*Metodología de las Ciencias del Comportamiento, Suplemento 2004,* 55-61 

# Reglas de selección de ítems en Tests Adaptativos Informatizados

Juan Ramón Barrada[1], Julio Olea y Vicente Ponsoda
*Universidad Autónoma de Madrid*

**Resumen**

El mecanismo de selección de ítems utilizado en la mayor parte de los Tests Adaptativos Informatizados (TAIs) se basa en el máximo en la función de información de Fisher. Frente a esta regla de selección se han propuesto alternativas. En este estudio se compara el funcionamiento de la regla del máximo de información de Fisher frente al de la función de información de Fisher por intervalo, la función de información de Fisher ponderada por la función de verosimilitud, la función Kullback-Leibler por intervalo y la función Kullback-Leibler ponderada por la función de verosimilitud. Las reglas basadas en la función de verosimilitud proporcionan mejores valores de RMSE. Las ganancias de estas reglas alternativas se producen fundamentalmente en los primeros ítems aplicados y para los niveles de rasgo bajos ($\theta < 0$).

**Abstract**

The commonly used rule for the selection of items in Computerized Adaptive Test employs the maximum in the Fisher's information function. Others rules have been proposed. In this study, we compare the rule of the maximum in the Fisher information function with the Fisher information function by interval, the Fisher information function weighted by the likelihood function, the Kullback-Leibler function by interval and the Kullback-Leibler information weighted by the likelihood function. Results show a better performance of the rules based in the likelihood function with RMSE as dependant measure. The improvements are achieved especially when few items have been presented and, also, for lower trait levels ($\theta < 0$).

Uno de los objetivos que se persiguen al aplicar cualquier test es la evaluación precisa y eficiente de los niveles de rasgo ($\theta$) de los examinados. Una forma de conseguirlo es mediante los denominados Tests Adaptativos Informatizados (TAIs), propuestos inicialmente por Lord (1971) y Owen (1975). A diferencia de los tests clásicos lineales (informatizados o no), en los TAIs no todos los evaluados reciben los mismos ítems. El ítem *(n+1)*-ésimo que se presentará a un examinado dependerá del patrón de respuesta a los *n* items anteriores. En general, puede describirse un TAI (Olea & Ponsoda, 2003; Wainer, 1990) como un proceso iterativo de [estimación del nivel de rasgo ($\hat{\theta}$) – selección del ítem óptimo para $\hat{\theta}$] hasta satisfacer algún criterio fijado con anterioridad (por ejemplo, obtener un error de medida por debajo de un cierto nivel o responder un determinado número de ítems).

Para ambos pasos del ciclo se han propuesto diferentes procedimientos. La mayor parte de los *procedimientos de estimación* de $\theta$ consisten en encontrar el nivel de rasgo para el que es máxima la probabilidad del patrón de respuestas. Los métodos bayesianos (Bock & Mislevy, 1982; Owen, 1975; Samejima, 1969) incluyen supuestos sobre la distribución de probabilidad de la población de examinados, mientras que los máximo-verosímiles (Birnbaum, 1968) no lo hacen. Para una comparación entre ellos, puede consultarse a Wang & Vispoel (1998).

Se han propuesto también varias *reglas de selección de ítems* (RSIs). Entre ellas, la más habitual consistente en seleccionar aquel ítem todavía no presentado que ofrece la máxima información de Fisher[2] (MIF) para $\hat{\theta}$ (Lord, 1977). Según aumenta el valor de la función de información de Fisher (FIF) para $\hat{\theta}$ se reduce el error de medida del nivel de rasgo. Asintóticamente, según se incrementan el número de items presentados, el error de medida de $\theta$ es $I(\hat{\theta})^{-1/2}$ (Bradley & Gart, 1962). Por tanto, la regla MIF parecería, en principio, óptima, puesto que reduce el intervalo de confianza en el que situar $\theta$ dada $\hat{\theta}$. Ahora bien, este procedimiento presenta limitaciones en la medición precisa del nivel de habilidad, que podemos describirlas como sigue:

[1] Dirección para correspondencia: Juan Ramón Barrada. Dpto. de Psicología Social y Metodología. Facultad de Psicología. Universidad Autónoma de Madrid. 28049 Madrid.
Email: juanra.barrada@estudiante.uam.es

[2] Por limitaciones de espacio, se han suprimido algunas de las fórmulas del artículo.

1. La selección de ítems basada en MIF será tanto más inadecuada cuanto mayor sea $|\hat{\theta}_n - \theta|$, ya que se elegirán ítems informativos para niveles de rasgo alejados del verdadero. Dado que $E\left(|\hat{\theta}_{n+1} - \theta|\right) < E\left(|\hat{\theta}_n - \theta|\right)$, la incidencia de este problema se va reduciendo según aumenta el número de ítems presentados (Lord, 1983).
2. Cuando se aplica el modelo logístico de tres parámetros (Birnbaum, 1968) es posible encontrar varios máximos locales para la función de verosimilitud (Samejima, 1977). Sin embargo, MIF sólo evalúa la FIF para un valor puntual $\hat{\theta}$. Esto supone no incorporar en el algoritmo de selección esta posible multiplicidad de máximos. Al igual que en el caso precedente, la posible incidencia de este problema se va reduciendo según aumenta el número de ítems administrados.
3. La información de Fisher puede ser descrita como la capacidad de un ítem para discriminar entre dos puntos adyacentes de nivel de rasgo. Sin embargo, para la correcta estimación de $\theta$, es deseable discriminar también entre niveles distantes, especialmente cuando se presenta un número reducido de ítems o nos encontramos en los primeros ítems de un TAI.

**RSIs alternativas a MIF centradas en la mejora de la precisión**

Con el fin de superar estos inconvenientes, en los últimos años se han propuestas otras RSIs alternativas, todas las cuales pueden describirse como casos particulares de una regla general (Veerkamp & Berger, 1997):

$$\max_{i \in B_n} \int_{\theta_{\min}}^{\theta_{\max}} V_i(\theta) W(\theta, x, g)\, d(\theta) \tag{1}$$

El *n*-ésimo ítem seleccionado será aquel que, de entre los todavía no presentados ($B_n$), ofrezca el valor máximo para la integral criterio. *W* es la función de ponderación, que viene condicionada al vector de ítems anteriormente presentados (*x*), al acierto o error en cada ítem (*g*) y al valor de $\theta$, y $V(\theta)$ es la función de valoración. Así, por ejemplo, para el caso de MIF:

$$W(\theta) = \begin{cases} 1, & \theta = \hat{\theta} \\ 0, & \theta \neq \hat{\theta} \end{cases} \tag{2}$$

$$V(\theta) = I(\theta) \tag{3}$$

MIF evalúa la función de información de Fisher para un único punto, $\hat{\theta}$. Sin embargo, en los primeros ítems de un TAI, parece conveniente considerar la información de niveles de rasgo relativamente separados de $\hat{\theta}$. Así, las reglas alternativas obtendrán la información para un intervalo de valores de $\theta$, con diferentes funciones de valoración (FIF y función Kullback-Leibler – KL, en este estudio) y distintos criterios de ponderación (función de verosimilitud e intervalo, en este trabajo). Esto da lugar a cuatro RSIs alternativas, que son las que vamos a evaluar a lo largo del trabajo, si bien no son las únicas propuestas (por ejemplo, Van der Linden -1998- expone otras RSIs que ofrecen resultados prometedores). Una descripción de las mismas, a modo de resumen, puede encontrarse en la tabla 1.

*RSIs basadas en la función de información de Fisher:*

Estas RSIs emplean como base la misma función de información que se utiliza en MIF. La diferencia radica en la función de ponderación, que no se considera únicamente $I(\hat{\theta})$. Se han propuesto dos criterios para *W*. El primero es la función de verosimilitud (FV), que daría lugar a la función de información de Fisher por función de verosimilitud – IF*FV (Veerkamp & Berger, 1997). La segunda propuesta la denominaremos función de información de Fisher por intervalo – IF*I (Veerkamp & Berger, 1997). El intervalo es el intervalo de confianza de θ dada $I_n(\hat{\theta})$.

*Tabla 1. Descripción de las distintas RSIs para la selección del ítem n+1-ésimo. ($\Phi$ es la distribución normal estándar acumulativa y el resto de símbolos corresponden a los descrito en el texto).*

| | IF*FV | KL*FV | IF*I | KL*I |
|---|---|---|---|---|
| $V(\theta_j)$ | $I(\theta)$ | $KL(\theta \| \hat{\theta})$ | $I(\theta)$ | $KL(\theta \| \hat{\theta})$ |
| $W_n(x_i, g_i, \theta_j)$ | $FV(\theta_j, x, g)$ | $FV(\theta_j, x, g)$ | 1 | 1 |
| $\theta_{\min}$ | -4 | -4 | $\min\left(\hat{\theta}-3, \hat{\theta}-\frac{\Phi^{-1}(.975)}{\sqrt{I(\hat{\theta})}}\right)$ | $\hat{\theta}-\frac{\Phi^{-1}(.975)}{\sqrt{n-1}}$ |
| $\theta_{\max}$ | 4 | 4 | $\min\left(\hat{\theta}+3, \hat{\theta}+\frac{\Phi^{-1}(.975)}{\sqrt{I(\hat{\theta})}}\right)$ | $\hat{\theta}+\frac{\Phi^{-1}(.975)}{\sqrt{n-1}}$ |

*RSIs basadas en la función Kullback- Leibler :*
La función de valoración adoptada para estas RSIs es la función KL, que pretende solventar el tercer problema que hemos descrito, ya que permite conocer la capacidad de discriminar entre cualquier par de niveles de rasgo. Para una descripción más detallada acerca de la función KL aplicada a los TAIs puede consultarse Chang & Ying (1996) y Eggen (1999).
Los criterios de ponderación que se han sugerido son los mismos que para las RSIs basadas en FIF, lo cual da origen a dos nuevas RSIs: función KL por función de verosimilitud -KL*FV- y función KL por intervalo -KL*I- (Chang & Ying, 1996).

De todos los estudios realizados hasta el momento sobre el tema (Chang & Ying, 1996; Chen, Ankenmann & Chang, 2000; Cheng & Liou, 2000; Veerkamp & Berger, 1997), únicamente el de Chen *et al.* (2000) incluye todas estas RSIs. Estos autores encuentran resultados similares a los que parcialmente ya ofrecían los trabajos previos: (a) ligeras ventajas de las RSIs alternativas frente a MIF, que se van reduciendo según se incrementan el número de ítems presentados; (b) las mejoras tienden a situarse, especialmente, en niveles bajos de $\theta$. Recientemente, se ha comenzado a poner a prueba el funcionamiento de algunas de estas RSIs para ítems politómicos, sin que parezcan ofrecer, por el momento, mejoras frente a MIF (van Rijn, Eggen, Hemker & Sanders, 2002).

**Objetivo del presente estudio**
Este estudio fue desarrollado para orientar la decisión sobre la RSI más conveniente para un TAI de evaluación del conocimiento del inglés escrito (Olea, Abad, Ponsoda & Ximénez, enviado). Pese a los resultados convergentes de los trabajos previos, las especificidades de este TAI (relativamente escaso número de ítems, estimación máximo-verosímil, distribución de parámetros) hace aconsejable evaluar los efectos de cada una de las RSIs.

## Estudios de simulación

*Método*
Cinco fueron las RSIs que se compararon: MIF, IF*FV, IF*I, KL*FV, KL*I. Se realizaron dos estudios, el primero para obtener un valor global de precisión de estas RSIs, el segundo para establecer en qué niveles de rasgo resultaba más eficaz cada una de ellas..

*Banco de ítems:*
Se empleó un banco de ítems actualmente en funcionamiento como base de un TAI aplicado a través de Internet (Olea *et al.*, enviado), formado por 197 ítems de opción múltiple, con 4 alternativas de respuesta, ideado para evaluar el conocimiento del inglés escrito. La distribución de sus parámetros es la siguiente: $a \sim N(1.3, 0.32)$; $b \sim N(0.23, 1)$; $c \sim N(0.21, 0.03)$.

*Arranque:*
La simulación comenzaba con una nivel de rasgo inicial elegido al azar dentro del intervalo (-0.5, 0.5). Puesto que para aplicar RSIs distintas a MIF se requiere un mínimo de un ítem presentado, el primer ítem siempre era seleccionado mediante MIF.

*Estimación/asignación del nivel de rasgo:*
Como es conocido, la estimación máximo-verosímil no proporciona valores reales cuando el patrón de respuestas es constante (sólo aciertos o sólo errores). Por ello, hasta el momento en el que había un mínimo de un acierto y un error, no se estimaba $\theta$ sino que era asignada mediante el método propuesto por Dodd (1990). En el caso de que todo sean aciertos, $\hat{\theta}$ se incrementa en ($b_{max}$ – $\hat{\theta}$)/2; de ser todo errores, $\hat{\theta}$ se reduce en ($\hat{\theta}$ – $b_{min}$)/2. Desde el momento que se produce variabilidad en las respuestas se aplica la estimación máximo-verosímil.

*Longitud del test:*
En cada simulación se aplicaron un total de 20 ítems. Los datos para su posterior análisis fueron recogidos cada cinco ítems. Resultados anteriores han mostrado que, para esta longitud total del test, la diferencia en precisión entre RSIs es ya muy reducida.

*Criterios de evaluación:*
Se empleo como variable dependiente la RMSE.

$$RMSE = \left( \sum_{i=1}^{r} \left( \hat{\theta}_i - \theta_i \right)^2 \Big/ r \right)^{(1/2)} \tag{11}$$

donde *r* es el número de replicas-evaluados.

*Estudio 1:*
Mediante este estudio, se evaluó, de un modo general, el funcionamiento de las distintas RSIs. Para ello, se generaron aleatoriamente 3000 niveles de rasgo, extraídos de una población N(0, 1).

*Estudio 2:*
En este estudio, se procedió a un análisis del funcionamiento de las distintas RSIs condicionado a los niveles de rasgo. Para ello, se simularon 450 examinados para siete niveles distintos de $\theta$, desde $\theta$ = -3 a $\theta$ = 3, con incrementos de +1.

## Resultados y discusión

En la figura 1 se muestran los resultados de los cinco RSIs para el estudio 1. Tal y como era de esperar, el RMSE, para todas las RSIs, se va reduciendo según aumenta el numero de ítems presentados. Dependiendo del RMSE medio que deseemos en una aplicación concreta del TAI, así habremos de establecer la longitud del test. Con 20 items administrados, por ejemplo, conseguimos un RMSE medio en torno a .30.

Al igual que en estudios previos, las RSIs basadas en la función de verosimilitud presentan en general mejoras en la precisión en comparación con MIF, si bien estas diferencias son escasas y se reducen progresivamente a medida que se aumenta el numero de ítems presentados. Los resultados ofrecidos por las RSIs basadas en evaluación por intervalo son peores a los de MIF. Especialmente pobre es el desempeño conseguido con IF*I.

En la figura 2 se presentan los resultados para los siete niveles de rasgo evaluados. Vemos cómo las medidas mas precisas, para cualquier número de ítems presentados, se corresponden con los niveles de rasgo medios-altos y altos ($\theta \cdot 0$). La precisión para todos los niveles de rasgo se incrementa según aumenta la longitud del test.

Podemos ver cómo las diferencias entre las distintas RSIs no se distribuyen igualmente para todos los niveles de rasgo. Estas diferencias se producen principalmente en los niveles bajos ($\theta < 0$), reproduciendo el patrón encontrado para el funcionamiento global: las reglas basadas en FV ofrecen mejores resultados, mientras que las que usan intervalos tienen una peor precisión. Las diferencias son mayores que las encontradas para el estudio 1, si bien debe considerarse que se producen en niveles de rasgo infrecuentes para la mayoría de los objetivos de evaluación del TAI.

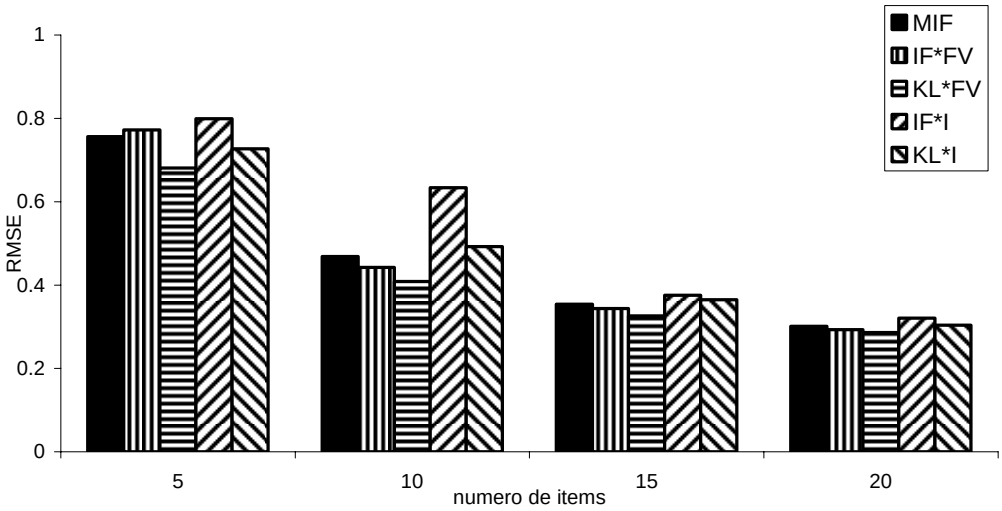


*Figura 1. RMSE de las distintas RSIs según número de items administrados.*

Encontramos, pues, datos concordantes y dispares con lo previamente publicado (Chang & Ying, 1996; Chen *et al*., 2000; Cheng & Liou 2000;. Veerkamp & Berger, 1997). Como similar: (a) la convergencia entre las distintas RSIs según se incrementan el número de ítems administrados; (b) la superioridad en cuanto a precisión de las RSIs basadas en FV; (c) los diferencias en precisión se encuentran en niveles bajos de rasgo. Una diferencia notable es el pobre funcionamiento de las reglas basadas en intervalos, especialmente IF*I. Esto apoya la necesidad de realizar simulaciones específicas para cada banco de ítems, puesto que algunos resultados publicados (y no ampliamente replicados) pueden deberse a características propias de los bancos de ítems. Parece que las reglas basadas en FV son más robustas que las fundamentadas en intervalos, con mejores resultados en una variedad de condiciones de aplicación, mientras que las de intervalo pueden depender de la presencia de algunas condiciones específicas para conseguir una precisión superior a MIF. A modo tentativo, creemos que pueden ser especialmente las distribuciones de los parámetros y las covarianzas entre los mismos.

Aunque las RSIs que emplean la FV como criterio de ponderación ofrecen mejoras en la precisión, esto no las convierte siempre en las más aconsejables. Estas RSIs han de ser evaluadas, igualmente, con respecto a otros criterios de interés en el contexto de los TAIs. Uno es el coste computacional que supone la aplicación de cada una de estas RSIs. Las RSIs basadas en FV requieren, tal y como fueron programas, en torno a unas diez veces mas tiempo que MIF en la selección de ítems. Según la condiciones de aplicación de la prueba, esto puede ser algo de mayor o menor relevancia. Por otro lado, en otros trabajos (Barrada, Olea y Ponsoda, 2003) se ha mostrado que las RSIs alternativas a MIF empleadas en este estudio ofrecen un peor control de la exposición, con lo cual se pone en tela de juicio la seguridad del banco (Stocking & Lewis, 2000; Way, 1998).

La RSI a emplear para un TAI concreto dependerá de: (a) el funcionamiento de las RSIs; (b) las características del banco; (c) las características y objetivos de la evaluación. De este modo, resulta complicado responder de un modo categórico a la pregunta sobre qué RSI es la que conviene aplicar. En general, en aquellos contextos en los que el evaluador sólo haya de tener en cuenta la maximización de la precisión de las estimaciones de rasgo, parece adecuado emplear RSIs basadas en FV, especialmente KL*FV.

## Referencias

Barrada, J. R., Olea, J., & Ponsoda, V. (2003) *Comparación en la tasa de exposición entre diferentes mecanismos de selección de ítems.* Comunicación presentada en el VIII Congreso de Metodología de las Ciencias Sociales y de la Salud, Valencia, España.

Birnbaum, A. (1968). Some latent ability models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.) *Statistical theories of mental test scores.* (pp. 392-479) Reading, MA: Addison-Wesley.

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

Bradley, R. A., & Gart, J. J. (1962). The asymptotic properties of ML estimators when sampling from associated populations. *Biometrika, 4*, 205-214.

Dodd, B. G. (1990) The effect of item selection procedure and stepsize on computerized adaptive attitude measurement using the rating scale model. *Applied Psychological Measurement, 14*, 355-366.
Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*, 213-229.
Chen, S.-Y., Ankenmann, R. D., & Chang, H.-H. (2000). A comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement, 24*, 241-255.
Cheng, P. E., & Liou, M. (2000). Estimation of trait level in computerized adaptive testing. *Applied Psychological Measurement, 24*, 257-265.
Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23*, 249-261.
Lord, F. M. (1971). Robbins-Monro procedures for tailored testing. *Educational and Psychological Measurement, 31*, 3-31.
Lord, F. M. (1977). A broad-range tailored test of verbal ability. *Applied Psychological Measurement, 1*, 95-100.
Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and their parallel-form reliability. *Psychometrika, 48*, 233-245.
Olea, J., Abad, F.J., Ponsoda, V, & Ximénez, M.C. (enviado). Un test adaptativo informatizado para evaluar el conocimiento del inglés escrito: Diseño y comprobaciones psicométricas.
Olea, J., & Ponsoda, V. (2003). Tests adaptativos informatizados. Madrid: UNED.
Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70*, 351-356.
Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, 34*, 100.
Samejima, F. (1977). A use of the information function in tailored testing. *Applied Psychological Measurement, 1*, 233-247.
Stocking, M. L., & Lewis, C. L. (2000). Methods of controlling the exposure of items in CAT. In W. J. van der Linden & C. A. W. Glas (Eds.) *Computerized adaptive testing: Theory and practice* (pp. 163-182). Dordrecht, The Netherlands: Kluwer Academic.
van der Linden, W. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika, 63*, 201-216.
van Rijn, P. W., Eggen, T. J. H. M., Hemker, B. T., & Sanders, P. F. (2002). Evaluation of selection procedures for computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 26*, 393-411.
Veerkamp, W. J. J., & Berger, M. P. F. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational & Behavioral Statistics, 22*, 203-226.
Wainer, H. (Ed.) (1990). *Computerized adaptive testing: Primer.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
Wang, T., Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement, 35*, 109-135.
Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice, 17*, 17-27.

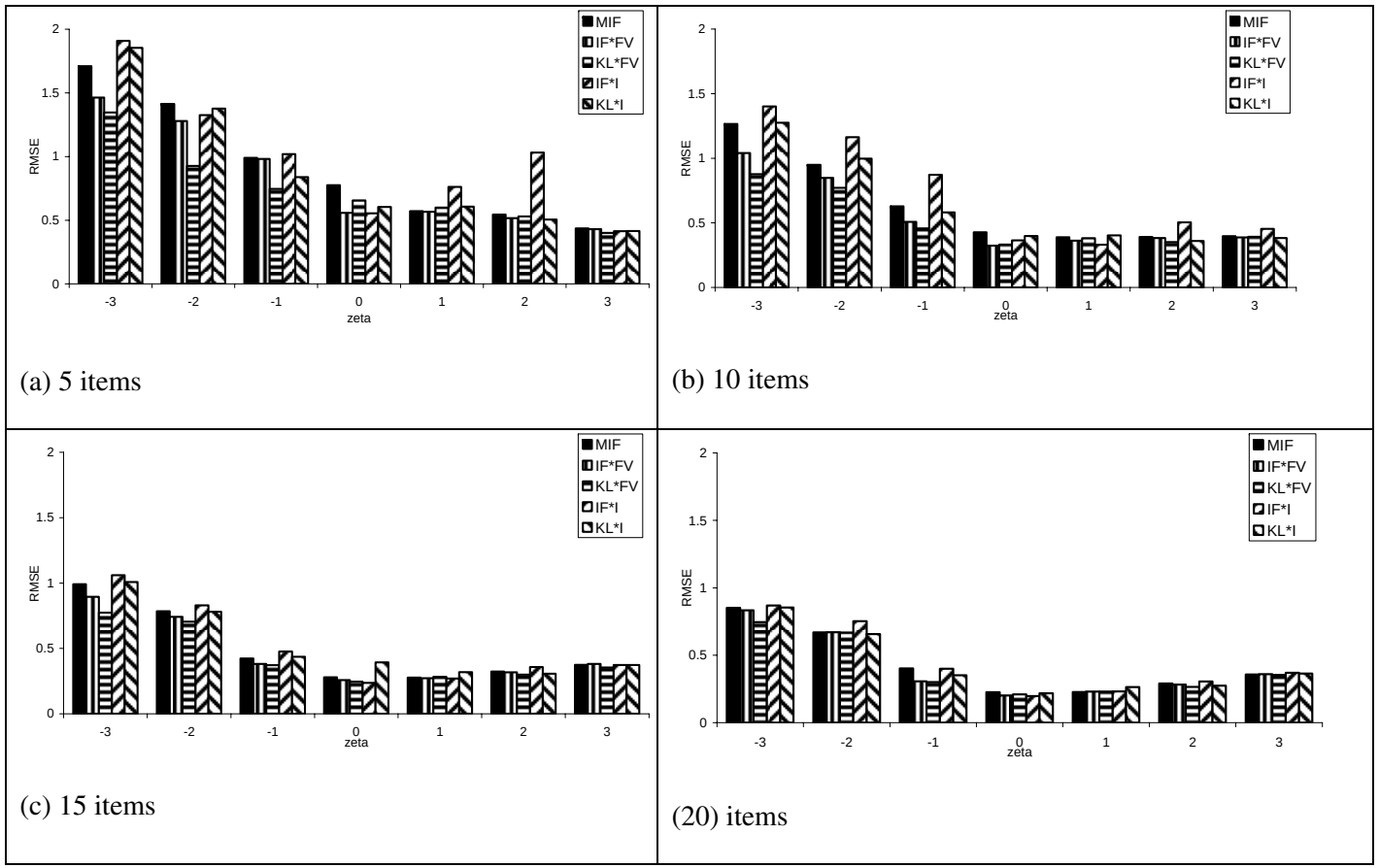


*Figura 2. RMSE de las distintas RSIs según nivel de rasgo y número de items administrados.*