TEORÍA CLÁSICA DE MEDICIÓN O TEORÍA DE RESPUESTA AL ÍTEM

LA EXPERIENCIA SUECA

Christina Stage

En este trabajo se intenta determinar si la aplicación de la teoría de respuesta al ítem (IRT, por sus siglas en inglés) permitiría mejorar la calidad de la prueba de admisión a las universidades en Suecia (SweSAT). Conforme a la evidencia analizada, la autora concluye que el uso de IRT no proporciona ninguna ventaja por sobre el método clásico en una prueba como la SweSAT. Con todo, señala que la teoría de respuesta al ítem está aún en desarrollo y que, por cierto, hay áreas específicas en que ella pudiera resultar muy promisoria. Por último, la autora destaca el caso de los tests adaptativos computarizados (TAC) o "tests a la medida", donde IRT no sólo proporciona un marco teórico adecuado, sino el único factible en esas circunstancias. Si en el futuro la SweSAT —o alguna versión de la misma— llegara a transformarse en un TAC, entonces forzosamente tendría que usarse IRT. Pero de mantenerse la SweSAT en su esquema actual, no se recomienda el uso de IRT en su confección.

CHRISTINA STAGE. Profesora del Departamento de Medición Educacional, Universidad de Umeå, SE 901 87 Umeå, Suecia. christina.stage@edmeas.umu.se

Este trabajo describe los esfuerzos realizados para dilucidar si la teoría de respuesta al ítem (IRT por sus siglas en inglés) sería aplicable a la Prueba de Admisión a las Universidades en Suecia (SweSAT). El objetivo ha sido determinar si un cambio desde la teoría clásica de la medición (TCM) hacia la IRT en el proceso de creación de ítems, diseño de pruebas, asignación de puntajes o "equating"* permitiría mejorar la calidad de la prueba. El trabajo consta de tres partes. La primera parte, "La aplicabilidad de modelos de IRT a los subtests de SweSAT", es un resumen de cinco informes (Stage, 1996, 1997, 1997b, 1997c y 1997d) que describen las diferentes etapas que se han seguido para investigar si un modelo de IRT podría ser ajustado a los cinco subtests de SweSAT por separado. La segunda parte, "Comparación entre análisis de ítems basándose en IRT y TCM", contiene una síntesis de tres informes anteriores (Stage, 1988a, 1988b y 1999), en los cuales se efectuaron comparaciones entre índices de dificultad y discriminación dentro de la TCM y parámetros de dificultad y discriminación dentro de la IRT en los tres subtests ERC, READ y WORD. En la tercera parte, "Aplicabilidad de la IRT al SweSAT: el test total", se describe un intento por ajustar un modelo de IRT a la SweSAT total. La conclusión fue que como el ajuste del modelo a los datos resultó algo dudoso, especialmente para el test total, no se obtenía ninguna ventaja al cambiar de TCM a IRT.

La Prueba de Admisión a las Universidades en Suecia (SweSAT) es un test referido a normas, que se emplea en el proceso de selección para ingresar a la educación superior en Suecia. Se aplica dos veces al año, una durante el trimestre de primavera y otra durante el trimestre de otoño. Luego de ser rendida, cada prueba en particular pasa a ser del dominio público, de modo que es preciso elaborar una nueva versión de ésta cada vez que se vuelve a administrar. Puesto que los resultados de la prueba tienen una validez de 5 años, es importante que los resultados de diferentes aplicaciones de la prueba sean comparables.

Desde 1996 la prueba está constituida por 122 ítems de selección múltiple, divididos en 5 subtests:

1. DS: Subtest de suficiencia de datos que mide razonamiento matemático y consta de 22 ítems.
2. DTM: Subtest que mide la habilidad para interpretar diagramas, tablas y gráficos, que consta de 20 ítems.

* Proceso por el cual los puntajes de diferentes pruebas se hacen comparables. (N. del E.)

3. ERC: Subtest de comprensión de lectura en inglés que consta de 20 ítems.
4. READ: Subtest de comprensión de lectura en sueco, que consta de 20 ítems.
5. WORD: Subtest de vocabulario, que consta de 40 ítems.

Desde que se aplicó la primera versión de SweSAT en 1977, la elaboración y la compilación del test, lo mismo que el "equating" de las formas de éste entre una administración y la próxima, han estado basados en la teoría clásica de medición (TCM).

En TCM, que comenzó a evolucionar con el test de Binet hace casi cien años, se considera que el puntaje del test está constituido por dos elementos, un "puntaje real" y un error. El puntaje real y el error aparecen como factores completamente independientes. El puntaje real es visto como un componente que no varía entre una forma de test y una forma paralela alternativa, y entre una y otra ocasión. Se considera que el error es una característica única de la medición específica, y es enteramente independiente del error que cabría esperar que surgiera en otra medición del mismo constructo. El puntaje real nunca se puede observar directamente, y sólo puede inferirse a partir de la coherencia en el desempeño teniendo en cuenta el puntaje obtenido de un test a otro.

La TCM ha sido un modelo productivo que condujo a la formulación de una serie de relaciones útiles:

- La relación entre la longitud del test y su precisión (confiabilidad).
- Estimaciones de la precisión de diferencias de puntajes y cambios de puntajes.
- La estimación de las propiedades de compuestos de dos o más medidas.
- La estimación del grado en que los índices de relación entre distintas mediciones son atenuados por el error de medición en cada una de ellas.

Aun cuando el principal centro de interés de la TCM es la información a nivel del test, las estadísticas de los ítems (es decir el nivel de dificultad de los ítems y el nivel de discriminación de los mismos) también son importantes. A nivel de ítems la TCM es relativamente sencilla, ya que no existen modelos teóricos complejos para relacionar la habilidad o el éxito de un examinado en un ítem específico. La proporción de un grupo claramente definido de examinados que responden un ítem correctamente (determinada de manera empírica) —valor *p*— se emplea como índice de

dificultad del ítem (en verdad se trata de un indicador inverso de dificultad, pues los valores más altos señalan que los ítems son más fáciles). La capacidad de un ítem para discriminar entre examinados de alta habilidad y de baja habilidad se expresa estadísticamente como el coeficiente de correlación entre los puntajes logrados en el ítem y los puntajes obtenidos en el test total.

Con frecuencia se alude a los modelos de TCM como modelos "débiles", porque sus supuestos son fácilmente confirmados en los datos del test.

La TCM adolece, sin embargo, de algunas deficiencias. Una de ellas es que los índices de dificultad y de discriminación de los ítems son dependientes del grupo; en efecto, los valores de estos índices dependen de los grupos de examinados en que se han obtenido. Otra limitación es la dependencia entre los puntajes observados y reales de los tests. Los puntajes observados y reales suben y bajan de acuerdo con los cambios en la dificultad de los tests. Otro punto débil tiene que ver con el supuesto de que los errores de medición son los mismos para todos los examinados. Las estimaciones de habilidad son, en realidad, menos precisas para alumnos de baja y alta habilidad que para los alumnos de habilidad promedio.

Durante las últimas décadas se ha desarrollado un nuevo sistema de medición, llamado teoría de respuesta al ítem (IRT por sus siglas en inglés), que ha llegado a transformarse en un importante complemento de la TCM en el diseño, la construcción y la evaluación de pruebas. Dentro del marco de IRT es posible obtener características de los ítems que *no* dependen de grupos; puntajes de habilidad que *no* dependen de los tests; y una medición de la precisión para cada nivel de habilidad.

De acuerdo con Hambleton y otros (1991):

> IRT se funda en dos postulados básicos: a) el desempeño de un examinado en el ítem de un test puede predecirse (o explicarse) mediante una serie de factores llamados rasgos, rasgos latentes o habilidades; y b) la relación entre el desempeño de un examinado en un ítem y la serie de rasgos implícitos en el desempeño en el ítem puede describirse con una función que aumenta monótonamente, llamada función característica del ítem o curva característica del ítem. Esta función especifica que a medida que el nivel del rasgo aumenta, la probabilidad de responder correctamente también aumenta (p. 7).

Existen diversos modelos de IRT, pero todos tienen en común el uso de una función matemática para especificar la relación entre el desempeño observable del examinado en un test y los rasgos o habilidades no observables que se supone están implícitos en el desempeño en el test. En cualquier

aplicación práctica de los modelos de rasgos latentes es preciso especificar la forma matemática de las curvas características del ítem y obtener estimaciones de los parámetros del ítem necesarios para describir las curvas. En el modelo de tres parámetros éstos son:

1. Dificultad del ítem “b”.
2. Discriminación del ítem “a”.
3. Un parámetro de pseudoadivinanza “c”.

En el modelo de dos parámetros no se supone que exista alguna adivinanza, mientras que en el modelo de un parámetro se supone que la discriminación del ítem es la misma para todos los ítems.

Los modelos de IRT se denominan modelos “fuertes”, ya que los supuestos pueden resultar difíciles de confirmar por los datos del test. Un importante supuesto incluido en los modelos más comunes de IRT es el de la unidimensionalidad, que significa que los ítems que constituyen el test sólo miden una habilidad. Lo que se requiere para que el supuesto de unidimensionalidad se confirme adecuadamente es la presencia de un factor predominante que influya en el desempeño en el test. Otro supuesto, relacionado con el anterior, es el de la independencia local, la cual indica que cuando las habilidades que influyen en el desempeño en el test se mantienen constantes, las respuestas del examinado a cualquier par de ítems son estadísticamente independientes.

Una vez que se especifica un modelo de rasgos latentes, la precisión con que éste estima la habilidad del examinado se puede determinar para distintos niveles de habilidad. La información varía con el nivel de habilidad, lo que hace posible determinar el error estándar de estimación para distintos niveles de habilidad. La función de información del ítem entrega información sobre la utilidad del ítem para medir la habilidad en un determinado nivel.

En la actualidad, las entidades que elaboran tests están prestando creciente atención a IRT para el diseño de los tests, la selección de los ítems, el método para afrontar el sesgo en los ítems, y el equating y notificación de los puntajes. El potencial de IRT para resolver este tipo de problemas es considerable. Sin embargo, para que puedan obtenerse los eventuales beneficios de un modelo de IRT resulta esencial que exista un ajuste entre el modelo y los datos del test que sean de interés. Un modelo de IRT que no se ajuste adecuadamente no generará parámetros invariantes.

> En muchas aplicaciones de IRT de las que se informa en la bibliografía, el ajuste entre datos y modelo y las consecuencias de su

> desajuste no han sido investigados adecuadamente. Como resultado de lo anterior se sabe menos de lo que podría suponerse, si se tiene en cuenta la voluminosa bibliografía existente sobre IRT, acerca de la conveniencia de determinados modelos de IRT para varias aplicaciones (Hambleton y otros, 1991, p. 53).

Hambleton y otros (1991) nos alertan aun más del peligro de confiarnos demasiado en los tests estadísticos, ya que éstos adolecen de un grave defecto: su sensibilidad al tamaño de la muestra de examinados. En lugar de ello los autores recomiendan que los juicios sobre el ajuste del modelo a los datos del test se basen en tres tipos de evidencia:

1. Validez de los supuestos del modelo para los datos del test.
2. Grado en que se obtienen las propiedades esperadas del modelo (por ejemplo, invarianza de los parámetros del ítem y de los parámetros de habilidad).
3. Exactitud de las predicciones del modelo empleando datos de tests reales y —si procede— simulados.

En las siguientes secciones de este trabajo se presentan los resultados de diversos tipos de análisis. El objetivo de estas investigaciones ha sido encontrar distintos tipos de evidencia en favor o en contra del ajuste de un modelo de IRT a los datos de la SweSAT.

1. LA APLICABILIDAD DE LOS MODELOS DE IRT A LOS SUBTESTS DE LA SWESAT

> IRT tiene un gran potencial para resolver muchos problemas en la aplicación de tests y en medición. Con todo, el éxito de determinadas aplicaciones de esta teoría no está asegurado por el solo hecho de procesar los datos del test por medio de uno de los programas computacionales (…). Los beneficios de los modelos de respuesta al ítem pueden obtenerse sólo cuando el ajuste entre el modelo y los datos de interés del test es satisfactorio (Hambleton y otros, 1991, p. 53).

Para investigar si un modelo de IRT podría ajustarse satisfactoriamente a cada uno de los 5 subtests DS, DTM, ERC, READ y WORD, se empleó una muestra aleatoria del 3% de los 85.506 examinados que rindieron la SweSAT en el trimestre de primavera de 1996. La muestra estaba compuesta de 2.461 alumnos: 1.349 mujeres y 1.112 hombres. Los resulta-

dos que obtuvieron estos examinados en cada subtest por separado proporcionaron los datos que se analizaron de distintas maneras.

La primera etapa consistió en realizar un análisis estándar de ítem por teoría clásica, cuyo resultado se presenta más abajo.

Análisis de ítem clásico

El análisis de ítem del subtest DS dentro del marco de la TCM arrojó un rango de valores p que iba de 0,40 a 0,81, y un rango de correlaciones biseriales que iba de 0,25 a 0,70. El coeficiente de confiabilidad, alfa, fue r = 0,82.

El análisis de ítem del subtest DTM dentro del marco de la TCM arrojó un rango de valores p que iba de 0,28 a 0,82, y un rango de correlaciones biseriales[1] que iba de 0,19 a 0,56. El coeficiente de confiabilidad, alfa, fue de r = 0,72.

El análisis de ítem del subtest ERC dentro del marco de la TCM arrojó un rango de valores p que iba de 0,28 a 0,82, y un rango de correlaciones biseriales que iba de 0,19 a 0,56. El coeficiente de confiabilidad, alfa, fue de r = 0,72.

El análisis de ítem del subtest READ dentro del marco de la TCM arrojó un rango de valores p que iba de 0,34 a 0,84, y un rango de correlaciones biseriales que iba de 0,21 a 0,45. El coeficiente de confiabilidad, alfa, fue de r = 0,68.

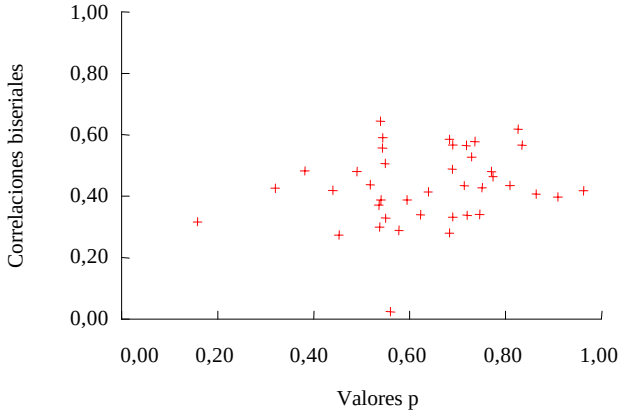
Por último, el análisis de ítem del subtest WORD dentro del marco de la TCM arrojó un rango de valores p que iba de 0,16 a 0,96, y un rango de correlaciones biseriales que iba de 0,02[2] a 0,64. El coeficiente de confiabilidad, alfa, fue de r = 0,85.

Los rangos de correlaciones biseriales indican que existe una considerable variación en el poder de discriminación de los ítems en la totalidad de los 5 subtests. A veces, sin embargo, el rango puede resultar engañoso debido a un par de "valores atípicos". Además, altas correlaciones biseriales se asocian a veces a ítems muy fáciles. Esos índices de discriminación no muestran realmente ítems eficaces, y por tanto los valores p fueron graficados contra las correlaciones biseriales para todos los ítems en cada subtest. En la Figura 1 se muestra como ejemplo el gráfico para el subtest WORD.

[1] Las correlaciones biseriales se calculan como la correlación entre el ítem y el puntaje total sin ese ítem.

[2] Había un ítem que se desviaba, el cual no operó adecuadamente; de ahí que la correlación biserial haya sido tan baja.

FIGURA 1: CORRELACIONES BISERIALES GRAFICADAS CONTRA VALORES P DE LOS 40 ÍTEMS DEL SUBTEST WORD



Los gráficos, que fueron similares para la totalidad de los 5 subtests, respaldaron el supuesto de que en realidad hubo variación en el poder discriminador de los ítems en todos los subtests. No pareció que existiera alguna conexión entre ítems fáciles y correlaciones biseriales altas. La conclusión fue que aparentemente se necesitaba un parámetro de discriminación de los ítems, y que por tanto un modelo IRT de un solo parámetro parecía inadecuado para los resultados de todos los subtests.

Con el fin de determinar si en los tests algunos alumnos habían adivinado las respuestas correctas, se estudió a los examinados que obtuvieron los resultados más bajos, escogiéndose a todos los que quedaron por debajo del primer percentil en cada subtest; los ítems difíciles se definieron como aquellos con valores p inferiores a 0,50.

Se estudiaron los resultados obtenidos por estos examinados de bajo desempeño en los ítems más difíciles de cada subtest, y se determinó que en el subtest DS los valores p para dichos alumnos en los 8 ítems más difíciles fueron:

p = 0,11; 0,30; 0,08; 0,14; 0,20; 0,13; 0,11; y 0,17

en el subtest DTM:
p = 0,26; 0,14; 0,11; 0,06; 0,17; 0,18; 0,11 y 0,20

en el subtest ERC:
p = 0,21; 0,24; 0,12; 0,22; 0,19; 0,15; y 0,35

en el subtest READ:
p = 0,17; 0,16; 0,12; y 0,15

en el subtest WORD:
p = 0,13; 0,01; 0,14; 0,01; 0,11; y 0,22.

Estos resultados indicaron que en cualquiera de los subtests difícilmente se puede excluir la posibilidad de que los alumnos hayan adivinado las respuestas, por lo que tampoco el modelo de dos parámetros parecía adecuado para ajustarse a los datos.

Análisis factorial

Un supuesto común a todos los modelos de IRT es que el grupo de ítems de los tests es unidimensional. Una medición bruta de la unidimensionalidad es el coeficiente de confiabilidad, alfa, ya que éste mide la coherencia interna de los ítems en un test. El coeficiente alfa varió entre 0,68 y 0,85 para los subtests. El coeficiente r = 0,68 indica que el subtest no es muy homogéneo, pero se trata sólo de una medición bruta. Un método más adecuado para evaluar la unidimensionalidad de un test es el análisis factorial (Hambleton y Rovinelli, 1986). Si el análisis factorial revela la existencia de sólo un factor dominante, sirve como respaldo para el argumento de la unidimensionalidad. Los resultados de los análisis factoriales fueron:

Para el subtest DS los análisis arrojaron tres factores con valores propios (*eigenvalues*) de 4,77; 1,21; y 1,09, respectivamente. La varianza explicada por el primer factor fue de 21,7%, y todos los ítems asignaban considerables cargas en el primer factor (entre 0,24 y 0,64).

Para el subtest DTM el resultado fue de 4 factores con valores propios de 3,3; 1,2; 1,1 y 1,0, respectivamente. La varianza explicada por el primer factor fue de 16,4%.

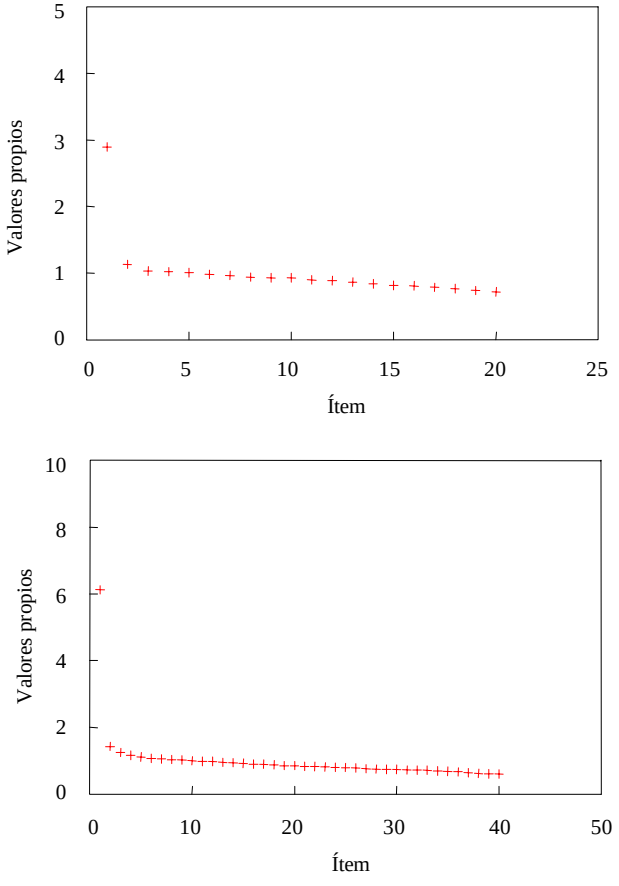
Para el subtest ERC el resultado fue también de 4 factores con valores propios de 3,8; 1,1; 1,0 y 1,0, respectivamente. La varianza explicada por el primer factor fue de 19,4 %.

Para el subtest READ el resultado fue de 5 factores con valores propios de 2,9; 1,1; 1,0; 1,0; y 1,0 respectivamente. La varianza explicada por el primer factor fue de 14,5%.

Para el subtest WORD, por último, el análisis factorial no rotado arrojó 9 factores con valores propios de: 6,1; 1,4; 1,2; 1,2; 1,1; 1,1; 1,0; 1,0 y 1,0 respectivamente. La varianza explicada por el primer factor fue de 15,4%.

Todos los valores propios fueron graficados, y los gráficos para los dos subtests con los primeros valores propios más pequeños, es decir, los subtests READ y WORD, aparecen en la Figura 2.

FIGURA 2: GRÁFICO DE VALORES PROPIOS PARA LOS SUBTESTS READ (ARRIBA) Y WORD (ABAJO)



En la Figura 2 se demuestra que, después de todo, parece haber un primer factor dominante en ambos subtests, ya que de acuerdo con Hambleton y Rovinelli (1986):

> La cantidad de factores “significativos” se determina buscando el “codo” en el gráfico. La cantidad de valores propios a la izquierda del codo es normalmente interpretada como el número de factores significativos implícitos en el desempeño en un test (p. 289).

Aun cuando habría sido preferible que el valor de la varianza explicada por el primer factor fuera generalmente mayor, no resulta improbable o irracional suponer que existe un único factor en los datos de test para cualquiera de los subtests.

El modelo logístico de IRT con 3 parámetros

Se intentó ajustar los resultados de cada subtest al modelo logístico de IRT con 3 parámetros por medio del programa BILOGW (Mislevey y Bock, 1990). Cuando la cantidad de ítems es de 20 ó más, se incluyen como resultado del programa estadísticas aproximadas de chi-cuadrado para la calidad del ajuste de cada ítem. Para estos efectos los casos en la muestra de calibración se clasifican en intervalos sucesivos del continuo latente de acuerdo con las estimaciones de habilidad reescalada a media = 0 y desviación estándar = 1. Lo anterior permite obtener una prueba de ajuste razonable si la cantidad de ítems es lo suficientemente grande como para que una asignación de casos resulte precisa, y si el tamaño de la muestra es lo suficientemente grande como para contener 3 ó más intervalos. En estos estudios el menor número de ítems fue de 20, y la muestra de examinados fue extensa. La cantidad de intervalos utilizados fue de 10 para la mayoría de los ítems. Los siguientes fueron los resultados de los análisis de calidad del ajuste:

En cuanto al subtest DS, el resultado fue que para 11 ítems hubo un desajuste entre los datos y el modelo a nivel de $\alpha = 0{,}05$, y para 4 de esos ítems el desajuste fue considerable a nivel de $\alpha = 0{,}01$. El índice de confiabilidad fue de r = 0,84, que puede compararse con el coeficiente alfa en el análisis clásico, que fue de r = 0,82.

En cuanto al subtest DTM, el resultado fue que para 10 ítems hubo un considerable desajuste entre los datos y el modelo, y para 8 de esos ítems el desajuste fue significativo a nivel de $\alpha = 0{,}01$. El índice de confiabilidad fue de r = 0,74, en comparación con el coeficiente alfa, que fue de r = 0,72.

En cuanto al subtest ERC, el resultado fue de 14 ítems con un considerable desajuste entre los datos y el modelo, y para 7 de esos ítems el desajuste fue significativo a nivel de $\alpha = 0{,}01$. El índice de confiabilidad fue de r = 0,80, mientras que el coeficiente alfa fue de r = 0,76.

En cuanto al subtest READ, el resultado fue de 9 ítems con un considerable desajuste entre los datos y el modelo, y para 7 de esos ítems el desajuste fue significativo a nivel de $\alpha = 0{,}01$. El índice de confiabilidad reportado fue de r = 0,72, y el coeficiente alfa fue de r = 0,68.

Y finalmente el resultado de los análisis de calidad de ajuste correspondiente al subtest WORD fue que para 8 de los ítems hubo un considerable desajuste entre los datos y el modelo a nivel de $\alpha = 0,05$, y para uno de estos 8 ítems el desajuste fue significativo a nivel de $\alpha = 0,01$. El índice de confiabilidad registrado fue de r = 0,87, mientras que el coeficiente de confiabilidad, alfa, fue de r = 0,85.

Análisis residuales

Se realizaron otros tipos de análisis de calidad del ajuste mediante el programa RESID (Rogers, 1994). Al efectuarse estos análisis los examinados se clasifican primero en categorías de habilidad. La cantidad de niveles de habilidad fue especificada en 8, y se calcularon las proporciones observadas de examinados en cada categoría de habilidad que respondieron el ítem correctamente. Las proporciones esperadas de respuestas correctas para cada intervalo de habilidad se obtuvieron calculando la probabilidad de éxito en el ítem en cada nivel de habilidad. Luego se determinaron los valores residuos (observados menos esperados) y los residuos estandarizados. El programa contiene asimismo estadísticas de ajuste chi-cuadrado para cada ítem como resultado.

Los siguientes fueron los resultados de los análisis RESID:

En cuanto al subtest DS, las diferencias entre valores observados y esperados fueron estadísticamente significativas para 2 ítems, ambos en el nivel 0,05.

En cuanto al subtest DTM, las diferencias fueron estadísticamente significativas para 6 ítems, 5 de ellos en el nivel 0,01.

En cuanto al subtest ERC, las diferencias fueron estadísticamente significativas para 7 ítems, 2 de ellos en el nivel 0,01.

En cuanto al subtest READ, las diferencias fueron estadísticamente significativas para 5 ítems, 2 de ellos en el nivel 0,01.

Por último, en cuanto al subtest WORD, las diferencias fueron estadísticamente significativas para 6 ítems, uno de ellos en el nivel 0,01.

Los residuos permiten comparar el desempeño pronosticado y el real. Los residuos brutos corresponden a las diferencias entre desempeño esperado y observado en un ítem en un nivel de desempeño especificado. Los residuos estandarizados (RE) tienen en cuenta el error muestral asociado a cada nivel de desempeño, al igual que la cantidad de examinados en ese nivel de desempeño en particular. Cuando el modelo se ajusta a los datos podría esperarse que los RE fueran pequeños y se distribuyeran alea-

toriamente alrededor de 0. Dentro del marco de la teoría de la regresión resulta común suponer que la distribución de los RE es aproximadamente normal. En la Tabla 1 se entrega un resumen de los RE de los análisis de calidad de los ajustes.

TABLA 1: VALORES ABSOLUTOS DE LOS RESIDUOS ESTANDARIZADOS (%) PARA LOS 5 SUBTESTS

RE	DS	DTM	ERC	READ	WORD
| 0-1 |	72,16	65,63	62,50	68,75	70,31
| 1-2 |	22,16	25,00	31,25	24,38	26,25
| 2-3 |	5,11	7,50	5,00	6,88	3,13
| >3 |	0,57	1,88	1,25	0,00	0,31

Los resultados que aparecen en la Tabla 1 demuestran que aun cuando las distribuciones para los subtests DTM y ERC son hasta cierto punto demasiado uniformes, todas las distribuciones de RE se aproximan bastante a la distribución normal que se supone proporciona un sólido respaldo al ajuste del modelo a los datos.

Hambleton y otros (1991) han formulado las siguientes recomendaciones en cuanto a la evaluación de ajuste del modelo a los datos:

> Al evaluar el ajuste del modelo a los datos, el mejor enfoque implica: a) diseñar y realizar una diversidad de análisis destinados a detectar los tipos de desajuste esperados; b) considerar cuidadosamente la serie completa de resultados; y c) emitir un juicio acerca de la conveniencia del modelo para la aplicación proyectada. Los análisis deberían incluir investigaciones de supuestos del modelo, del grado en que se obtienen las características deseadas del modelo, y de las diferencias entre predicciones del modelo y datos reales. Se pueden realizar tests estadísticos, pero es preciso tener cuidado al interpretar la información estadística. El número de investigaciones que pueden llevarse a cabo es casi ilimitado (p. 74).

Para el subtest DS se efectuó una comparación entre los parámetros de los ítems, estimados en dos muestras distintas de alumnos que rindieron el examen. La correlación entre valores *b* fue de r = 0,95, y la correlación entre valores *a* fue de r =0,72.

Para el subtest DTM se realizó una comparación entre parámetros estimados dentro de la IRT e índices calculados al interior de la TCM. La

correlación entre los valores *b* estimados y los valores *p* calculados fue de r = -0,94, y la correlación entre valores *a* estimados y correlaciones biseriales calculadas (r_{bis}) fue de r = 0,82. La correlación entre puntajes del test y parámetros estimados de habilidad fue de r = 0,95. Se efectuó otra comparación entre parámetros estimados en alumnos y alumnas que rindieron el test, y el resultado fue que la correlación entre valores *b* estimados en hombres y mujeres examinados fue de r = 0,89, y entre valores *a* de r = 0,91.

Se efectuaron comparaciones análogas para el subtest ERC. La correlación entre valores *b* estimados y valores p calculados en este subtest fue de r = -0,88, y entre valores *a* estimados y r_{bis} calculada fue de r = 0,73. La correlación entre valores *b* estimados en hombres examinados y valores *b* estimados en mujeres examinadas fue de r = 0,93, y la correlación entre valores *a* fue de r = 0,83.

Las mismas comparaciones se efectuaron para los subtests READ y WORD. Para el subtest READ la correlación entre valores p y *b* fue de r = -0,98, y para el subtest WORD dicha correlación fue de r = -0,74. Para el subtest READ la correlación entre r_{bis} y valores *a* fue de r = 0,84, mientras que para el subtest WORD fue de r = 0,82. La correlación entre valores *b* estimados en hombres y valores *b* calculados en mujeres en el subtest READ fue de r =0,92, y en el subtest WORD esta correlación fue de r = 0,85. La correlación entre valores *a* estimados en hombres examinados y los valores *a* calculados en mujeres examinadas fue de r = 0,76 en el subtest READ y de r = 0,77 en el subtest WORD. Los resultados de estas comparaciones entre grupos se presentan en la Tabla 2.

TABLA 2: RELACIONES ENTRE PARÁMETROS DE LOS ÍTEMS ESTIMADOS DENTRO DE LA IRT E ÍNDICES DE ÍTEMS CALCULADOS DENTRO DE LA TCM, Y ENTRE PARÁMETROS DE LOS ÍTEMS DE IRT ESTIMADOS EN HOMBRES Y MUJERES EXAMINADOS

Correlación	DTM	ERC	READ	WORD
valores p y *b*	-0,94	-0,88	-0,98	-0,74
r_{bis} y valores *a*	0,82	0,73	0,84	0,82
M y H est. *b*	0,89	0,93	0,92	0,85
M y H est. *a*	0,91	0,83	0,76	0,77

M: mujeres; H: hombres.

1.1. Análisis

Los resultados de estos intentos por ajustar separadamente un modelo logístico de IRT de 3 parámetros a cada uno de los 5 subtests de la SweSAT son algo variados. Los resultados de los análisis iniciales con la TCM respaldaron la necesidad de aplicar un modelo de 3 parámetros en la totalidad de los 5 subtests. Los análisis factoriales respaldaron la tesis de la unidimensionalidad en el subtest DS, para el cual el primer factor podía explicar el 21,7% de la varianza del test. En cuanto a los otros subtests el respaldo de los análisis factoriales para la unidimensionalidad fue más débil: 19,4% de la varianza explicada por el primer factor para el subtest ERC; 16,4% para el subtest DTM; 15,4% para el subtest WORD; y 14,5% para el subtest READ. Cuando la varianza explicada es inferior al 20% no se tiene la certeza para poder afirmar que el test es unidimensional.

El supuesto de independencia local también resulta problemático al menos para 3 de los subtests del SweSAT. El subtest READ consta de 4 textos con 5 preguntas en relación con cada texto. Pese a que estos 5 ítems son independientes entre sí, pertenecen al mismo texto. El subtest DTM está compuesto por 10 figuras, tablas y mapas con dos preguntas para cada representación gráfica. El subtest ERC está formado por una cantidad variable de textos con 2 a 5 preguntas para cada uno. Existen dudas en cuanto a si los ítems son en verdad localmente independientes en los tests de este formato.

Otro problema se observa en los distintos resultados obtenidos en las 2 pruebas estadísticas. Se descubrió que por lo general la cantidad de ítems que presentaban un desajuste estadísticamente significativo entre los datos y el modelo era mayor aplicando el programa BILOG que aplicando el programa RESID, y ello podría explicarse, al menos en parte, por el hecho de que el BILOG divide en más grupos de habilidad que el RESID. Sin embargo, se descubrió que mientras algunos de los ítems exhibían un considerable desajuste entre los datos y el modelo empleando el programa RESID, ello no ocurría al aplicar el BILOG. Este fenómeno se verificó en 2 ítems del subtest DS, 2 del subtest DTM, uno del subtest ERC, 2 del subtest READ, y 3 del subtest WORD.

Hasta ahora los análisis efectuados no han confirmado totalmente ni rechazado de plano el ajuste entre el modelo logístico de IRT de 3 parámetros y los datos del SweSAT.

2. COMPARACIÓN ENTRE LOS ANÁLISIS DE ÍTEMS BASADOS EN LA TEORÍA DE RESPUESTA AL ÍTEM Y EN LA TEORÍA CLÁSICA

Tal como ocurre en todas las evaluaciones altamente decisivas, la serie de pruebas piloto o pretests de los ítems del SweSAT constituyen una etapa crucial del proceso de elaboración de este test. El pretest tiene diversos objetivos (Henrysson, 1971), de los cuales los más importantes para el SweSAT son:

- Determinar la dificultad de cada ítem de modo que se pueda hacer una selección, la cual indicará el nivel de dificultad del subtest, que es similar a anteriores versiones de éste.
- Identificar ítems débiles o defectuosos con distractores que no funcionan.
- Determinar el poder de cada ítem para discriminar entre examinados de buen y mal desempeño en la variable de logro medida.
- Identificar ítems sesgados (por género).

Desde que la SweSAT se aplicó por primera vez en 1977, la elaboración y la compilación del test, lo mismo que el equating de las diferentes formas de éste entre una administración y la próxima se ha basado en la TCM. Sobre la base de los datos obtenidos en los pretests, los ítems son rechazados o seleccionados para el test definitivo, y las estadísticas que se usan en el análisis de ítems son:

Valores p de los ítems.
Valores p de los distractores.
Correlaciones biseriales (r_{bis}).
Valores p de hombres y mujeres.
La regresión del test por ítem.

La principal limitación de la TCM a este respecto es que la estadística de la persona (es decir el puntaje obtenido en el test) depende de la muestra de ítems (es decir, del test), y las estadísticas de los ítems dependen de la muestra de examinados. El argumento esencial para preferir el uso de modelos de IRT al empleo de procedimientos de TCM es que la IRT debería generar como resultado medidas independientes de las muestras. Con IRT una persona debería teóricamente recibir la misma estimación de habilidad, independientemente del test rendido, y las estadísticas de los ítems deberían mantenerse estables en distintos grupos de individuos. De allí que la gran

ventaja de IRT sea la invarianza de los parámetros de los ítems. Una desventaja de la IRT es que para estimar los parámetros los tamaños de las muestras tienen que ser grandes.

IRT ha sido objeto de intensas investigaciones por parte de los psicometristas, y además se han publicado numerosos libros y artículos (Fan, 1988). No obstante, los estudios empíricos disponibles se han centrado principalmente en diversas aplicaciones de IRT, y muy pocos han comparado efectivamente la TCM e IRT en cuanto al análisis de ítems y el diseño de tests. Continúa Fan (1998):

> Resulta hasta cierto punto sorprendente que los estudios empíricos en que se examinan y/o comparan las características de invarianza de estadísticas de ítems dentro de ambos marcos de medición sean tan escasos. Tal parece que en la comunidad de medición se ha dado por sentada la superioridad de IRT sobre TCM a este respecto, y no se ha estimado necesaria ninguna indagación empírica. El silencio empírico en torno a este tema parece ser anómalo (p. 361).

Desde el segundo trimestre de 1996 el pretest de los ítems para la SweSAT se ha realizado en conexión con la administración regular de la prueba, lo cual significa que la muestra de examinados a la cual se aplica el pretest ha sido extraída del verdadero universo de examinados y está compuesta por 1.000 individuos como mínimo. Este nuevo procedimiento para realizar el pretest permitiría emplear la IRT para analizar ítems y compilar nuevas versiones de la prueba.

2.1. Objetivo

La finalidad del presente estudio fue comparar las estadísticas de los ítems dentro del marco de la TCM con los parámetros de los ítems dentro del marco de IRT, y analizar la estabilidad de las dos series de características de los ítems desde la etapa de pretest hasta la del test regular. Específicamente se formularon las siguientes preguntas:

1. ¿De qué manera se comparan los índices de dificultad de los ítems determinados dentro del marco de la TCM con los parámetros de dificultad de los ítems calculados por IRT?:

a) ¿para los datos del pretest?,

b) ¿para los datos del test regular?

2. ¿De qué manera se comparan los índices de discriminación de los ítems determinados dentro del marco de la TCM con los parámetros de discriminación de los ítems estimados por IRT?:

a) ¿para los datos de pretests?,

b) ¿para los datos del test regular?

3. ¿Cuál es el grado de estabilidad de los índices de ítems dentro de la TCM desde los datos de pretest hasta los datos del test regular?

4. ¿Cuál es el grado de estabilidad de los parámetros de ítems dentro de IRT desde los datos del pretest hasta los datos del test regular?

2.2. Método

Teoría clásica de la medición

En el test regular aplicado durante el segundo trimestre de 1997 se incluían 20 ítems de WORD, 16 de READ y 14 de ERC, los que ya habían sido pretesteados durante el segundo trimestre de 1996. Para estos ítems se calcularon los valores p y las correlaciones biseriales. Se estimaron los mismos índices para los correspondientes ítems en los datos del pretest, y se compararon los valores.

Teoría de respuesta al ítem

Las 5 combinaciones de pretests para el WORD, en las cuales habían sido diseminados los ya mencionados 20 ítems de WORD, se ejecutaron con el programa BILOGW junto con el test regular de WORD desde el segundo trimestre de 1996, y se estimaron los parámetros *a*, *b* y *c*. A partir del segundo trimestre de 1997 el subtest WORD también se ejecutó en BILOGW y se calcularon los parámetros de los ítems. Por último, se compararon los parámetros estimados para los 20 ítems comunes.

Las 8 versiones de pretests para el READ se ejecutaron en BILOGW a contar del segundo trimestre de 1996 junto con el subtest regular de READ, y se estimaron los parámetros *a*, *b* y *c*. El subtest de READ se ejecutó en BILOGW desde el segundo trimestre de 1997, y se estimaron los parámetros de los ítems. Se registraron y compararon las estimaciones de los parámetros para los 16 ítems comunes.

Las 4 versiones de pretests de ERC fueron ejecutadas en BILOGW a partir del segundo trimestre de 1996 junto con el subtest regular de ERC, y se calcularon los parámetros *a*, *b* y *c*. A partir del segundo trimestre de 1997 el subtest de ERC se ejecutó en BILOGW y se estimaron los parámetros de los ítems. Se registraron y compararon las estimaciones de los parámetros de los ítems para los 14 ítems comunes.

2.3. Resultados

Un problema que se presenta al analizar la estabilidad de los parámetros de los ítems es que el pretest tiene dos objetivos. Uno es el de obtener información sobre el nivel de dificultad y el poder de discriminación de los ítems para así estar en condiciones de compilar tests de dificultad análoga. El otro objetivo es asegurarse de que todos los ítems funcionen de manera satisfactoria. Si un ítem no está funcionando con la suficiente eficacia, será modificado o excluido. Si se introducen cambios importantes, el ítem será pretesteado una vez más antes de ser puesto en práctica, pero si los cambios son menores será aplicado en el test regular. Tales cambios, sin embargo, dan a entender que los ítems no son exactamente los mismos en la versión del pretest que en la del test regular. Otro problema es que los ítems pueden ser dispuestos en distinto orden en el cuadernillo del pretest y en el del test regular, y que los ítems tienden a tornarse más difíciles al final del cuadernillo. En las Tablas 3 a 5 se entrega la ubicación del ítem en el cuadernillo del pretest al igual que en el del test regular, y los ítems en los cuales se introdujeron cambios menores entre ambas etapas están marcados con un *.

El subtest WORD

En la Tabla 3 se muestran los índices de dificultad y discriminación calculados dentro del marco de la TCM para 20 ítems de WORD en el pretest, al igual que en test regular. En la misma tabla se presentan los parámetros de dificultad y discriminación estimados dentro del marco de IRT.

La correlación entre valores p calculada en datos de pretests y los valores p calculados en datos del test regular fue de r = 0,93.

La correlación entre valores b estimados en datos de pretests y valores b estimados en datos del test regular fue de r = 0,92.

La correlación entre valores p y *b* fue de r = -0,93, tanto para los datos de pretests como para los del test regular.

La correlación entre r_{bis} calculadas en datos de pretests y r_{bis} calculadas en datos del test regular fue de r = 0,81.

La correlación entre valores *a* estimados en datos de pretests, y de valores *a* estimados en datos del test regular fue de r = 0,74.

La correlación entre la r_{bis} de discriminación de los ítems y el parámetro de discriminación de los ítems *a* fue de r = 0,65 para los datos de pretests y de r = 0,64 para los datos del test regular.

TABLA 3: CARACTERÍSTICAS DE LOS ÍTEMS PARA 20 ÍTEMS DE WORD CALCULADAS DENTRO DEL MARCO DE LA TCM Y ESTIMADAS DENTRO DEL MARCO DE IRT

Ítem N°		Dificultad TCM		Dificultad IRT		Discrimin. TCM		Discrimin. IRT	
pretest	test reg.	pretest	test reg.	pretest	test reg.	pretest	test reg.	pretest	test reg.
8	1*	0,73	0,71	-0,43	-0,43	0,60	0,58	1,29	1,14
20	4*	0,79	0,72	-0,96	-0,64	0,46	0,41	0,73	0,62
39	5*	0,78	0,74	-1,94	-1,19	0,25	0,33	0,32	0,43
18	9	0,68	0,71	-0,25	-0,59	0,44	0,43	0,71	0,63
36	10*	0,75	0,72	-0,92	-0,81	0,50	0,53	0,72	0,78
27	11*	0,80	0,82	-1,49	-1,79	0,35	0,35	0,48	0,46
36	15*	0,71	0,58	-0,55	0,11	0,40	0,37	0,59	0,55
14	16*	0,65	0,70	-0,02	-0,65	0,44	0,48	0,81	0,72
5	19	0,46	0,42	0,46	0,71	0,42	0,37	0,55	0,50
16	23	0,65	0,62	0,11	0,08	0,35	0,40	0,58	0,72
38	24	0,58	0,56	0,08	0,16	0,47	0,30	0,75	0,40
12	25	0,51	0,59	0,17	-0,02	0,58	0,58	0,97	1,13
24	27	0,69	0,66	0,37	0,33	0,36	0,35	1,07	0,83
4	28*	0,53	0,44	0,35	0,51	0,56	0,52	1,55	1,04
4	29	0,42	0,42	1,16	1,08	0,33	0,26	0,71	0,61
5	35*	0,31	0,38	1,59	0,89	0,32	0,46	0,45	0,95
37	36*	0,71	0,62	-0,37	-0,07	0,43	0,44	0,70	0,70
6	38*	0,41	0,46	1,25	0,49	0,31	0,37	0,70	0,48
28	39*	0,27	0,40	1,61	1,22	0,28	0,32	1,13	1,18
39	40*	0,31	0,31	2,27	2,45	0,23	0,21	0,42	0,45

El subtest READ

La correlación entre valores p de los ítems en las versiones de pretests, y valores p de los ítems correspondientes en el test regular fue de r = 0,78.

La correlación entre valores *b* calculada en datos de pretests y valores *b* estimados en datos del test regular fue de r = 0,55.

La correlación entre valores p calculados en datos del pretest y valores *b* estimados en datos de pretests fue de r = -0,90.

La correlación entre TCM e IRT en cuanto a los índices de dificultad en datos del test regular fue de r = -0,92.

La correlación entre la r_{bis} de los ítems en las versiones de pretests y la r_{bis} de los ítems correspondientes en el test regular fue de r = 0,66.

La correlación entre valores *a* estimados en datos de pretests y valores *a* estimados en datos del test regular fue de r = 0,54.

La correlación entre los índices de discriminación de TCM e IRT en los datos de pretests fue de r = 0,35. La correlación entre índices de discriminación de CTT e IRT en datos del test regular fue de r = 0,78.

TABLA 4: CARACTERÍSTICAS DE 16 ÍTEMS DE READ CALCULADAS DENTRO DEL MARCO DE LA TCM Y ESTIMADAS DENTRO DEL MARCO DE IRT

Ítem N°		Dificultad TCM		Dificultad IRT		Discrimin. TCM		Discrimin. IRT	
pretest	test reg.	pretest	test reg.	pretest	test reg.	pretest	test reg.	pretest	test reg.
5	5	0,74	0,74	-0,61	-0,69	0,30	0,32	0,50	0,59
7	6*	0,30	0,68	2,16	-0,52	0,23	0,36	0,46	0,64
6	7	0,78	0,71	-1,16	-0,45	0,37	0,38	0,54	0,73
8	8	0,80	0,81	-1,31	-1,09	0,40	0,35	0,59	0,63
17	9	0,64	0,81	0,37	-0,88	0,37	0,43	0,85	0,90
20	10	0,52	0,69	1,32	0,27	0,25	0,22	0,64	0,42
17	11	0,61	0,75	1,42	-1,20	0,18	0,11	0,52	0,37
20	12	0,45	0,52	1,92	0,71	0,17	0,33	0,72	0,86
14	13	0,59	0,66	0,48	-0,14	0,35	0,36	0,83	0,69
15	14*	0,36	0,50	1,85	0,74	0,24	0,30	0,53	0,72
14	15*	0,24	0,30	1,67	1,08	0,34	0,43	0,87	1,23
16	16*	0,28	0,57	2,54	0,33	0,17	0,28	0,59	0,54
17	17*	0,35	0,53	1,33	0,51	0,35	0,36	0,72	0,69
19	18	0,63	0,76	0,82	-0,77	0,29	0,32	0,76	0,60
19	19	0,56	0,65	0,84	0,04	0,29	0,31	0,69	0,70
20	20	0,57	0,60	0,38	0,22	0,34	0,34	0,57	0,91

El subtest ERC

TABLA 5: CARACTERÍSTICAS DE 14 ÍTEMS DE ERC CALCULADAS DENTRO DEL MARCO DE LA TCM Y ESTIMADAS DENTRO DEL MARCO DE IRT

Ítem N°		Dificultad TCM		Dificultad IRT		Discrimin. TCM		Discrimin. IRT	
pretest	test reg.	pretest	test reg.	pretest	test reg.	pretest	test reg.	pretest	test reg.
1	1	0,33	0,38	1,65	1,31	0,30	0,33	0,59	0,79
2	2	0,72	0,66	0,20	0,34	0,34	0,33	0,83	0,76
3	3	0,35	0,29	1,70	1,74	0,28	0,29	-0,72	0,71
4	4	0,41	0,47	0,83	0,54	0,45	0,47	0,76	0,83
5	5*	0,62	0,73	0,78	0,35	0,16	0,54	0,27	1,05
1	6	0,77	0,78	-0,90	-0,62	0,62	0,54	1,00	1,10
2	7	0,54	0,50	0,09	0,49	0,48	0,46	0,67	0,81
3	8	0,53	0,53	0,77	0,63	0,37	0,42	0,90	0,99
11	9*	0,41	0,60	0,81	-0,07	0,41	0,38	0,56	0,55
5	10	0,67	0,58	-0,52	0,19	0,57	0,51	0,84	1,02
14	12	0,60	0,51	-0,13	-0,10	0,51	0,46	0,71	0,73
13	13	0,68	0,62	-0,27	-0,03	0,56	0,56	0,96	1,10
14	14	0,65	0,65	-0,40	-0,19	0,57	0,52	0,85	0,91
10	15	0,77	0,74	-0,85	-0,52	0,52	0,52	0,80	0,95

El promedio de los valores p fue de 0,58 para los datos de pretests, al igual que para los del test regular. La correlación entre valores p de los ítems en las versiones de pretests y de valores p de los correspondientes ítems en el test regular fue de r = 0,86.

El promedio de los valores *b* fue de 0,27 para los ítems del pretest y de 0,29 para los ítems del test regular. La correlación entre valores *b* estimados en datos del pretest y valores *b* estimados en datos del test regular fue de r = 0,88.

La correlación entre la r_{bis} de los ítems en las versiones del pretest y la r_{bis} de los ítems correspondientes en el test regular fue de r = 0,57.

La correlación entre valores *a* estimados en datos del pretest y valores *a* estimados en datos del test regular fue de r = 0,34.

Para los 12 ítems, que no habían sido modificados entre los pretests y el test regular, la correlación fue de r = 0,95 para los valores p, y r = 0,96 y ρ = 0,94 para r_{bis}.

Para los 12 ítems invariables la correlación entre valores *b* del pretest y del test regular fue de r = 0,96, y entre valores *a*, de r = 0,82.

La correlación entre dificultades de TCM y dificultadores de IRT fue de r = -0,90 para los datos del pretest, al igual que para los datos del test regular. La correlación entre discriminaciones calculada dentro del marco de la TCM y estimada dentro del marco de la IRT fue de r = 0,74 para los datos del pretest y de r = 0,76 para los datos del test regular.

Los resultados para los 3 subtests se resumen en la Tabla 6.

TABLA 6: ESTABILIDAD DE ÍNDICES DE ÍTEMS DE TCM Y DE PARÁMETROS DE ÍTEMS DE IRT DESDE VERSIONES DEL PRETEST HASTA LA VERSIÓN DEL TEST REGULAR. RELACIONES ENTRE ÍNDICES DE ÍTEMS DE TCM Y PARÁMETROS DE ÍNDICES DE IRT EN EL PRETEST Y EN EL TEST REGULAR

subtest	pretests y test reg.		pretests y test reg.		valores p y *b*		r_{bis} y valores *a*	
	p	*b*	r_{bis}	*a*	pre	reg.	pre	reg.
ERC	0,86 (0,95)	0,88 (0,92)	0,57	0,74	-0,90	-0,90	0,74	0,76
READ	0,78	0,55	0,66	0,54	-0,90	-0,92	0,35	0,68
WORD	0,93	0,92	0,81	0,74	-0,93	-0,93	0,64	0,65

2.4. Análisis

Para los subtests WORD y READ (el modelo logístico de 3 parámetros), en ninguno de los ítems de pretests se detectó un desajuste considerable con respecto al modelo logístico de 3 parámetros. Sin embargo, para 3 ítems del pretest de ERC se observó un desajuste, el cual era significativo a nivel de $\alpha = 0{,}01$. En los subtests regulares hubo un ítem en el subtest WORD, uno en el subtest READ, y 2 en el subtest ERC, en los cuales se observó un desajuste entre modelo y los datos, que era considerable a nivel de $\alpha = 0{,}01$. Estos ítems eran el Nº 10 en el subtest regular de WORD, el Nº 11 en el subtest regular de READ, y los Nºs 9 y 14 en el subtest ERC.

La conclusión general de los estudios es que la concordancia entre los resultados de los análisis de ítems dentro de los dos marcos distintos, TCM e IRT, fue razonablemente aceptable. La correlación entre dificultades de los ítems para las versiones del test regular fue de r = -0,93 para el subtest WORD, r = -0,92 para el subtest READ, y r = -0,90 para el subtest ERC.

En cuanto a la estabilidad de los datos desde la etapa de pretests hasta la del test regular no hubo grandes diferencias entre ambas teorías. Para el subtest WORD la concordancia entre dificultades en los pretests y en el test regular fue de r = 0,93 dentro de la TCM, y de r = 0,92 dentro de IRT. Para el subtest READ la correlación entre dificultades de TCM fue de r = 0,78, y entre dificultades de IRT la correlación fue de r = 0,55. Para el subtest ERC la correlación al interior de TCM fue de r = 0,87 y al interior de IRT de r = 0,88. En términos generales, las correlaciones entre dificultades de ítems en los datos de pretests y del test regular fueron en realidad más altas para los índices de TCM que para los parámetros de IRT.

> Puesto que en teoría la IRT difiere considerablemente de la TCM, y posee algunas ventajas competitivas cruciales con respecto a esta última, parece razonable esperar que existan apreciables diferencias entre las estadísticas sobre personas e ítems basadas en IRT y en la TCM. En teoría esas relaciones no son del todo claras, excepto que los dos tipos de estadísticas deberían estar monótonamente relacionadas bajo ciertas condiciones (Crocker y Algina, 1986; Lord, 1980), pero dichas relaciones rara vez han sido objeto de investigaciones empíricas, y por ende en gran parte son desconocidas (Fan, 1998, p. 360).

La conclusión general derivada de estas comparaciones es que la predicción desde los datos de pretests hasta los datos del test regular es aceptable, pero que ello se puede aplicar tanto a la TCM como a IRT.

A decir verdad, las predicciones formuladas dentro del marco de la TCM fueron por lo general más correctas que las planteadas dentro del marco de IRT. Los parámetros de ítems de IRT no fueron completamente invariantes. Puesto que los grupos en los cuales se aplicaron los pretests constituían muestras amplias y representativas del universo real de examinados, era dable esperar este resultado. Con todo, el dilema es que para lograr estimar los parámetros de ítems dentro del marco de IRT resulta imperioso contar con muestras amplias, y cuando éstas son lo suficientemente extensas, los índices de ítems dentro del marco de la TCM son asimismo muy estables.

Lo que suele considerarse la principal desventaja de la TCM es que las estadísticas de ítems, tales como la dificultad y la discriminación, dependen de la muestra particular de examinados de la cual se obtienen. Con frecuencia se estima que la principal ventaja teórica de IRT radica en la invarianza de las correspondientes estadísticas de los ítems. La invarianza de los parámetros de ítems en todos los grupos es una de las características más importantes de la teoría de respuesta al ítem (Lord, 1980, p. 35). En los estudios citados en este trabajo los parámetros de los ítems estimados dentro del marco de la IRT no fueron superiores, en cuanto a la invarianza entre grupos, a las estadísticas derivadas dentro del marco de la TCM. El problema puede ser que para lograr esta invarianza de parámetros de IRT, el ajuste entre el modelo y los datos debe ser perfecto. Desgraciadamente no existen criterios objetivos sobre el ajuste del modelo a los datos, pero de acuerdo con Hambleton y otros (1991) "…invarianza y ajuste del modelo a los datos son conceptos equivalentes" (p. 24).

3. APLICABILIDAD DE LA IRT A LA SWESAT: EL TEST TOTAL

La SweSAT es puntuada de acuerdo con la teoría clásica de la medición (TCM); el puntaje bruto para cada examinado corresponde al número de ítems respondidos correctamente. Todos los ítems tienen el formato de selección múltiple y son puntuados de manera dicotómica, es decir "1" para una respuesta correcta y "0" para una incorrecta. Se ha descubierto que en este tipo de ítems los examinados difieren en cuanto a su tendencia a adivinar respuestas, o a omitir ítems cuya respuesta correcta ignoran, lo cual puede generar una varianza irrelevante. Si bien se han inventado varios métodos para rectificar el efecto de adivinar las respuestas, los estudios empíricos no han respaldado el empleo de métodos de corrección (Crocker y Algina, 1986, p. 403). A quienes rinden la SweSAT se los insta a marcar una respuesta en todos los ítems. El puntaje total del test se ecualiza y se

transforma en un puntaje normado, el cual se utiliza en el proceso de selección para ingreso a la educación superior.

La SweSAT consta de 5 subtests que miden el conocimiento de vocabulario (WORD), el pensamiento lógico (DS), la comprensión de lectura en sueco (READ), la comprensión de lectura en inglés (ERC), y la habilidad para interpretar diagramas, tablas y mapas (DTM). En estudios anteriores se investigó la aplicabilidad de un modelo logístico de IRT de 3 parámetros a cada subtest (Stage, 1996, 1997a, 1997b, 1997c, 1997d). El resultado de esos estudios no fue una confirmación ni una refutación del ajuste entre el modelo logístico de IRT de 3 parámetros y los datos de la SweSAT.

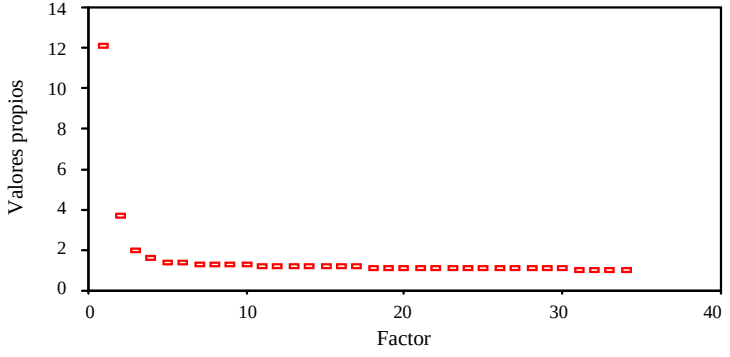
Puesto que es el puntaje total del test lo que indica el resultado obtenido por quienes rindieron la SweSAT, la aplicabilidad de un modelo de IRT adecuado al test total resulta importante. En el presente estudio el objetivo era investigar si un modelo logístico de IRT de 3 parámetros podía aplicarse exitosamente al test total.

Se escogió el modelo de 3 parámetros porque había resultado ser el más adecuado para cada uno de los subtests. No había motivos para creer que el test total debiera diferenciarse de sus partes en cuanto a la necesidad de un parámetro de discriminación y de un parámetro de pseudoadivinanza.

Unidimensionalidad

Un análisis factorial no rotado de los puntajes de los tests totales arrojó un primer factor de 12,1, un segundo factor de 3,7, un tercer factor de 2,0, y 34 factores por sobre 1,0. El primer factor explicaba el 9,9% de la varianza, el segundo factor el 3,0%, y los restantes factores entre el 1,7% y el 0,8% cada uno. En la figura 3 se presenta un gráfico de los valores propios.

FIGURA 3: GRÁFICO DE VALORES PROPIOS PARA LA TOTALIDAD DE LA SWESAT



En la Figura 3 puede apreciarse que aun cuando existe una especie de "codo" en el gráfico, y el primer factor es dominante, el supuesto de unidimensionalidad resulta incierto, ya que el segundo factor es hasta cierto punto demasiado fuerte.

También se efectuó un análisis factorial a nivel de subtests, el cual arrojó un primer factor con valor propio de 3,05; un segundo factor con valor propio de 0,83; un tercero con 0,41; un cuarto con 0,38, y un quinto con 0,35. El primer factor explicaba el 61% de la varianza. Este análisis proporcionó mayor respaldo al supuesto de la unidimensionalidad.

El modelo logístico de 3 parámetros

El programa BILOGW se empleó para ajustar los resultados de la SweSAT al modelo logístico de IRT de 3 parámetros. El resultado del análisis de calidad del ajuste de los ítems fue que para 67 de ellos había un desajuste entre datos y modelo que era significativo a nivel de $\alpha = 0,05$, y para 44 de ellos el desajuste era considerable a nivel de $\alpha = 0,01$. El programa se ejecutó en 28.505 resultados del test, y puesto que las pruebas estadísticas de ajuste del modelo son muy sensibles al tamaño muestral, tales resultados eran esperables. Esto es lo que Hays (1969) llama la falacia de evaluar un resultado teniendo en cuenta sólo la significación estadística:

> Es posible lograr que prácticamente cualquier estudio arroje resultados significativos si empleamos suficientes sujetos, sin importar cuán absurdo pueda ser el contenido (p. 326).

El programa también fue ejecutado en dos muestras aleatorias distintas de 1.000 resultados de tests. Para la primera muestra la cantidad de ítems en los cuales se observó un significativo desajuste había disminuido a 7, y en 4 de dichos ítems el desajuste era considerable a nivel de $\alpha = 0,01$. Para la segunda muestra la cantidad de ítems con un importante grado de desajuste fue de sólo 6, pero en ninguno de ellos el desajuste fue considerable a nivel de $\alpha = 0,01$.

Se efectuó un análisis de residuos en todo el universo usando el programa RESID, el cual reveló la existencia de sólo 10 ítems con un significativo grado de desajuste entre los datos y el modelo, tres de ellos a nivel de $\alpha = 0,01$. No obstante, de acuerdo con el programa BILOGW, en 6 de dichos ítems el desajuste no era acentuado.

Los residuos estandarizados se distribuyeron de la siguiente manera (véase p. 196 para una explicación):

RE	porcentaje
| 0 – 1 |	73,91
| 1 – 2 |	23,91
| 2 – 3 |	1,88
| > 3 |	0,31

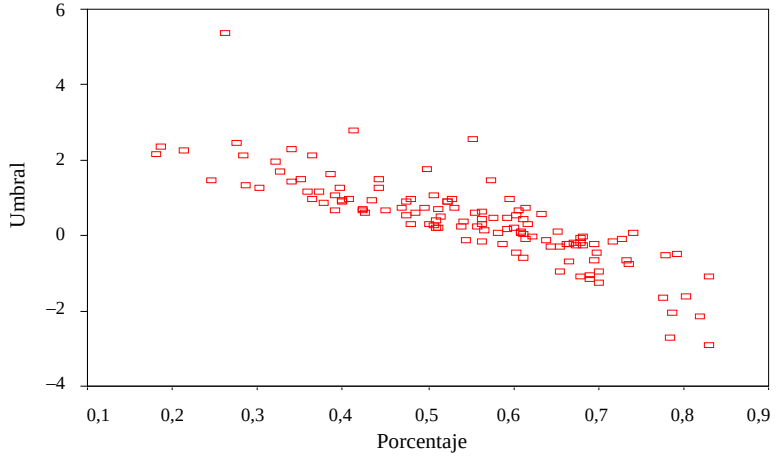
Como la distribución de los RE se acerca mucho a la distribución normal, el análisis de residuos respalda la tesis del ajuste del modelo a los datos.

Comparación entre estadísticas de ítems

La correlación entre valores p de TCM y valores *b* de IRT fue de r = -0,84, y entre la r_{bis} de TCM y los valores *a* de IRT la correlación fue de r = 0,63. La correlación entre los puntajes de tests de TCM y las estimaciones de habilidad de IRT fue de r = 0,96. En la Figura 4 se grafican los valores *b* contra los valores p.

En esta figura parece haber una relación curvilínea entre valores *b* y p. Puesto que los valores p expresan en realidad la dificultad de los ítems en una escala ordinal, se efectuó una transformación a una escala de intervalos por normalización (véase, por ejemplo, Aiken, 1991, p. 96). Sin embargo, esta normalización no hizo más que aumentar la correlación entre valores *b* y p a r = 0,85.

FIGURA 4: GRÁFICO DE DISPERSIÓN DE VALORES *b* DE IRT CONTRA VALORES p DE TCM



Con el objeto de investigar más a fondo la invarianza de los parámetros se tomaron 2 muestras aleatorias distintas, de 1.000 individuos cada una, en el universo de personas que rindieron la prueba. Para cada muestra se calcularon las estadísticas de ítems de TCM y se estimaron los parámetros de ítems de IRT. Se realizaron comparaciones entre valores p y correlaciones biseriales de ambas muestras y entre valores *b* y valores *a*. También se realizaron comparaciones con los mismos valores tomados del universo. El resultado de estas comparaciones aparece en la Tabla 7.

TABLA 7: ESTABILIDAD DE LAS ESTADÍSTICAS DE ÍTEMS DENTRO DE AMBOS MARCOS DE MEDICIÓN. CORRELACIONES ENTRE DIFICULTADES DE ÍTEMS Y DISCRIMINACIONES DE ÍTEMS BASADAS EN LA TCM Y EN IRT

Correlación entre	Universo–muestra A	Universo–muestra B	Muestra A – muestra B
valores p	0,995	0,996	0,989
valores *b*	0,940	0,946	0,960
r_{bis}	0,931	0,957	0,882
valores *a*	0,764	0,775	0,683

3.1. Análisis

Los intentos por ajustar un modelo logístico IRT de 3 parámetros al test SweSAT rendido en el último trimestre de 2002 no fueron muy fructíferos. Es difícil obtener una respuesta inequívoca en cuanto al ajuste del modelo a los datos, ya que no existen criterios objetivos. Con todo, el análisis factorial a nivel de ítems fue desalentador, lo mismo que el test BILOGW chi-cuadrado aplicado al universo total de alumnos que rindieron la prueba. Por otra parte, el test RESID resultó alentador, al igual que los tests BILOGW chi-cuadrado aplicados a dos muestras de 1.000 examinados cada una. Sin embargo, los parámetros de ítems no fueron invariantes, y de acuerdo con Hambleton y otros (1991) "…la invarianza y el ajuste del modelo a los datos son conceptos equivalentes" (p. 24). Por tanto, la conclusión debe ser que el modelo logístico de 3 parámetros no se ajustó a los datos del test SweSAT.

Por otra parte, los resultados de los análisis de TCM fueron muy estimulantes. Los índices de dificultad de los ítems, lo mismo que sus índices de discriminación, fueron muy estables entre ambas muestras aleatorias, como también entre el universo y las dos muestras.

3.2. A modo de conclusión

Puesto que el SweSAT ha sido elaborado dentro del marco de la TCM, ha parecido razonable comparar los resultados de los análisis efectuados dentro de IRT con los resultados correspondientes obtenidos en el contexto de la TCM. Las comparaciones empíricas entre los resultados derivados de ambas teorías no son muy habituales. Lawson (1991) comparó el modelo de un parámetro y la TCM en tres series de datos y encontró "…notables similitudes ente los resultados obtenidos a través de métodos de medición clásicos y los resultados obtenidos a través de métodos de rasgo latente[3] de un parámetro" (p. 163). Fan (1998) examinó las estadísticas de personas e ítems derivadas de IRT y la TCM en una base de datos de evaluación en gran escala de nivel estatal. Sus resultados indican que las estadísticas de personas e ítems derivadas de ambos marcos de medición eran bastante comparables. También descubrió que la invarianza de las estadísticas de ítems en todas las muestras, característica donde por lo general se considera que radica la superioridad de los modelos de IRT, parecía ser similar dentro de ambos marcos de medición. Tanto Lawson (1991) como Fan (1998) concluyen sus artículos citando pasajes de un discurso de apertura pronunciado por Thorndike en 1982 en una conferencia australiana centrada en modelos de IRT:

> Dudo que vaya a observarse un cambio significativo en la mayor parte de los tests, tanto en los elaborados a nivel local como en los estandarizados. Los ítems que escogeremos para un test no diferirán mucho de los que habríamos seleccionado con procedimientos anteriores, y el test resultante seguirá conservando en gran medida las mismas propiedades (p. 12).

En los estudios citados en este trabajo los índices de ítems de TCM eran no sólo comparables con los parámetros de ítems de IRT, sino que además eran por lo general más invariantes entre distintas muestras de examinados. Una posible explicación para estos resultados es que el modelo de IRT no se ajustaba a los datos del test. Pero incluso si los resultados se deben a un deficiente ajuste entre modelo y los datos, la única conclusión razonable es que para los datos de la SweSAT la TCM parece funcionar mejor que IRT.

La SweSAT tiene buena acogida entre los examinados y las universidades de Suecia. La exigencia más importante que se le formula es que

[3] Los modelos de IRT se denominan en ocasiones modelos de rasgo latente.

clasifique a los examinados de la manera más justa posible con respecto a su éxito previsto en los estudios. Otros requisitos con que debe cumplir el test son los siguientes:

- Debe estar en concordancia con los objetivos y el contenido de la educación superior.
- No debe tener efectos negativos en la educación secundaria de ciclo superior.
- Debe ser posible puntuar la prueba con rapidez, a bajo costo y objetivamente.
- No debe ofrecer la ocasión para que un individuo mejore sus resultados por medio de ejercicios mecánicos o asimilando principios especiales de resolución de problemas.
- Los examinados deben considerar el test como una experiencia significativa y apropiada.
- Es preciso observar la exigencia de que el reclutamiento esté libre de sesgos. Ningún grupo debe ser discriminado debido a razones tales como género, clase social, etc.

Vale la pena empeñarse en introducir todas las modificaciones que permitan mejorar el test en cualquiera de sus aspectos. Ello requeriría, sin embargo, efectuar cambios que aumentarían el grado de validez de dicho test. Ese debe ser con certeza el principal objetivo de todo cambio: aumentar la validez del test y no hacer que se ajuste a una determinada teoría de medición.

La IRT consiste en una familia de modelos que, según se ha sostenido, son útiles en el diseño, la construcción y la evaluación de tests educacionales. Es de esperar que a medida que se profundicen las investigaciones se vayan resolviendo los restantes problemas técnicos asociados a la aplicación de los modelos. También se espera que en los años venideros se desarrollen modelos más recientes y más aplicables, permitiendo que IRT proporcione soluciones aun mejores a importantes problemas de medición (Hambleton y otros, 1991). Como se mencionó anteriormente, otro importante supuesto de los modelos de IRT es la unidimensionalidad, lo cual significa que los ítems de un test miden una sola habilidad. Hay modelos en los que se supone que se requiere más de una habilidad para explicar el desempeño en un test. Estos así llamados modelos multidimensionales son, a pesar de todo, más complejos y no han sido tan bien desarrollados como los modelos unidimensionales.

Aun cuando en la actualidad IRT no parece ser aplicable a la construcción y el diseño de la SweSAT, se encuentran en marcha algunos trabajos

destinados a investigar si IRT puede emplearse para ecualizar las distintas versiones del test. Emons (1998) realizó un exhaustivo estudio: "Nonequivalent groups IRT observed score equating. Its applicability and appropriateness for the Swedish Scholastic Aptitude Test" ("Ecualización de puntajes observados en IRT para grupos no equivalentes. Su aplicabilidad y conveniencia para la Prueba de Admisión a las Universidades en Suecia"), en el cual utiliza preguntas que habían sido incluidas como ítems ancla en pretests aplicados en el segundo trimestre de 1996, para así ecualizar el test del segundo trimestre de 1997. Puede que en ese estudio, sin embargo, la cantidad de ítems ancla haya sido demasiado exigua para constituir un vínculo que permitiera la ecualización. En algunos subtests el vínculo constó de sólo 2 a 3 ítems. Estudios similares se realizan continuamente, y en la actualidad los resultados obtenidos con el método tradicional de "ecualización equipercentil de grupos equivalentes" se comparan permanentemente con la ecualización de la IRT empleando vínculos de diferentes magnitudes.

Otra área de estudios sobre SweSAT en que se emplean hoy en día modelos de IRT es el funcionamiento diferencial de los ítems (FDI). Las curvas características del ítem ilustran de un modo muy apropiado el problema del FDI, ya que las curvas muestran la probabilidad de responder correctamente un ítem, dado un cierto nivel de habilidad. La comparación de las curvas para diversos grupos de examinados (en el caso de SweSAT, principalmente hombres y mujeres) corresponde exactamente a la definición más aceptada de FDI. Para estos tipos de estudios puede usarse el puntaje del subtest, y a nivel de subtests la unidimensionalidad parece ser un supuesto aceptable.

Por último, en el caso de los tests adaptativos computarizados (TAC) o tests a la medida, IRT proporciona el único marco teórico adecuado. En los TAC, los ítems que le toca responder a un examinado dependen de su desempeño en los anteriores ítems en el test. Sólo los ítems que entregan más información sobre el examinado son los que se aplican. Los examinados con un alto nivel de habilidad no necesitan responder los ítems más fáciles, y a su vez los examinados con bajo nivel de habilidad no necesitan contestar los ítems de alta dificultad. De esta manera el test puede acortarse considerablemente y aun así entregar la misma información y la misma precisión en la medición que el test convencional más largo. Si en el futuro la SweSAT —o alguna versión de la misma— llegara a transformarse en un TAC entonces resultará imperioso aplicar IRT.

REFERENCIAS BIBLIOGRÁFICAS

Aiken, L. R. *Psychological Testing and Assessment*. Massachusetts: Allyn & Bacon, 1991.

Crocker, L. L. y J. Algina. *Introduction to Classical and Modern Test Theory*. New York: Holt, Rinehart & Winston, 1986.

Emons, W., H., M. (1998). "Nonequivalent Groups IRT Observed Score Equating. Its Applicability and Appropriateness for the Swedish Scholastic Aptitude Test". *Educational Measurement* N° 32 (1998), Umeå, Umeå University, Department of Educational Measurement.

Fan, X. "Item Response Theory and Classical Test Theory: An Empirical Comparison of their Item/Person Statistics". *Educational and Psychological Measurement*, Vol. 58 N° 3 (1998), p. 357-381.

Hambleton, R. K. y R. J. Rovinelli. "Assessing the Dimensionality of a Set of Test Items". *Applied Psychological Measurement*, 10 (1986), pp. 287-302.

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

Hambleton, R. K. y R. W. Jones. "Comparison of Classical Test Theory and Item Response Theory and their Applications to Test Development". *Educational Measurement: Issues and Practice*, 12 (3) (1993), pp. 535-556.

Hays, W. L. *Statistics*. Londres: Holt, Rinehart and Winston, 1969.

Henrysson, S. "Gathering, Analyzing, and Using Data on Test Items". En R. L. Thorndike (ed.), *Educational Measurement.* Washington DC: American Council on Education, segunda edición, 1971.

Lawson, S. *One-Parameter Latent Trait Measurement: Do the Results Justify the Effort?* The Annual Series of the Southwest Educational Research Association, Vol. 1 (1991), pp. 159-168.

Lord, F. M. *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale NJ: Lawrence Erlbaum, 1980.

Mislevey, R. L. y R. D. Bock BILOG-W. Manual for BILOG-W. Item Analysis and test Scoring with Binary Logistic Models. Chicago: Scientific Software International, 1990.

Rogers, J. RESID. Assessment of Fit for Unidimensional IRT Models. Programa desarrollado en University of Massachusetts School of Education, 1994.

Stage, C. "An Attempt to Fit IRT Models to the DS Sub-test in The SweSAT". *Educational Measurement* N° 19 (1996), Umeå, Umeå University, Department of Educational Measurement.

Stage, C. "The Applicability of Item Response Models to the SweSAT. A Study of the DTM Sub-test". *Educational Measurement* N° 21 (1997a), Umeå: Umeå University, Department of Educational Measurement.

Stage, C. "The Applicability of Item Response Models to the SweSAT. A Study of the ERC Sub-test". *Educational Measurement* N° 24 (1997b), Umeå: Umeå University, Department of Educational Measurement.

Stage, C. "The Applicability of Item Response Models to the SweSAT. A Study of the READ Sub-test". *Educational Measurement* N° 25 (1997c), Umeå, Umeå University, Department of Educational Measurement.

Stage, C. "The Applicability of Item Response Models to the SweSAT. A study of the WORD Sub-test". *Educational Measurement* N° 26 (1997d), Umeå, Umeå University, Department of Educational Measurement.

Stage, C. “A Comparison Between Item Analysis Based on Item Response Theory and Classical Test Theory. A Study of the SweSAT Sub-test WORD”. *Educational Measurement* N° 29 (1998a), Umeå, Umeå University, Department of Educational Measurement.

Stage, C. “A Comparison Between Item Analysis Based on Item Response Theory and Classical Test Theory. A Study of the SweSAT Sub-test ERC”. *Educational Measurement* N° 30 (1998b), Umeå, Umeå University, Department of Educational Measurement.

Stage, C. “A Comparison Between Item Analysis Based on Item Response Theory and Classical Test Theory. A Study of the SweSAT Sub-test READ”. *Educational Measurement* N° 33 (1999), Umeå, Umeå University, Department of Educational Measurement.

Thorndike, R. L. (1982). “Educational Measurement: Theory and Practice”. En D. Spearritt (ed.), *The Improvement of Measurement in Education and Psychology: Contributions of Latent Trait Theory*. Princeton, NJ: ERIC Document Reproduction Service N° ED 222545. □