

Didáctica

TEST ADAPTATIVOS INFORMATIZADOS

Fabio Alexander Sierra–Matamoros, Juliet Rocío Valdelamar–Jiménez,
Félix Adrián Hernández–Tamayo & Luz Miriam Sarmiento–García
Universidad Nacional de Colombia, Colombia*

Un Test Adaptativo Informatizado (TAI) se define como un instrumento conformado por un banco de ítems, calibrado según principios de la Teoría de Respuesta al Ítem (TRI), que implica un procedimiento para la estimación del nivel de habilidad del examinado, y otro para seleccionar el ítem más adecuado de acuerdo con dicho nivel (Olea & Ponsoda, 1996); y cuya elaboración, aplicación, calificación y actualización se realiza por medio de un soporte informático (Hambleton, Zaal, & Pieters, 1991). El supuesto que soporta a los TAI es la construcción de pruebas hechas a la medida, es decir, la posibilidad de presentar ítems al examinado de acuerdo con el desempeño que ha mostrado en los ítems precedentes. En otras palabras, las pruebas se configuran a partir de la respuesta que el examinado da a un ítem y su correspondencia con un nivel estimado de habilidad que permite la escogencia y presentación de uno de los ítems restantes del banco (Olea & Ponsoda, 1996). Este procedimiento implica, por una parte, la posibilidad de presentar distintos ítems y pruebas a diferentes examinados, pues se supone que poseen diferentes niveles de habilidad y, por otra parte, el uso de pocos ítems, ya que se aplican aquellos que más informan acerca del nivel de habilidad (Muñiz, 1997).

Aunque las pruebas adaptativas podrían remontarse a Binet (Gregory, 2001) y al uso de métodos psicofísicos clásicos (Muñiz, 1997), se considera a Fred Lord como pionero en el establecimiento de procedimientos adaptativos basados en la TRI (Olea & Ponsoda, 1996). Los primeros intentos de aplicación de TAIs se realizaron a partir de los años setenta gracias a los desarrollos de modelos teóricos de la TRI y de la informática, que se conjugaron con necesidades de índole aplicada en el campo de la evaluación psicológica. Así, el primer sistema integrado para aplicar bancos de ítems de forma adaptativa fue editado por la *Assesment Systems Corporation* en 1984 (Olea & Ponsoda, 1996). Los primeros TAIs elaborados se usaron en la milicia y fueron desarrollados por ASVAB en los Estados Unidos y MicroPAT en Europa. En la actualidad, el uso de TAIs se ha extendido a varios test estandarizados como el *Test of English as a Foreign Language* (TOEFL) del ETS y a distintos campos de evaluación como aptitudes, personalidad y conocimientos.

Cuando la selección y presentación de ítems se realiza de acuerdo con los principios de la Teoría de Respuesta al Ítem (TRI) resulta necesaria la utilización de un soporte informático (Olea & Ponsoda, 1996). Dicho soporte permite llevar a cabo los procedimientos matemáticos y estadísticos requeridos, acceder al banco y presentar el ítem de forma inmediata. El uso del soporte informático establece, además, un cambio en la aplicación del test, puesto que el examinado recibe el test, lo responde a través de un computador, obtiene su calificación de forma inmediata y sólo responde los ítems más informativos (Hambleton et al., 1991). El software que se utiliza en los TAIs se caracteriza por constar de una serie de módulos que procesan diferentes

* Trabajo realizado en el marco de la asignatura “Psicometría”, dictada por la docente Olga Rosalba Rodríguez en el II semestre de 2005. Correos de los autores: fasierram@unal.edu.co, jrvaldelamarj@unal.edu.co, fahernandezt@unal.edu.co, lmsarmientog@unal.edu.co

pasos de la prueba de forma independiente y se encuentran en una relación jerárquica. Así, se encuentran módulos que permiten la construcción de ítems, la presentación de los mismos, la finalización de la prueba, el cálculo del nivel de habilidad, el almacenamiento de los resultados, la baremación, la actualización del banco de ítems y la presentación de un informe escrito al examinado sobre su desempeño (Olea & Ponsoda, 1996). Entre el software utilizado se encuentra el MicroCAT de la Assessment System Corporation (Hambleton et al., 1991; Olea & Ponsoda, 1996; Muñiz, 1997), el DEMOTAC y APT-System, elaborado en España (Olea & Ponsoda, 1996).

Estructura de los TAIs

Un TAI se compone de un banco de ítems calibrado según alguno de los procedimientos derivados de la TRI, un procedimiento de estimación del nivel de habilidad y un procedimiento de selección sucesiva de ítems (Olea & Ponsoda, 1996). La calibración de los ítems bajo uno de los modelos de la TRI, tiene como propósito la estimación de los parámetros de los ítems, teniendo en cuenta la invarianza de la medida respecto del instrumento y del grupo de examinados. Dado que la TRI intenta buscar mediciones invariantes respecto del instrumento y de los examinados, su empleo en la construcción de los TAIs es fundamental, por lo que a continuación se expone de forma sucinta los supuestos que la fundamentan: *Curva característica del ítem, unidimensionalidad e independencia local*.

Según la TRI, la probabilidad de acertar un ítem depende de los valores en la variable medida por el ítem, es decir, del nivel de habilidad del sujeto (θ). La relación entre esa probabilidad $p(\theta)$ y el nivel de habilidad (θ) se puede explicar mediante una función matemática denominada *Curva Característica del Ítem* (CCI). La CCI posee tres propiedades a las que se denomina *parámetros*: Índice de discriminación (parámetro a); índice de dificultad (parámetro b); y probabilidad de acertar al azar cuando no hay nivel de habilidad (parámetro c) (Muñiz, 1997; Hambleton & Swaminathan, 1985; Lord, 1980). El nivel de habilidad (θ) puede asumir valores entre $-\infty$ y $+\infty$, aunque en las estimaciones hechas en los TAIs se emplea un rango de -3 a 3 , en el que 0 se refiere a un nivel de habilidad medio, -3 a un nivel de habilidad mínimo (la probabilidad de acertar el ítem con este nivel se acerca al azar), y 3 , un nivel de habilidad alto. La curva puede explicarse por medio de una de dos tipos de funciones matemáticas: función logística y curva normal acumulada. Cada función puede incluir uno, dos o tres parámetros. En los TAIs suele utilizarse el modelo logístico de tres parámetros (Muñiz, 1997), por lo que se recomienda que el banco a usar tenga ítems con un parámetro a elevado (Hambleton et al., 1991), una distribución uniforme de frecuencias que caracterice al parámetro b y una probabilidad de acierto aleatorio en el parámetro c (Olea & Ponsoda, 1996).

El supuesto de *unidimensionalidad* implica que los ítems constituyen una sola dimensión, es decir, miden un mismo y único rasgo, de lo cual se deduce que la probabilidad de acertar un ítem depende únicamente de un sólo factor, el nivel de habilidad (θ). El supuesto de *independencia local* entre los ítems se deriva del anterior e implica que la respuesta que un sujeto da a un ítem no viene influida por sus respuestas a otros, es decir, la probabilidad de acertar un número de ítems es igual al producto de las probabilidades de acertar cada uno de ellos (Muñiz, 1997). Estos dos supuestos resultan imprescindibles en un TAI, pues al presentar un número reducido de ítems, la estimación final de (θ) no puede depender de que los ítems se refieran a uno u otro rasgo (Olea & Ponsoda, 1996).

En relación con el procedimiento de estimación del nivel de habilidad, en los TAIs se tienen en cuenta las respuestas que el examinado da a los ítems presentados y se lleva a cabo utilizando las características psicométricas de los ítems, principalmente el parámetro b , o éstas unidas con valores de habilidad previos a través de algoritmos matemáticos. La forma en la que dicha

estimación se realiza varía de acuerdo con el tipo de estrategias de funcionamiento de un TAI. El valor numérico (estimación de θ) que se obtiene está comprendido entre el rango de -3 a 3. En la medida en que se presentan más ítems, el nivel de habilidad estimado se acerca de forma más precisa al nivel de habilidad real, lo que quiere decir que el error de medición disminuye (Olea & Ponsoda, 1996).

El procedimiento de selección sucesiva de ítems consiste en la aplicación de estrategias para seleccionar un ítem aún no presentado, para lo cual se tienen en cuenta las características psicométricas del ítem o el nivel de habilidad estimado. Así, se selecciona entre los ítems restantes del banco aquél que resulte más informativo para un cierto nivel de habilidad. La selección varía de acuerdo con la estrategia de funcionamiento empleada (Olea & Ponsoda, 1996).

Funcionamiento del TAI

Para empezar a responder un TAI se presenta un primer ítem al examinado, y de acuerdo con su respuesta se le presenta un segundo ítem; la respuesta al segundo ítem se usa para seleccionar un tercer ítem y así sucesivamente. Cuando ha logrado un criterio establecido de antemano (cierta cantidad de ítems o nivel de precisión), se finaliza la presentación de ítems. Se plantean aquí las estrategias empleadas en tres momentos de ejecución del TAI: *Arranque, selección y presentación sucesiva de ítems, y parada*.

Estrategias de arranque: Se debe establecer un criterio de acuerdo con el cual se presenta el primer ítem. Hay varias posibilidades para establecer dicho criterio: (a) Conocer el nivel de habilidad del examinado, a partir de información previa; (b) asignar al examinado un nivel de habilidad; o (c) asignar un ítem de acuerdo con un valor determinado en el parámetro b . En el primer caso se puede considerar el rendimiento del examinado en pruebas similares, su desempeño académico en un área determinada si el TAI mide esa variable ó se puede aplicar una prueba corta con un número fijo de ítems para hacer una estimación de habilidad inicial (Hambleton et al, 1991; Olea & Ponsoda, 1996). En el segundo caso se puede asumir que el nivel de habilidad del examinado se encuentra en un rango medio (por ejemplo entre -1 y +1), y de acuerdo con ese rango se selecciona un ítem apropiado (Olea & Ponsoda, 1996). En el tercer caso se presenta un ítem de dificultad media (Hambleton et al, 1991; Olea & Ponsoda, 1996).

Estrategias de selección y presentación sucesiva de ítems: Después de iniciar el test con la presentación de un primer ítem y una primera respuesta del examinado, se requiere una estrategia para seleccionar los ítems restantes de manera sucesiva. Hay dos formas generales de hacer dicha selección y presentación: (a) Estrategias de dos niveles; y (b) estrategias de múltiples niveles. La estrategia de dos niveles implica la presentación de un primer test, denominado rutina, con un pequeño número de ítems preestablecidos de dificultad heterogénea, que responden todos los examinados. De acuerdo con el desempeño en este test se hace una primera estimación de su nivel de habilidad y se les asigna a uno de varios test de clasificación, conformados por ítems de dificultad homogénea; estos test de clasificación difieren entre sí por el parámetro b de sus ítems (Hambleton et al, 1991; Olea & Ponsoda, 1996). La aplicación del test de clasificación garantiza una mayor precisión del nivel de habilidad (Muñiz, 1997), aunque en dicha estimación suelen usarse ambos test (Hambleton et al, 1991).

La estrategia de múltiples niveles inicia con la presentación de un sólo ítem, y de acuerdo con la respuesta dada al mismo, se presenta un segundo ítem adecuado cuya respuesta lleva a un tercer ítem, y así sucesivamente. Esta segunda estrategia plantea una especie de ramificación o vías de conexión de los ítems, que puede estar establecida de antemano o que puede irse creando; la primera forma se denomina *ramificación fija* y la *segunda ramificación variable*.

En las estrategias de *ramificación fija*, la respuesta dada a cada ítem lleva necesariamente a uno de dos ítems consecutivos, que puede presentar mayor o menor dificultad, según la estimación del parámetro b . La ramificación fija posee a su vez varias posibilidades: (a) test de forma piramidal; (b) test de nivel flexible; y (c) test estratificado adaptativo. El primero consiste en la forma básica de ramificación fija: Una respuesta correcta lleva a un ítem más difícil ya establecido con anterioridad, y una respuesta incorrecta conduce a un ítem más fácil, establecido de antemano. El test de nivel flexible consiste en la selección de un ítem restante del banco, cuyo grado de dificultad es el siguiente mayor al del ítem presentado si se respondió bien o el siguiente menor si se respondió mal; en este caso no hay un ítem preestablecido para respuestas correctas o incorrectas, sino que existen varios ítems para escoger si tienen un mismo parámetro b . El test estratificado adaptativo, en lugar de presentar un ítem, presenta un grupo de ítems con determinada dificultad, que pueden ser de mayor o menor dificultad de acuerdo con el desempeño del examinado en el grupo de ítems anterior que era más fácil o más difícil (Hambleton et al, 1991; Olea & Ponsoda, 1996).

Las estrategias de *ramificación variable* permiten una mejor aproximación a la idea del test adaptativo, pues en cada momento de la prueba se selecciona un ítem adecuado no sólo de acuerdo con su parámetro b sino con su parámetro a , y con el nivel de habilidad previo estimado. Por lo tanto la estrategia implica realizar una estimación del nivel de habilidad provisional del examinado después de su respuesta a cada ítem y seleccionar el ítem más adecuado para presentarlo inmediatamente. La estimación se hace usando uno de dos procedimientos estadísticos: Estimación condicional por el método de máxima verosimilitud, o procedimientos de selección bayesianos. La selección del ítem más adecuado se lleva a cabo por uno de los siguientes métodos: Método de máxima información, o el método de máxima precisión esperada (Hambleton et al, 1991; Olea & Ponsoda, 1996).

Estrategias de parada: La presentación de ítems se puede terminar cuando: (a) se ha alcanzado un número determinado de ítems; (b) se ha logrado un valor de precisión determinado; ó (c) se ha alcanzado un criterio de ubicación dentro de un grupo. La medida de precisión usada generalmente es el error de medición. En el primer caso se obtiene una medida de precisión específica para cada examinado. En el segundo caso, que es el más deseable, se obtiene una medida de precisión única para todos los examinados aunque el número de ítems presentado no sea el mismo para todos. En el tercer caso, propio de test calificados con referencia a criterio, la presentación de ítems finaliza cuando el examinado ha sido asignado a uno de los grupos (grupo con habilidad o grupo sin habilidad) con un nivel de confianza preestablecido (por ejemplo 95%) (Hambleton et al, 1991; Olea & Ponsoda, 1996).

Desarrollos investigativos de los TAI

Los TAIs se han convertido en una herramienta útil para la estimación de la habilidad de los sujetos, por lo que la investigación en este campo ha crecido para el mejoramiento de su estructura y de las condiciones de aplicación. Una de las líneas de investigación está enfocada a los efectos de la ansiedad sobre el desempeño del examinado. En un estudio que compara los TAIs de alto y bajo nivel de dificultad con los TAAIs, se ha encontrado que hay un mejor desempeño en estos últimos al disminuir los niveles de ansiedad, aunque también la longitud del test jugó un papel importante, pues entre más corto sea el test, la ansiedad es menor (Rocklin & O'Donnell, 1987). El hecho de que el examinado pueda tener control sobre el nivel de dificultad de los ítems tiene mejores resultados que cuando se presentan en un mismo orden (Rocklin, O'Donnell & Holst, 1995).

Siguiendo la revisión que realiza Ponsoda, Hontangas, Olea, Revuelta, Abad, & Ximénez, (2004) sobre la investigación de los TAIs, se han propuesto nuevos métodos para controlar la

exposición de los ítems, tales como: Método de información específica, de Davey & Fan (2000); el método de “el vecino más próximo” (Nearest – Neighbours method), de Chen & Liou (2003); y el método de composición de mini bancos de ítems a elegir, de Stocking & Swanson. (1998)

Las investigaciones sobre la elaboración, el mantenimiento y la renovación de los bancos de ítems se han enfocado, por un lado, a la estimación de parámetros de los ítems, ya sea a partir de los índices psicométricos de la TCT, la reestimación de los ítems (ítems operativos) y la estimación de nuevos ítems (ítems pretest) y, por otro lado, a la generación automática de ítems, los cuales pueden ser isomorfos, es decir, que posean contenido y propiedades psicométricas similares. Así mismo, se han propuesto TAIs de razonamiento cuantitativo, por Bejar, Lawless, Morley, Wagner, Bennett & Revuelta (2002, citados por Ponsoda et al., 2004) y TAIs basados en un modelo de respuesta multinivel, por Glass & van der Linden (2003, citados por Ponsoda et al., 2004).

También se han investigado las propiedades de los TAIs por medio de procedimientos de simulación (Chang & Ying, 2004; Olea & Ponsoda, 1996; Abad, Olea, Ponsoda, Ximénez & Mazuela, 2004). Así, por ejemplo, en el estudio de Abad et al (2004) se encontró que si un banco de ítems calibrado trata las omisiones como respuestas fraccionalmente correctas produce valores estimados de habilidad más parecidos a la habilidad verdadera del sujeto, a diferencia de cuando se tratan las omisiones como errores, que conducen a la sobreestimación de los niveles de habilidad.

Olea & Ponsoda (1996) sugieren que la investigación sobre TAIs se llevará a cabo en torno a tres temas principales: Condiciones de aplicación, implementación de nuevos modelos de TRI y extensión de los contenidos a evaluar. Estas temáticas se refieren a asuntos como el uso de bancos cortos o la posibilidad de revisar o diferir las respuestas, superar la restricción de unidimensionalidad y evaluar procesos psicológicos básicos (memoria, percepción, atención) respectivamente.

Ventajas y limitaciones

Un TAI manifiesta ciertas ventajas frente a los test clásicos, algunas de ellas son ventajas de la TRI sobre la Teoría Clásica de Test (TCT), y otras se derivan de la estructura y funcionamiento del TAI. En primer lugar, los TAIs son precisos y eficientes (Gregory, 2001). La precisión de la medida se puede establecer para cada examinado (Olea & Ponsoda, 1996). La eficiencia implica la reducción del tiempo de aplicación así como el número de preguntas. El tiempo puede reducirse hasta un 50% del empleado en un test clásico (Gregory, 2001; Hambleton et al, 1991); dicha reducción tiene que ver con el empleo de pocos ítems, esos que resultan más informativos del nivel de habilidad (Chang & Ying, 2004). El uso de estos pocos ítems tiene algunos beneficios colaterales tales como la disminución de los efectos de fatiga y el aumento de la motivación, ya que los examinados responderán ítems adaptados a su nivel, es decir, no muy fáciles ni muy difíciles (Hambleton et al, 1991; Muñiz, 1997). Los TAIs ofrecen también un alto grado de seguridad para el banco de ítems, ya que no se van a presentar los mismos ítems a los diferentes examinados (Olea & Ponsoda, 1996).

El uso del PC ofrece la posibilidad de presentar ítems complejos tales como mediciones de tiempos de respuesta, sonido, simulación, conducta interactiva (Muñiz, 1997). Otra ventaja práctica es la agilidad en la calificación, gracias a la utilización de un software especializado; dicha calificación no tiene en cuenta el número de aciertos y errores sino los valores de los distintos parámetros de los ítems de acuerdo con el modelo de TRI y la estrategia de selección empleados. Sin embargo la aplicación de los TAIs puede resultar más costosa que la aplicación de los test clásicos de lápiz y de papel debido al requerimiento de equipos y de software

especializado (Muñiz, 1997). Además los examinados pueden experimentar cierto nivel de ansiedad al responder el TAI, bien sea porque no sienten control sobre el proceso o porque no pueden realizar ciertas correcciones en respuestas de las que no están seguros; ante lo cual se pueden emplear tests autoadaptativos informatizados (TAAIs), en los que el examinado va determinando su nivel de dificultad para reducir la ansiedad (Olea & Ponsoda, 1996).

Referencias

- Abad, F., Olea, J., Ponsoda, V., Ximénez, C. & Mazuela, P. (2004). Efecto de las omisiones en la calibración de un test adaptativo informatizado. *Metodología de las Ciencias del Comportamiento, Suplemento Especial*.1-6
- Chang, Y. & Ying, Z. (2004). Sequential estimation invariable length Computerized adaptive testing. *Journal of Statistical: Planning and Inference*, 121, 249-264.
- Chen, P. E. & Liou, M. (2003). Computerized adaptive testing using the Nearest-Neighbours criterion. *Applied Psychological Measurement*, 27, 204-216.
- Davey, T. & Fan, M. (2000). *Specific information item selection for adaptive testing*. NCME . New Orleans.
- Gregory, R. (2001). *Evaluación psicológica: Historia, principios y aplicaciones*. (3a. ed.). México: Manual Moderno.
- Hambleton, R. & Swaminathan, H. (1985). *Item response theory. Principles and applications*. Boston: Kluwer Nijhoff
- Hambleton, R., Zaal, J. & Pieters, J. (1991). Computerized adaptive testing: Theory, applications and standards. En R. Hambleton & J. Zaal (Eds.), *Advances in educational and psychological testing: Theory and applications*. Boston: Kluwer Academic.
- Lord, F. (1980). *Application of item response theory to practical testing problems*. Hillsdale, N, NJ: LEA
- Muñiz, J. (1997). *Introducción a la Teoría de Respuesta a los Ítems*. Madrid: Pirámide.
- Olea, J. & Ponsoda, V. (1996). Test adaptativos informatizados. En J. Muñiz (Ed.), *Psicometría* (pp. 729-783). Madrid: Universitas, S. A.
- Ponsoda, V., Hontangas, P., Olea, J., Revuelta, J., Abad, F. & Ximénez, C., (2004). Los tests adaptativos informatizados: Investigación actual. *Metodología de las Ciencias del Comportamiento, Suplemento Especial*. 505-510
- Rocklin, T. & O'Donnell, A (1987). Self – adapted testing: A performance – improving variant of computerized adaptive testing. *Journal of Educational Psychology*, 79 (3), 315 – 319
- Rocklin, T., O'Donnell, A. & Holst, P. (1995). Effects and underlying mechanisms of self-adaptive testing. *Journal of Educational Psychology*, 87 (1), 103 – 116
- Stocking, M.L. & Swanson, L. (1998). Optimal design of item banks for computerized adaptive tests. *Applied Psychological Measurement*, 22, 271-280.