analysis*. Londres: Charles Griffin, 1987.

Birnbaum, A. "On the Foundations of Statistical Inference (with Discussion)". *Journal of the American Statistical Association* 57 (1962), pp. 269-326.

Bock, D., Gibbons, R., Schilling, S., Muraki, E., Wilson, D., y Wood, R. TESTFACT 4.0. Test scoring, item statistics, and item factor analysis [Programa computacional]. Chicago: Scientific Software International, 2003.

Camilli, G., y Shepard, L. A. *Methods for Identifying Biased Test Items.* Thousand Oaks, CA: Sage, 1994.

Cantor, G. "Beiträge Zur Begründung der Transfiniten Mengenlehre". *Math. Ann.* 46 (1895), pp. 481-512.

Cox, D. R. "Role of Models in Statistical Analysis". *Statistical Sciences* 5 (1990), pp. 169-174.

Crocker, L., y Algina, J. *Introduction to Classical and Modern Test Theory.* Philadelphia: Harcourt Brace Jovanovich College Publishers, 1986.

Cronbach, L., Gleser, G., y Rajaratnam, N. "Theory of Generalizability. A Liberation of Reliability Theory". *British Journal of Mathematical and Statistical Psychology*, 16 (1963), pp. 137-173.

Dusaillant, F. "Técnicas de Medición en Pruebas de Admisión a las Universidades". *Estudios Públicos* 90 (otoño 2003).

Ellis, B. *Basic Concepts of Measurement.* Cambridge: Cambridge University Press, 1968.

Ellis, B. "Differential Item Functioning: Implications for Test Translations". *Journal of Applied Psychology*, 74 (1989), pp. 912-921.

Embretson, S. E., y Reise, S. P. *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

Fisher, R. A. "On the Mathematical Foundations of Theoretical Statistics". *Philosophical Transaction of the Royal Society of London, Series A* 222 (1922), pp. 309-368.

Fischer, G. H. "Applying the Principles of Specific Objectivity and Generalizability to the Measurement of Change". *Psychometrika* 52 (1987), pp. 565-587.

Fischer, G. H. "Some Neglected Problems in IRT". *Psychometrika* 60 (1995a), pp. 459-487.

Fischer, G. H. "Derivations of the Rasch Mode". En G. H. Fischer e I. W: Molenaar (eds.), *Rasch Models. Foundations, Recent Developments and Applications.* New York: Springer, 1995b.

Fraley, R., Waller, N., y Brennan, K. "An Item Response Theory Analysis of Self-Report Measures of Adult Attachment". *Journal of Personality and Social Psychology*, 78 (2000), pp. 350-365.

Goldberger, A. S. "Econometrics and Psychometrics: A Survey of Communalities". *Psychometrika* 36 (1971), pp. 83-107.

Goldberger, A. S. "Structural Equation Methods in the Social Sciences". *Econometrica* 40 (1972), 979-1001.

Godber, T., Anderson, V., y Bell, R. "The Measurement and Diagnostic Utility of Intrasubtest Scatter in Pediatric Neuropsychology". *Journal of Clinical Psychology*, 56 (2000), pp. 101-112.

Gray-Little, B., Williams, V., y Hancock, T. "An Item Response Theory Analysis of the Rosenberg Self-Esteem Scale". *Personality and Social Psychology Bulletin*, 23 (1997), pp. 443-451.

Gulliksen, H. *Theory of Mental Tests*. New York: Wiley, 1950.

Guttman, L. "A Basis for Analyzing Test-Retest Reliability". *Psychometrika* 10 (1945), pp. 255-273.

Guttman, L. "Reliability Formulas that do not Assume Experimental Independence". *Psychometrika* 18 (1953), pp. 225-239.

Hambleton, R. K., Swaminathan, H., y Rogers, H. J. *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage, 1991.

Hand, D. J. "Statistics and the Theory of Measurement (with Discussion)". *Journal of the Royal Statistical Society, Series A* 159 (1996), pp. 445-492.

Helmholtz, H. V. "Zählen und Messenerkenntnis-Theoretish Betrachet". *Philosophische Aufsätze Eduard Zeller gewidmet*, Leipzig, 1895.

Holland, P. W., y Rosenbaum, P. R. "Conditional Association and Unidimensionality in Monotone Latent Variable Models". *The Annals of Statistics* 14 (1980), pp. 1523-1543.

Hölder, O. "Die Axiome der Quantität und Die Lehre von Mass". *Ber. Verh. Kgl. S\"achsis. Ges. Wiss. Leipzig, Math.-Phys. Classe* 53 (1901), pp. 1-64.

Junker, B. "On the Interplay Between Nonparametric and Parametric IRT, with some Thoughts about the Future". En A. Boomsma *et al.* (eds.), *Essays on Item Response Theory*. New York: Springer, 2001.

Koopmans, T. C., y Reiersøl, O. "Identification of Structural Characteristics". *The Annals of Mathematical Statistics* 21 (1950), pp. 165-181.

Krantz, D. H., Luce, R. D., Suppes, P., y Tversky, A. *Foundations of Measurement. Volume I. Additive and Polynomial Representations*. New York: Academic Press, 1971.

Lazarsfeld, P. F. "The Logical and Mathematical Foundation of Latent Structure Analysis". En S. A. Stouffer *et al.* (eds.), *Measurement and Prediction*. New York: Wiley, 1959.

Lindley, D. V., y Novick, M. R. "The Role of Exchangeability in Inference". *The Annals of Statistics* 9 (1981), pp. 45-58.

Lord, F. M., y Novick, M. R. *Statistical Theories of Mental Test Scores*. Massachusetts: Addison Wesley, 1968.

Maller, S. "Item Invariance in Four Subtests of the Universal Nonverbal Intelligence Test (UNIT) Across Groups of Deaf and Hearing Children". *Journal of Psychoeducational Assessment*, 18 (2000), pp. 240-254.

Manski, C. F. *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press, 1995.

McCullagh, P. "What is a Statistical Model? (with Discussion)". *The Annals of Statistics* 30 (2002), pp. 1225-1310.

Michell, J. *An Introduction to the Logic of Psychological Measurement*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1990.

Michell, J. "Bertrand Russell's 1897 Critique of the Traditional Theory of Measurement". *Synthese* 110 (1997a), pp. 257-276.

Michell, J. "Quantitative Science and the Definition of Measurement in Psychology (with Discussion)". *British Journal of Psychology* 88 (1997b), pp. 355-383.

Mislevy, R. J. "Recent Developments in the Factor Analysis of Categorical Variables". *Journal of Educational Statistics*, 11 (1986), pp. 3-31.

Mouchart, M., y San Martín, E. "Specification and Identification Issues in Models Involving a Latent Hierarchical Structures". *Journal of Statistical Planning and Inference* 111 (2003), pp. 143-163.

Muraki, E., y Bock, D. PARSCALE 4.1: IRT item analysis and test scoring for rating-scale data [Programa computacional]. Chicago: Scientific Software International, 2003.

Muthén, L., y Muthén, B. *Mplus: Statistical Analysis with Latent Variables* [programa computacional]. Los Angeles: Muthen y Muthen, 1998.

Novick, M. R., y Lewis, C. "Coefficient Alpha and the Reliability of Composite Measurements". *Psychometrika* 32 (1967), pp. 1-13.

Novick, M. R. "The Axioms and Principal Results of Classical test Theory". *Journal of Mathematical Psychology* 3 (1968), pp. 1-18.

Novick, M. R. "Statistics as Psychometrics". *Psychometrika* 45 (1980), pp. 411-424.

Panter, A., Swygert, K., Dahlstrom, W., y Tanaka, J. "Factor Analytic Approaches to Personality Item-Level Data". *Journal of Personality Assessment*, 68 (1997)., 561-589.

Pfanzagl, J. *Theory of Measurement*. Physica-Verlag, Würzburg, 1968.

Rouse, S., Finger, M., y Butcher, J. "Advances in Clinical Personality Measurement: An Item Response Theory Analysis of the MMPI-2 PSY-5 Scales". *Journal of Personality Assessment*, 72 (1999), pp. 282-307.

San Martín, E. "Modeling Problems Motivated by the Specification of Latent Linear Structures". Aceptado para publicación en el *Journal of Mathematical Psychology*, 2003.

Sobel, M. E. "Measurement, Causation and Local Independence in Latent Variable Models". En M. Berkane (ed.), *Latent Variable Modeling and Applications to Causality*. New York: Springer, 1997.

Swistak. P. "Paradigms of Measurement". *Theory and Decision* 29 (1990), pp. 1-17.

Thissen, D., y Wainer, H. *Test Scoring*. Mahwah, NJ: Lawrence Erlbaum Association, 2001.

Van der Linden, W., y Hambleton, R. *Handbook of Modern Item Response Theory*. New York: Springer, 1997.

Waller, N., Thompson, J., y Wenk, E. "Using IRT to Separate Measurement Bias from True Group Differences on Homogeneous and Heterogeneous Scales: An Illustration with the MMPI". *Psychological Methods*, 5 (2000), pp. 125-146.

Woods, C. M. "Factor Analysis of Scales Composed of Binary Items: Illustration with the Maudsley Obsessional Compulsive Inventory". *Journal of Psychopathology and Behavioral Assesment*, 24 (2002), pp. 215-223. □