TÉCNICAS DE MEDICIÓN EN PRUEBAS DE ADMISIÓN A LAS UNIVERSIDADES*

Francisca Dussaillant

Las pruebas de admisión a la educación superior en Chile han sido objeto de un intenso debate en los últimos doce meses. Entre otros aspectos, se ha sugerido aplicar en estas pruebas la teoría de respuesta al ítem (IRT, por sus siglas en inglés), en reemplazo de la teoría clásica de medición (TCM) que se usó en Chile en el marco de la PAA y que sigue utilizándose en Estados Unidos (SAT), en Suecia (SweSat), Israel (PET) y en el mundo entero en las pruebas de admisión universitaria.

En este trabajo se describen los elementos centrales de IRT, los modelos más utilizados dentro de esta teoría y algunas de sus aplicaciones, y se introduce el concepto de información de un ítem. A su vez, se presenta una recopilación de estudios e investigaciones que examinan en profundidad cuán robusta es IRT a las violaciones de los supuestos que la sustentan, cuán reales son las propiedades que la distinguen de la teoría clásica, y se analizan algunas de las conse-

FRANCISCA DUSSAILLANT. Ingeniera Civil Industrial, P. Universidad Católica de Chile. Master en Educación, University of North Carolina-Chapel Hill. Investigadora del Centro de Estudios Públicos.

* Una versión preliminar de este trabajo apareció como *Documento de Trabajo* N° 347 (febrero de 2003), Centro de Estudios Públicos, Santiago.

La autora agradece a Carmen le Foulon y Harald Beyer por sus importantes aportes y sugerencias durante la preparación de este trabajo.

cuencias prácticas que traería su implementación en una prueba a gran escala, especialmente en el caso de que ésta tenga consecuencias para el examinado.

Del estudio realizado se concluye que IRT es una teoría promisoria, pero que *no* es la panacea que vendrá a solucionar todos los problemas que existen en medición educacional. Se sugiere, por lo tanto, que ella sea aplicada con cautela y como complemento a la teoría clásica, sobre todo en el caso de pruebas a gran escala con consecuencias para los estudiantes. Así lo recomendaría también la experiencia internacional, que muestra que se ha preferido mantener la teoría clásica como modelo dominante a partir del cual se calculan los puntajes en pruebas de admisión a la universidad, dejando IRT solamente para análisis secundarios.

1. INTRODUCCIÓN

Las teorías de medición sirven como marco teórico en el diseño e implementación de pruebas. Estas teorías entregan la metodología para la asignación de puntajes, proveen mecanismos para determinar las características de las preguntas o ítems, y a partir de ellas se derivan métodos para realizar otros análisis de interés. En este trabajo se analizarán dos de las principales teorías que se utilizan en el ámbito de medición educacional: la teoría clásica y la teoría de respuesta al ítem. El objetivo principal del artículo es analizar en particular la teoría de respuesta al ítem, sus posibles ventajas, aplicabilidad y limitaciones. Las siguientes líneas de esta introducción presentan de manera sucinta los principales análisis y conclusiones del trabajo.

Antes que nada, se hace necesario definir algunos términos. Distintas pruebas miden diferentes características de los examinados, por ejemplo: conocimiento matemático, razonamiento científico, capacidad de memorización, vocabulario, etc. En medición educacional, por "habilidad" se entiende la característica del examinado que va a ser medida a través de la prueba y, por consiguiente, ésta se utiliza para representar de manera genérica cualquiera de esas características. Por otro lado, en lo que respecta a los ítems de una prueba, hay parámetros (o descriptores) que permiten describir sus atributos particulares. Los descriptores de un ítem que suelen ser más frecuentemente utilizados y, por consiguiente, calculados o estimados son: a) nivel de "dificultad" (o sea cuán complicada es la pregunta) y b) nivel de "discriminación" de un ítem (o sea la capacidad que tiene éste para distinguir a los alumnos más aptos de los menos aptos).

1.1. Teoría Clásica

Una forma de medición que se utiliza con mucha frecuencia en pruebas a gran escala, como las de admisión a la universidad (por ejemplo, en el SAT de Estados Unidos y en la PAA en Chile), es la llamada teoría clásica (TC). En teoría clásica, el indicador de la *habilidad* de un estudiante corresponde al puntaje de éste en la prueba, construido a partir del número de respuestas correctas (o número de respuestas correctas netas) que obtuvo. Como indicador de la *dificultad* de una pregunta, en este sistema se utiliza la proporción de personas que la contestaron correctamente. Por otro lado, el índice de *discriminación* de un ítem se calcula como la correlación entre la respuesta a éste y el puntaje en la prueba total. Ambos descriptores de los ítems pueden calcularse ya sea en el contexto de una prueba piloto o experimental, o en la prueba definitiva u operacional.

Como se puede apreciar, en teoría clásica el grado de habilidad de una persona depende del grupo de ítems (vale decir, de su nivel de dificultad y discriminación) que contiene la prueba. Por ejemplo, si la prueba es fácil, un mismo alumno tendrá un puntaje mayor que si la prueba es difícil. Con esto resulta difícil hacer comparaciones entre estudiantes que han rendido pruebas diferentes. A su vez, los índices de dificultad y de discriminación de los ítems dependen del grupo de personas que rinden la prueba. Así, un mismo ítem puede ser catalogado como fácil si el grupo que rindió la prueba es excepcionalmente hábil, pero como difícil si el grupo que rindió la prueba es desaventajado. Con respecto a la discriminación, un ítem puede aparecer muy discriminatorio en el contexto de un grupo con nivel heterogéneo de habilidades, pero poco discriminatorio si el grupo que rindió la prueba es muy homogéneo (es decir, si todos los estudiantes tienen un nivel de habilidad similar). Esta dependencia de la habilidad de un estudiante con respecto al grupo de ítems de la prueba, junto con la dependencia de los descriptores de los ítems con respecto a las características del grupo, la llamaremos *dependencia circular*.

Otra debilidad de la teoría clásica es que supone que la precisión con que se hace la medición es igual para todos los examinados, independientemente de su nivel de habilidad. Este supuesto es bastante discutible. Intuitivamente es claro que una prueba en que, por ejemplo, la mayor parte de sus preguntas son difíciles, va a distinguir más finamente entre dos personas con habilidad superior a la media que entre dos personas menos hábiles. Los que tienen habilidades inferiores obtendrán una estimación menos precisa de su habilidad, ya que son pocas o ninguna las preguntas de la prueba que responderán correctamente, y que, por lo tanto, servirían para

distinguirlos. Esta debilidad, junto con la dependencia circular que se genera en el cálculo de la habilidad de los examinados y la dificultad y discriminación de los ítems, han llevado a buscar un método que permitiese obtener una medida de la habilidad de los examinados que sea independiente de los ítems a que éstos se han enfrentado, una caracterización de los ítems independiente de la población a la que se aplican, y al mismo tiempo una medida más fiel de la precisión con que se está midiendo la habilidad. La satisfacción de estos requerimientos se ha intentado encontrar en la teoría de respuesta al ítem o IRT.

1.2. Teoría de respuesta al ítem (IRT, por sus siglas en inglés)

Los modelos IRT se centran en los ítems e intentan establecer, para cada uno de ellos, la *probabilidad* de ser respondidos correctamente. Esta probabilidad depende de la habilidad del examinado y de ciertas características de los ítems, entre las que pueden contarse su grado de dificultad y discriminación, y la probabilidad de ser respondido correctamente, por azar, por un individuo de muy baja habilidad. Hay varios modelos IRT de distinta complejidad. El modelo más simple es aquel que sólo diferencia los ítems según su grado de dificultad. Sin embargo, otros modelos permiten además incluir otros descriptores (o parámetros del ítem), como su grado de discriminación y la probabilidad de responderlo correctamente al azar.

Aparte de lo anterior, IRT hace posible conocer el nivel de "certeza" o "precisión" que un ítem aporta a la estimación para cada nivel de habilidad. En términos técnicos esto se llama "información" de un ítem. *Mientras mayor es la información que aporta el ítem a un determinado nivel de habilidad, mayor es la precisión en la estimación de ese nivel de habilidad.* Esto permite construir pruebas "a la medida" del objetivo educacional que se persigue.

La principal ventaja teórica de IRT es que mediante su utilización se lograría que un estudiante obtuviese siempre la misma estimación de su habilidad, independientemente de las preguntas (del banco de preguntas testeadas) que le tocó responder[1]. También, con IRT un ítem tendría siempre los mismos parámetros que lo describen (dificultad, discriminación, etc.), independientemente del grupo que rindió la prueba. Esta notable propiedad se llama *invarianza* y es la piedra angular de la teoría de respuesta al ítem. A su vez, es la principal ventaja que distingue a IRT de la teoría

[1] Lo único que podría variar es el error de medición.

clásica. Sin embargo, es preciso destacar que la invarianza se cumple siempre y cuando se satisfagan ciertos supuestos y requisitos que se enunciarán a continuación.

Supuestos de IRT. (1) *Unidimensionalidad*: este supuesto consiste en que en una prueba todos los ítems están midiendo una y sólo una característica de los examinados. Esta propiedad está íntimamente ligada al supuesto número (2) *independencia local*, que postula que, dado un nivel de habilidad, las respuestas a los ítems no pueden estar correlacionadas entre sí. En otras palabras, si hay correlación entre preguntas, ésta sólo se explica por habilidad.

Otros supuestos importantes son: (3) que todos los alumnos que rindan la prueba hayan tenido *experiencias educacionales similares*; (4) que la prueba *no haya sido apurada*, y que (5) no haya "*efectos de contexto" no controlados*. Los "efectos de contexto" se refieren a que algunas preguntas se comportan de modo diferente según la posición que tengan en la prueba. Estos efectos se controlan adjudicando a la pregunta la misma posición en el pretest que en el test operacional.

Requisito básico de IRT. Aparte de los supuestos mencionados, para que se cumpla la propiedad de invarianza, y para que no haya problemas con la estimación de habilidad de los estudiantes, es de gran importancia que las predicciones del modelo se ajusten a los datos reales. Es decir, el modelo tiene que ser capaz de predecir con la mayor exactitud posible el comportamiento de los estudiantes frente a las distintas preguntas.

Ventajas y desventajas de IRT. La teoría de respuesta de ítem presenta una serie de potenciales ventajas sobre la teoría clásica. La principal de ellas es la *invarianza* de los puntajes de la prueba y de las características de las preguntas. También surge, gracias a las curvas de información, herramienta exclusiva de IRT, la posibilidad de optimizar el proceso de selección de preguntas según el objetivo educacional que se persigue[2]. A su vez, con la teoría de respuesta al ítem se hace posible implementar Pruebas Adaptativas de Computador (CAT). En estas pruebas, cada examinado rinde una prueba "a su medida", maximizando así la precisión en la medición. Otra ventaja de IRT es que presenta métodos alternativos para realizar ciertos análisis secundarios, como la detección de sesgos[3], y presenta un

[2] Con IRT se hace posible controlar el error para los niveles de habilidad que se quieren medir con mayor precisión, al seleccionar los ítems más idóneos para este objetivo.

[3] "Sesgos" corresponden a diferencias de desempeño para grupos de igual habilidad. Por ejemplo, un ítem estaría sesgado si los hombres con un cierto nivel de habilidad lo contestan de manera diferente que las mujeres que tienen ese mismo nivel de habilidad. En ese caso, un factor distinto a la "habilidad" que pretende medir la prueba, estaría jugando un rol importante en el desempeño de los individuos.

método alternativo de "equating" (proceso por el cual dos pruebas se hacen comparables).

Por cierto, todas las ventajas anteriores se pierden cuando los supuestos y requisitos de IRT no se cumplen. Para que haya invarianza, piedra angular de IRT y principal ventaja de ésta sobre la teoría clásica, es fundamental que exista, como señalamos anteriormente, unidimensionalidad e independencia local y que las predicciones del modelo se ajusten bien a los datos reales. Sin embargo, en muchas ocasiones la naturaleza de las disciplinas mismas les impide someterse a tales restricciones. Es por ello que en la práctica los supuestos y requisitos de IRT se transgreden a menudo. Este incumplimiento de los supuestos lleva no sólo a perder la invarianza sino que afecta directamente a la estimación de habilidad e introduce errores en aplicaciones secundarias de la teoría.

Veamos a continuación algunos ejemplos de transgresiones típicas de los supuestos y requisitos de la teoría.

De partida, se ha verificado que ciertas disciplinas son claramente multidimensionales, lo que atenta contra el requisito de unidimensionalidad de las pruebas. Por ejemplo, hay estudios que han demostrado que algunas pruebas de ciencias presentan varias dimensiones relevantes. Un estudio del NELS (National Educational Longitudinal Study), prueba de evaluación de la educación en Estados Unidos, detectó que las preguntas de la prueba apuntaban a tres dimensiones diferentes, claramente identificables. Éstas se denominaron "razonamiento y conocimientos básicos", "razonamiento científico cuantitativo" y "razonamiento espacial-mecánico" (Hamilton *et al*. 1997, pp. 181-200, y Nussbaum *et al.*, 1997, pp. 151-173). Es esperable que en otras disciplinas, como ciencias sociales, también haya multiplicidad de dimensiones. Son pocas las soluciones que existen para corregir este problema. Una posibilidad sería seleccionar preguntas que representen sólo a una dimensión, pero esto implicaría descartar preguntas potencialmente valiosas para los objetivos de la prueba. Otra posibilidad es la de crear subpuntajes, uno por cada una de las dimensiones testeadas. Con esto se incurre en una significativa disminución de la precisión en la medición, ya que cada subtest constará de muy pocas preguntas.

También se dan violaciones al supuesto de independencia local, las que se producen cuando las preguntas de una prueba están organizadas en torno a un estímulo común (Kolen y Brennan, 1995). Ejemplos de este tipo de preguntas son las comprensiones de lectura, donde a veces varios ítems se refieren a un mismo pasaje extenso. Lo mismo sucede cuando varios ítems se refieren al mismo gráfico o al mismo diagrama. En esos casos, es muy probable que las preguntas estén correlacionadas entre sí, aun cuando

se haya controlado por habilidad. Prescindir de este tipo de preguntas (comprensiones de lectura de textos largos, análisis de gráficos, etc.) puede ser muy costoso. Por ejemplo, hay tareas que sólo se pueden medir en torno a un texto largo, como la capacidad para seleccionar la idea principal, la capacidad de síntesis, la capacidad de jerarquizar y la capacidad de ordenar una secuencia.

En otras palabras, *si se quiere imponer el cumplimento estricto de los supuestos de la teoría de respuesta al ítem se restringiría el campo de lo preguntable*, obligando a sacrificar buenas preguntas y a crear otras con contenidos ad-hoc. Esto puede redundar en preguntas de poca calidad, es decir, se deterioraría la validez de contenido de la prueba. *Por otro lado, si no se cumplen los supuestos de la teoría, se pierden todas las ventajas que ésta presenta frente a la teoría clásica*, y más aún, se puede incurrir en errores en la estimación de los puntajes de los estudiantes.

Continuando con el tema de las violaciones a los supuestos de IRT, cabe referirse al supuesto de las experiencias educativas similares. Es muy probable que en pruebas de gran escala diferentes estudiantes hayan sido sometidos a experiencias educativas muy distintas. Por ejemplo, si la habilidad que se va a medir son "conocimientos en biología", dos estudiantes que tienen la misma habilidad en ese aspecto pueden haber asistido a clases muy diferentes. El profesor de María le dio mucho énfasis a la dimensión humana de la biología: María debió estudiar el cuerpo humano en detalle, las enfermedades más importantes, etc. Andrea en cambio tenía un profesor que le dio mucho más énfasis a la botánica y a todo lo que tiene que ver con el reino vegetal. Supongamos que Andrea y María son ambas muy buenas alumnas y saben el detalle de lo que se les ha enseñado. Ambas son igualmente "hábiles" en biología. Sin embargo, lo más seguro es que una pregunta sobre polinización, por ejemplo, le va a parecer más fácil a Andrea que a María, y una pregunta sobre poliomielitis le parecerá más fácil a María que a Andrea. Lo anterior indicaría que, al momento de establecer la dificultad de una pregunta, va a ser importante en qué grupo de estudiantes se está testeando este nivel de dificultad. Si el grupo piloto tiene muchas Andreas la pregunta tendrá una dificultad diferente que si el grupo piloto tiene muchas Marías. Con ello, la dificultad de la pregunta deja de ser independiente de la muestra sobre la que se mide. En el fondo, IRT funciona en el supuesto de experiencias educacionales similares. Si no, la invarianza (o independencia de los descriptores de los ítems a la muestra) deja de ser real[4].

[4] Ver, por ejemplo, Miller y Linn (1988) citados en Fan (1998); Masters (1988) y Traub (1983) citados en Linn (1990); y Yen, Green y Burket (1987) citados en Green, Yen y Burket (1989), pp. 297-312.

Aun si se han cumplido todos los supuestos del modelo, es posible que el ajuste de éste a los datos no sea óptimo. Es decir, es posible que el modelo no sea capaz de predecir el comportamiento real de los estudiantes. Un pobre ajuste implica falta de invarianza, con lo que se pierde la principal ventaja de IRT por sobre la teoría clásica. Otro posible efecto de un modelo que no se ajuste a los datos es una estimación errónea de la habilidad de los estudiantes.

Otro problema de IRT tiene relación con la existencia de efectos de contexto. La dificultad y discriminación de una pregunta puede depender de la posición que ésta tiene en la prueba. Ha sido documentado que cierto tipo de preguntas presentan un efecto de "fatiga", es decir, se hacen más difíciles mientras más avanzada es la posición que ocupan. Otras preguntas han demostrado tener un efecto de "práctica", es decir, se hacen más fáciles si ha habido preguntas similares con anterioridad en la misma prueba. Estos efectos deben controlarse manteniendo siempre la posición de los ítems constante, tanto en las pruebas experimentales como en la prueba definitiva.

En el caso de pruebas de admisión universitaria, IRT trae consigo una serie de dificultades prácticas. Por una parte, la comprensión de los puntajes por parte de los estudiantes se dificulta. Con IRT puede suceder que estudiantes que tienen el mismo número de respuestas correctas, incorrectas y omitidas tengan puntajes diferentes, lo que suscitaría recelo e incomprensión por parte de los afectados.

Otro problema práctico tiene relación con la preparación de las pruebas por parte de los liceos. Al usar IRT en una prueba de admisión universitaria, se dificulta la creación de pruebas de ensayo, ya que los establecimientos educacionales no tienen la tecnología para estimar la dificultad y discriminación de las preguntas, ni los puntajes de los estudiantes.

En general, de implementarse IRT, la opinión pública se verá enfrentada a dificultades para verificar si el proceso de asignación de puntajes y de calibración de preguntas ha sido realizado de manera óptima.

El presente trabajo pretende introducir al lector en el tema de las teorías de medición, y en particular en la teoría de respuesta al ítem (IRT). La segunda sección describe de manera somera los fundamentos de la teoría clásica. La tercera sección trata en mayor detalle la teoría de respuesta al ítem, sus modelos y conceptos. En esta sección se estudiarán los distintos modelos de respuesta al ítem, se examinarán los supuestos y requisitos que necesita la teoría para funcionar de manera adecuada, se explicará cómo se estiman los parámetros del examinado (habilidad) y de las preguntas (dificultad y discriminación, entre otros), y se introducirá al lector en las funciones de información, herramienta de IRT que permite adecuar la selección de las preguntas de la prueba al objetivo educacional especifico que ésta persi-

gue. La cuarta sección presenta un recuento y discusión sobre las consecuencias de las violaciones de los supuestos en IRT (unidimensionalidad, independencia lineal, entre otros). En esta sección también se examina la evidencia empírica que existe sobre la propiedad de invarianza. El objetivo de este examen es responder, al menos en parte, la interrogante de si en condiciones de medición reales se da realmente esta invarianza, tanto en las estimaciones de la habilidad del examinado como en la estimación de la dificultad y discriminación de los ítems. La cuarta sección, además, toca el tema de la calidad de las preguntas y la comprensión de los puntajes por parte de la población. Finalmente, la última sección presenta las conclusiones de este trabajo.

2. TEORÍA CLÁSICA DE MEDICIÓN

Esta teoría ha servido a la comunidad durante la mayor parte del siglo XX. Su mayor ventaja es que no depende de supuestos teóricos importantes y es un modelo relativamente simple. En esta teoría, cuando las preguntas de la prueba son dicotómicas[5], la medida de habilidad del estudiante corresponde simplemente al número de respuestas correctas que obtuvo en la prueba (o en algunos casos a este número menos una fracción de las respuestas incorrectas)[6].

La teoría clásica de medición presenta una alta dependencia de la muestra. Las estadísticas que describen los ítems de la prueba (p. ej., dificultad y discriminación) dependen del grupo de estudiantes que la rindieron. Como medida del nivel de dificultad de un ítem se utiliza, en el caso de preguntas con puntaje dicotómico, el cociente entre el número de estudiantes en la muestra que responden correctamente el ítem y el número total de estudiantes (es decir, la proporción de respuestas correctas). La discriminación de un ítem se expresa estadísticamente como la correlación entre los puntajes en el ítem y los puntajes totales de la prueba. Así, un ítem tendrá una alta discriminación cuando los estudiantes que lo responden correctamente son aquellos que tienen los puntajes más altos en la prueba y los que lo responden incorrectamente son aquellos que obtienen los menores puntajes.

[5] Es decir, preguntas en que los puntajes son 0 ó 1, sin graduaciones intermedias.

[6] Esta teoría también permite preguntas en las que se entregue puntaje parcial por una respuesta que no está completa, y permite que las distintas preguntas tengan un peso diferente en el puntaje total, peso determinado a priori según algún criterio definido. Sin embargo, para mayor simpleza, en este documento solamente se analizará la aplicación de la teoría clásica a preguntas dicotómicas con el mismo peso en el puntaje final.

La teoría clásica de medición también presenta una alta dependencia de los ítems, es decir, los puntajes que describen el desempeño de los alumnos dependen del grupo de ítems que se les entregó.

Quizá la principal dificultad que presenta la teoría clásica es que las características de los individuos no pueden separarse de las características de la prueba: ambas sólo pueden ser interpretadas en el contexto de la otra. La característica de los individuos que nos interesa la llamaremos la "habilidad" medida por la prueba, que en teoría clásica está definida por el "valor esperado del desempeño observado en la prueba de interés" (Hambleton, Swaminathan y Rogers, 1991). Como podemos ver, la habilidad de un examinado está definida solamente en términos de una prueba en particular. Cuando la prueba es difícil, el examinado aparecerá con una menor habilidad que cuando ésta es fácil. La dificultad de un ítem está definida como "la proporción de los examinados en el grupo de interés que contestan el ítem incorrectamente". Así, que el ítem sea difícil o fácil depende de la habilidad de los integrantes del grupo. La discriminación de los ítems y la confiabilidad de los puntajes de la prueba también están definidas a partir del grupo de examinados. Así, las características de la prueba y de los ítems cambian cuando el grupo de los examinados cambia, y las características de los examinados cambia cuando el grupo de ítems que conforman la prueba cambia. Lo anterior hace difícil (aunque se han creado métodos[7]) comparar examinados que toman diferentes pruebas, y comparar ítems cuyas características se han obtenido usando diferentes grupos de examinados.

Otro problema de la teoría clásica de medición es que supone que el error de medición es constante para todos los examinados. El anterior es un supuesto poco plausible, ya que, como se explicó en la sección introductoria, la precisión de los puntajes varía de manera desigual según la habilidad de los examinados. Por ejemplo, consideremos a un estudiante que obtiene un puntaje cero. Este puntaje dice que el estudiante tiene una habilidad baja pero no queda claro cuán baja. Por otro lado, cuando un examinado contesta algunos ítems correctamente y otros de manera incorrecta, el puntaje de la prueba provee información más precisa sobre lo que el examinado puede o no puede hacer. Las medidas de confiabilidad[8] en teoría clásica dejan también mucho que desear. Los diversos coeficientes de confiabilidad que

[7] Hay algunos métodos para hacer "equating" con teoría clásica. De hecho, en la PAA Verbal se ecualizan las diferentes formas de ésta, y el SAT I utiliza métodos de "equating" por teoría clásica, complementándolos con IRT.

[8] Medida de estabilidad de los puntajes que responde a la pregunta: ¿Se mantienen los puntajes cuando factores que no tienen relación con los propósitos de medición (contexto) varían?

una prueba entrega son un límite inferior para la confiabilidad o estimadores sesgados de ésta. Finalmente, la teoría clásica es una teoría orientada a la prueba y no al ítem. Así, con esta teoría se hace imposible predecir cómo se va a comportar en un ítem particular un examinado o un grupo de examinados. Preguntas como: ¿cuál es la probabilidad de que un determinado examinado conteste correctamente un cierto ítem?, son necesarias para el desarrollo de una serie de aplicaciones de pruebas. Este tipo de preguntas no se pueden responder desde la teoría clásica.

Por todo lo anterior se han buscado métodos alternativos de medición en los cuales las características de los ítems no sean dependientes del grupo examinado y los puntajes de los examinados no dependan del grupo de ítems utilizados en la prueba. Son deseables métodos que se expresen a nivel de los ítems y no a nivel de las pruebas, y que provean una medida de la precisión en la estimación de cada nivel de habilidad. IRT es uno de estos métodos, que ha experimentado un crecimiento exponencial en las décadas recientes.

3. IRT, MODELOS Y CONCEPTOS[9]

La teoría de respuesta al ítem descansa en dos postulados principales:

- El desempeño de un examinado en un ítem puede ser explicado por un grupo de factores llamados rasgos latentes o habilidades.
- La relación entre el desempeño de un examinado y el grupo de rasgos latentes se puede describir utilizando una función monotónicamente creciente llamada la "curva característica del ítem" (CCI). Esta función describe, para cada ítem, la probabilidad de responder correctamente según el nivel de habilidad del examinado. En general, es de esperar que mientras el nivel de habilidad aumenta, la probabilidad de responder correctamente a un ítem también aumente.

Hay muchos modelos posibles de respuesta al ítem, que se diferencian en la forma matemática de su CCI. Los modelos IRT contienen una serie de parámetros que describen un ítem y uno o más parámetros para describir al examinado. El primer paso en la aplicación de IRT es la estimación de estos parámetros. En este documento nos centraremos en los mode-

[9] La estructura de este capítulo está basada principalmente en Hambleton, Swaminathan y Rogers (1991).

los más utilizados para describir ítems de respuesta dicotómica (es decir, que consideran sólo dos opciones: respuesta correcta o respuesta incorrecta). A pesar de que estos modelos no consideran la posibilidad de omitir una pregunta, hay una serie de posibilidades para incluir este tipo de respuestas en ellos[10].

Los modelos matemáticos empleados por IRT se basan en una serie de supuestos sobre los datos en los que se aplica el modelo. El principal de estos supuestos es que sólo una habilidad o rasgo latente se mide por los ítems que conforman la prueba. El supuesto anterior se denomina *unidimensionalidad*[11]. La *independencia local* es un requisito necesario pero no suficiente para que se cumpla la unidimensionalidad. Es por ello que muchos autores (p. ej. Hambleton *et al.* 1991) mencionan esta propiedad, que describiremos más adelante, como otro de los supuestos de IRT. Otro supuesto que se hace en todos los modelos IRT es que la CCI refleja la verdadera relación entre las variables no observables (habilidad) y las observables (respuestas a los ítems).

Se desprende del párrafo anterior que un requisito para que IRT funcione es que los supuestos enunciados se cumplan adecuadamente. Otro requisito fundamental tiene que ver con el ajuste del modelo a los datos. Un determinado modelo de respuesta al ítem puede ser o no adecuado para un conjunto particular de ítems. Si el patrón real de comportamiento de los estudiantes frente al ítem se diferencia significativamente de las predicciones que el modelo hace, entonces se puede afirmar que el ajuste no es el adecuado. En cualquier aplicación de IRT es esencial examinar el ajuste del modelo a los datos.

Cuando un determinado modelo de IRT se ajusta adecuadamente a éstos, y cuando no hay una violación importante del supuesto de unidimensionalidad, aparecen varias características deseables de IRT:

- Los estimadores de la habilidad de un examinado serán siempre iguales (excepto por errores de medición), independientemente del grupo que rinde la prueba, y los estimadores de los parámetros de los ítems obtenidos para diferentes grupos de examinados serán los mismos. La propiedad anterior, piedra angular de IRT, es la invarianza tanto de los parámetros de los ítems como del rasgo latente del examinado que se pretende medir;

[10] El tratamiento de las respuestas omitidas está desarrollado, por ejemplo, en Lord (1980), pp. 225-231.

[11] Hay modelos multidimensionales para IRT, pero aún no están suficientemente desarrollados. En general, los modelos que se han aplicado en pruebas de gran escala son modelos unidimensionales.

- Con IRT los errores estándares de estimación[12] de habilidad son diferentes para cada nivel del rasgo latente.

A continuación se discutirán en detalle los supuestos de unidimensionalidad e independencia local.

Unidimensionalidad

Un supuesto común de IRT es que sólo una habilidad es medida por el grupo de ítems que conforman una prueba. Este supuesto no se cumple estrictamente en ningún caso debido a que varios aspectos cognitivos, de personalidad y operacionales siempre afectan de una u otra manera al desempeño en las pruebas. Entre estos aspectos se puede contar la ansiedad, nivel de motivación, rapidez, tendencia a responder al azar cuando no se sabe la respuesta, y otras destrezas cognitivas diferentes de la que supuestamente se está midiendo. Lo que de verdad importa es que entre todos los rasgos que afectan al desempeño en una prueba, uno de ellos sea claramente dominante sobre el resto. Existen modelos IRT multidimensionales, pero no están suficientemente desarrollados como para su implementación.

Independencia local

Este supuesto implica que no existen dependencias[13] entre ítems que no sean atribuibles al rasgo latente que se está midiendo, lo que en términos estadísticos se traduce en que, para un individuo o un grupo de individuos de la misma habilidad, la probabilidad de que presenten un determinado patrón de respuestas en un grupo de ítems es igual al producto de las probabilidades de respuesta para cada ítem individual. Cuando se cumple el supuesto de unidimensionalidad, se obtiene también independencia local: en ese sentido los dos conceptos son equivalentes. Sin embargo, es posible que exista independencia local sin haber unidimensionalidad, cuando todos los rasgos latentes que influencian el desempeño son tomados en consideración. Por ejemplo, un ítem de una prueba de matemáticas que requiere un alto nivel de comprensión de lectura va a ser respondido incorrectamente por estudiantes que tienen poca habilidad lectora, sin importar sus destrezas

[12] En IRT el concepto “error estándar de estimación” de la habilidad es similar al de “error estándar de medición” de la teoría clásica.

[13] En lenguaje estadístico, que —dado el nivel de habilidad— no exista correlación entre las respuestas a los diferentes ítems.

matemáticas. En ese caso, la independencia local no se cumpliría. Sin embargo, si todos los examinados tienen el nivel lector adecuado[14], entonces solamente la destreza matemática tendrá efecto en el desempeño. La independencia local no se cumple cuando el ítem de una prueba contiene una pista sobre la alternativa correcta, o cuando provee información relevante sobre la respuesta a otra pregunta. La habilidad para detectar la pista es una habilidad diferente de la que está siendo examinada, por lo que no habrá independencia local.

3.1. Modelos más populares de IRT

La curva característica de un ítem (CCI) es una expresión matemática que relaciona la probabilidad de responder correctamente un ítem con la habilidad medida por la prueba y las características del ítem. Es posible concebir un número infinito de modelos de IRT, pero hoy en día sólo unos pocos se usan normalmente. En general los distintos modelos se diferencian por el número de parámetros que describen los ítems. Los tres modelos unidimensionales más populares son los modelos logísticos de uno, dos y tres parámetros que describiremos a continuación.

3.1.1. Modelo logístico de un parámetro

Es uno de los modelos de IRT más utilizados. Su CCI está dada por

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \qquad i = 1, 2, \ldots n \qquad [1]$$

donde

$P_i(\theta)$ es la probabilidad de que un examinado con habilidad θ conteste el ítem i correctamente.

b_i es el parámetro de dificultad del ítem i.

n es el número de ítems en la prueba.

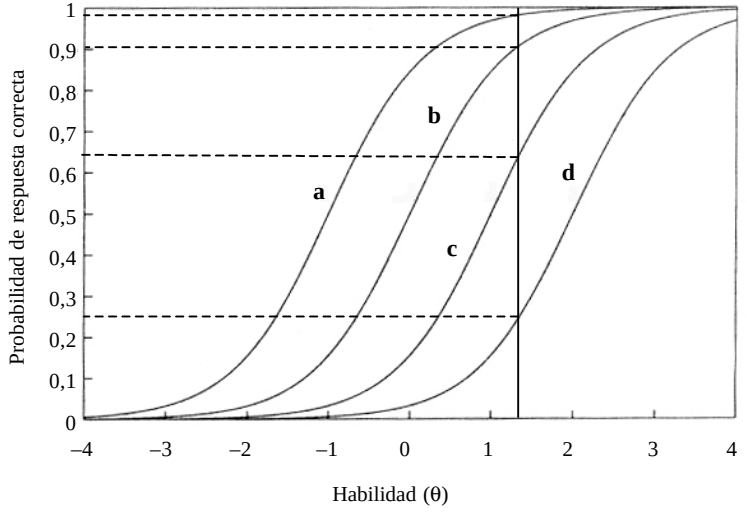
En este modelo, b_i corresponde al punto en la escala de habilidad en el cual la probabilidad de responder correctamente es 0,5. Cuanto mayor es el valor de b_i, tanto más difícil es el ítem, es decir, mayor la habilidad que se requiere para que un examinado tenga una probabilidad del 50% de responder correctamente.

[14] En otras palabras, si controlamos por comprensión lectora.

Cuando los valores de habilidad[15] de los estudiantes se transforman para tener una media de 0 y una desviación estándar de 1, los valores de b_i varían típicamente entre –2,0 y 2,0. Ítems con b_i cercanos a –2,0 corresponden a ítems muy fáciles para el grupo de examinados a partir de los cuales se realizó la estandarización de la escala, e ítems con b_i cercano a +2,0 son ítems muy difíciles.

A continuación se presentan las curvas características para cuatro ítems de distinta dificultad que han sido modelados de acuerdo al modelo logístico de un parámetro. Como se puede observar en la Figura 1, las curvas de este modelo son muy similares una a otra; la única diferencia entre ellas está dada por su posición en el eje horizontal.

FIGURA 1: CURVAS CARACTERÍSTICAS DE ÍTEMS SEGÚN EL MODELO LOGÍSTICO DE UN PARÁMETRO[16]



En la Figura 1, la curva **a** corresponde a un ítem de dificultad –1,0, la curva **b** a un ítem de dificultad 0,0, la curva **c** a un ítem de dificultad 1,0 y la curva **d** a un ítem de dificultad 2,0. Por lo tanto el ítem **d** es el más difícil, mientras que el **a** es el más fácil. La línea vertical corresponde a las personas con habilidad de aproximadamente 1,3. Como podemos ver, una persona que pertenezca a ese nivel de habilidad contestará correctamente la

[15] Los valores de habilidad θ están entre $+\infty$ y $-\infty$.

[16] Figura adaptada de Hambleton, Swaminathan y Rogers (1991), p. 14.

pregunta **d** con una probabilidad inferior a 0,3. Sin embargo, esta persona responderá **c** correctamente con una probabilidad cercana a 0,65, y responderá **b** correctamente con una probabilidad superior a 0,9. La misma persona, casi con certeza, responderá **a** de manera correcta (la probabilidad en este caso se acerca a 1).

En este modelo logístico de un parámetro, también llamado a veces modelo Rasch[17], se supone que la dificultad es la única característica del ítem que influencia en el desempeño del examinado. Este modelo no presenta ningún parámetro que corresponda al índice de discriminación de la teoría clásica. Lo anterior es equivalente a suponer que todos los ítems tienen el mismo grado de discriminación. Este modelo tampoco toma en cuenta que los examinados de baja habilidad van a tender a responder al azar, acertándole a veces a la alternativa correcta en el caso de pruebas con preguntas de selección múltiple.

Claramente el modelo de un parámetro se basa en supuestos bastante restrictivos, por lo que no siempre se ajusta de manera adecuada a los datos.

3.1.2. Modelo logístico de dos parámetros

La CCI para este modelo es menos restrictiva ya que relaja una de las suposiciones del modelo de un parámetro: la suposición de que todos los ítems tienen igual discriminación. En el modelo de dos parámetros aparece a_i, que correspondería a una medida de la discriminación del ítem. A continuación se presenta la forma matemática de la CCI según este modelo.

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \qquad i = 1, 2, \ldots n \qquad [2]$$

Como podemos ver, este modelo se parece mucho al modelo de un parámetro, excepto por la presencia de dos elementos nuevos. El factor D es una constante que se introduce para que la función logística se acerque lo más posible a una normal. Normalmente el valor de D utilizado es 1,7.[18]

Como ya se mencionó, el otro elemento adicional del modelo de dos parámetros es a_i, que corresponde a la discriminación del ítem. La discriminación es una medida de la capacidad que tiene el ítem de distinguir entre

[17] Aunque en estricto rigor el modelo Rasch es diferente matemáticamente al modelo logístico de un parámetro, muchas veces se utiliza el nombre del primero para denominar al segundo.

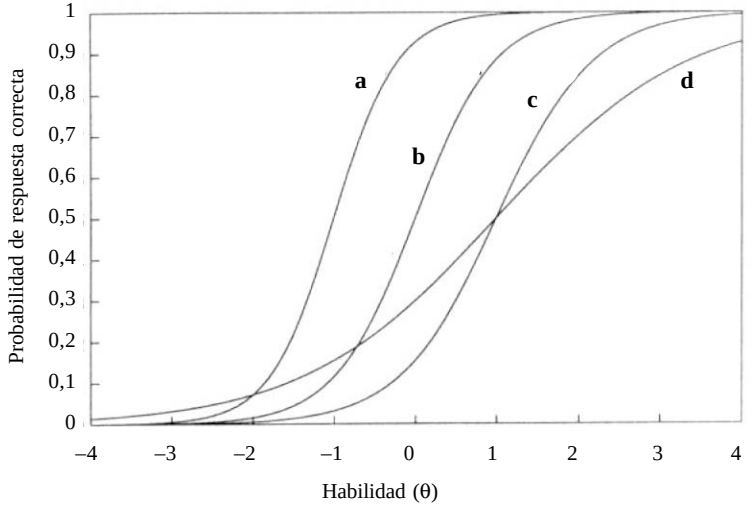
[18] Se ha demostrado que cuando D=1,7, los valores de $P_i(\theta)$ para la ojiva normal de dos parámetros y el modelo logístico de dos parámetros difieren el valor absoluto en menos de 0,01 para todos los valores de θ. (Hambleton, Swaminathan y Rogers, 1991, p. 15).

un estudiante hábil y uno menos hábil. Si la probabilidad de contestar correctamente el ítem es similar para todos los niveles de habilidad, entonces estamos frente a un ítem poco discriminatorio. Si, al contrario, un estudiante con habilidad 1 tiene una probabilidad significativamente inferior de contestar correctamente el ítem que un estudiante con habilidad 2, entonces el ítem está discriminando adecuadamente a los estudiantes cuyas habilidades se ubican en esos rangos. El parámetro de discriminación a_i es proporcional a la pendiente de la CCI en el punto donde $\theta = b_i$. Los ítems con pendiente más pronunciada son más útiles para separar a los estudiantes en los diferentes niveles de habilidad. De hecho, la utilidad de un ítem para separar a los estudiantes con habilidades superiores a un cierto nivel θ de aquellos con habilidad inferior a θ, está dada por la pendiente de la CCI en θ.

Teóricamente, el parámetro de discriminación a_i está definido en la escala $(-\infty, +\infty)$. Los ítems con discriminación negativa (aquellos en que a menor nivel de habilidad del examinado, mayor la probabilidad de que conteste la pregunta correctamente) son eliminados automáticamente de la prueba. Es inusual que un ítem presente discriminaciones superiores a 2, por lo que el rango usual de estos parámetros es (0,2).

A continuación se presentan las curvas características de cuatro ítems que han sido modelados según el modelo de dos parámetros (Figura 2).

FIGURA 2: CURVAS CARACTERÍSTICAS DE ÍTEMS SEGÚN EL MODELO LOGÍSTICO DE DOS PARÁMETROS[19]



[19] Figura adaptada de Hambleton, Swaminathan y Rogers (1991), p. 16.

Los valores de los parámetros para los ítems de la figura son los que siguen:

- Ítem **a**: Discriminación (a) = 1,5 Dificultad (b) = –1,0
- Ítem **b**: Discriminación (a) = 1,2 Dificultad (b) = 0,0
- Ítem **c**: Discriminación (a) = 1,0 Dificultad (b) = 1,0
- Ítem **d**: Discriminación (a) = 0,5 Dificultad (b) = 1,0

Como se puede apreciar en la Figura 2, el hecho de que las curvas presenten distintas pendientes indica que los valores del parámetro de discriminación a_i varían. Por ejemplo, podemos comparar dos ítems con dificultades idénticas pero diferente discriminación (ítems **c** y **d** en la figura) y notar que las curvas se comportan de manera significativamente diferente. La curva **c** es claramente más discriminante que la curva **d**, lo que se traduce en que en la primera es más fácil distinguir a los alumnos de habilidad inferior a 1,0 de aquellos con habilidad superior a 1,0. Como vemos, un estudiante de habilidad 0,5 responde el ítem **c** correctamente con una probabilidad cercana a 0,3, mientras un estudiante de habilidad 1,5 responde correctamente al mismo ítem con una probabilidad cercana a 0,7. Por otro lado, el estudiante de habilidad 0,5 responde correctamente **d** con una probabilidad de alrededor de 0,4, mientras el estudiante de habilidad 1,5 responde correctamente **d** con una probabilidad aproximada de 0,6. Con ello queda claro que la pregunta **c** distingue mejor entre estos dos estudiantes que la pregunta **d**, ya que, como vimos, la diferencia entre las probabilidades de contestar correctamente para estos dos estudiantes es de alrededor de 0,4 para la pregunta **c** mientras que para el ítem **d** esta diferencia se reduce a 0,2.

De la Figura 2 se desprende también que en este modelo de dos parámetros la probabilidad de contestar correctamente el ítem, para una persona de bajo nivel de habilidad es, en todos los casos, cero. Esto indica que el modelo de dos parámetros no considera la posibilidad, para un individuo de habilidad baja, de adivinar la respuesta correcta. Este supuesto puede ser plausible en el caso de preguntas de respuesta abierta pero no en el caso de preguntas de selección múltiple. Sólo si la prueba de selección múltiple es muy fácil es posible que un modelo de dos parámetros se ajuste adecuadamente (Hambleton, Swaminathan y Rogers, 1991), debido probablemente a que en ese caso los estudiantes de bajo nivel de habilidad no necesitarán adivinar, ya que conocerán la respuesta correcta.

3.1.3. Modelo logístico de tres parámetros

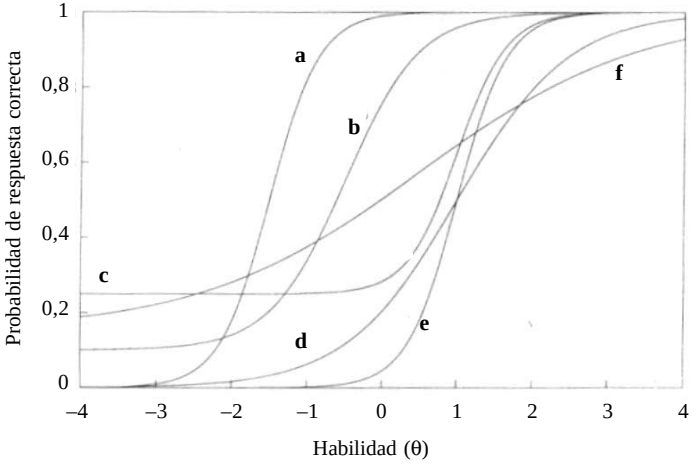
El modelo de tres parámetros incluye la posibilidad de que un estudiante de baja habilidad adivine la respuesta correcta. Con ello se relaja un supuesto del modelo de dos parámetros ($P(\theta) = 0$ para θ muy bajo), que es poco viable en el caso de preguntas de selección múltiple. La expresión matemática para este modelo es

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \qquad i = 1, 2, \ldots n \quad [3]$$

El parámetro adicional que presenta este modelo y que lo diferencia del de dos parámetros es c_i. Este parámetro provee, en caso de ser necesario, una asíntota diferente de cero a la CCI, que corresponde a la probabilidad de que un estudiante con bajo nivel de habilidad adivine la alternativa correcta. Es común que c_i tome valores inferiores que el valor que resultaría si los estudiantes respondiesen al azar el ítem (0,2 en el caso de preguntas con cinco alternativas).

Es interesante notar que cuando hacemos $c_i = 0$, estamos volviendo al modelo de dos parámetros. Si adicionalmente fijamos la discriminación en $a_i = 1$, estamos volviendo al modelo de 1 parámetro.

FIGURA 3: CURVAS CARACTERÍSTICAS DE ÍTEMS SEGÚN EL MODELO LOGÍSTICO DE TRES PARÁMETROS[20]



[20] Figura adaptada de Hambleton, Swaminathan y Rogers (1991), p. 18.

En la Figura 3, que presenta las curvas características de seis ítem, los valores de los parámetros para los ítems son los que siguen:

- Ítem a: Discriminación (a) = 1,8 Dificultad (b) = –1,5 Parámetro c = 0,0
- Ítem b: Discriminación (a) = 1,2 Dificultad (b) = –0,5 Parámetro c = 0,1
- Ítem c: Discriminación (a) = 1,8 Dificultad (b) = 1,0 Parámetro c = 0,25
- Ítem d: Discriminación (a) = 0,8 Dificultad (b) = 1,0 Parámetro c = 0,0
- Ítem e: Discriminación (a) = 1,8 Dificultad (b) = 1,0 Parámetro c = 0,0
- Ítem f: Discriminación (a) = 0,4 Dificultad (b) = 0,5 Parámetro c = 0,15

Como se puede apreciar, las curvas son bastante diferentes unas de otras según el valor de los parámetros que las definen.

3.2. Propiedad de invarianza

Como vimos anteriormente, la principal ventaja que diferencia a IRT de la teoría clásica es la independencia de la muestra, es decir, el que los parámetros de los ítems sean independientes del grupo de examinados para los cuales se estimaron, y los parámetros de habilidad sean independientes del grupo de ítems que contestó el individuo[21]. Es por ello que esta propiedad de "invarianza" es considerada en la literatura como la piedra angular de IRT (Hambleton *et al.*, 1991; Dorans 1990, entre otros).

Todas las ventajas de IRT sobre la teoría clásica se fundan en el supuesto de que la invarianza es una propiedad real y efectiva que en la práctica se cumple. El proceso de "equating"[22] por IRT que incluso permite preecualizar[23] los parámetros de las preguntas, el proceso de detección de sesgos, el cálculo de errores de estimación de habilidad y el proceso de selección de preguntas idóneas para lograr un objetivo dado, son todos aplicaciones que con IRT se logran de mucho mejor manera que por teoría clásica, y que se fundamentan directamente en esta propiedad de invarianza de los modelos de respuesta al ítem.

En términos más técnicos, la invarianza significa que los parámetros que caracterizan a un ítem (a_i, b_i, c_i) no dependen de la distribución de

[21] Excepto el error de estimación, que puede variar según los ítems que se eligen para estimar la habilidad.

[22] Es decir, el proceso por el cual se logra que dos pruebas diferentes sean comparables.

[23] El pre-equating es un proceso de calibración de preguntas que permite que ellas estén calibradas en un banco de ítems incluso antes de ser utilizadas en una situación operacional.

habilidad del grupo de examinados con los que se estimaron, y el parámetro que caracteriza al examinado *(θ)* no depende del grupo de ítems utilizados en la prueba. Cuando el modelo IRT se ajusta a los datos, la invarianza implica que la misma CCI se obtiene independientemente de la distribución de habilidades de los estudiantes del grupo utilizado para estimar los parámetros de los ítems. Sin embargo, para la apropiada estimación de los parámetros del modelo es necesario que la muestra utilizada sea heterogénea. Lo anterior significa que en el grupo de individuos que conforman la muestra con la que se estimarán los parámetros debe haber personas de todas las habilidades, aunque no necesariamente distribuidas de la misma manera que la población objetivo.

Es importante hacer notar sin embargo ciertos puntos respecto de la propiedad de invarianza: esta propiedad rige solamente si el ajuste del modelo a los datos es perfecto. Incluso si el ajuste del modelo a los datos es perfecto, debido a que las CCI describen un comportamiento probabilístico, es muy improbable que se observe invarianza estricta[24].

Es por lo tanto de extrema importancia evaluar hasta qué punto se cumple la propiedad de invarianza, ya que cada aplicación de IRT capitaliza en esta propiedad de los modelos. Aunque la invarianza nunca podrá ser observada en el sentido estricto, es importante evaluar el grado en que se cumple esta propiedad al usar diferentes subgrupos de examinados. Por ejemplo, si dos muestras de distintas habilidades son extraídas de la población y los parámetros de los ítems son calculados para cada una de ellas, la congruencia entre los dos sets de estimadores es una medida del grado de cumplimiento de la propiedad de invarianza. Hasta la fecha, no existen criterios objetivos para la evaluación de esta propiedad, por lo que la decisión de si se cumple o no depende del criterio de quien aplica el modelo.

Es difícil sobrestimar la importancia de la invarianza de los parámetros en IRT. Como se dijo anteriormente, esta propiedad es la piedra angular de la teoría, y su cumplimiento hace posible aplicaciones tan importantes como el equating, creación de bancos de ítems, investigación de sesgo en los ítems, etc.

3.3. Estimación de los parámetros

El primer y más importante paso en la aplicación de IRT a los datos de una prueba es la estimación de los parámetros que caracterizan el modelo elegido. De hecho, la aplicación exitosa de la teoría de respuesta al ítem

[24] Es decir que en todas las instancias en las que se estimen los parámetros se llegue exactamente al mismo estimador.

depende de la disponibilidad de procedimientos satisfactorios para la estimación de los parámetros del modelo.

Tanto la habilidad de los examinados como los parámetros que describen a cada ítem son en principio desconocidos. Lo único que se conoce son las respuestas de los estudiantes a los distintos ítems. Por lo tanto, el problema de la estimación es determinar tanto el valor de θ para cada examinado, como el valor de los parámetros descriptores de cada ítem testeado. Si θ fuese una variable observable o conocida, la tarea de la estimación de los parámetros de los ítems sería bastante más sencilla de lo que de hecho es. El problema es que en IRT el parámetro de habilidad θ es un rasgo latente del examinado que no se conoce a priori.

La estimación de los parámetros se puede realizar de diversas maneras. Se busca encontrar el valor de los parámetros que produzca la curva de mejor ajuste a los datos. Esta estimación de los parámetros para lograr el mejor ajuste posible a los datos se realiza utilizando el criterio de máxima verosimilitud. A continuación se explicará a grandes rasgos en qué consiste este proceso de estimación.

Cuando los parámetros de los ítems son conocidos

En este caso lo que se pretende es encontrar el valor de habilidad (θ) que maximiza la probabilidad de tener un determinado patrón de respuestas[25]. La función por maximizar es la función de probabilidad conjunta que explica el patrón de respuesta a los ítems que tuvo el estudiante. Debido a que rige el supuesto de independencia local, esta probabilidad conjunta no es más que la multiplicación de las probabilidades de respuesta de ese estudiante a cada uno de los ítems. Así, para un determinado estudiante, su habilidad θ se encuentra al maximizar

$$L(u_1,u_2,u_3,...,u_n|\theta)=\prod_{j=1}^{n}P_j^{u_j}Q_j^{1-u_j}=(P_1^{u_1}Q_1^{1-u_1})\times(P_2^{u_2}Q_2^{1-u_2})\times...\times(P_n^{u_n}Q_n^{1-u_n}) \quad [4]$$

Donde u_j corresponde a la respuesta que el estudiante dio a cada uno de los *n* ítems, y toma el valor de 1 ó 0 (pregunta correcta o incorrecta). P_j es la probabilidad de responder correctamente el ítem *j* para el estudiante que tiene habilidad θ, y $Q_j = 1–P_j$ corresponde a la probabilidad de contestar incorrectamente[26]. Nótese que esta función, llamada función de verosimilitud, se define sobre los *n* ítems respondidos por el examinado.

[25] El patrón de respuesta es la secuencia de respuestas que dio el alumno a las preguntas. Por ejemplo, el patrón {1,0,0,1,1,1,0,0,1,0} indica que el estudiante se enfrentó a diez ítems, respondiendo correctamente los #1, 4, 5, 6 y 9.

[26] Recordar que la forma de P_j está dada por la curva característica del ítem descrita en [1], [2] ó [3], según el modelo que se esté utilizando.

Por ejemplo, si tenemos un estudiante que rindió una prueba que constaba de tres ítems, y contestó correctamente el primero (u_1 = 1), incorrectamente el segundo (u_2 = 0) y correctamente el tercero (u_3 = 1), su *habilidad* se obtendría maximizando

$$P_1^1 Q_1^0 \times P_2^0 Q_2^1 \times P_3^1 Q_3^0 = P_1 \times Q_2 \times P_3 \quad [5]$$

Como los parámetros de las preguntas son conocidos, la única incógnita que tiene [5] es el nivel de habilidad θ del estudiante. P_i (y por lo tanto también Q_i) está definido en la CCI del modelo que se ha elegido para los ítems (ya sea [1], [2] ó [3]). Encontrar θ en este caso es directo, basta con encontrar el valor de la habilidad con la cual [5] es máximo.

Cuando los valores de θ son conocidos para cada uno de los estudiantes de la muestra, pero los parámetros de los ítems no se conocen

En este caso el proceso para encontrarlos es similar, pero no idéntico al descrito en el caso anterior. Se deben encontrar ahora los parámetros de los ítems que maximizan, para cada ítem, su función de verosimilitud. Esta función está dada por

$$L(u_1, u_2, u_3, ..., u_K | \theta, a, b, c) = \prod_{i=1}^{K} P_i^{u_i} Q_i^{1-u_i} = (P_1^{u_1} Q_1^{1-u_1}) \times (P_2^{u_2} Q_2^{1-u_2}) \times ... \times (P_K^{u_K} Q_K^{1-u_K}) \quad [6]$$

Donde u_i es la respuesta del estudiante *i* al ítem en cuestión. P_i es la probabilidad que el estudiante *i* tiene de responder correctamente el ítem, y Q_i = 1–P_i es la probabilidad de responder incorrectamente. Nótese que ahora la función de verosimilitud se define sobre los *K* estudiantes que respondieron el ítem.

En este caso, el ejemplo es también levemente diferente. Supongamos que tenemos tres estudiantes que respondieron un ítem. El primer estudiante lo respondió correctamente, el segundo también los respondió correctamente y el tercero lo respondió incorrectamente. Si conocemos las habilidades de los tres estudiantes, debemos encontrar los parámetros (uno, dos o tres, dependiendo del modelo elegido) que maximizan

$$P_1^1 Q_1^0 \times P_2^1 Q_2^0 \times P_3^0 Q_3^1 = P_1 \times P_2 \times Q_3 \quad [7]$$

Como las habilidades de los estudiantes son conocidas, las incógnitas de [7] corresponden a los parámetros del ítem. P_i (y por lo tanto también

Q_i) está definido en la CCI del modelo que se ha elegido para el ítem (ya sea [1], [2] ó [3]). Para encontrar los parámetros, basta con encontrar el valor de ellos con el cual [7] es máximo.

Sin embargo, hay que considerar que en algún momento del proceso ni los parámetros que definen el ítem ni la habilidad de los examinados se van a conocer. Este problema se resuelve considerando simultáneamente todos los ítems y todos los examinados, es decir, maximizando la función de verosimilitud conjunta para ítems y examinados.

Cuando no se conocen los valores de θ ni los parámetros que describen los ítems

En este caso la estimación de ellos se complica. Ahora hay que encontrar los parámetros que maximizan la función de verosimilitud conjunta para todos los ítems y todos los examinados que conforman la muestra. Esta función por maximizar, cuando la independencia local se cumple, está dada por

$$L(u_1, u_2, u_3, ..., u_K | \theta, a, b, c) = \prod_{i=1}^{K} \prod_{j=1}^{n} P_{ij}^{u_{ij}} Q_{ij}^{1-u_{ij}} \qquad [8]$$

Como podemos ver, esta función considera el patrón de respuestas de *K* examinados a *n* ítems. El número de parámetros de habilidad que se busca es *K* (uno por cada examinado), y el número de parámetros de ítems es 3*n*, 2*n* o *n*, dependiendo de si se pretende usar el modelo de tres, dos o un parámetro. Así, para el modelo de tres parámetros se deben encontrar los 3*n*+*K* parámetros que maximizan la ecuación [8].

Hay que notar que la estimación conjunta de los parámetros[27] genera un problema de indeterminación de éstos. Si en la CCI de, por ejemplo, el modelo logístico de tres parámetros se reemplaza θ por $\theta^* = \alpha\theta + ß$, se reemplaza *b* por $b^* = \alpha b + ß$ y *a* por $a^* = \alpha a + ß$, la probabilidad de responder correctamente se mantiene sin cambiar, es decir, $P(\theta) = P(\theta^*)$. Como α y ß son constantes arbitrarias, no existe un único conjunto de parámetros que maximice la función de verosimilitud [8]. Este problema puede ser eliminado eligiendo una escala arbitraria de habilidad o de dificultad (recordar que θ y *b* están en la misma escala). Por ejemplo se puede fijar la media y desviación estándar de los *K* valores de habilidad (o los *n* valores de dificultad), en 0 y 1 respectivamente. No debe olvidarse eso sí que se está eligiendo una escala arbitraria. Los parámetros de nuevas preguntas testeadas en una nueva muestra estarán por lo tanto en una nueva escala. Es

[27] Es decir, la estimación al mismo tiempo de los parámetros que describen a los ítems (*a*, *b* y *c*), y los parámetros que describen a las personas (θ).

por ello que éstos deben calibrarse de manera que sean comparables con los de la muestra original.

Una vez que se ha eliminado la indeterminación al fijar la escala, se calculan los valores de los parámetros que maximizan [8]. Para ello se puede utilizar el procedimiento de *estimación conjunta de máxima verosimilitud*, que corresponde a un proceso en dos etapas. En la primera etapa se eligen valores iniciales para los parámetros de habilidad. Usualmente se utiliza como valor inicial de θ el logaritmo de la razón entre número de respuestas correctas y el número de respuesta erradas del examinado. Estos valores son estandarizados para eliminar la indeterminación y con ellos se calculan los parámetros de los ítems.

En la segunda etapa, utilizando los parámetros de los ítems recién calculados, se estiman los parámetros de habilidad. Este proceso se repite reiterativamente hasta que los valores de los estimadores obtenidos en dos etapas consecutivas no cambien significativamente. El procedimiento de *estimación conjunta de máxima verosimilitud*, a primera vista atractivo, tiene una serie de desventajas. Primero, el método no puede estimar habilidad de un estudiante en el caso de tener todas las respuestas correctas, o todas incorrectas[28]. Segundo, no hay estimadores para los parámetros de los ítems que han sido respondidos correctamente (o incorrectamente) por todos los examinados. Los ítems y los examinados que presenten estos patrones deben ser eliminados de la estimación. Tercero, para los modelos de dos y tres parámetros, el procedimiento no produce estimadores consistentes de los parámetros de habilidad y discriminación, aunque cuando la muestra es suficientemente grande este problema se reduce. Finalmente, en el caso del modelo de tres parámetros, el procedimiento numérico[29] para encontrar los estimadores puede fallar.

Por todo lo anterior, se han desarrollado métodos alternativos para la estimación de los parámetros. Uno de ellos es el de *estimación de máxima verosimilitud marginal*. Con este método, los parámetros de los ítems se estiman sin necesidad de hacer referencia a los parámetros de habilidad (θ), y los estimadores obtenidos son consistentes. Sin embargo, para utilizar este método se recurre a una suposición adicional: la muestra de estudiantes utilizada para hacer la estimación es representativa de la población de interés, y por lo tanto sus habilidades siguen una distribución conocida. Este método de estimación debe utilizarse con muestras suficientemente grandes de examinados.

[28] Algunos investigadores consideran ésta una seria desventaja del método. Otros no lo perciben como un tema problemático. En general, no hay consenso al respecto.

[29] Normalmente el procedimiento numérico utilizado en la estimación corresponde a la versión multivariable del método de Newton-Raphson.

En algunas ocasiones, incluso el método de *estimación de máxima verosimilitud marginal* puede fallar, es decir el proceso puede no converger a un resultado satisfactorio incluso después de muchas iteraciones. Esta falla sucede en especial en la estimación del parámetro *c*, en el modelo de tres parámetros. Malos estimadores de *c* a su vez resultan en la degradación de los estimadores de los otros parámetros de los ítems, y de la habilidad. Existe un proceso de estimación bayesiana que solucionaría este problema.

3.4. Evaluación del ajuste del modelo a los datos

La teoría de respuesta al ítem (IRT) tiene un gran potencial para resolver muchos de los problemas relativos a la medición mediante pruebas. Sin embargo, el éxito de las aplicaciones específicas de IRT no está asegurado por el solo hecho de calcular los parámetros de las preguntas según alguno de los modelos antes mencionados. Las ventajas de IRT se obtienen solamente cuando el ajuste de los datos al modelo es adecuado. Cuando los datos se ajustan pobremente al modelo, los parámetros de los ítems y los estimadores de habilidad de los examinados no tendrán la propiedad de la invarianza.

Hambleton, Swaminathan y Rogers (1991) recomiendan que los juicios sobre el ajuste del modelo a los datos deben estar basados en tres tipos de evidencia:

1. validez de los supuestos del modelo elegido;
2. nivel en el que se cumplen las propiedades esperadas del modelo (principalmente la invarianza de los parámetros);
3. precisión en las predicciones del modelo utilizando datos reales y, si es necesario, datos simulados;

La modelación de los datos con más de un modelo, para poder comparar resultados de ajuste, es una actividad especialmente útil para elegir el modelo apropiado. A continuación se detalla cada uno de los tres puntos anteriores.

3.4.1. Revisión de los supuestos

El principal supuesto para los modelos IRT aquí presentados es la unidimensionalidad. Otro supuesto importante es que la administración de la prueba no fue apurada (es decir, hubo el tiempo suficiente para contestar

las preguntas). Si la prueba fue apurada, entonces los parámetros de las preguntas dependerán no sólo de la habilidad latente del estudiante, sino de su rapidez. Además, cuando la prueba es apurada entra en juego fuertemente un factor de contexto: la dificultad estimada de una pregunta comienza a depender de la posición en la que está en una prueba, ya que, de encontrarse ésta al final, muchos individuos la omitirán o contestarán erróneamente, no por falta de conocimientos sino por falta de tiempo.

El modelo de dos parámetros tiene un supuesto adicional a los anteriormente enunciados: la probabilidad de que un estudiante de habilidad baja adivine la respuesta correcta es mínima o nula. En el modelo de un parámetro se supone además que los índices de discriminación son iguales para todos los ítems.

Hambleton *et al.* (1991) p. 57), presenta un recuento de métodos para examinar cada uno de estos supuestos. Entre ellos, se cita el análisis de factores como uno de los métodos que se utilizan normalmente en la evaluación de la unidimensionalidad.

3.4.2. Revisión de la invarianza

La invarianza se puede medir utilizando varios métodos directos. Por ejemplo, para revisar la invarianza del parámetro de habilidad (θ) se puede entregar a los examinados dos pruebas con niveles de dificultad marcadamente diferentes. Estimadores de la habilidad de los examinados se obtienen a partir de cada una de las pruebas. Si la invarianza se cumple, los estimadores de habilidad no debieran depender de la prueba a partir de la cual se calcularon.

Para medir la invarianza de los parámetros de los ítems se pueden calcular éstos a partir de dos grupos distintos de estudiantes. Si los estimadores son razonablemente similares, entonces la invarianza de los parámetros de los ítems se estaría cumpliendo.

3.4.3. Revisión de las predicciones

Hay variados métodos para chequear las predicciones del modelo. Uno de los más prometedores es el análisis de residuos. En este método se elige el modelo de respuesta al ítem, se calculan los parámetros y se realizan predicciones de desempeño para cada nivel de habilidad. Luego se comparan los valores predichos con los valores de los datos reales. La

diferencia entre el desempeño observado y el predicho es el "residuo bruto". Este residuo se estandariza para tomar en cuenta el error de muestreo. El estudio de estos residuos puede entregar información valiosa. Por ejemplo, es importante que el tamaño de los residuos sea independiente del nivel de habilidad. Es decir, si los residuos en los niveles de habilidad superiores son grandes y los de niveles de habilidad inferiores son pequeños, eso puede ser indicio de que el ajuste no es el adecuado. Una manera de evaluar el ajuste a los datos es examinando la distribución de los residuos. Se considera que el modelo se ajusta adecuadamente cuando los residuos son pequeños y se distribuyen de manera similar que los residuos generados a partir de una simulación del modelo[30].

Pruebas estadísticas, generalmente chi-cuadrado, también se pueden aplicar para evaluar el ajuste del modelo a los datos. Para más detalle sobre estos métodos ver Hambleton, Swaminathan y Rogers, 1991, pp. 57-61.

3.5. La escala de habilidad (o del rasgo latente que se está midiendo)

El propósito fundamental de las pruebas es la asignación al examinado de un puntaje que refleje su nivel de logro de un objetivo medido por la prueba. Este puntaje debe ser interpretado con cuidado y debe ser válido en el contexto del uso que se le dará. En el marco de la teoría clásica, el puntaje asignado corresponde al número de respuestas correctas[31]. En teoría de respuesta al ítem, se supone que cada examinado tiene un nivel de habilidad θ que determina las probabilidades de responder correctamente un ítem. Así, basándose en el patrón de respuestas del examinado a la prueba, un puntaje de habilidad θ es asignado a cada examinado. Desafortunadamente[32], la relación entre θ y el número total de respuestas correctas no es directa. Como fue descrito anteriormente, θ es independiente del grupo particular de ítems administrados y de la población a la que pertenece el examinado. Esta propiedad de invarianza es la que distingue a θ de los

[30] Los residuos generados a partir de la simulación de un modelo muestran cómo se distribuirían éstos si el modelo tuviese un ajuste perfecto. Por lo tanto, si los residuos reales se asemejan a los simulados, el ajuste del modelo es el adecuado.

[31] En el caso de preguntas con puntaje dicotómico (0 ó 1). En algunos casos se agrega una penalización por cada respuesta incorrecta.

[32] El problema que se genera al no existir una relación directa entre el número de respuestas correctas y el puntaje obtenido, es que se hace muy difícil para el estudiante entender el origen de su puntaje y muy difícil para los profesores simular una prueba de ensayo. Así, por ejemplo, es posible (y altamente probable) que dos individuos con el mismo número de respuestas correctas tengan diferentes puntajes finales en las pruebas.

puntajes asignados mediante teoría clásica. La escala de θ debe ser considerada como una escala absoluta con respecto al rasgo latente o habilidad que se está midiendo.

La escala θ original puede ser alterada mediante transformaciones lineales del parámetro que no alteran el cálculo de las probabilidades de responder correctamente los ítems (si es que los parámetros de los ítems son transformados también a su vez mediante el mismo proceso). Otra transformación de la escala de θ, y quizás una de las más importantes y utilizadas, es la *transformación a puntaje verdadero* (τ). En pocas palabras, el puntaje verdadero de un examinado con habilidad θ corresponde a la suma de las probabilidades de responder correctamente los ítems, evaluada en el valor de θ correspondiente. Si graficamos los *puntajes verdaderos* (τ) para todo el rango de posibles θ, lo que estamos haciendo es graficar la suma de las CCI utilizadas en las pruebas. La curva resultante se llama *curva característica de la prueba* (CCP). Mediante esta curva se mapea la relación existente entre la habilidad θ del examinado y su *puntaje verdadero* τ. El *puntaje verdadero* (τ) puede ser considerado como una relación no lineal de θ. Ya que $P_j(\theta)$ está entre 0 y 1, entonces el rango para el puntaje verdadero está entre 0 y n (con n = número total de ítems que conforman la prueba). La transformación de θ a *puntaje verdadero* (τ) tiene implicancias importantes. Primero que nada, se eliminan los puntajes negativos[33]. La nueva escala, además, se mueve en el rango de 0 a n, que es más fácil de interpretar, aunque el puntaje τ obtenido por este método no corresponde a número de respuestas correctas, con lo que su interpretación se puede prestar a malos entendidos. La última implicancia de la transformación de θ a τ es que el τ de un examinado cuya habilidad θ es conocida puede ser computado para un set de ítems que nunca le han sido administrados. Esta característica es utilizada para las pruebas adaptadas al examinado[34] (hoy en día ampliamente utilizadas, por ejemplo, en Estados Unidos, en el GRE y el TOEFL, entre otros).

3.6. Funciones de información para los ítems y las pruebas

La teoría de respuesta al ítem (IRT) entrega una poderosa herramienta para describir los ítems y las pruebas, que mejora y facilita el proceso de selección de preguntas según los objetivos deseados. Éstas son las funcio-

[33] Recordar que θ se mueve en el rango $[-\infty, +\infty]$.

[34] Corresponde a pruebas en las que el computador entrega preguntas más fáciles o más difíciles, de acuerdo al desempeño que ha tenido el individuo en los ítems anteriores. De esa manera se hace posible estimar con menos preguntas y con mayor precisión la habilidad del examinado.

nes de información. Antes que nada, introduciremos un nuevo concepto ampliamente utilizado en IRT, el concepto de *información* de un ítem y de una prueba.

Cuando se habla normalmente de "tener información", se evidencia que se tiene conocimiento de algo sobre un tema u objeto en particular. En estadística y psicometría, el término *información* tiene un significado similar, aunque un poco más técnico. En estadística, *información* se define como el recíproco[35] de la *precisión* con la que un parámetro es estimado (Baker, 2001). Así, al estimar un parámetro con precisión se obtendrá más conocimiento (y por ende información) de él que cuando el parámetro se estima de manera menos precisa. En IRT es posible determinar una curva de información para cada ítem, y una curva de información para cada prueba (que corresponde a la suma de las curvas de los ítems que la integran). La forma de las curvas de información de los ítems, $I_i(\theta)$, es

$$I_i(\theta) = \frac{\left[\frac{dP_i}{d\theta}\right]^2}{P_i(\theta)Q_i(\theta)} \qquad i = 1, 2, ..., n \qquad [9]$$

donde $I_i(\theta)$ es la información que entrega el ítem i en el nivel de habilidad θ, $P_i(\theta)$ es la probabilidad de responder correctamente al ítem i dada por su CCI, y $Q_i = 1 - P_i$ es la probabilidad de responder de manera incorrecta el ítem i. En el caso del modelo de tres parámetros, la ecuación [9] puede simplificarse a

$$I_i(\theta) = \frac{2{,}89a_i^{\,2}(1-c_i)}{[c_i + e^{1.7a_i(\theta-b_i)}][1+e^{-1.7a_i(\theta-b_i)}]^2} \qquad i = 1, 2, ..., n \qquad [10]$$

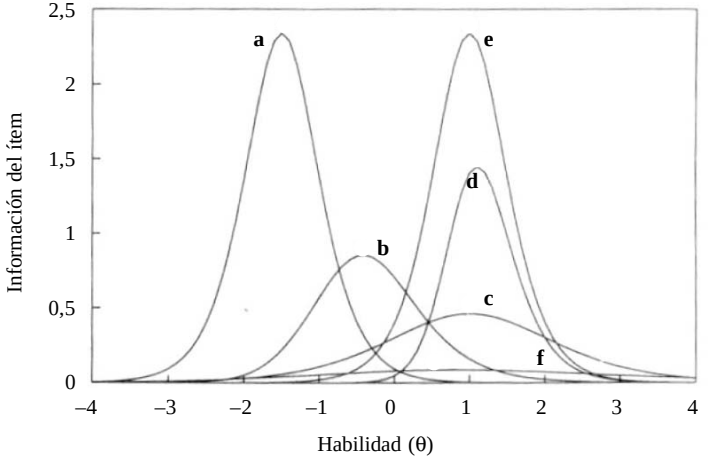
A continuación se presenta la Figura 4, que muestra las curvas de información correspondientes a los seis ítems que se presentaron hacia el final de la sección 3.1.

A partir de [10] y del análisis de la figura, se puede inferir el rol de los parámetros de los ítems en la función de información:

- un ítem entrega información máxima para $\theta = b$, es decir, cuando la dificultad del ítem se iguala a la habilidad del estudiante al que se está evaluando (esto es efectivo en los modelos de 1 y 2 parámetros,

[35] Por "recíproco" queremos decir que precisión e información están en relación inversa. La medida de "precisión" que se usa aquí es la desviación estándar de los estimadores. A medida que la "precisión" (desviación estándar) aumenta, menos precisa es la estimación.

FIGURA 4: EJEMPLOS DE CURVAS DE INFORMACIÓN DE ÍTEMS[36]



en el modelo de tres parámetros, cuando $c > 0$, la máxima información se obtiene para θ levemente superior a b);

- la información de un ítem es generalmente alta cuando el valor del parámetro de discriminación (a) es alto. Esto se puede observar en el caso de los ítems **a** y **e** en la Figura 4, que corresponden a aquellos con mayores discriminaciones;
- la información de un ítem aumenta si c se acerca a 0 (esto se puede corroborar comparando las curvas de información para **c** y **e**, curvas que difieren solamente en el valor de c.

Como se aprecia en la Figura 4, un ítem con niveles bajos de discriminación es prácticamente inútil, ya que aporta muy poca información (observar el caso del ítem **f**). Sin embargo, es importante notar que incluso los ítems con niveles altos de discriminación entregan muy poca información en algunos rangos de habilidad[37]. Podemos ver que **a**, por ejemplo, casi no entrega información sobre los rangos de habilidad superiores a 1, aunque entrega considerable información en el rango de habilidades que van entre –2,5 y –0,5 (recordar que la dificultad (b) del ítem **a** es –1,5, por lo que el

[36] Figura adaptada de Hambleton, Swaminathan y Rogers (1991), p. 93.

[37] Un ítem muy discriminante entregará mucha información para el rango de θ situado en los alrededores de $\theta=b$. Para rangos de θ muy diferentes al nivel de dificultad de la pregunta, ésta entregará información mínima o nula.

ítem estaría entregando más información en los alrededores de $\theta = b$, como es de esperar).

Las funciones de información juegan un rol importantísimo en el desarrollo de pruebas y evaluación de ítems, ya que muestran la contribución que los ítems hacen a la estimación de habilidad para los diferentes puntos del eje θ. Estas curvas permiten elegir el ítem que mejor se ajusta a los requerimientos de la prueba. Como se observó más arriba, un ítem tendrá poco valor si presenta un parámetro *a* bajo y un parámetro *c* alto, ya que en esos casos la información que entregaría sería mínima. La utilidad de un ítem específico va a depender de las necesidades específicas de la prueba que se está desarrollando. Un ítem puede proveer considerable información en un extremo del continuo de habilidad, pero si se requiere información de otro lugar de la escala, su valor se pierde. Sin embargo, se debe tener claro que si el modelo IRT elegido se ajusta pobremente a los datos, la curva de información puede inducir a conclusiones equivocadas.

Claramente, las funciones de información de los ítems proveen un nuevo enfoque para juzgar la utilidad de los ítems y construir las pruebas.

Función de información de una prueba

Esta función está dada por:

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta) \qquad i = 1, 2, ..., n \qquad [11]$$

Es decir, la información que entrega una prueba en un nivel de θ dado corresponde a la suma de las funciones de información de los ítems que conforman la prueba, evaluadas para ese valor de θ.

La cantidad de información provista por una prueba en un determinado θ está inversamente relacionada con la precisión con la cual la habilidad es estimada en ese punto.

$$SE(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}} \qquad [12][38]$$

donde $SE(\hat{\theta})$ es el *error estándar de estimación*[39]. Cuando se conoce la información de una prueba en un determinado θ, puede establecerse una

[38] La ecuación [12] es válida siempre que los parámetros del modelo IRT hayan sido calculados utilizando los métodos de máxima verosimilitud.

[39] El error estándar de estimación es similar a lo que en teoría clásica se denomina error estándar de medición.

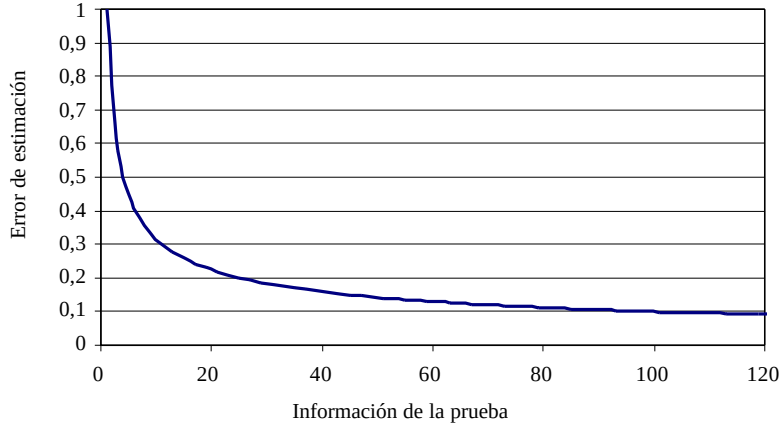
banda de confianza, dada por el error estándar de estimación, que ayuda a la interpretación del valor de θ encontrado. Es importante notar que el valor que toma $SE(\hat{\theta})$ varía con el nivel de habilidad, mientras que en la teoría clásica el error estándar de medición se considera constante.

La magnitud del error estándar depende en general de:

- El número de ítems en la prueba. A mayor el número de ítems, menor el error.
- La calidad de éstos. Mientras mayor es la información que entrega cada ítem, se considera que tiene mejor calidad, ya que aporta de mejor manera a disminuir el error de estimación de la prueba. (Mayor información se obtiene cuando el ítem presenta altas discriminaciones y valores bajos del parámetro *c*).
- La correspondencia entre la dificultad de los ítems y la habilidad del examinado. Pruebas demasiado fáciles o demasiado difíciles para el grupo que se va a examinar resultarán en errores estándares grandes, ya que el máximo de información que entrega la prueba estará en un rango de habilidad diferente al del grupo objetivo.

El valor de los errores estándares se estabiliza para niveles importantes de información. Así, aumentos de la información que entrega una prueba a, por ejemplo, más de 25, tiene sólo un leve impacto en la magnitud del error de estimación de θ. Lo anterior se puede comprobar en el siguiente gráfico que muestra cómo varía el error en distintos niveles de información (Figura 5).

FIGURA 5: INFORMACIÓN DE UNA PRUEBA VS. ERROR ESTÁNDAR DE ESTIMACIÓN DE LA HABILIDAD



Como muestra la Figura 5, el beneficio (en términos de disminución del error de estimación) que se obtiene al agregar una medida de información a una prueba es cada vez menor. En otras palabras, agregar el décimo ítem a una prueba de nueve ítems trae mayores beneficios que agregar el sexagésimo ítem a una prueba de 59 (suponiendo que todos los ítems son igualmente informativos). Sin embargo, no hay que olvidar que a pesar de lo anterior, con la prueba de 60 ítems se entregará una estimación de habilidad más precisa que con una prueba con menos preguntas. La última pregunta aporta cada vez menos, pero de todas maneras aporta algo a la precisión de la medición.

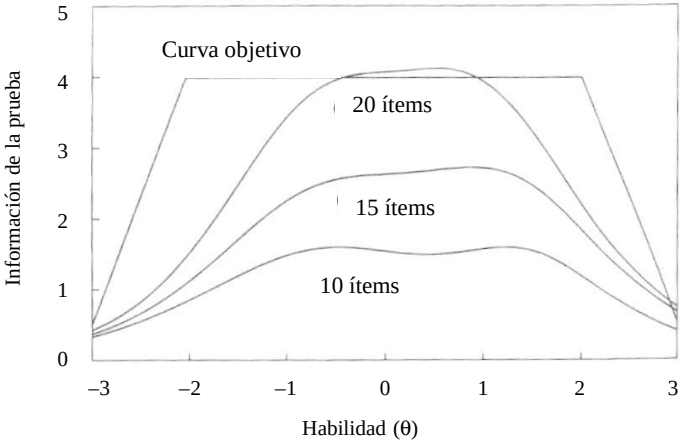
3.7. Construcción de pruebas

La teoría de respuesta al ítem ofrece un método de selección de ítems más poderoso que lo que ofrece la teoría clásica, principalmente debido a la propiedad de invarianza de los parámetros. Además el hecho de que la habilidad θ se mida en la misma escala que la dificultad b, hace posible seleccionar los ítems que son más útiles en ciertas regiones de la escala de habilidad. Quizá la principal ventaja de IRT es que permite seleccionar los ítems de acuerdo a la cantidad de información que aportan a la información total de la prueba, según las especificaciones de ésta. Ya que la información está relacionada con la precisión en la medición, es posible elegir los ítems que produzcan una prueba que mida con determinada precisión un cierto nivel de habilidad de interés.

El primer paso en la construcción de una prueba es determinar la forma deseada de su función de información. Para una prueba que apunta a un amplio rango de habilidades, la función de información objetivo debe ser relativamente plana, de manera de proveer información igualmente precisa en toda la escala de habilidad. Para una prueba en la que se establece un puntaje de corte (por ejemplo para separar expertos de novicios), la función de información objetivo debe tener un claro máximo cerca del puntaje de corte en la escala de habilidad.

La Figura 6 presenta la curva de información objetivo para una prueba en la que el interés es tener información sobre un amplio espectro de habilidades. En esta figura también se muestra el proceso a través del cual se intenta producir, con los ítems disponibles en el banco, una curva que se acerque lo mejor posible al objetivo.

FIGURA 6: CURVAS DE INFORMACIÓN PARA UNA PRUEBA CUYO OBJETIVO ES MEDIR UN AMPLIO ESPECTRO DE HABILIDADES[40]

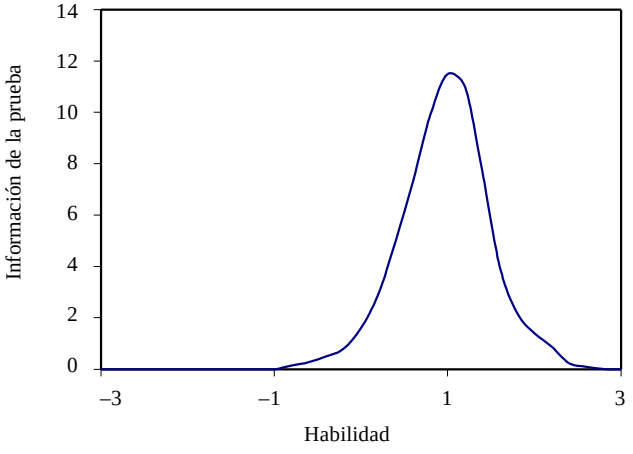


Para construir pruebas referidas a un criterio[41], la forma de la curva de información objetivo será totalmente diferente, buscándose que el peak de información esté alrededor del nivel de habilidad de corte. Por ejemplo, en una prueba de razonamiento matemático queremos distinguir a aquellos que cumplen con un cierto mínimo de aquellos que no lo alcanzan. La habilidad que medirá esta prueba será "razonamiento matemático" y los estudiantes que superen un cierto nivel de corte serán aprobados y el resto será reprobado. Supongamos que el nivel de corte está en $\theta = 1{,}0$, entonces el objetivo de la prueba es distinguir con la mayor precisión posible a aquellos que superan ese criterio de aquellos que no lo hacen. No existe interés en distinguir, por ejemplo, a un estudiante de habilidad 2,0 de uno de habilidad 2,5, ni tampoco importa ser capaces de distinguir a uno de habilidad 0,0 de uno de habilidad –1,0. En este caso, la curva objetivo para la prueba será una curva cuyo máximo de información se ubicará en $\theta = 1{,}0$, y que entregará poca información sobre rangos de habilidad distantes del punto de corte. La Figura 7 ilustra una posible curva de información objetivo para una prueba como la recién descrita.

[40] Figura adaptada de Hambleton, Swaminathan y Rogers (1991), p. 104.

[41] Como se explicó anteriormente, en estas pruebas se busca diferenciar a las personas que satisfacen un determinado criterio de aquellas que no. Para ello, se establece un punto de corte en el continuo de habilidad, alrededor del cual lo deseable es que exista un máximo de información.

FIGURA 7: POSIBLE CURVA DE INFORMACIÓN OBJETIVO PARA UNA PRUEBA REFERIDA A CRITERIO



A continuación presentamos las curvas de información para la prueba de admisión a las universidades en Suecia (SweSAT). Es importante dejar en claro que en el caso sueco, IRT no es utilizado para el desarrollo de las pruebas. La SweSAT todavía usa teoría clásica, aunque respalda a veces los análisis con IRT, por lo que fue posible conocer su curva de información.

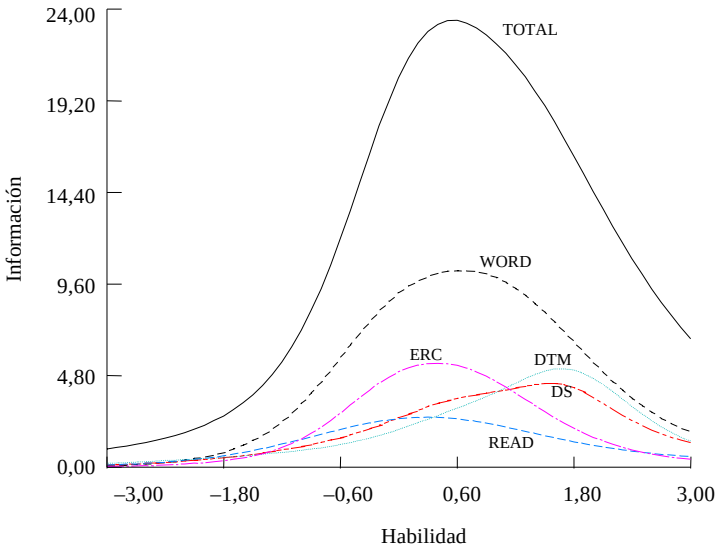
La SweSAT consta de cinco subtests:

- DS: Suficiencia de datos. Mide razonamiento matemático y consta de 22 ítems.
- DTM: Subtest de 20 ítems que miden la capacidad de interpretar diagramas, tablas y gráficos.
- ERC: Comprensión de lectura en inglés. 20 ítems.
- READ: Comprensión de lectura en sueco. 20 ítems.
- WORD: 40 ítems de vocabulario.

La Figura 8 presenta las curvas de información para los cinco subtests y para la prueba total.

Como podemos ver, el subtest WORD es el que más información estaría aportando. Lo anterior se debe a que éste corresponde al subtest con un mayor número de preguntas. Además, podemos ver que los peaks de

FIGURA 8: CURVAS DE INFORMACIÓN PARA EL SWESAT 2002



Fuente: Cristina Stage, Ph. D. Universidad Umeå, Departamento de Medición Educacional, Suecia.

información de todas las curvas están levemente desplazados hacia la derecha, en especial en el caso de los subtests DTM y DS. Este desplazamiento es comprensible, ya que lo que busca una prueba de admisión a las universidades es distinguir con precisión a aquellos que serán seleccionados para estudiar en ellas, que en general corresponden a estudiantes con niveles de habilidad superior a la media. Como podemos ver, en la curva de información TOTAL de la prueba, casi no hay información sobre estudiantes de habilidades inferiores a –1, ya que ese rango de habilidades no interesa para los objetivos de la prueba. Sin embargo, la curva entrega considerable información en el rango de habilidades que va entre 0 y 2, ya que es ése el rango que más interesa.

No existe una curva de información "óptima" en sentido general, ya que cada prueba persigue objetivos diferentes y por lo tanto las curvas de información que satisfacen mejor sus requerimientos son también distintas unas de otras. En el caso chileno, queda en evidencia la incompatibilidad entre una prueba que evalúe adecuadamente a la educación escolar, y una prueba que descubra de manera más precisa a aquellos estudiantes que

tendrían un buen desempeño en la universidad. Para establecer la curva de información óptima para un determinado test se deben definir los objetivos de éste y determinar, mediante una discusión en la que deben participar integrantes de todos los grupos involucrados[42], el segmento en el continuo de habilidad sobre el que se necesita tener mayor información. El uso de funciones de información permite al encargado de desarrollar la prueba diseñar un test que se ajuste estrechamente a las especificaciones (suponiendo que se cuenta con un banco de ítems suficientemente grande y con ítems de calidad).

Sin embargo, existen algunos problemas con este enfoque (el de información de la prueba) que no pueden perderse de vista. Uno de ellos es que el uso de criterios estadísticos para seleccionar los ítems no asegura por sí solo la validez de contenido de la prueba. Es muy fácil sobrevalorar los criterios estadísticos y no tomar en cuenta el importante rol que juega el contenido de los ítems en el desarrollo de las pruebas. Más adelante se presenta una discusión detallada sobre este punto fundamental.

La estimación final y definitiva de los parámetros se hará en la prueba operacional, pero a esas alturas ya es tarde para seleccionar los ítems, por lo que las estimaciones hechas a priori en el estudio piloto servirán para este objetivo. Un problema de utilizar las curvas de información de los ítems en el desarrollo de pruebas es que valores altos de *a* (parámetro de discriminación) obtenidos a partir de grupos experimentales muchas veces son sobreestimaciones[43], resultando en curvas de información sesgadas (Hambleton, Swaminathan y Rogers, 1991). Es muy posible que una prueba construida con ítems de alto coeficiente *a* resulte, en la realidad, diferente de lo esperado. Una solución al problema de sobreestimación de los parámetros de discriminación es agregar algunos ítems extras a la prueba para que sirvan de compensación y la prueba finalmente entregue una información adecuada a los objetivos iniciales. Otra solución es utilizar en el estudio piloto muestras de estudiantes lo suficientemente numerosas en la estimación de los parámetros, de manera de obtener estimadores de éstos lo más correctos posible.

---

[42] En el caso de una prueba de admisión a las universidades, entre los grupos interesados estarían las universidades que utilizarían los resultados de la prueba como instrumento de selección.

[43] En general, la propiedad de invarianza no se cumple estrictamente y muchas veces los parámetros, especialmente b y c y en menor grado a, sí varían según el grupo a partir del cual son estimados.

4. DISCUSIÓN Y ALGUNOS AVANCES DE LA INVESTIGACIÓN ACADÉMICA EN IRT

En la sección anterior se expuso a grandes rasgos lo que es la teoría de respuesta al ítem (IRT). A partir de lo expuesto surgen una serie de interrogantes que vale la pena estudiar más a fondo. Entre ellas está todo lo que tiene que ver con el cumplimiento de los supuestos sobre los que se funda la teoría, especialmente la unidimensionalidad y las implicancias que tiene en el modelo cualquier desviación de ella. Otro tema por profundizar es el que tiene que ver con la invarianza de los parámetros y hasta qué punto ésta se cumple en la práctica. También es de interés la discusión sobre el impacto práctico que puede tener la implementación de puntajes IRT en la población.

4.1. Unidimensionalidad (e independencia local)

Varios estudios académicos se refieren a éste el principal supuesto sobre el que se construye IRT[44]. Al respecto, Childs y Oppler (2000, pp. 939-955), demuestran que el tema de la unidimensionalidad es un asunto de grados. Pequeñas desviaciones del supuesto tienen un impacto ínfimo en la aplicación del modelo. Los estudios de Childs y Oppler se basaron en un grupo de ítems que presentaban indicios evidentes de desviaciones del supuesto. Al comparar parámetros obtenidos en diferentes administraciones de los ítems (a diferentes grupos de personas), observaron que los parámetros de dificultad se mantuvieron estables, aunque los de discriminación no tanto, especialmente cuando se ajustó el modelo de tres parámetros. Los parámetros *c* fueron los que presentaron menor estabilidad. También observaron que los parámetros no variaban significativamente según si la calibración se hacía a partir de submuestras unidimensionales de ítems, o si se hacía a partir del pool completo de preguntas. Finalmente, los investigadores estudiaron la ordenación de los estudiantes al calcular un puntaje único a partir de los datos multidimensionales, y a partir de una combinación de subpuntajes calculados para los subgrupos unidimensionales de ítems. Las ordenaciones obtenidas se correlacionaron entre sí con valores de ρ superiores a 0,99. Childs y Oppler concluyen por tanto que pequeñas desviaciones de la unidimensionalidad no tienen un impacto significativo en la orde-

[44] Debe quedar claro que el hecho de que las desviaciones de la unidimensionalidad no sean detectadas al usar teoría clásica no significa que ésta sea una teoría superior en ese sentido: en teoría clásica también rige el supuesto de unidimensionalidad, pero en el caso de violaciones de ésta, las consecuencias prácticas son mucho menores que en IRT.

nación de los estudiantes que rinden la prueba. Green, Yen y Burket (1989) también sostienen que no es necesario que la unidimensionalidad se cumpla estrictamente para poder sacar provecho de los beneficios que traen consigo los modelos IRT. Otros estudios, que corroboran lo encontrado por Childs y Oppler tienden a demostrar que IRT es robusta a pequeñas desviaciones de la unidimensionalidad (Drasgow y Parsons, 1983; Harrison, 1986; Reckase, 1979)[45]. Sin embargo, Ansley y Forsyth (1985), Doody (1987) y Reckase (1987)[46] muestran que la violación de la unidimensionalidad afecta seriamente a la estimación de los parámetros. Esto se complica más aun con la controversia sobre cuán apropiado es el análisis de factores[47] para analizar la dimensionalidad de IRT. No está claro si las dimensiones definidas por este método corresponden efectivamente a aquellas definidas por los rasgos latentes (Ansley, 1984)[48].

Krisci, Hsu y Yu (2001) examinan cuán robustos son los modelos de IRT a violaciones de la unidimensionalidad y concluyen que en el caso de aplicar modelos IRT a datos potencialmente multidimensionales se debe establecer cuál es la estructura de esa multidimensionalidad. Si hay una dimensión dominante y varias dimensiones menores, posiblemente IRT podrá ser aplicado. Se sugiere una varianza explicada de por lo menos 20% para el factor dominante. Cuando se dan varias dimensiones con predominancias similares, se debe evaluar la correlación entre ellas. Si las diferentes dimensiones están altamente correlacionadas (> 0,4) y las correlaciones son las mismas entre pares de dimensiones, es factible la aplicación de modelos IRT (Ackerman, 1989; Drasgow y Parsons, 1983; Harrison, 1986)[49]. Si el análisis empírico de las preguntas lleva a la conclusión de que existe una fuerte multidimensionalidad y que las distintas habilidades medidas son más bien estadísticamente independientes, una posible solución es reconsiderar la definición del dominio de contenidos de la prueba, dividiéndolo en dos o más dominios relativamente amplios pero más homogéneos. Cada uno de estos subdominios debe tener un puntaje diferente. En el caso de uso de "subpuntajes", se recomienda enfáticamente el uso de bandas de error tanto a nivel individual como a nivel grupal, de manera de evitar que exista una sobreinterpretación de las diferencias entre los diferentes subpuntajes debidas solamente al azar (Tate, 2002).

Del recuento anterior queda claro que IRT podría ser robusto a pequeñas desviaciones de la unidimensionalidad. Sin embargo, desviaciones

[45] Citados en Krisci, Hsu y Yu (2001).
[46] Citados en Krisci, Hsu y Yu (2001).
[47] Método ampliamente utilizado para evaluar la dimensionalidad.
[48] Citado en Krisci, Hsu y Yu (2001).
[49] Citados en Krisci, Hsu y Yu (2001).

"más que leves" del supuesto pueden afectar severamente a las principales aplicaciones[50] de IRT. Por ello es interesante estudiar cuáles son las fuentes más comunes de multidimensionalidad y de dependencia local (ambos supuestos, como se explica en la primera parte de este estudio, están altamente interrelacionados), para así facilitar la detección de estos potenciales problemas.

Kolen y Brennan (1995) explican que la independencia local implica que no hay dependencias entre los ítems que conforman una prueba que no sean atribuibles al rasgo latente o habilidad que el ítem está midiendo. Un ejemplo en el que esta característica probablemente no se cumpliría es cuando las pruebas están compuestas por conjuntos de preguntas basadas en un estímulo común (pasaje de lectura, gráfico, etc.). En este caso, el supuesto de independencia local sería violado porque los ítems asociados a un mismo estímulo probablemente estarán más interrelacionados entre ellos que con un ítem asociado a un estímulo diferente.

Lord (1980) considera que probablemente pruebas de ortografía, vocabulario, comprensión de lectura, razonamiento aritmético, analogías, series de números y pruebas que midan habilidades espaciales serán aproximadamente unidimensionales. Sin embargo previene que es muy fácil imaginar pruebas que violen este supuesto. Una prueba de conocimientos en química, por ejemplo, puede por una parte requerir entrenamiento matemático o destrezas aritméticas y por otra puede requerir manejo de conocimientos no matemáticos. Loyd (1988, pp. 135-143), declara que los ítems de las pruebas basadas en currículum probablemente se desviarán de la unidimensionalidad. Shavelson y Lau (2002) hablan del carácter multidimensional de las pruebas de ciencias. Al respecto Hamilton *et al.* (1997) y Nussbaum *et al.* (1997) encontraron evidencia de la estructura multidimensional del desempeño en ciencias, estableciendo tres dimensiones latentes a las que llamaron "conocimientos y razonamiento básicos", "razonamiento científico cuantitativo" y "razonamiento espacial-mecánico". Los ítems clasificados como de "razonamiento científico cuantitativo" correspondían a preguntas sobre contenidos de química, problemas que requerían computaciones matemáticas o ambos. Muchos incluían términos o procedimientos que se enseñan en cursos de química o física, como el cálculo de densidades o el balance de ecuaciones químicas. La mayoría de las preguntas de "razonamiento espacial-mecánico" incluían diagramas y requerían del uso de visualización del problema. Los ítems de "conocimientos y razonamiento básicos" preguntaban sobre conocimientos fácticos y razonamiento prin-

[50] "Equating", construcción de pruebas, análisis de sesgos, construcción de bancos de ítems.

cipalmente en las áreas de biología y astronomía. Lau *et al.* (2002)[51] encontraron que variables motivacionales predecían el desempeño en ciencias incluso después de controlar por habilidad cognitiva y características demográficas. Kupermintz y Roeser (2002)[52] encontraron evidencia similar a la de Lau *et al.* (2002). Loyd (1988) explica que, en un caso como el anterior, una manera de cumplir con los supuestos de IRT implicaría eliminar los ítems que estuviesen creando la multidimensionalidad. El problema potencial de esto es que la eliminación de ítems puede generar sesgos sistemáticos de contenido o de nivel de dificultad en las pruebas. Traub y Wolfe (1981)[53] son aun más enfáticos al sostener que existe una disparidad entre los objetivos de una prueba de contenidos y el concepto de unidimensionalidad. Advierten que los esfuerzos por lograr la unidimensionalidad seguramente van a restringir el rango de conocimientos examinado.

Otra fuente de violación de la unidimensionalidad son los efectos de contexto[54]. Existe una extensa literatura al respecto, y las conclusiones a las que se han llegado son variadas. En general, se ha observado que en algunas situaciones de examinación (pruebas con límite de tiempo, por ejemplo) y en ciertos tipos de preguntas son más propensos a sufrir estos efectos. Leary y Dorans (1985, pp. 387-413), observan que en pruebas con límite de tiempo los estudiantes presentan un desempeño significativamente mejor cuando los ítems están ordenados de más fácil a más difícil. Yen (1980)[55] exploró los efectos del contexto en el cálculo de los parámetros de los ítems mediante IRT. Su estudio revela caídas consistentes en las correlaciones entre parámetros de discriminación calculados en contextos diferentes (en comparación con las correlaciones entre parámetros calculadas en contextos similares). También observó una baja en la estabilidad de los parámetros de dificultad. En particular, Yen encontró que los ítems de comprensión de lectura eran en general más difíciles cuando aparecían al final de una prueba que cuando aparecían al principio. Kingston y Dorans (1982)[56] realizaron un estudio similar y encontraron que algunas preguntas de la sección analítica del GRE presentaban "efecto de práctica" (es decir, su dificultad disminuía cuanto más cerca del fin de la prueba se ubicasen), mientras que los ítems de comprensión de lectura presentaban "efecto de

[51] Citado en Shavelson y Lau (2002).

[52] Citado en Shavelson y Lau (2002).

[53] Citado en Linn 1990.

[54] Los efectos de contexto tienen que ver con la ordenación de las preguntas en la prueba, y con el contexto en que se rinde la prueba (variaciones en el nivel de ansiedad, extensión total de la prueba, etc.).

[55] Citado en Dorans (1990).

[56] Citado en Dorans (1990).

fatiga" (es decir, la dificultad del ítem era mayor mientras más cerca del fin de la prueba se ubicase). Otros tipos de ítems estudiados por los investigadores no presentaron efectos significativos. Eignor y Cook (1983)[57] también observaron la falta de invarianza de los parámetros de los ítems de comprensión de lectura. Wightman (1981)[58] nota que, cuando la prueba tiene dos pasajes de comprensión de lectura, el nivel de desempeño en el segundo tiende a ser inferior al del primero. Una consecuencia potencialmente seria de la dependencia de la posición de los pasajes de comprensión de lectura es que la precisión real de un puntaje de prueba podría ser apreciablemente menor que la precisión aparente o nominal determinada por la teoría clásica o la teoría IRT. Como resultado de esto, una decisión aparentemente precisa sobre si un estudiante ha logrado un cierto objetivo puede de hecho estar equivocada debido a que el error de medición es mayor que lo calculado (Tate, 2002.)

Siguiendo con el tema de los efectos de contexto, Leary y Dorans (1985) notan que los ítems que miden aptitudes tienden a ser más sensibles a los cambios de orden que los ítems de pruebas de conocimientos. En una nueva investigación, Kingston y Dorans (1984) profundizan en el tema de estos efectos. En esta oportunidad descubren un leve efecto de práctica en los ítems de antónimos y un leve efecto de fatiga en los ítems de comprensión de lectura (GRE). En la sección analítica del examen GRE se observó que los ítems de análisis de explicaciones y los de diagramas lógicos presentaban efectos de práctica importantes y consistentes. Con lo anterior queda claro que los efectos de contexto, especialmente los que se refieren a la ubicación de las preguntas en la prueba, son específicos al tipo de ítem estudiado.

El diseño inapropiado de las condiciones de administración de la prueba puede resultar en la aparición de factores de multidimensionalidad. Por ejemplo, cuando el tiempo para contestar las preguntas es inadecuado para una proporción significativa de la población examinada, entra en juego el factor velocidad. Lo mismo sucede si la prueba es demasiado larga, en que entraría el factor "resistencia a la fatiga". Incluso cuando el procedimiento de administración de la prueba es el adecuado para la prueba operacional, puede haber dudas respecto de los datos recogidos en las experiencias piloto utilizadas para evaluar empíricamente las pruebas. Por ejemplo, existe la preocupación de que la motivación de los examinados en una prueba experimental que "no cuenta" es menor que en la prueba operacional. Esta diferencia puede producir conclusiones inválidas en lo que

[57] Citado en Dorans (1990).
[58] Citado en Leary y Dorans (1985).

respecta a variadas interrogantes empíricas, incluyendo la de la dimensionalidad de la prueba[59] (Tate, 2002).

En el trasfondo de la discusión sobre unidimensionalidad está la siguiente cuestión: podría ser que una determinada prueba, muy apropiada desde el punto de vista pedagógico y de los objetivos educacionales que persigue, presente multidimensionalidad. En ese caso, es importante distinguir que los requerimientos para que IRT funcione con esa prueba no son siempre compatibles con los requerimientos reales de ésta. Por ejemplo, el hecho de que un pasaje largo de comprensión de lectura pueda ser fuente de multidimensionalidad no es razón suficiente para tomar la decisión de no hacer preguntas de este tipo en una prueba. El método estadístico a utilizar en el análisis de una prueba no debe primar por sobre consideraciones más importantes relacionadas con el contenido a medir y el cumplimiento de los objetivos reales que persigue la prueba en cuestión.

4.2. Invarianza

La propiedad de invarianza es la piedra angular de IRT, y lo que la distingue de la teoría clásica de medición. Es bastante sorprendente que estudios empíricos que examinen o comparen la invarianza de los parámetros en el contexto de ambos modelos sean tan escasos. Al respecto, Fan (1998, pp. 357-381), argumenta que "Al parecer la superioridad de IRT sobre la TCM (teoría clásica de medición) en este sentido ha sido dada por sentada por la comunidad de medición, y ningún tipo de escrutinio empírico ha sido considerado necesario. El silencio empírico en esta materia parece ser una anomalía".

Stage (1998a) comparó los estadísticos de la sección de Comprensión de Lectura en inglés de la prueba SweSAT calculados en el marco de la teoría clásica y en el de IRT. Descubrió que los estadísticos que definen discriminación y dificultad tenían alta correlación entre los modelos. Stage también estudió la correspondencia entre los datos obtenidos en los pretests y los datos de la prueba operacional, no encontrando diferencias importan-

[59] Es posible que una prueba, en el contexto experimental (menos presionado), muestre, por ejemplo, dos dimensiones, pero que estas dos dimensiones se fundan en una en el contexto operacional. Por ejemplo, si la prueba es de ciencias, es muy posible que en la prueba piloto los estudiantes sepan o biología o física, según sus preferencias particulares, transformándose éstas en dimensiones diferentes. Sin embargo, en el contexto operacional, los estudiante tienen más incentivo para rendir y estudiar la materia que no conocen. Por lo tanto, es posible que los que saben biología aprendan física y viceversa, transformándose entonces la medición en una medición unidimensional del rasgo latente "conocimiento científico".

tes en lo que se refiere a dificultad de las preguntas[60], pero sí en la discriminación. En teoría clásica la correlación entre las discriminaciones calculadas en el pretest y las del test operacional fue de 0,57, mientras que en IRT fue de 0,34. La conclusión final de Stage fue que la supuesta superioridad de IRT sobre la teoría clásica no se verificaba con sus datos. La investigadora obtuvo resultados similares al estudiar la sección de Comprensión de Lectura en sueco del SweSAT (Stage, 1998b), y de la sección de Vocabulario (Stage 1999).

Miller y Linn (1988)[61] reportaron también considerables diferencias en las curvas características de sus ítems, resultado de variaciones en la cobertura de instrucción de distintos grupos. Lo anterior sugiere falta de invarianza de sus parámetros. Masters (1988) y Traub (1983)[62] sostienen que en pruebas de conocimientos los parámetros de los ítems van a ser sensibles a diferencias en la experiencia educacional. Otro estudio de Cook, Eignor y Taft (1988)[63] también reporta falta de invarianza en los estimadores de dificultad de los ítems, tanto en el contexto de la teoría clásica como en el contexto de IRT. Yen, Green y Burcket (1987)[64] explican que algunos ítems pueden presentar parámetros marcadamente diferentes en ciertos establecimientos educacionales y distritos debido a patrones específicos de instrucción y aspectos curriculares. Por ejemplo, una prueba de aritmética puede ser esencialmente unidimensional entre grupos a los que se les ha enseñado prácticamente lo mismo. Sin embargo, si algún grupo de alumnos de tercer año ha pasado la unidad de multiplicación y otro grupo no, los estimadores de los parámetros para los ítems de multiplicación derivados de la muestra de estudiantes que conocen el tema van a ser sustancialmente diferentes de los parámetros derivados del grupo que no sabe multiplicar (Green, Yen y Burket, 1989).

Fan (1998) estudió la invarianza de los parámetros de dificultad y discriminación de los ítems utilizando para ello los datos del Texas Assessment of Academic Skills, que corresponden a pruebas basadas en currículum. Con respecto a los parámetros de dificultad, Fan encontró que los parámetros de teoría clásica eran levemente más invariantes que los de IRT. También observó que los parámetros de dificultad en ambas teorías eran más invariantes que los de discriminación, sin observarse una ventaja sistemática de un modelo por sobre otro. Cabe señalar que los resultados ante-

[60] Con teoría clásica, la correlación entre los estadísticos de dificultad del pretest y los del posttest fue de 0,87 y con IRT 0,88.

[61] Citado en Fan (1998).

[62] Citado en Linn (1990).

[63] Citado en Fan (1998).

[64] Citado en Green, Yen y Burket (1989).

riores no se deben a desviaciones de la unidimensionalidad ni a pobre ajuste entre modelo y datos. Ambos supuestos fueron chequeados por el autor y obtuvo resultados razonables de ajuste para los modelos de dos y tres parámetros, verificándose un nivel adecuado[65] de unidimensionalidad.

En general, existe evidencia mixta respecto de la consistencia de los parámetros de los ítems. Algunos investigadores han dicho que estos estimadores, al ser calculados a partir de diferentes muestras, han resultado razonablemente consistentes (Dorans y Kingston, 1985; Forsyth, Saisangjan y Gilmer 1981; Rentz y Barshaw, 1977)[66]. Otros han encontrado que los estimadores varían considerablemente (Cook, Eignor y Taft, 1984; Loyd y Hover, 1980; Slinde y Linn, 1978)[67]. Esta combinación de estudios sugiere que los modelos IRT no pueden ser aplicados sin un análisis cuidadoso de la situación. Cada situación de medición debe ser evaluada individualmente, para poder establecer la aplicabilidad del modelo (Loyd, 1988).

4.3. Impacto práctico de IRT

Las curvas características de los ítems y las curvas de información para las pruebas son herramientas valiosas y efectivas en la construcción de pruebas. Sin embargo, no debe nunca olvidarse que este tipo de herramienta no sirve de nada sin un buen banco de preguntas. Las herramientas estadísticas que IRT provee al análisis de los ítems son un complemento, pero no reemplazan ni permiten evadir el esfuerzo editorial inherente a la construcción de buenas preguntas (Green, Yen y Burket, 1989). Es importante que las preguntas no sólo exhiban los parámetros adecuados, sino que deben ser óptimas también en lo que se refiere a contenido y diseño, y ese tipo de evaluación no tiene nada que ver con las herramientas estadísticas aquí presentadas. Linn (1990) hace notar que en la descripción de los métodos de selección de ítems por IRT no se hace mención alguna del contenido de los ítems a ser seleccionados. No debe olvidarse que en pruebas de conocimientos las especificaciones de contenido son prioritarias. Al parecer, existe preocupación por parte de algunos académicos de que la sobrevaloración de la estadística lleve a distorsiones en las mediciones. Rudner (1998) enfatiza el punto: “los bancos de ítems y la teoría de respuesta al ítem no son la panacea que va a solucionar todos los problemas de medición. Sigue

[65] No está demasiado claro cuál es este “nivel adecuado”. La única sugerencia que se encontró fue la de Reckase (1979), citado en Krisci, Hsu y Yu (2001), que habla de que el factor dominante explica un 20% de varianza.

[66] Citados en Loyd (1988).

[67] Citados en Loyd (1988).

siendo importante el cuidado y esfuerzo en la creación de ítems. [...] Se debe hacer todo el esfuerzo posible para incluir sólo ítems de alta calidad en el banco de ítems. El mismo cuidado y esfuerzo debe ser dedicado a la creación de las preguntas. Los ítems obtenidos de fuentes externas deben ser evaluados cuidadosamente tanto en lo que se refiere a correspondencia con el currículum como a la calidad técnica de éstos."

Otro problema práctico importante que IRT presenta a los profesionales de educación es la complejidad en la estimación de los parámetros y en general del modelo matemático en que se funda IRT. Al contrario de la teoría clásica, la aplicación de IRT a los ítems requiere del manejo de un modelo matemático complejo, y genera puntajes difíciles de interpretar. Para llevar a cabo el proceso se requiere de software computacional sofisticado y de personal calificado, lo que implica que los profesores y personas encargadas de preparar a los estudiantes para las pruebas probablemente no tendrán acceso a los parámetros de las preguntas que diseñen. Incluso si tuviesen estos parámetros, el proceso de estimación del puntaje del estudiante no es trivial, requiriéndose para ello un software especial. Esto dificulta sobremanera la posibilidad de hacer estimaciones a priori de cuál será el desempeño final de los estudiantes. Además, ni el estudiante, ni sus familiares, ni los profesores tendrán las herramientas necesarias para hacer un juicio independiente sobre cuán razonables les parecen los resultados en las pruebas. Todo esto añade un áurea misteriosa a los resultados de IRT (Lloyd, 1988).

Este misterio que rodea a IRT también se aplica a la escala de puntajes obtenidos por este método. La manera en que los parámetros que describen los ítems y el parámetro de habilidad θ del examinado están expresados no tiene ninguna relación directa con datos observables (es decir, número de respuestas correctas). No sólo es posible, sino altamente probable, que dos alumnos con exactamente el mismo número de respuestas correctas, incorrectas y omitidas tengan puntajes diferentes. Con ello, la generación de puntajes mediante IRT parece, para quien no maneja el modelo, un proceso mágico, lo que impide al ciudadano común hacer una revisión y chequeo de los resultados (Lloyd, 1988).

En el caso de Estados Unidos, donde existe una amplia gama de pruebas con consecuencias para admisión universitaria de pre y postgrado, la visualización de los problemas prácticos que trae IRT ha hecho que la principal compañía encargada del diseño de pruebas ETS[68] decida evitar en lo posible la asignación de puntajes y corrección de pruebas utilizando la

[68] Educational Testing Service. Encargados de diseñar, entre otras, las pruebas SAT I, SAT II, GRE, TOEFL, GMAT, PRAXIS, etc.

teoría de respuesta al ítem. Es por ello que *los puntajes de todas las pruebas de consecuencias administradas en formato de papel y lápiz son asignados por teoría clásica*. En el caso de pruebas adaptativas al computador, donde la teoría clásica no es una posibilidad, ETS y otras compañías privadas (por ejemplo Kaplan, Barrons, Petersons) han desarrollado software especializado para su preparación[69]. Este software entrega ensayos computacionales de la prueba, de manera que el estudiante practique y obtenga estimaciones reales de su puntaje al final de cada práctica. Esto no sería posible mediante un facsímil impreso en papel, ya que estas pruebas utilizan IRT y por lo tanto la estimación de los puntajes no sería posible sin el software. En suma, los problemas prácticos que traen consigo los exámenes que utilizan IRT como principal teoría han llevado a que Estados Unidos haya hecho lo posible para mantener la teoría clásica y, en el caso de no ser esto posible, las instituciones encargadas del diseño de estas pruebas se han encargado de entregar a los estudiantes, muchas veces de manera gratuita, un software que les permita ensayar la prueba de manera adecuada.

5. CONCLUSIONES

Como hemos visto, IRT es una teoría que promete solucionar una serie de problemas que hasta ahora existían en la medición educacional. La principal ventaja que ofrece esta teoría es la invarianza de los parámetros que describen los ítems (dificultad, discriminación, probabilidad de responder correctamente al azar), y de los parámetros que describen a las personas (la habilidad θ en el caso de modelos unidimensionales).

La utilización de IRT en contextos de evaluación educacional es sin duda un aporte significativo que facilitará y perfeccionará la tarea de diseño e implementación de pruebas, especialmente aquellas de gran escala. IRT es una teoría emergente que está en continuo perfeccionamiento y es probable que en un futuro cercano las bondades de los modelos IRT sean aun mayores. Con todo, es importante poner una nota de cautela, ya que las bondades de la teoría pueden ser fácilmente sobredimensionadas.

De partida, no se debe olvidar que IRT no es necesariamente una alternativa a la teoría clásica de medición, sino que puede también ser visto como un complemento. Mientras más información se tenga de una prueba y

[69] En el caso de ETS, el Software POWERPREP se diseñó para preparar las versiones adaptativas al computador del GRE, GMAT y TOEFL. En los primeros dos casos POWERPREP es gratuito.

de sus ítems, mejor. Al usarse las dos teorías se están haciendo simultáneamente los análisis con ambas, con lo que, de fallar alguna, siempre estará la otra como respaldo.

Tampoco se debe olvidar que la teoría clásica tiene algunas ventajas significativas, especialmente en lo que se refiere a la compresión de los puntajes que entrega. En general, sobre todo en el caso de pruebas con consecuencias, con estudiantes altamente motivados y familias altamente involucradas, la importancia de comprender los puntajes se acrecienta. El proceso de generación de puntajes de la teoría clásica[70] tiene un fundamento sencillo y comprensible para la población en general. La utilización de la teoría clásica impediría que dos estudiantes con el mismo número de respuestas correctas, omitidas e incorrectas tengan diferentes puntajes en la prueba. La situación anterior sería pan de cada día en el caso de asignarse puntajes utilizando IRT, y sería incomprensible para quienes no están familiarizados con la teoría.

Hay quienes sostienen que el hecho de que con IRT sea posible asignar distinto "peso" a preguntas con distinta dificultad y discriminación es una ventaja que por sí sola justificaría el uso de la teoría de respuesta al ítem en la asignación de puntajes. Sin embargo, de ser ése el único objetivo buscado, no se justifica el cambio de teoría. Con la teoría clásica es posible asignar distintos pesos a diferentes preguntas según criterios previamente establecidos, y el proceso de asignación de pesos en este caso podría hacerse más transparente a la opinión pública que al utilizar IRT. Sin embargo, aún en este último caso los beneficios que traería el asignar pesos diferenciados a las preguntas son cuestionables.

La asignación de puntajes por teoría clásica facilita también la tarea de los profesores que van a proveer a los estudiantes de pruebas de ensayo. Cuando se utiliza IRT, es muy difícil para un profesor predecir el puntaje que tendrá un alumno en la prueba real, ya que lo más seguro es que no tendrá los parámetros de las preguntas que él mismo ha diseñado para el ensayo, ni las herramientas para estimar el puntaje a partir de éstos. Cuando se utiliza teoría clásica la tarea de puntuar correctamente pruebas de ensayo se facilita sobremanera. No debemos perder de vista la experiencia internacional[71] que nos enseña que estos problemas prácticos han sido una de las razones principales por las cuales se ha optado, siempre que ha sido posible, por mantener la asignación de puntajes por teoría clásica, especialmente en pruebas con consecuencias para los examinados. IRT se puede utilizar

[70] Porcentaje de respuestas correctas o número de correctas menos una fracción de las incorrectas, en el caso de preguntas de puntaje dicotómico.

[71] Por ejemplo, el caso sueco y el caso de Estados Unidos.

para otros procesos paralelos en los cuales presenta ventajas reales, como estudios de sesgos, creación de bancos de datos, selección de preguntas y, en algunos casos, "equating"[72].

Como se ha destacado anteriormente, IRT es un buen aporte a la evaluación educacional, pero no es la panacea. A continuación ahondaremos en este punto explicando los peligros de su sobrevaloración.

IRT es una teoría que se funda en una serie de supuestos que se cumplen rara vez en la realidad. Es importante evaluar que los supuestos se cumplan de manera adecuada, es decir, que si existen desviaciones del óptimo, éstas no sean tan importantes como para invalidar las aplicaciones de IRT y afectar significativamente la propiedad de invarianza de los parámetros. Para ello, es fundamental que expertos en el tema realicen las pruebas adecuadas para testear si efectivamente se está cumpliendo la unidimensionalidad y si el modelo se está ajustando a los datos experimentales. Si no es así, las medidas que se tomen al respecto deben ser compatibles con los objetivos primordiales de la prueba. Por ejemplo, si se detecta que un determinado tipo de pregunta es fuente de multidimensionalidad, se debe evaluar el efecto desde el punto de vista del contenido y propósitos de la prueba, y cuáles serían los efectos de su eliminación. Es posible que en muchos casos sea preferible cambiar el modelo estadístico a utilizar antes que desechar definitivamente el tipo de pregunta.

IRT es una teoría emergente que aún es debatida en el área académica. Existen muchos puntos a afinar y no hay consenso en cuáles son las pruebas óptimas para evaluar el cumplimiento de los supuestos y el ajuste del modelo a los datos. Es muy importante que los expertos encargados de implementar el modelo estén al tanto de las nuevas investigaciones y vayan avanzando conjuntamente con los progresos que se vayan dando en la teoría.

Es fundamental también que se estudie el efecto que tienen en los parámetros las variaciones significativas de experiencias educacionales de los estudiantes, variaciones de motivación y cambios en la ordenación de las preguntas en la prueba. De detectarse que hay efectos significativos en las estimaciones[73], deben tomarse medidas atinentes. Por ejemplo, si se detecta que un cierto tipo de preguntas presenta un marcado efecto de fatiga, debe cuidarse que la ubicación de éstas en las pruebas piloto sea similar a la ubicación que tendría en la prueba operacional. Debido a que es

[72] Sin embargo, debe evaluarse para cada caso en particular cuál es el método idóneo de equating. Puede ser algún método basado en IRT, o alguno de teoría clásica.

[73] Es decir, que la invarianza no rige para poblaciones que han tenido diferentes experiencias educacionales, o que las preguntas presentan distintos parámetros según la posición que tienen en la prueba.

común que los parámetros estimados a partir de estudios piloto varíen, es altamente recomendable recalcularlos nuevamente en el contexto de la prueba operacional. Los parámetros obtenidos en la prueba piloto serán útiles para el proceso de selección de preguntas para la prueba, situación en la que es deseable, pero no fundamental, que éstos estén bien estimados. Sin embargo, para el cálculo final de los puntajes es primordial que los parámetros utilizados no estén sesgados[74], de manera de obtener una estimación acertada y precisa de la habilidad de los estudiantes examinados.

Es muy importante nunca perder de vista el objetivo primordial de la prueba. La sobrevaloración de los datos estadísticos que describen el comportamiento de las preguntas puede conducir a un actitud miope en el diseño de las pruebas, y traer consigo consecuencias desastrosas en la enseñanza. Es muy importante no sacrificar aspectos primordiales de la educación en aras del cumplimiento de los supuestos de una teoría. Por ejemplo, es muy posible que en una prueba de ciencias sociales se observen desviaciones significativas de la unidimensionalidad. Es probable que las preguntas de geografía estén midiendo un rasgo latente (o habilidad) diferente del que miden las preguntas de historia universal, o de economía. La solución estadística a este problema es simplemente seleccionar las preguntas de manera que se mantenga la unidimensionalidad. Eliminar las preguntas de geografía podría ser una solución en este caso particular. Lo anterior es razonable desde el punto de vista de IRT, pero no es razonable desde el punto de vista de los objetivos primordiales de la prueba. No debe perderse de vista que en una situación de medición el modelo debe adecuarse a los objetivos de la prueba y no viceversa. Una posible solución al problema de la multidimensionalidad y que no entraña tener que desechar preguntas, es la entrega de subpuntajes para cada uno de los rasgos independientes de la prueba. En el caso anterior esto se traduciría en entregar el puntaje de geografía separado del de historia. El único problema que esto podría traer es que estos subdominios pueden contener pocas preguntas, con lo que aumentaría el error de estimación de cada rasgo latente.

El problema de las violaciones de la unidimensionalidad no se da solamente en casos como el anterior, en que los subdominios corresponden claramente a contenidos diferentes. Puede suceder también que los distintos niveles de dificultad de las preguntas sean una fuente de multidimensionalidad. Por ejemplo, las preguntas más fáciles pueden estar midiendo el rasgo "capacidad de memorizar datos", mientras las difíciles pueden estar midiendo "capacidad de razonamiento". Ambas habilidades posiblemente serán

[74] Es decir, que no se desvíen de su valor real.

independientes y, por lo tanto, fuente de multidimensionalidad. También existe evidencia de que las preguntas de bloque, es decir, una agrupación de preguntas que se basan en un estímulo común (pasaje de lectura, gráfico, etc.), presentan dependencia local y por lo tanto son fuente de multidimensionalidad. Existe evidencia documentada de la multidimensionalidad del conocimiento científico. Shavelson y Lau (2002) hablan de esto y citan varios autores que han verificado la existencia de múltiples dimensiones en pruebas de conocimiento en estas áreas. Nuevamente hay que tener cuidado de no pasar a llevar los objetivos primordiales de la prueba en aras de mantener el supuesto de unidimensionalidad (por ejemplo, eliminando por completo las preguntas de razonamiento, eliminando los bloques de preguntas, o eliminando preguntas que reflejen alguna de las dimensiones del conocer científico).

Otro tema de radical importancia tiene que ver con la validez de las preguntas y la idoneidad editorial de éstas. La validez de las preguntas es un concepto bastante amplio, que tiene que ver con el nivel en que las preguntas son relevantes y representan efectivamente el campo que pretenden representar. La validez también tiene que ver con el nivel con que una pregunta está midiendo el constructo que intenta medir (por ejemplo, que una pregunta que pretende medir razonamiento matemático no esté midiendo en vez comprensión de lectura), y que sirva para el objetivo para el que se va a utilizar (si el objetivo de la prueba es predecir rendimiento académico en la universidad, que la habilidad medida sirva efectivamente para eso –validez predictiva—). La idoneidad editorial tiene que ver con la claridad en la diagramación y redacción de la pregunta. Los aspectos recién mencionados son tanto o más importantes que las teorías de medición utilizadas, y son totalmente independientes de ellas, es decir, muchos de los defectos de las preguntas en lo que se refiere a su validez de contenido e idoneidad editorial no son detectables por los métodos estadísticos de análisis de ítems. La sobrevaloración de las teorías de medición, exacerbada en el caso de IRT por ser una teoría tan promisoria, lleva a que en muchos casos se olvide que no basta con que una pregunta tenga la dificultad y discriminación adecuadas, sino que se debe cuidar de estar midiendo un conocimiento relevante en el contexto de los objetivos de la prueba.

Algo similar a lo anterior sucede cuando se tiene un exceso de confianza en las bondades de IRT en lo que se refiere al proceso de la construcción de las pruebas[75]. Por ejemplo, en una prueba de currículum de biología las preguntas se escogerían de acuerdo al aporte que entregan en la

[75] Es decir, el proceso de seleccionar las preguntas desde un banco de ítems.

configuración de una función de información objetivo de la prueba[76]. Luego se extrae del banco de ítems el grupo de aquellas que generan el mejor ajuste a ésta. Las preguntas contenidas en este grupo no necesariamente conformarán una muestra equilibrada del currículum. Por ejemplo, es posible que todas la preguntas seleccionadas tengan como tema la célula y ninguna trate el organismo humano, o ecología, o genética. A pesar de generar una curva de información idónea, el grupo de preguntas está lejos de cumplir los objetivos que tenía la prueba. Es por ello que los expertos en construcción de pruebas tienen que establecer en lo posible las restricciones adecuadas al proceso de selección de preguntas, y deben revisar acuciosamente el producto final obtenido.

No se debe nunca olvidar que IRT no reemplaza el criterio y sentido común en lo que se refiere a diseño de pruebas, sino que tan solo facilita y agiliza el proceso. No debe perderse de vista que tras el sofisticado aparataje matemático y computacional que trae consigo la teoría, aun estamos frente a un proceso de alta subjetividad, donde el criterio de quien lo aplica juega un rol fundamental.

Referencias Bibliográficas

Baker, F. *The Basics of Item Response Theory*. Estados Unidos: ERIC Clearinghouse on Assessment and Evaluation, segunda edición, 2001.

Childs, R. y S. Oppler. "Implications of Test Dimensionality for Unidimensional IRT Scoring: An Investigation of a High Stakes Testing Program". *Educational and Psychological Measurement*, Vol. 60, N° 6 (2000).

Dorans, N. "Item Parameter Invariance: The Cornerstone of Item Response Theory". *Research in Personnel and Human Resources Management*, Vol. 3 (1990).

Fan, X. "Item Response Theory and Classical Test Theory: An Empirical Comparison of Their Item/Person Statistics". *Educational & Psychological Measurement*, Vol. 58, N° 3 (1998).

Green, D.; W. Yen y G. Burket. "Experiences in the Application of Item Response Theory in Test Construction". *Applied Measurement in Education*, Vol. 2, N° 4 (1989).

Hambleton, R.; H. Swaminathan y J. Rogers. *Fundamentals of Item Response Theory*. Sage Publications, 1991.

Hamilton, L. S.; E. M. Nussbaum, H. Kupermintz, J. I. M. Kerkhoven y R. E. Snow, "Enhancing the Validity and Usefulness of Large-scale Educational Assessments: II. NELS: 88 Science Achievement". *American Educational Research Journal*, N° 32 (1995).

[76] Esta función de información objetivo es diferente según el tipo de prueba que se está diseñando y los objetivos que persigue. Por ejemplo, las funciones de información para una prueba de competencias mínimas deberán ser muy diferentes a las de una prueba de admisión a las universidades (véase sección 3.7).

Hamilton, L. S.; E. M. Nussbaum y R. E. Snow. "Interview Procedures for Validating Science Assessments". *Applied Measurement in Education*, N° 10 (1997).

Kingston, N. y N. Dorans. "Item Location Effects and Their Implications for IRT Equating and Adaptive Testing". *Applied Psychological Measurement*, Vol. 8, N° 2 (1984).

Kirisci, L.; T. Hsu y L. Yu. "Robustness of Item Parameter Estimation Programs to Assumptions of Unidimensionality and Normality". *Applied Psychological Measurement*, Vol. 25, N° 2 (2001).

Kolen, M. y R. Brennan. *Test Equating*. Springer-Verlag, 1995.

Rudners, L. *Item Banking*. ERIC/AE Digest, 1998.

Leary, F. y N. Dorans. "Implications for Altering the Context in Which Test Items Appear: A Historical Perspective on an Immediate Concern". *Review of Educational Research*, Vol. 55, N° 3 (1985).

Linn, R. "Has Item Response Theory Increased the Validity of Achievement Test Scores?" *Applied Measurement in Education*, Vol. 3, N° 2 (1990).

Lord, F. *Applications of Item Response Theory to Practical Testing Problems*. LEA Publishers, 1980.

Loyd, B. "Implications of Item Response Theory for the Measurement Practitioner". *Applied Measurement in Education*, Vol. 1, N° 2 (1988).

Nussbaum, E. M.; L. Hamilton y R. E. Snow. "Enhancing the validity and usefulness of large scale educational assessments: IV. NELS: 88 science achievement to 12th grade". *American Educational Research Journal*, N° 34 (1997).

Shavelson, R. y S. Lau. "Multidimensional Validity Revisited: A Multidimensional Approach to Achievement Validation". *CSE Technical Report* 574. National Center for Research on Evaluation. University of California, Los Angeles, 2002.

Stage, C. *A. Comparison Between Item Analysis Based on Item Response Theory and Classical Test Theory. A Study of the SweSAT Subtest ERC*. Umeå University, Department of Educational Measurement, 1998a.

Stage, C. *A. Comparison Between Item Analysis Based on Item Response Theory and Classical Test Theory. A Study of the SweSAT Subtest WORD*. Umeå University, Department of Educational Measurement, 1998b.

Stage, C. *A. Comparison Between Item Analysis Based on Item Response Theory and Classical Test Theory. A Study of the SweSAT Subtest READ*. Umeå University, Department of Educational Measurement, 1999.

Tate, R. "Test Dimensionality". En G. Tindal y T. Haladyna (eds.), *Large-Scale Assessment Programs for All Students*. LEA Publishers, 2002 □.