Revista Electrónica de Metodología Aplicada
2002, Vol. 7 nº 1, pp. 1-20

Estimación de habilidad y precisión en tests adaptativos informatizados y tests óptimos: Un caso práctico

Francisco J. Abad [1], Julio Olea [1], Eulogio Real [2] y Vicente Ponsoda [1]
e-mail: fjose.abad@uam.es
[1] Universidad Autónoma de Madrid; [2] Universidad de Santiago de Compostela

RESUMEN.
Algunas veces, en contextos aplicados, se necesita escoger entre un Test Adaptativo Informatizado (TAI) o adaptado al sujeto y un Test Óptimo (TO) o adaptado al grupo. Este problema fue analizado para una prueba de Vocabulario Inglés. En un estudio anterior se encontraron algunas diferencias de habilidad entre grupos difíciles de explicar (Olea, Revuelta, Ximénez y Abad, 2000); nosotros intentamos explicar aquí tales diferencias como "error" (v.g.: sesgo) debido al procedimiento de estimación. Para ello, se desarrolló un programa para calcular el sesgo teórico como función del nivel de habilidad (Lord, 1983; Samejima, 1993a). En primer lugar, se analizó el sesgo teórico para el TO. En segundo lugar, se estudió la convergencia entre los resultados teóricos, simulados y empíricos considerado las estimaciones del TAI como las menos sesgadas. Los resultados muestran que las diferencias entre el TAI y el TO no pueden explicarse por el procedimiento de estimación y deben atribuirse a otras causas.

Palabras Clave: Tests Adaptativos Informatizados, Tests Óptimos, Teoría de la Respuesta al Ítem, Sesgo.

ABSTRACT.
Sometimes, in applied contexts, there is needed to choose among Computerized Adaptive Test (CAT) or adapted to the subject and Optimal Tests (OT) or adapted to the group. This problem was analyzed for an English Vocabulary Test. Some unexplained group ability differences were found in a previous study (Olea, Revuelta, Ximénez and Abad, 2000); here we try to explain such differences as "error" (v.g.: bias) from the estimation procedure. A program was developed to calculate the theoretical bias as function of the ability level (Lord, 1983; Samejima, 1993a). In the first place, the theoretical bias was analyzed for the OT. In second place, the convergence between the theoretical, simulated and empirical results was studied considering the CAT estimates as the less unbiased ones. The results show that the differences between the CAT and the OT cannot be explained by the estimation procedure and they should be attributed to other causes.

Key words: Computerized Adaptive Tests, Optimal Tests, Item Response Theory, Bias.

1.- Introducción.

Dos de las principales aplicaciones que se derivan de la Teoría de la Respuesta al Ítem (TRI) y de las aplicaciones informáticas en Psicometría son los Tests Adaptativos Informatizados (TAIs) y los Tests Óptimos (TOs). Un TAI consiste básicamente en presentar al evaluando los ítems de un banco calibrado que más contribuyen a la precisión de la estimación de su nivel de habilidad (Olea y Ponsoda, 1996). Por otro lado, un TO es un test fijo, que se aplica a todos los evaluandos, cuyos ítems se seleccionan de un banco calibrado para que cumpla determinadas condiciones psicométricas o restricciones respecto a los contenidos que debe incluir (Hambleton, Slater, Narayanan y Setiadi, 1996). Dependiendo de los objetivos de la aplicación, en la construcción del TO pueden enfatizarse aspectos diferentes como la precisión global del test, la precisión asociada a un punto concreto de la escala de habilidad o su validez de contenido.

Tanto en uno como en otro tipo de tests se requiere, entre otras cosas, un método estadístico para estimar el nivel de habilidad (θ). Los más utilizados son la estimación de máxima verosimilitud (ML) y dos procedimientos bayesianos: la estimación máxima a posteriori (MAP) y la esperada a posteriori (EAP). Mientras que la estimación ML se fundamenta únicamente en los datos empíricos, los métodos bayesianos incorporan información sobre la distribución a priori de los niveles de habilidad de la población. La eficacia de cualquier método de estimación de parámetros de habilidad se valora a partir del sesgo y del error típico de medida que generan para diferentes niveles de habilidad y para distintas condiciones de aplicación. Normalmente este tipo de valoración se plantea mediante estudios de simulación, para tests fijos de diferente longitud y dificultad. Este tipo de trabajos tienen ventajas indudables, si bien no siempre es fácil generalizar las conclusiones a situaciones reales que no se ajustan estrictamente a las condiciones simuladas. El planteamiento del presente trabajo es analizar el grado de equivalencia entre las estimaciones de habilidad y los errores de medida que proporcionan los 3 métodos estadísticos en una aplicación empírica intrasujeto de un TO y un TAI.

1.1.- Métodos de estimación de la habilidad, error típico y sesgo teórico

1.1.1.- Estimación de máxima verosimilitud (ML)

Para el modelo logístico de 3 parámetros, conocidos los parámetros de los ítems, la estimación del nivel de habilidad de un evaluando puede obtenerse a partir de la función de verosimilitud:

$$L = L(u \mid \theta) = \prod_{j=1}^{n} P_j^{u_j} Q_j^{1-u_j} \qquad (1)$$

donde *u* es el vector de respuestas a los ítems, P_j es la probabilidad de acertar el ítem *j* $(u_j=1)$ dado un nivel de rasgo θ y Q_j la probabilidad complementaria $(u_j=0)$. Mediante métodos numéricos (v.g.: Newton-Raphson) se obtiene el estimador máximo-verosímil, que es la solución de la ecuación:

$$\ln(L)' = \sum_j a_j \frac{P_j^*}{P_j}\left(u_j - P_j\right) = 0 \qquad (2)$$

donde ln(*L*)' es la derivada parcial de ln(*L*) con respecto a $\boldsymbol{q}$ y P_j^* es la probabilidad de acierto según el modelo de 2 parámetros $[(P_j-c_j)/(1-c_j)]$. Asintóticamente, el error típico de medida es el inverso de la raíz cuadrada de la información I(θ), siendo el valor de ésta:

$$I = I(\boldsymbol{q}) = \sum_{j=1}^{n} a_j^2 \left(P_j^*\right)^2 \frac{Q_j}{P_j} \qquad (3)$$

Lord (1983) derivó la función de sesgo teórico para el método ML, que fue generalizada por Samejima (1993a; 1993b) como:

$$BIAS(MLE(\boldsymbol{q})) \approx \frac{\sum_{j=1}^{n} a_j I_j \left(P_j^* - 0.5\right)}{I^2} \qquad (4)$$

1.1.2.- Estimación bayesiana MAP

La idea fundamental de los métodos bayesianos es incorporar a la función de verosimilitud información sobre la distribución a priori de la habilidad en la población. La estimación MAP (maximum a posteriori) o estimación bayesiana modal (Samejima, 1969) es el valor $\boldsymbol{q}$ que maximiza la probabilidad posterior:

$$P(\boldsymbol{q} \mid u) = \frac{g(\boldsymbol{q})L(u \mid \boldsymbol{q})}{L(u)} \qquad (5)$$

donde g(θ) es la distribución a priori de la habilidad y L(u) es la verosimilitud del patrón de respuestas *u* independientemente de θ. Siguiendo a Baker (1992), y puesto que el denominador de la ecuación (5) no contiene θ, puede demostrarse que siendo g(θ) una distribución normal, N(μ, σ), el máximo se encuentra igualando la derivada del logaritmo del numerador a 0, resolviendo la ecuación:

$$-\frac{(\theta-\mu)}{\sigma^2}+\ln(L)'=0 \qquad (6)$$

Para θ próximas a μ y/o para σ elevada (a priori débil) el componente izquierdo de la ecuación (6) se aproxima a 0. De nuevo, la varianza del estimador es asintóticamente el inverso de la función de información, que en este caso queda:

$$J(\theta)=\frac{1}{\sigma^2}+I \qquad (7)$$

El sesgo teórico para MAP se relaciona con el sesgo en ML (Lord, 1986):

$$BIAS\,(MAP(\theta)) \approx BIAS\,(MLE(\theta))-\frac{\theta}{I} \qquad (8)$$

donde se asume distribución a priori normal N(0, 1).

1.1.3.- Estimación bayesiana EAP

El estimador EAP es la media de la misma distribución posterior referida anteriormente y se expresa como (Bock y Aitkin, 1981; Bock y Mislevy, 1982):

$$E(\theta \mid u)=\int_{-\infty}^{\infty}\theta P(\theta \mid u)d\theta \qquad (9)$$

Usando los puntos de cuadratura de Gauss-Hermite (ver Stroud y Sechrest, 1966) esa integración pueden resolverse directamente, sin proceso iterativo:

$$E(\boldsymbol{q} \mid u) \cong \sum_{k=1}^{q} \boldsymbol{q}_k \frac{G(\boldsymbol{q}_k)L(u \mid \boldsymbol{q}_k)}{\sum_{l=1}^{q} G(\boldsymbol{q}_l)L(u \mid \boldsymbol{q}_l)} \qquad (10)$$

Donde, asumiendo distribución normal N(μ, σ), los valores de θ_k próximos a μ obtendrán mayor ponderación y en mayor medida si la σ es pequeña (a priori fuerte) y/o el número de ítems aplicados es pequeño (L más uniforme). La varianza del estimador es:

$$PSD(\boldsymbol{q} \mid u) \cong \sum_{k=1}^{q} (\boldsymbol{q}_k - \boldsymbol{q})^2 \frac{G(\boldsymbol{q}_k)L(u \mid \boldsymbol{q}_k)}{\sum_{l=1}^{q} G(\boldsymbol{q}_l)L(u \mid \boldsymbol{q}_l)} \qquad (11)$$

Donde, de nuevo, la varianza dependerá de la ponderación que se haga de los valores alejados de θ.

1.2.- Sesgo y error típico en estudios de simulación

Se han efectuado diversos estudios de simulación (Kim y Nicewander, 1993; Warm, 1989), bajo diferentes condiciones realistas de longitud del test, discriminación y dificultad de los ítems; generalmente, la distribución a priori es normal N(0,1), próxima a la media de la población simulada. Se concluye que bajo todos los métodos se produce cierto sesgo aunque en distinto grado y dirección. La dirección del sesgo se formaliza en las funciones del sesgo teórico para la estimación ML (ecuación 4) y para la estimación MAP (ecuación 8). El sesgo ML es "hacia fuera"; es decir, positivo si el nivel de habilidad del evaluando es sensiblemente superior a la media de la dificultad de los ítems que se le administran y negativo si es inferior. El sesgo MAP es "hacia dentro"; en la ecuación 8 puede verse que es el resultado de añadir un componente (positivo para niveles de habilidad bajos y negativo para niveles altos) al sesgo ML. Kim y Nicewander (1993) muestran en un estudio de simulación que el sesgo es más importante en la estimación ML. Sin embargo, si las estimaciones ML son relativamente insesgadas, la estimación MAP puede producir sesgo positivo para los niveles bajos de habilidad y

negativo para los niveles altos. En cuanto al error típico, las diferencias entre las estimaciones bayesianas EAP y MAP son pequeñas, produciéndose para ambas un menor error típico que para la estimación máximo verosímil.

Respecto a las propiedades de los diferentes métodos de estimación en los TAIs, algunos estudios de simulación (Warm, 1989; Vispoel, Wang y Bleiler, 1997; Wang y Vispoel, 1998) concluyen que los métodos bayesianos producen menor error típico pero mayor sesgo. En cuanto a lo primero, es lo esperable teóricamente. Para un mismo valor θ, *por definición*, la información es mayor en MAP (ecuación 7) que en ML (ecuación 3), y tanto mayor cuanto menor sea la varianza de la distribución a priori. Además, De Ayala, Shafer y Sava-Bolesta (1995) muestran mediante un estudio de simulación cómo la desviación típica de la distribución posterior se aproxima al error típico observado (v.g.: desviación típica para la distribución de las θ estimadas) si el número de puntos de cuadratura es suficientemente grande; por el contrario, el error típico teórico ML es generalmente inferior al empírico cuando se utilizan TAIs cortos (Warm, 1989). En cuanto al sesgo, si el TAI es informativo para cualquier valor de θ, las estimaciones ML resultan esencialmente insesgadas, mientras que las bayesianas producen cierta regresión a la media de la distribución a priori especificada. Esto hace que los métodos bayesianos sean recomendables cuando el objetivo sea ordenar a los sujetos (aunque con más de 30 ítems habrá pocas diferencias) mientras que máxima verosimilitud es preferible en situaciones en las que el objetivo es: a.) efectuar comparaciones de grupos por su media; b.) equiparar las estimaciones de la habilidad de diferentes tipos de tests (adaptativos y fijos); c.) situar con máxima precisión a un sujeto dentro de uno de dos grupos (mastery testing). En esos contextos, la estimación bayesiana puede ser inadecuada si no se ajusta bien la distribución a priori (Wang y Vispoel, 1998; Wang, 1997).

1.3.- Planteamiento del estudio

El objetivo del presente estudio es comparar las estimaciones que proporcionan los 3 métodos estadísticos descritos (ML, MAP y EAP) en un TO con las que se obtienen en un TAI que utiliza el método ML. La razón de introducir estimaciones bayesianas para el TO es que, en tests fijos, pueden ser más insesgadas. En general, el interés en este tipo de comparaciones tiene su origen en un estudio previo (Olea, Revuelta, Ximénez y Abad, 2000) en que se obtuvo mayor nivel medio de habilidad estimado (método ML) en una muestra a la que se aplicó el TO que en otra en que se aplicó un TAI. Este resultado, inicialmente inesperado, podría en principio deberse a que: a.) ambos grupos tuvieran diferente nivel de habilidad, algo posible pero improbable dada la asignación aleatoria de los sujetos a los grupos; b.) las propiedades de las estimaciones ML aplicadas a dos tipos de tests distintos (óptimo y adaptativo); c.) otras características de la situación o del test (p.e., que en un test fijo los ítems pueden hacerse públicos más fácilmente de sesión a sesión). En el presente estudio se planteó un diseño intrasujeto que permita establecer en qué medida las características de precisión (y sesgo) de los tests (TO y TAI) explican sus diferencias.

Desde un punto de vista aplicado, este tipo de comparaciones puede ser relevante para tomar decisiones en el proceso de diseño de un TAI a partir de sus ancestros convencionales; por ejemplo, todos los estudios realizados con el CAT-ASVAB para equiparar las estimaciones en ambos tipos de tests cuando se aplican a sujetos distintos (Segall y Moreno, 1999). En otros casos, existen situaciones en que se permite a los sujetos elegir el tipo de test, para intentar minimizar el nivel de ansiedad que puede generar alguno de ellos (Wise, 1999). En ambas situaciones conviene conocer bajo qué procedimiento de estimación y bajo qué condiciones y/u objetivos las estimaciones de un TO son comparables a las de un TAI. Esta cuestión es importante porque si ambas estimaciones son comparables un TO posee ventajas en términos de *eficiencia* (v.g.: coste de aplicación). Uno de nuestros objetivos ha sido evaluar en qué medida las diferencias entre las estimaciones en un TO y un TAI pueden venir explicadas por el procedimiento de estimación en el TO.

2.- Método

2.1.- Participantes

La muestra estuvo formada por 88 estudiantes de primer curso de Psicología de las Universidades Autónoma de Madrid y Santiago de Compostela. Las edades oscilaron entre 17 y 19 años.

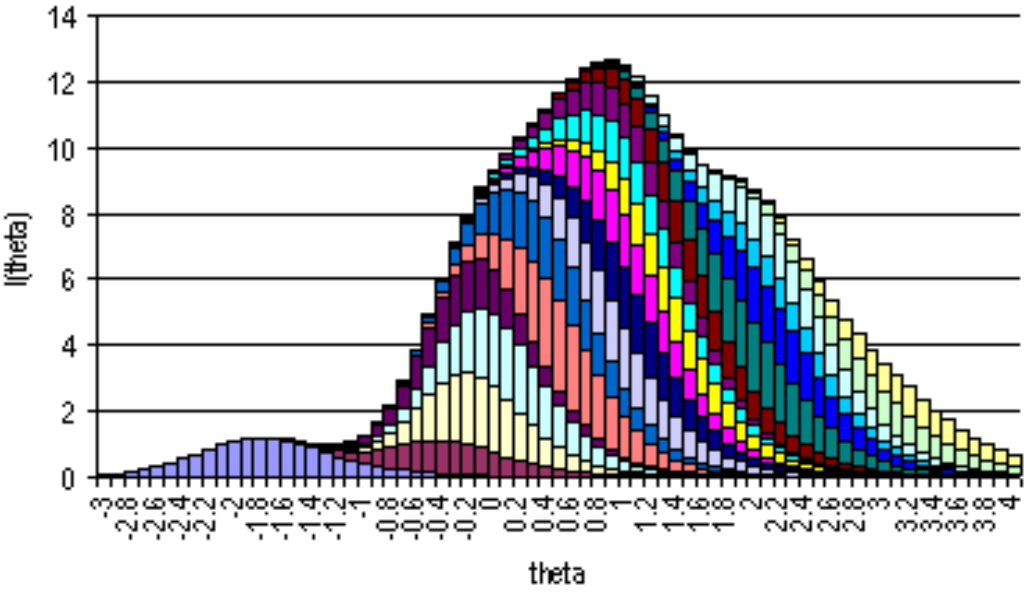
2.2.- Instrumentos

Cada sujeto respondió a un TO informatizado y a un TAI, ambos de 20 ítems. Desde un banco original de 221 ítems de vocabulario inglés, calibrados según el modelo logístico de 3 parámetros, se seleccionaron 20 ítems (en la forma en que se describe en el siguiente párrafo) para formar el TO, quedando el banco de referencia para el TAI formado por los 201 ítems restantes. Más datos sobre el cumplimiento de los supuestos del modelo, las propiedades psicométricas de los ítems, su precisión y capacidad predictiva pueden consultarse en Olea, Ponsoda, Revuelta y Belchí (1996) y Ponsoda, Wise, Olea y Revuelta (1997).

En el citado estudio previo (Olea et al., 1996), se obtuvo en una muestra de estudiantes universitarios una media de habilidad de 0.57 y una desviación típica de 0.92. Dado que esta era la población objetivo, para diseñar el TO se seleccionaron al azar 20 valores de una distribución normal con estos parámetros, y se eligieron los 20 ítems del banco que resultaban más informativos para estos niveles de habilidad. Los ítems se presentaron a los sujetos ordenados de menor a mayor dificultad. En la figura 1 se muestran los parámetros de los ítems y la función de información del test. Como puede verse, el TO discrimina razonablemente bien para niveles de habilidad entre -.30 y 2.40 (error típico inferior a .4).

FIGURA 1: Función de información separada por ítem (ordenados en dificultad) para el TO

	a	b	c
1	1.55	-1.91	.19
2	1.66	-.53	.35
3	2.10	-.23	.19
4	1.92	-.01	.10
5	1.86	-.32	.26
6	2.12	.45	.09
7	1.79	.34	.11
8	1.78	.83	.10
9	2.00	.93	.24
10	1.75	.89	.16
11	1.79	1.19	.28
12	1.63	.93	.21
13	1.53	.89	.24
14	1.52	1.43	.15
15	1.81	1.93	.07
16	1.80	2.05	.16
17	1.59	1.98	.34
18	1.49	2.11	.12
19	1.46	2.80	.22
20	1.44	3.00	.27



El algoritmo implementado en el TAI (Ponsoda, Olea y Revuelta, 1994) asume que se han respondido a dos ítems, uno con b=-4 que se acierta y otro con b=4 que se falla. Como procedimiento de arranque, se selecciona un nivel de habilidad aleatorio entre -0.4 y +0.4. Para evitar las estimaciones extremas después de responder a los primeros ítems, el programa aplica la solución propuesta por Revuelta y Ponsoda (1997). El procedimiento de selección progresiva de los ítems se basa en el criterio de máxima información de Fisher. Después de la respuesta a un ítem, se estima un nivel de habilidad provisional según el procedimiento ML. Como criterio de control de la exposición de los ítems, se aplica el método progresivo (Revuelta y Ponsoda, 1998).

2.3.- Procedimiento y análisis de datos

En primer lugar se realizaron, utilizando el programa BILOG (Mislevy y Bock, 1990), las siguientes estimaciones de habilidad y error típico de medida: a.) ML de las respuestas a los 20 ítems presentados de modo adaptativo, incluyendo los dos ítems previos ficticios (TAI-ML); b.) ML de las respuestas al TO (TO-ML); c.) EAP de las respuesta al TO, fijando una distribución a priori N(0, 1)(TO-EAP); d.) MAP en el TO, con esa misma distribución a priori (TO-MAP); e.) EAP en el TO,

con una distribución a priori N(0.57, 0.92) (TO-EAPP); f.) MAP en el TO, con esa misma distribución a priori (TO-MAPP).

En segundo lugar, se comprobó que la muestra obtenida era representativa de la población objetivo que se pretendía evaluar. En tercer lugar, se estableció el grado de sesgo de las estimaciones TAI. Comprobado esto, las estimaciones TAI-ML se utilizan para el resto de los análisis como mejor aproximación a las θ reales.

En cuarto lugar, se establecieron las funciones *teóricas* de sesgo y error típico (fórmulas 3, 4, 7 y 8) para el TO (TO-ML y TO-MAP). Adicionalmente se simularon 100 réplicas de esa muestra partiendo de los parámetros fijos para los sujetos (las $\boldsymbol{q}$ del TAI) y para los ítems (parámetros *a*, *b* y *c* conocidos). Las estimaciones para los datos simulados se realizaron utilizando PARSCALE (Muraki y Bock, 1997) para la estimación EAP y MULTILOG (Thissen, 1991) para las estimaciones ML y MAP. Para cada método de estimación se calculó el promedio en las 100 réplicas de:

a) la media de las θ estimadas ($\bar{\theta}$);
b) el RSME o raíz del error cuadrático medio entre la estimación de $\boldsymbol{q}$ con respecto a q

$$\sqrt{\sum^{100}\left(\hat{\boldsymbol{q}}-\boldsymbol{q}\right)^2/100}$$

c) el sesgo o la diferencia media entre $\boldsymbol{q}$ estimada y $\boldsymbol{q}$:

$$\sum^{100}\left(\hat{\boldsymbol{q}}-\boldsymbol{q}\right)/100\Big)$$

d) el SE o raíz del error cuadrático medio entre la $\boldsymbol{q}$ estimada y la $\boldsymbol{q}$ estimada como promedio en las 100 réplica:

$$\sqrt{\sum^{100}\left(\hat{\boldsymbol{q}}-\bar{\boldsymbol{q}}\right)^2/100}$$

Recordemos (Wang y Vispoel, 1998) que $RMSE^2 = SE^2 + sesgo^2$; esto quiere decir que el error total (RMSE) puede descomponerse en dos componentes, uno relacionado con el error sistemático (sesgo) y otro relacionado con el error aleatorio o de muestreo (SE).

En quinto lugar, se analizaron las diferencias psicométricas (en error típico y sesgo) entre el TAI y el TO (pruebas t). Diferencias en error típico entre distintas estimaciones TO con respecto al TAI se manifestarán en sus correlaciones con éste. Diferencias en sesgo de las estimaciones TO con respecto al TAI se manifestarán en diferencias "de nivel", ya sean globales (medias) o locales (por nivel de habilidad). En la misma línea se estudió los efectos de utilizar distribuciones a priori diferentes según reflejen o no las características de la muestra de forma realista.

Finalmente, se analizó la convergencia entre los resultados teóricos, simulados y empíricos. Esta comparación es clave ya que si convergen los resultados empíricos y simulados (patrón de sesgo, error típico, etc. para el TAI y las distintas estimaciones TO) las diferencias del TO con el TAI podrían explicarse por sus características de precisión y sesgo.

3.- Resultados

3.1.- Características y distribución de las estimaciones TAI

Los errores típicos que se obtuvieron para las estimaciones TAI oscilaron entre .25 y .34. Estableciendo la función de sesgo teórico que se corresponde con los ítems aplicados a cada sujeto y que se definió en (4) se obtienen sesgos entre .08 y -.05 (media=.0097). Por lo tanto, aún sin conocer el nivel de habilidad verdadero, puede asumirse que para la estimación TAI-ML el sesgo es pequeño.

La distribución de las estimaciones TAI no difiere significativamente de la distribución de la muestra original ni en media ($t_{87} = 1.670$; $p = .098$) ni en variabilidad ($\chi^{287} = 64.15$; $p = .968$). Por lo tanto, las características de la muestra son adecuadas ya que no hay sujetos fuera del rango de habilidad de la población objetivo a la que el TO va dirigido.

3.2.- Sesgo y error típico teóricos del TO (estimaciones MAP y ML)

Los valores medios de sesgo teórico pueden observarse en la Tabla 1. En las figuras 2 y 3, se representan los sesgos teóricos para las estimaciones TO-MAP y TO-ML, respectivamente. Como puede verse en las figuras, se obtienen valores razonables de sesgo. Más concretamente, el sesgo-ML teórico (figura 3) presenta valores entre -.14 y .13 (media= 0.01). Por otro lado, el sesgo teórico-MAP (figura 2) tiende a ser negativo hasta un sesgo máximo de -.40 (media = -0.06). Así,

teóricamente, para la estimación TO-MAP pueden producirse sesgos negativos (por debajo de -.20) para niveles de rasgo por encima de 2.10. En el otro extremo del continuo de habilidad el sesgo es inapreciable para cualquiera de los procedimientos de estimación. En este sentido son esperables pequeñas diferencias de nivel para las estimaciones bayesianas con respecto a TO-ML.

FIGURA 2: Sesgo por nivel de habilidad para las estimaciones bayesianas teóricamente (TO-MAP(t)) y en la simulación (TO-MAP(s), TO-EAP(s)). Diferencias en el estudio empírico entre las estimaciones MAP y EAP del TO y la estimación del TAI (TO-MAP(e), TO-EAP(e))

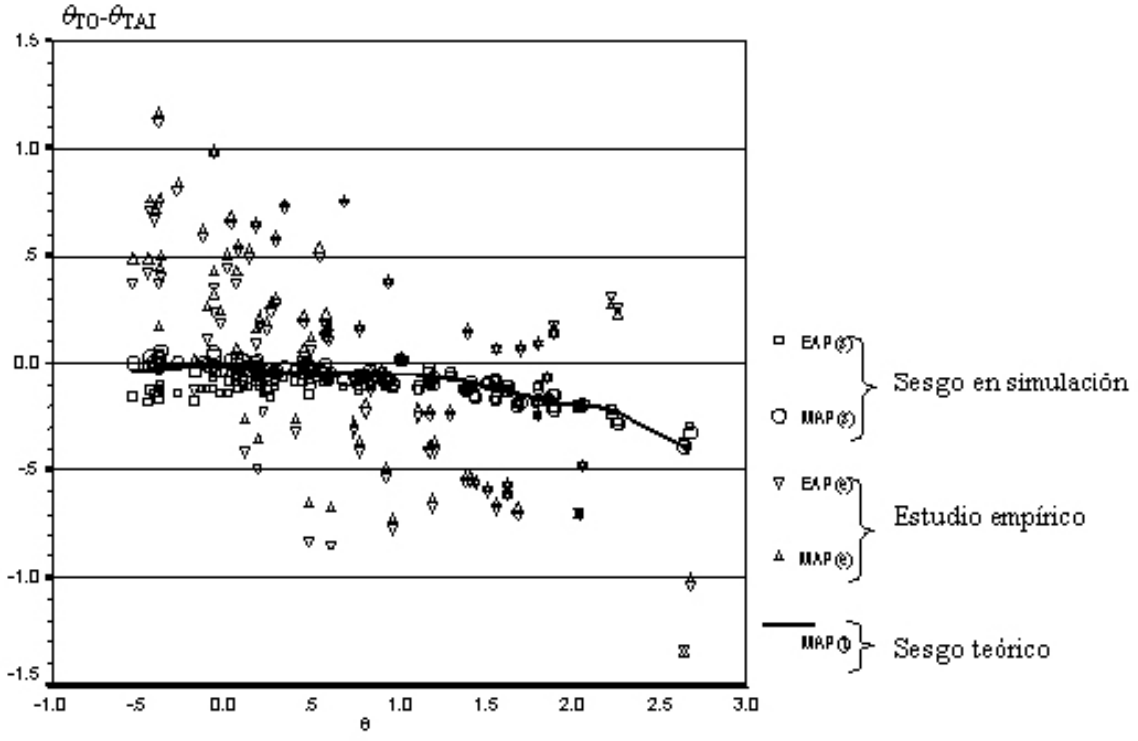
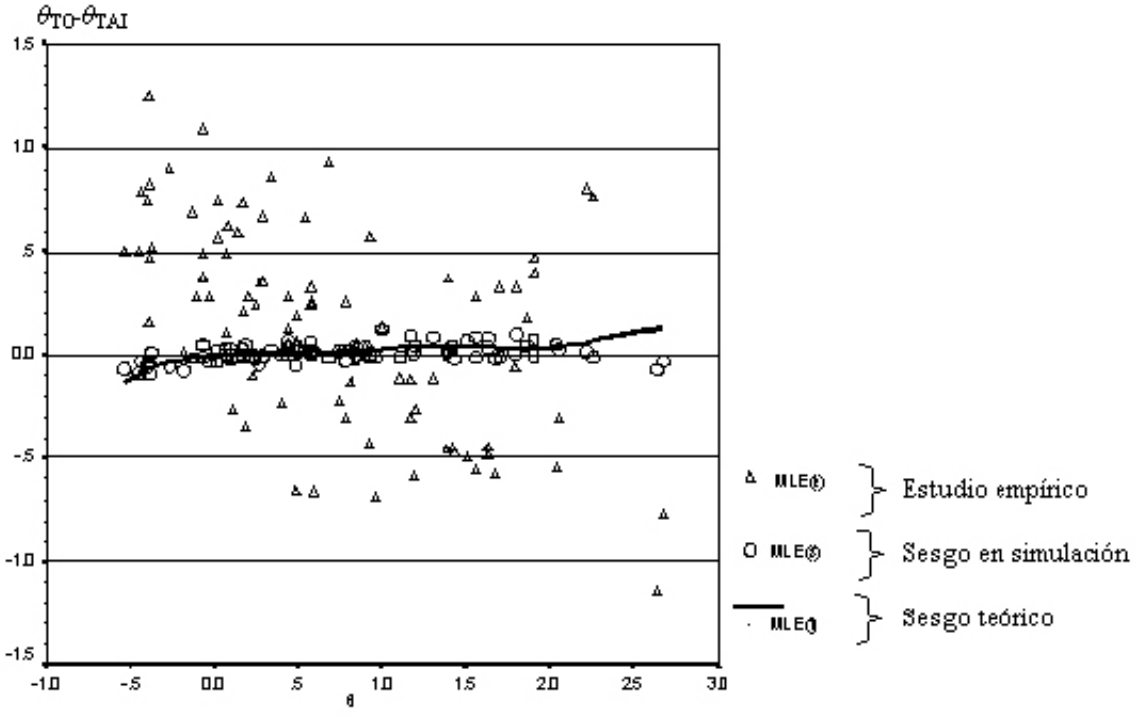


FIGURA 3: Sesgo por nivel de habilidad para la estimación máximo-verosimil teórica (TO-ML(t)) y en el estudio de simulación (TO-ML(s)). Diferencias en el estudio empírico entre las estimaciones ML del TO y la estimación del TAI (TO-MLE(e))



En cuanto al error típico *teórico* (ver figuras 2 y 3) no existen grandes diferencias entre los 2 procedimientos (ML: media = .32; MAP: media = .31). Esta precisión es lógicamente inferior a la obtenida en el TAI (.28). Dados estos resultados, cabría esperar pocas diferencias en cuanto al ordenamiento de los sujetos entre los distintos procedimientos del TO, aunque sí entre estos y el TAI.

3.3.- Sesgo y SE del TO obtenidos en la simulación (EAP, MAP y ML)

En las mismas figuras 2 y 3, se representan los sesgos "por simulación" (promedio del sesgo obtenido en las 100 réplicas para cada nivel de habilidad) para las estimaciones TO-ML, TO-MAP y TO-EAP. En la parte derecha de la tabla 1 se presentan los valores medios de sesgo, SE y RMSE para estos procedimientos. Más concretamente, para el sesgo-ML simulado se obtuvo una media de 0.01. En algunas réplicas, los sujetos de muy baja (inferior a -.13) o muy alta habilidad (superior a 2.05) se obtenían un patrón constante de respuestas (15 % de los niveles de habilidad). Puesto que para esos sujetos no existe estimación no se incluyeron para establecer el promedio de la estimación. Para los mayores niveles de habilidad (2.62 y 2.67) se obtuvo la pérdida máxima (13 y 23% de las réplicas respectivamente). Para el resto de los niveles de habilidad en los que hubo pérdida ésta resultó pequeña (menos del 5 % de las réplicas). Eso puede explicar que se obtuviera un sesgo menor del esperado en los sujetos de alta habilidad, mientras que los resultados obtenidos para el resto de los sujetos son los esperados. La correlación entre el sesgo teórico y el obtenido mediante simulación eliminando los 2 niveles de habilidad con perdida mayor fue positiva ($r = .59$; $p < .000$). El hecho de que la correlación no sea más alta puede explicarse por el valor prácticamente nulo del sesgo a lo largo del continuo de habilidad.

Por otro lado, el sesgo simulado-MAP es negativo (media =-.05). Así pues los resultados de la simulación (MAP) convergen con los teóricos ($r = .92$; $p<.000$). El sesgo-simulado-EAP es ligeramente superior (media = -.11) aunque correlaciona altamente con el sesgo-simulado-MAP ($r = .76$; $p < .000$).

En cuanto al SE (error típico) no existen grandes diferencias entre los 2 procedimientos bayesianos (ver tabla 1); la estimación ML tiene un SE algo superior a ambos. La precisión para el TO es lógicamente inferior a la obtenida en el TAI. Dados estos resultados, cabría esperar pocas diferencias en cuanto al ordenamiento de los sujetos *entre* los distintos procedimientos del TO, siendo el peor funcionamiento para la estimación máximo verosímil.

3.4.- Resultados empíricos:

3.4.1.- Ordenamiento de los sujetos

En la tabla 1 se incluyen las correlaciones con el TAI entre los niveles de habilidad estimados por cada uno de los métodos/tests empleados. Las correlaciones entre las estimaciones realizadas en el TAI y las obtenidas en el TO descienden considerablemente (entre 0.79 y 0.80) con respecto a las correlaciones, no mostradas, entre los distintos procedimientos de estimación aplicados al TO (>.997). Asumiendo como referencia el TAI, no parece preferible un procedimiento de estimación u otro en el TO; esto implica que las estimaciones no difieren de forma importante en cuanto al error típico de

estimación, pues de lo contrario esas diferencias se manifestarían en las correlaciones. Los resultados obtenidos son coherentes con los obtenidos a partir de los estudios de simulación que muestran correlaciones que difieren poco entre sí. Las correlaciones TAI-TO empíricas (en torno a .80) están por debajo de las obtenidas en los estudios de simulación (correlaciones en torno a .90) lo que puede venir en parte explicado porque en las simulaciones las θ de los sujetos son conocidas.

TABLA 1: Medias y desviaciones típicas (entre paréntesis) para las estimaciones[+] de habilidad y de precisión. En la fila inferior: correlaciones entre las distintas estimaciones de habilidad con la estimación TAI.

	Resultados empíricos						Estudios de Simulación		
	TAI	TO-ML	TO-EAP	TO-MAP	TO-EAPP	TO-MAPP	TO-ML(S)	TO-EAP(S)	TO-MAP(S)
Habilidad	.71 (.79)	.85 (.68)	.73 (.63)	.77 (.60)	.79 (.61)	.81 (.59)	.72(.81)	.61(.77)	.66(.72)
SE	.28 (.02)	.31 (.04)	.32 (.05)	.29 (.02)	.31 (.04)	.29 (.02)	.34(.05)	.31(.04)	.29(.03)
Sesgo		.14(.49)	.02(.48)	.06(.48)	.07(.48)	.10(.48)	.01(.04)	-.11(.06)	-.05(.08)
RMSE							.34(.05)	.34 (.05)	.31(.04)
Correl. con TAI		.79	.79	.80	.80	.80	.92*	.92*	.93*

+ La estimación ML se basa en aquellas muestras que dan lugar a estimaciones finitas.

* Este valor es el promedio de las correlaciones en las 100 replicas.

3.4.2.- Error típico

En la tabla 1 se incluyen también las medias y desviaciones típicas de los niveles de habilidad y errores típicos (en base a fórmulas) estimados en cada test por los diferentes métodos. El menor error típico se obtiene para la estimación TAI, luego le siguen algunas estimaciones bayesianas (MAP), después la estimación ML. Todas esas diferencias resultaron significativas ($p < .01$). Finalmente los peores resultados se obtuvieron para la estimación EAP. El error típico medio obtenido para las EAP es mayor (TO-EAP: t_{87}=-2.683; $p = .009$) o indiferenciable (TO-EAPP: t_{87}=.801; $p = .425$) del de la estimación TO-ML. En cualquier caso, la precisión fue significativamente mayor para las estimaciones MAP ($p<.01$).

3.4.3.- Diferencias de nivel

En cuanto al nivel medio de habilidad estimado, la diferencia más clara se observa entre la estimación TAI y la TO-ML (t_{87}=-2.639; $p < .05$), siendo ésta última mayor, tal como ocurría en trabajos anteriores (Olea et al., 2000). Esto no puede explicarse como un problema de sesgo de la estimación ML (p.e., los sujetos de alta habilidad se estiman con sesgo positivo) puesto que algunos sujetos con más puntuación en el TAI obtienen puntuaciones *menores* en el TO-ML y, más importante,

muchos con puntuaciones bajas obtienen puntuaciones *mayores* (ver figura 3). Además, en el análisis teórico y el estudio de simulación se muestra que las estimaciones ML deben ser en términos *globales* insesgadas (t_{87}=-1.748; p=.084).

Por el contrario, no existen diferencias (p > .05), en general, entre las estimaciones TAI y cualquiera de las bayesianas (TO-MAP y TO-EAP). Sin embargo, de las propiedades psicométricas de los tests analizadas a partir del estudio de simulación lo razonable sería esperar que esas estimaciones estuvieran globalmente *sesgadas* (EAP; t_{87} = 17.676; p < .000; MAP; t_{87} = 5.881; p < .000).

En el patrón general anterior de resultados se observan claras discrepancias con lo observado en los estudios de simulación. En resumen, las estimaciones obtenidas empíricamente son superiores a las obtenidas como promedio de las 100 replicas (ML; t_{87} = 2.426; p < .017; MAP; t_{87} = 2.326; p < .022; EAP; t_{87} = 2.492; p < .015).

Con respecto a las comparaciones de las estimaciones TO entre sí, todas resultan significativas (p < .001). Los valores más altos se corresponden con las estimaciones TO-ML, a los que siguen las estimaciones TO-MAPP, TO-EAPP, TO-MAP y TO-EAP, respectivamente. Este ordenamiento revela los efectos de: a.) fijar una distribución a priori con distinta media (0 ó .52); b.) establecer un procedimiento MAP o EAP. En cualquier caso, es claro que el hecho de fijar una distribución a priori "más adecuada" a las características de la población objetivo no incrementa la eficacia o similaridad de las estimaciones bayesianas con las estimaciones del TAI.

4.- Discusión

Los resultados en cuanto a medias de habilidad estimada (superior en el TO que en el TAI) coinciden con los obtenidos en un diseño inter-sujeto (Olea et al., 2000). En el estudio actual, dado el diseño intrasujeto utilizado, este resultado no puede atribuirse a diferencias en el nivel de habilidad de los grupos, por lo que las diferencias se deben definitivamente a diferencias en las características de los tests. Teóricamente el TAI debe constituir la mejor referencia para la habilidad verdadera del sujeto puesto que se adapta al sujeto, frente al TO que se adapta al grupo. Las ventajas de un TO se plantean siempre en términos de *eficiencia* en algunos aspectos (vg.: información que proporcionan dado el coste de aplicación). Uno de nuestros objetivos ha sido evaluar en qué medida las diferencias entre las estimaciones en un TO y un TAI pueden venir explicadas por el procedimiento de estimación en el TO.

Por un lado, existen diferencias en cuanto al ordenamiento de los sujetos. Las correlaciones entre las estimaciones del TAI y el TO distan de ser perfectas; sin embargo, el hecho de utilizar distintos

procedimientos de estimación para el TO (ML, MAP ó EAP) afecta escasamente a su eficiencia frente al TAI. Una diferencia llamativa entre el estudio de simulación y el estudio empírico es el descenso sustancial de la correlación entre las estimaciones TO y TAI en la aplicación empírica (de .9 a .8). Este descenso puede ser explicado: a.) en parte porque en la simulación una de las variables (la q real o del TAI) se mide con perfecta precisión. Por ejemplo, si corregimos por atenuación la correlación (.8) utilizando una estimación de la fiabilidad del TAI a partir del error típico medio de estimación de la habilidad [$r_{xx}=1-(SE/S_q)^2$] obtenemos una correlación de .856 [.8/0.874]; b.) en parte también por la presencia de diferencias entre los procedimientos más allá de los procesos de estimación.

En términos de precisión también existen algunas diferencias. Por ejemplo, en el estudio de simulación las estimaciones bayesianas tienen menor error típico (MAP<EAP<ML). Aunque los errores típicos obtenidos en el estudio empírico (y basados en fórmulas) muestran un ordenamiento diferente en precisión (MAP>ML>EAP). Esto puede deberse en parte a que en la estimación ML se subestima en mayor medida el verdadero error (Warm, 1989; Wang y Vispoel, 1998), especialmente en condiciones en las que los ítems son altamente discriminativos, como es nuestro caso.

Por otro lado, existen también diferencias de nivel entre las estimaciones del TO y las estimaciones del TAI. Esas diferencias no son independientes del método de estimación utilizado. Aparentemente las estimaciones bayesianas del estudio empírico en el TO resultan comparables en mayor medida a las estimaciones del TAI. Un objetivo de este estudio era analizar si esas diferencias de nivel para las estimaciones ML podían ser atribuidas al proceso de estimación. El estudio de simulación muestra las diferencias entre las estimaciones ML del TO y del TAI no pueden explicarse por el diferente sesgo de las distintas estimaciones.

Dada esta ausencia de convergencia entre lo encontrado en el estudio de simulación y los resultados empíricos es difícil evaluar en qué medida los otros métodos de estimación bayesianos (EAPP y MAPP) para el TO pueden resultar más efectivos. Existen 2 variantes en el proceso de estimación bayesiano. La primera variante se refiere a la utilización de la moda (MAP) o la media de la distribución posterior (EAP). Para nuestra muestra, TO-EAP tiende a ser inferior a TO-MAP. Esto es porque la media de la distribución posterior tiende a quedar por debajo de la moda de esa distribución; esto sólo puede ocurrir porque la función de verosimilitud posea asimetría negativa. Bajo otras condiciones los resultados podrían ser distintos. La segunda variante se refiere a la distribución a priori que se fija. En nuestro caso, el efecto de situar una distribución realista aleja (aunque *no* significativamente) las medias obtenidas para el TO de las obtenidas en el TAI (de forma similar para EAP y para MAP). Podemos reforzar la conclusión, a la luz de los resultados obtenidos, de que el uso de parámetros previos "conocidos" en la estimación bayesiana no resulta recomendable. Diversos autores aconsejan la utilización de distribuciones a priori con una desviación amplia, de forma que la influencia de ese conocimiento previo influya únicamente para aquellos sujetos para los que realmente no existe ninguna información. Según Lord (1984), la regresión bayesiana hacia la media es mayor cuanto más homogéneo sea el grupo (lo que se fijará en la desviación típica de la distribución a priori)

y cuanto más corto y menos fiable sea el test. En nuestro caso, es claro que: a.) en las estimaciones TO-MAPP se generan menores errores típicos "teóricos" pero que no se corresponden con mayores correlaciones con las estimaciones TAI; b.) las estimaciones TO-EAPP y TO-MAPP no se aproximan significativamente más a las estimaciones TAI que las estimaciones TO-EAP y TO-MAP; c.) establecer una distribución general es seguramente más justificable como procedimiento en nuestro caso, al establecer que la distribución está estandarizada bajo esos parámetros (0,1) y la consideración de que valores fuera del rango delimitado por ellos son poco probables.

A pesar de estas dificultades, el presente estudio permite ejemplificar el análisis del patrón de sesgo de los distintos procedimientos de estimación para un TO y permiten concluir que en ocasiones el uso de las estimaciones ML puede ser más adecuado para las comparaciones de grupos a los que se aplican distintas formas de un mismo test. Por otro lado, hemos podido descartar dos hipótesis para explicar las diferencias inesperadas entre grupos asignados aleatoriamente a un TAI y un TO como un efecto del procedimiento de estimación en un estudio previo (Olea et al., 2000) y solo queda abierta la posibilidad de que otras características de la situación o del test (p.e., que en un test fijo los ítems pueden hacerse públicos más fácilmente de sesión a sesión) puedan explicar esas diferencias.

Para terminar, señalar que una limitación de este trabajo tiene que ver con el posible sesgo de la estimación obtenida en el TAI. En realidad, el sesgo teórico obtenido para el TAI es sólo orientativo, pues se calcula para las puntuaciones estimadas que de hecho podrían estar sesgadas. Para analizar esa alternativa debería utilizarse un TAI de mayor longitud o utilizar como criterio de parada el error típico de forma que nos aseguráramos que todos los sujetos son medidos con alta y, en el segundo caso, similar precisión.

5.- Referencias.

Baker, F.B. (1992): *Item Response Theory. Parameter estimation techniques.* New York: Marcel Dekker.

Bock, R.D. & Aitkin, M. (1981): Marginal maximum likelihood estimation of item parameters: an application of an EM algorithm. *Psykometrika*, 46, 443-459.

Bock, R.D. & Mislevy, R.J. (1982): Adaptative EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement,* 6, 431-444.

De Ayala, R.J., Schafer, W.D. & Sava-Bolesta, M. (1995): An investigation of the standard errors of expected a posteriori ability estimates. *British Journal of Mathematical and Statistical Psychology, 47,* 385-405.

Hambleton, R.K., Slater, S.C., Narayanan, P. y Setiadi, H. (1996): Construcción automatizada de los tests: conceptos básicos, avances técnicos y aplicaciones. En J. Muñiz (Coor.). *Psicometría*. Madrid: Universitas.

Kim, J.K. & Nicewander, A. (1993). Ability estimation for conventional tests. *Psychometrika, 58 (4),* 587-599.

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika, 48, 2,* 233-245.

Lord, F. M. (1984). *Maximum likelihood and bayesian parameter estimation in item response theory*. Educational Testing Service, Princeton, N.J.

Lord, F.M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement, 23,* 157-162.

Mislevy, R.J. & Bock, R. D. (1990). BILOG 3: *item analysis and test scoring with binary logistic models.* Mooresville, IN: Scientific-Software International.

Muraki, E. & Bock, R. D. (1997). PARSCALE: IRT *item analysis and test scoring for rating-scale data.* Chicago, IL: Scientific-Software International.

Olea, J. & Ponsoda, V. (1996). Tests Adaptativos Informatizados. En J. Muñiz (Coor.). *Psicometría*. Madrid: Universitas.

Olea, J.; Ponsoda, V.; Revuelta, J.; Belchí, J. (1996). Propiedades psicométricas de un test adaptativo de vocabulario inglés. *Estudios de Psicología, 55,* 61-73.

Olea, J., Revuelta, J., Ximénez, C. y Abad, F.J. (2000). Psychometric and psychological effects of review on computerized fixed and adaptive tests. *Psicológica, 21(1-2),* 157-173.

Ponsoda, V.; Olea, J.; Revuelta, J. (1994). ADTEST: A computer-adaptive test based on the maximum information principle. *Educational and Psychological Measurement, 54 (3),* 680-686.

Ponsoda, V.; Wise, S.L.; Olea, J.; Revuelta, J. (1997).An investigation of self-adapted testing in a Spanish high school population. *Educational and Psychological Measurement, 57 (2),* 210-221.

Revuelta, J. y Ponsoda,V. (1997). Una solución a la estimación inicial en los tests adaptativos informatizados, *R.E.M.A.*, 2(2), 1-6.

Revuelta, J.; Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, *35 (4)*, 311-327.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, 17.

Samejima, F. (1993a). An approximation for the bias function of the maximum likelihood estimate of a latent variable for the general case where the item responses are discrete. *Psychometrika, 58*, 119-138.

Samejima, F. (1993b). The bias function of the maximum likelihood estimate of ability for the dichotomous response level. *Psychometrika, 58,* 195-209.

Segall, D.O.; Moreno, K.E.(1999). Development of the Computerized Adaptive Testing version of the Armed Services Vocational Aptitude Battery. En Drasgow, F.; Olson-Buchanan, J. (Eds.) *Innovations in computerized assessment*. Mahwah, NJ, USA: Lawrence Erlbaum Associates.

Stroud, A.H. & Secherst, D. (1966). *Gaussian quadrature formules.* Englewood Cliffs, NJ: Prentice-Hall.

Thissen, D. (1991). *MULTILOG user's guide: multiple, categorical item analysis and test scoring using item response theory*. Chicago, IL: Scientific-Software International.

Vispoel, W.P., Wang, T. & Bleiler, T. (1997): The efficiency, reliability, and concurrent validity of adaptative and fixed-item music listening tests. *Journal of Educational Measurement, 34,* 43-63.

Wang, T. (1997). Essentially unbiased estimates in computerized adaptive testing. *Paper presented at the annual meeting of the AERA, Chicago*.

Wang, T.; Vispoel, W.P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement*, *35(2)*, 109-135.

Warm, A. W. (1989). Weigthed likelihood estimation of ability in item response theory with tests of finite length. *Psychometrika, 54*, 427-450.

Wise, S.L. (1999). Tests autoadaptados informatizados: fundamentos, resultados de investigación e implicaciones para la aplicación práctica. En Olea, J., Ponsoda, V. & Prieto, G. (Eds): *Tests Informatizados. Fundamentos y aplicaciones.* Madrid: Pirámide.