Psicothema
2001. Vol. 13, nº 4, pp. 685-690

ISSN 0214 - 9915 CODEN PSOTEG

Pasado, presente y futuro de los Tests Adaptativos Informatizados: entrevista con Isaac I. Bejar

Antonio J. Rojas Tejada
Universidad de Almería

En este artículo se presenta el resultado de una entrevista con Isaac I. Bejar. El Dr. Bejar es actualmente Investigador Científico Principal y Director del Centro para el Diseño de Evaluación y Sistemas de Puntuación perteneciente a la División de Investigación del Servicio de Medición Educativa (*Educational Testing Service*, Princeton, NJ, EE.UU.). El objetivo de esta entrevista fue conversar sobre el pasado, presente y futuro de los Tests Adaptativos Informatizados. En la entrevista se recogen los inicios de los Tests Adaptativos y de los Tests Adaptativos Informatizados y últimos avances que se desarrollan en el *Educational Testing Service* sobre este tipo de tests (modelos generativos, isomorfos, puntuación automática de ítems de ensayo…). Se finaliza con la visión de futuro de los Tests Adaptativos Informatizados y su utilización en España.

Past, present and future of Computerized Adaptive Testing: Interview with Isaac I. Bejar. In this paper the results of an interview with Isaac I. Bejar are presented. Dr. Bejar is currently Principal Research Scientist and Director of Center for Assessment Design and Scoring, in Research Division at Educational Testing Service (Princeton, NJ, U.S.A.). The aim of this interview was to review the past, present and future of the Computerized Adaptive Tests. The beginnings of the Adaptive Tests and Computerized Adaptive Tests, and the latest advances developed at the Educational Testing Service (generative response models, isomorphs, automated scoring…) are reviewed. The future of Computerized Adaptive Tests is analyzed, and its utilization in Spain commented.

Isaac I. Bejar es actualmente *Investigador Científico Principal (Principal Research Scientist)* y Director del *Centro para el Diseño de Evaluación y Sistemas de Puntuación (Center for Assessment Design and Scoring),* perteneciente a la *División de Investigación (Research Division)* del *Servicio de Medición Educativa (Educational Testing Service,* a partir de ahora ETS).

El Dr. Isaac I. Bejar comienza su carrera investigadora como Asistente de Investigación (*Research Assistant*) en la Universidad de Minnesota (año 1970), universidad donde tiene lugar la lectura de su tesis doctoral (1975). Posteriormente, obtiene una beca postdoctoral del *National Institute of Education* para continuar los estudios de Investigación en medición educativa en la *Northwestern University.* En el año 1978 entra a formar parte de la plantilla investigadora del ETS, organización donde ha tenido lugar prácticamente toda su producción científica.

El ETS está localizado a pocos kilómetros de Princeton (N.J., EE.UU.). En 340 acres de arbolado y verde paraje se sitúan, al estilo de campus universitario, un conjunto de edificios con nombres de personas emblemáticas en la historia de la medición educativa y psicológica estadounidense (J. Bryant Conant, Carl Brigham, Ben Wood, L.L. Thurstone, S. Messick, y el más reciente que se dedicará a Frederick Lord…). Estos edificios albergan a 2100 empleados, lo que la convierte en la más grande institución de EE.UU. dedicada a medición psicológica y educativa. El ETS, creado en 1947, es una organización de carácter privado, sin ánimo lucrativo, que desarrolla y aplica instrumentos de medida a distintos niveles: tests psicológicos y educativos a escala internacional (TOEFL), federal y estatal (NAEP y SAT), en empresas, etc. El ETS administra anualmente 11 millones de tests en Estados Unidos y en otros 180 países (http://www.ets.org). En los proyectos que desarrolla el ETS colaboran tanto el gobierno federal como los gobiernos estatales, así como los distritos escolares locales e iniciativas privadas preocupadas por los temas de evaluación psicológica y educativa. Desde que en 1947 se fundara el ETS, la investigación ha sido el tema central de esta organización.

El centro que dirige Isaac I. Bejar, *Center for Assessment Design and Scoring,* es un centro perteneciente a la División de Investigación (*Research Division*) donde se combinan los principios de la psicología cognitiva, los avances sobre medición y las nuevas tecnologías para hacer avanzar la evaluación educativa y profesional. Concretamente, son temas de investigación de este centro: nuevos diseños de evaluación, generación de ítems, elaboración de modelos para evaluar dominios de contenidos, procesamiento del lenguaje natural, formas alternativas de puntuación automática, y, en general, una amplio abanico de temas relacionados con los sistemas de puntuación. Los trabajos y avances realizados por los investigadores de este centro en los últimos años en el campo de los Tests Adaptativos Informatizados (TAI) le han conferido un gran prestigio internacional.

Correspondencia: Antonio J. Rojas Tejada
Facultad de Psicología
Universidad de Almería
04120 Almería (Spain)
E-mail: arojas@ual.es

Actualmente Isaac I. Bejar es reconocido, dentro del campo de los TAI, si no como el progenitor de los Modelos de Respuesta Generativos, si como una de las figuras más importantes en ello (Bennett, 1998). Por todo esto, podemos decir que hoy en día Isaac I. Bejar es uno de los investigadores con mayor proyección internacional del ETS.

Tests adaptativos y tests adaptativos informatizados

—Si bien las aplicaciones iniciales de Tests Adaptativos se pueden remontar a los trabajos sobre inteligencia que en 1908 realizó Binet (p.e. Weiss, 1985; Reckase, 1989), sabemos que fue Frederick M. Lord en el ETS, quien en los años 80 empezó un amplio programa de investigación sobre los tests adaptativos (p.e. Lord, 1971a, 1971b, 1971c). ¿Cómo ha visto ud. evolucionar en ETS esta idea desde los Tests Adaptativos hasta los actuales programas de TAI?

—Podemos suponer que la idea proviene de los trabajos de Binet, entre otros. Pero tengo entendido que la idea de ajustar (*taylor*) el test a un individuo de una forma automatizada viene de una sugerencia de Bill Turnbull, que fue el segundo presidente de ETS. Él se acerca a Fred Lord, allá por los años 60, con el propósito de implementar la idea de adaptar un test según las respuestas del examinado. Es interesante notar que en aquella época de los 60, no se daba por sentado que los ordenadores se convertirían en lo que son hoy en día, prácticamente un electrodoméstico más. Es por eso que algunas de las investigaciones que llevaba a cabo Fred Lord tenían que ver con la adaptación sin computadora. Fue mucho después, a principios de los 70, cuando se consolidó la idea de que el ordenador sería el medio de administración de tests adaptativos.

El trabajo de Fred fue subvencionado, primordialmente, por la ONR (*Office of Naval Research*) y esta misma agencia subvencionó las investigaciones de David Weiss, quien formó un grupo de trabajo que produjo las primeras implementaciones de Tests Adaptativos Informatizados en Minnesota. Uno de éstos grupos tenía que ver con vocabulario y otro con rendimiendo académico. Este último grupo estuvo a mi cargo y publiqué algo relacionado con ese proyecto (Bejar, 1983). Ambos sistemas fueron implementados en lo que se llamaba una mini-computadora, de Hewlett Packard, si mal no recuerdo. Lo que sí recuerdo muy bien, sin embargo, es que tendía a caerse (*crash*) con bastante frecuencia.

Uno de los retos que presentó el sistema adaptativo para la medida del rendimiento era cómo presentar gráficas y símbolos. Los terminales de aquella época no tenían la capacidad de los equipos de hoy día y solo podían trabajar con texto. Ese reto lo resolvimos sencillamente creando unos folletos que contenían las gráficas. Claro que esto era engorroso para el examinado, pero la tecnología de aquella época no permitía más. Los sistemas que se desarrollaron en Minnesota eran experimentales.

El primer test operacional se desarrolló aquí en ETS y estuvo a cargo de Bill Ward (p.e. Ward, 1984). La versión actual se llama *Accuplacer,* que es un test de clasificación (*Placement Tests*) que permite asignar a los examinados a cursos que se ajusten a su experiencia y sus destrezas educativas (College Board, 2000).

—En ETS se desarrollan actualmente varios programas de TAI: por ejemplo el Scholastic Aptitude Tests I: Reasoning Test *(SAT I), el* Graduate Management Admisions Test *(GMAT), el* Graduate Record Examination *(GRE), el* Test Of English for Foreigner Language *(TOEFL), el* National Council of Architectural Registration Boards *(NCARB) o el* National Council Licensure EXamination for registered nurses *(NCLEX). ¿Cuál es su valoración de lo que se ha dado en llamar primera generación de Tests Informatizados (Bennett, 1998)?*

—La primera generación de tests informatizados constituyen un logro, por un lado, y una decepción, por otro. Constituyen un logro si tenemos en cuenta la complejidad de los sistemas que son necesarios para desarrollar, administrar, y puntuar tales tests. Por otro lado, esta generación no constituye un adelanto en cuanto a mejorar nuestro entendimiento sobre el significado de las puntuaciones. Bob Linn expresó una similar decepción en referencia a la Teoría de Respuesta a los Ítems (TRI), cuando concluye que si bien la TRI es un adelanto metodológico no ha mejorado la validez de las puntaciones (Linn, 1990).

La primera generación de tests informatizados nos ha decepcionado también al no cumplir con las promesas que se habían hecho. Por ejemplo, no parece haberse disminuido el tiempo necesario para administrar el test de fiabilidad comparable con un test lineal. Además los costes de desarrollo parecen haber aumentado de forma que, hoy en día, por ejemplo el *Graduate Record Examination* (un test para la admisión en Centros de Postgrado) representa un coste mayor para el estudiante (tanto en dinero como en esfuerzo), ya que se debe desplazar al *tests center* (centro encargado de administrar los tests), sin que midamos algo fundamentalmente diferente o superior a lo que se medía anteriormente con tests lineales.

Esta primera generación ha presentado anomalías inesperadas. Por ejemplo, la relación entre la dificultad de un ítem y el tiempo de respuesta que se requiere es bastante compleja, y, si no se controla, puede repercutir negativamente en algunos estudiantes. O, también, el hecho de que el test resulta un poco 'acelerado' (*speeded*) en ciertos casos, y esto podría resultar en una puntuación incierta para los examinados.

—En esta primera generación de Tests Informatizados se ha combinado los conocimientos en psicometría (tests adaptativos) con la tecnología informática. Pero estos Tests Informatizados son sustancialmente los mismos que los que se administraban en papel: miden las mismas habilidades, se usan los mismos diseños comportamentales y se usa el mismo tipo de tareas (Bennett, 1998). ¿Cuándo veremos aparecer la próxima generación de tests informatizados y cuáles serán sus características?

—Se está trabajando fuertemente para darle mejor uso al ordenador como medio de administración de tests. Ya Bennett bosquejó las etapas que él percibe como necesarias para llegar a generaciones más adelantadas (Bennett, 1998). Me parece esencial, para conseguir esas futuras generaciones de tests informatizados, llegar a dominar mejor la utilización del ordenador, especialmente abaratar el proceso de desarrollo, sin que ello suponga un gran impacto para la validez, y, quizás, incluso para mejorarla (Bejar y Bennet, 1999).

El problema central de todo esto es que el modelo adaptativo basado en ordenadores ha partido de la forma tradicional de elaborar bancos de ítems. Además, la introducción del ordenador en este proceso, lejos de lograr abaratar el proceso, lo ha encarecido. Pensemos que cuando se hace una administración, los ítems de un banco tienen que ser desarmados, para que las preguntas no se expongan demasiadas veces, cosa que suele ocurrir cada nueve meses aproximadamente. Ello ha supuesto que se inviertan muchos recursos (tanto humanos como materiales) en elaborar nuevos ítems para mantener el banco. Si logramos reducir los costes de producción y hacerlo de forma más eficiente, todos estos recursos

que se dedican ahora a producir ítems para el banco, se podrían utilizar para abordar la investigación y desarrollo de tests fundamentalmente distintos a los actuales.

—Y, ¿qué procedimientos se podrían probar para conseguir reducir esos altos costes?

—Una posible estrategia para reducir estos costes puede ser incrementar la perdurabilidad de un banco de ítems. Creo que esto se podría lograr, en parte, con el uso de modelos generativos.

Modelos generativos de ítems

—Como ha dicho su amigo y compañero Randy Bennett, ud. realmente está considerado si no como el progenitor de los Modelos Generativos de Ítems (Bejar, 1993, 1996), si como una de las figuras más importantes en ello (Bennett, 1999). ¿Qué se entiende por Modelos Generativos de Ítems y cómo contribuye a abaratar costes?

—La idea esencial de estos modelos es poder captar el conocimiento que los expertos en la elaboración de ítems han vertido en los bancos de ítems que se han realizado. Así, no solo es posible a través de un modelo generativo producir ítems, sino que también lo hacemos con un control de los atributos psicométricos de los ítems que se van a generar. Es decir, podremos generar ítems donde conocemos por anticipado sus parámetros (p.e. su dificultad), siempre y cuando tengamos un modelo psicológico de cómo los sujetos responden a los ítems. En este sentido el proceso adaptativo se aumenta con un proceso generativo.

Esta idea me parece que representa una mejor utilización de la computadora ya que va más allá de extraer ítems de una base de datos, ya que el ordenador no es meramente un sistema para administrar los ítems de un banco, sino que toma un papel activo en la composición de los mismos; el ordenador asiste en el proceso de crear ítems de acuerdo con ciertas especificaciones.

En este caso la «promesa» es que será posible rebajar el coste de producción. Pero esto es solo una parte, que por supuesto ayudará, pero hay que empezar a pensar en elaborar bancos de ítems teniendo en mente el modelo adaptativo, y abandonar la forma clásica de elaborarlos.

—Relacionado con la aplicación de la idea de Modelos Generativos, podemos comentar que, en 1999, su equipo de trabajo recibió el Premio de Aplicación Tecnológica que otorga el National Council on Measurement in Education *(NCME) a contribuciones significativas en el campo de Teoría de la Medida. Y en enero de este año, ha conseguido el Premio de Investigación de ETS, donde igualmente se le reconoce sus contribuciones científicas. Ambos premios se han basado, sobretodo, en sus recientes trabajos en el desarrollo de la versión informatizada adaptativa del Examen de Licenciatura de Arquitectos (*National Council of Architectural Registration Boards –NCARB–*). Nos puede explicar en qué ha consistido el proyecto del Examen de Licenciatura de Arquitectos (NCARB) y qué ha supuesto de novedoso desde el punto de vista de la investigación psicométrica.*

—El proyecto NCARB representa otro ejemplo de mejor utilización de la computadora para administrar tests. El proyecto comienza a finales de la década de los 80. El cliente NCARB, una organización nacional a la que pertenecen las juntas de certificación de cada estado, le interesaba la idea de emplear medios informáticos para medir mejor las destrezas de arquitectos principiantes (en EE.UU. es necesario terminar la carrera, hacer un internado y pasar el test NCARB para obtener una licencia que le permite al arquitecto/a ejercer independientemente). La idea de automatizar la calificación de diseños arquitectónicos fue parte del plan desde un principio por dos razones. Se reconocía que un tribunal tiene ciertas limitaciones para calificar ciertos aspectos, por lo menos de una forma eficiente. La otra razón eran los altos costes de la calificación por tribunales, que suponían varios millones de dólares. A mí me interesó mucho el proyecto porque permitía explorar maneras de extraer más información de una respuesta a través de un proceso objetivo. Nos dimos cuenta pronto de que desarrollar software para calificar diseños de una forma automatizada era muy costoso y de esa necesidad surge la idea de extender la aplicabilidad del software a una gran variedad de respuestas y la idea de implementar los Modelos Generativos toma cuerpo. En este proyecto empezamos a crear formas paralelas aleatorizadas a partir de modelos de ítem e isomorfos.

—¿Qué son y qué ventajas aportan las formas paralelas aleatorizadas a partir de modelos de ítems e isomorfos (p.e. Bejar 1991, 1995; Bejar y Yocom, 1991)?

—Estos dos conceptos están muy relacionados. El término isomorfo lo tomamos prestado de Herb Simon (Simon y Hayes, 1976). En nuestro caso se refiere a ítems que son psicométricamente equivalentes e intercambiables. Es decir, tantos los atributos de contenido como los parámetros en un modelo de Teoría de Respuesta a los Ítems, son, en principio, idénticos a través de un conjunto de isomorfos. Claro que en la práctica los isomorfos no resultan idénticos y la cuestión es, metodológicamente hablando, cuánto se puede apartar uno del isomorfismo total sin que hayan repercusiones negativas en la medición.

Recientemente hemos estado utilizando el término 'modelo de ítem' (*ítem model*) para referirnos a un conjunto de isomorfos y a los principios y restricciones que controlan la producción de esos isomorfos. El término modelo de ítem lo hemos tomado prestado de Tony LaDucca (LaDucca et al., 1986).

Esto supone que a la hora de crear un test, y teniendo en cuenta los componentes o especificaciones que lo definen, se puede generar un gran número de versiones de ítems (modelos de ítems) que midan cada una de dichas especificaciones. Tenemos por un lado, la idea de realizar formas paralelas aleatorizadas a través de los distintos conjuntos de modelos de ítems y, además, cada modelo de ítem genera diferentes isomorfos aleatoriamente. Por lo tanto, cada cual responde a un test diferente pero, si todo ha ido bien, todos los tests producidos de esta manera son equivalentes tanto psicométricamente como desde el punto de vista de la puntuación del examinado. Es decir, no es necesario llevar a cabo equiparación de puntuaciones.

Desde un principio supusimos que si bien el ordenador ofrecía muchas ventajas, también sería interesante hacer las cosas de otra manera para evitar sus posibles desventajas, como por ejemplo la sobreexposición de ítems. Hay que pensar que un test informatizado se administra constantemente en el *tests center*, y no dos o tres veces al año como en el formato de lápiz y papel. Desde un principio sabíamos que la gente iba a salir del *tests center y* que al instante alguien recogería toda la información disponible sobre el test para preparar a futuros tomadores del mismo. Eso se da por sentado que va a ocurrir.

En parte, la motivación para implementar formas paralelas aleatorizadas a partir de modelos de ítems e isomorfos era lograr una mayor seguridad (respecto a la exposición de los ítems). Con el uso de modelos de ítems existen un sin fin de formas que se generan aleatoriamente a la hora de elaborar el test, y pienso que esto

es mucho más seguro que la que se logra con los tests tradicionales, donde se puedan utilizar dos o tres formas fijas equivalentes. Es verdad que se podrían hacer formas paralelas aleatorizadas de antemano si tuviéramos un banco inmenso de ítems, pero esto sería muy costoso y además, en el caso del NCARB, esto no funcionaba porque este test conllevaba calificación automatizada. En este sentido, las formas paralelas aleatorizadas creadas en el momento de administrar el test a partir de modelos de ítems, que en este caso incluye las especificaciones para calificar las respuestas a cada isomorfo, supuso la solución.

—Además de implementar los Modelos Generativos, el NCARB llevaba un sistema de puntuación automática de ítems de ensayo, ¿cómo se logró esto y por qué se pensó en ello?

—Cada test NCARB es como unos 15 tipos de tareas de diseños de arquitectura, donde cada tipo requiere un programa informático aparte. Lo que se quería conseguir es que esos programas sirvieran para codificar un gran número distinto de preguntas, porque si no no rendía. Hubo un programa que costó casi medio millón de dólares, solo en lo concerniente a programadores, ya que el tiempo que invirtieron los arquitectos era por voluntad propia y no se cobraba.

Todo esto tenía cierto sentido porque ellos hacían un test administrado de forma grupal dos veces al año y les costaba casi dos millones de dólares, y decidieron invertir para no gastar tanto dinero y para medir un poco mejor. Existe cierta controversia sobre si se logró la segunda meta de medir mejor.

Desde el punto de vista metodológico parece claro que sí, pero no todo el mundo está de acuerdo. El dilema es el siguiente: ellos tenían antes un examen que tomaba 12 horas, era un diseño arquitectónico completo. Entonces, ¿qué pasaba?, pues que a la hora de calificarlo, tenían tanto que calificar que no podían estudiar todos sus componentes, y se calificaba globalmente. Todo el tiempo que invertía el estudiante en hacer el diseño se perdía en la calificación. Por lo que se optó en el test NCARB, ante la incapacidad de desarrollar un programa informático para calificar el diseño completo, fue dividir el diseño en componentes.

Ahora, la controversia viene porque algunos alegan que al hacer eso se ha perdido la integración. Habría que investigar si se perdió la integración, pero si ahora es así, antes también se perdía porque no se tomaban en cuenta todos los elementos por separado. Esto sin duda ha quedado por investigar.

Futuro de los tests adaptativos informatizados

—¿Qué claves piensa Ud. que decidirán el futuro de los Tests Adaptativos Informatizados?

—Yo estoy seguro que el futuro de los tests informatizados dependerá mucho del mejor aprovechamiento de la tecnología, por un lado, y de la incorporación de los avances en la psicología cognitiva en el proceso de diseño de pruebas, por otro. Hay que incorporar la psicología a la psicometría, que desde un principio estuvieron incorporados, recordemos los trabajos de Thurstone. Pero pienso que pasó algo a partir de los años 50 donde se separaron, y la metodología se volvió en la meta en sí misma.

Por supuesto, será necesario tener gran creatividad para apartarse de los patrones y diseños anteriores cuando esto sea necesario. Por ejemplo, un componente importante del costo de los tests informatizados es el precio del *tests center*. El costo es tan alto que hace prohibitivo diseños que requieran varias horas. Sería conveniente ingeniarse algún método que permita realizar las mediciones, aprovechándonos de los ordenadores que muy pronto estarán en el escritorio de todos los estudiantes. Internet, sin duda, abrirá las puertas a ello. Pero el problema vendrá en la certificación de que la persona que está tomando el test es el autor de ello. Habrá que ingeniarse algo.

Internet abre grandes posibilidades. Con Internet ya existen experiencias novedosas en ETS, por ejemplo existe un proyecto con el test *Graduate Record Examination* (GRE) llamado '*on the fly adaptive testing*', donde se administrará experimentalmente un componente del examen a través de Internet. Habrá un servidor en ETS y las distintas universidades harán uso del test a forma de catálogo: uno enviará el tipo de tests que se requiere y el servidor seleccionará los ítems adecuados para hacer la evaluación. Ya no será solo adaptativo el test sino el lugar para tomarlo.

También se están probando nuevos lenguajes para escribir páginas web, como el formato XML, que permite mejorar el HTML, ya que desprende el contenido del formato, y es una forma muy versátil de presentar preguntas en la web. Esto es especialmente útil en los Modelos Generativos donde es un algoritmo el que genera varias versiones y todas salen bien formateadas. ETS es parte de un consorcio que tiene como meta establecer *interoperability standards* respecto a ítems y, en general, a materiales educativos electrónicos. El lenguaje XML es clave en este proceso (http://www.imsproject.org/question/index.html).

Debemos aprender también de los métodos que resultan de la 'nueva economía', donde muchos procesos se llevan a cabo de forma distributiva y en colaboración. Existe un posible ejemplo que tengo en mente aquí en ETS. Los tests de certificación del *National Board* (Consejo Nacional) requieren que el profesor presente un vídeo donde se demuestre sus destrezas pedagógicas en el aula. La puntuación en este componente se integra a la puntuación de tests estandarizados. En este caso, el video recoge información que quizás no se pueda obtener de otra forma, ya que simula una observación directa del maestro en acción. Este diseño ilustra el concepto de distribuir el proceso de medición de forma que se reduzca el tiempo necesario en el *tests center*, que reducirá un tipo de costo, y a la misma vez recogería información que no se podría obtener en el *tests center* (claro que en este caso, el costo de la producción de video no necesariamente reduce el precio final).

—Y en este futuro, ¿cómo ve ud. al ETS, seguirá teniendo el mismo papel predominante en el desarrollo de Tests Adaptativos Informatizados, como lo ha tenido hasta ahora en el ámbito de la medición mundial?

—Se considera que ETS es el líder actualmente en el campo de medición educativa. Todo lo que se está haciendo en ETS es con miras a seguir manteniendo ese liderato, y parte de ello se debe a haber comercializado por primera vez el modelo adaptativo. Pero en ese proceso nos dimos cuenta de una serie de limitaciones de ese modelo. En ETS parece haber un desencanto general por el modelo adaptativo; puede que ETS descarte el modelo adaptativo tal y como lo conocemos actualmente, a no ser que encontremos una mejor forma de utilizarlo.

Puede que en un futuro, nos enfoquemos más en la integración de la medición en la educación. En ese caso se puede hablar de modelos adaptativos, pero es otro tipo de modelo adaptativo, es un modelo donde, por ejemplo, el ordenador escoge una pregunta que incrementa la certeza sobre el estado del conocimiento de los examinados, que esto no es lo mismo que medir una sola dimensión, es algo más generalizado. En esto destaca el trabajo de Bob Mislevy (p.e. Almond y Mislevy, 1999), que está trabajando en un mo-

delo donde la medida no se limita al modelo unidimensional, sino que representa lo que se quiere medir del estudiante con una serie de variables, donde puede que, en cierto tipo de instrucción, queramos tener mucha certeza sobre una destreza, pero no sobre otra. En este sentido será adaptativo, es un proceso que requiere cierta inteligencia, pero no es solo con miras a maximizar la información de una única dimensión como lo hacen los tests adaptativos convencionales.

La meta de ETS es mantener el liderato, pero no necesariamente en los Tests Adaptativos Informatizados, se quiere ir más alla de esa tecnología, que en realidad empezó en esa década (años 80) donde los tests de admisión a las universidades eran el principal producto de medición y la relación con el proceso de educación era lejano. Hoy día se va más por la integración de la evaluación dentro del proceso educativo.

Tests adaptativos informatizados en España

—Ud. conoce los trabajos de Psicometría que se hacen España. No hace mucho estuvo como conferenciante invitado en el VI Congreso Nacional de Metodología *y pudo conocer qué se está haciendo en este País. Además ha participado en el texto sobre* Tests Adaptativos Informatizados *que publicaron Olea, Ponsoda y Prieto (1999). ¿Qué piensa sobre los trabajos que se realizan en España sobre TAIs y sobre psicometría en general?*

—Yo estoy muy impresionado con el nivel de psicometría que se practica en España, por lo menos en cuanto a la investigación. He visitado varios países y me consta que en España se está produciendo investigación psicométrica a los más altos niveles. Ha sido muy recientemente cuando he tomado conciencia de este hecho. Aunque he visitado España de turista con alguna frecuencia, mis contactos profesionales fueron muy escasos. Conocí a Miguel Ángel Pérez en los años 80 y mucho más recientemente he estado intercambiando correo electrónico con Vicente Ponsoda. Poco a poco me he ido empapando sobre el trabajo que se lleva allí. Incluso, un profesor del equipo de Vicente, Javier Revuelta, visitó ETS hace un par de años en calidad de becado y seguimos en contacto. Vicente me pidió que participara en el libro sobre tests adaptativos que mencionaste, y esto me sirvió para conocer a los psicómetras más destacados de España. Tuve suerte de que saliera el libro antes del congreso porque pude conocer personalmente a muchos de los autores. Por cierto, este libro recoge una impresionante colección de trabajos. Es una lástima que sólo esté en español, ya que me parece que sería de interés a una audiencia mucho más amplia.

—Ud. sabe que en España no existe una tradición de aplicación de tests estandarizados, ni siquiera en la medición educativa. Sin embargo, la investigación en Tests Adaptativos Informatizados no se ha visto mermada y sigue los últimos avances y desarrollos en este campo. En este país se están investigando prácticamente los mismos problemas que existen en la medición mediante tests en EE.UU. (p.e. Aguado, Santa Cruz, Dorronsoro y Rubio, 2000; Revuelta, 2000; Revuelta y Ponsoda, 1998). En este sentido, la investigación en este campo no está teniendo repercusiones prácticas, así como la práctica no está determinando la agenda de los problemas a investigar. Quizás deberíamos decir que la investigación en Test Adaptativos Informatizados en España está reflejando más la práctica de medición de EE.UU. ¿Qué piensa ud. sobre esta agenda de investigación española y qué aspectos en común ve entre la práctica de medición en ambos países?

—Precisamente, el contenido de la investigación psicométrica en España, tanto el que aparece en revistas internacionales como la investigación que aparece en fuentes nacionales, se preocupa de temas muy relacionados con la agenda estadounidense. No hay que ir muy lejos para darse cuenta que esa agenda se ajusta a unas necesidades muy específicas de EE.UU. Ingenuamente había pensado que si estamos trabajando una agenda similar es por que los problemas prácticos son los mismos en ambos países, pero tu pregunta sugiere, y también me he enterado por mi cuenta, que esto no es así.

La agenda de investigación que se refleja en las revistas dedicadas a este tema está relacionada, en gran medida, por un lado, con los tests de admisiones a la universidad, lo que para ustedes sería la selectividad; y por otro lado, tenemos la problemática asociada con tests de certificaciones y licencias, es decir, la determinación de si el examinado consta de cierto nivel de conocimientos, en el caso de estudiantes, o pericia, en el caso de profesionales. Más recientemente han aparecido las encuestas educativas que requieren su propia agenda de investigación. A esto tenemos que añadir que, desde hace mucho tiempo, la tecnología que se utiliza para administrar tests más eficientemente o para medir con más precisión, o para medir constructos que no se prestan al papel y lápiz, se ha incorporado a la agenda de investigación.

Si dividimos la práctica psicométrica de esta forma, exámenes de admisiones, certificaciones y licencias y las encuestas educativas, quizá podríamos esclarecer un poco qué aspecto de la práctica existe en común entre España y Estados Unidos. Por ejemplo, con respecto a los tests de admisiones, los tests adaptativos en EE.UU. se empezaron a concebir en ETS por Fred Lord, como comenté anteriormente. Los beneficios que se vislumbraban eran más precisión y mayor eficiencia. Esto a su vez respondía al volumen de tests de este tipo con relación a la selección de estudiantes. No conozco bien el sistema universitario de España para determinar la aplicabilidad de esta línea de investigación allí. Te puedo adelantar que la incorporación de la computadora al proceso de medición no siempre ha resultado en mayores eficiencias o precisión como se había esperado, por lo menos no de primera instancia. Sin embargo, existe poca duda de que al fin y al cabo los tests se administrarán por ordenador, sencillamente porque todo se hará a través de ese medio. Las décadas que se lleva estudiando este tema nos facilitarán poner los puntos sobre la ies.

Con relación a los tests de certificación, licencias y encuestas educativas me parece que son mucho más aplicables en todo el mundo. Los tests de certificación y licencias se caracterizan por un diseño diferente a los tests de admisiones. En los tests de certificación se trata de determinar si el estudiante rinde mas allá de un nivel mínimo establecido previamente por un proceso de fijación de estándares. Lo mismo se aplica en el caso de los tests que se le administran a principiantes en una profesión como parte del proceso de dar credenciales (tests de licencias). Este diseño parece ser aplicable más generalmente. Tengo entendido que en México, por ejemplo, un grupo está desarrollando estándares para tests de licencias.

Los tests de admisiones surgen debido a la variedad de sistemas educativos que son parte del sistema estadounidense y que no permite utilizar exclusivamente el rendimiento escolar como medida de selección. En países donde esa variabilidad no exista se puede concebir una prueba de certificación (al estilo de la selectividad española, pero de forma estandarizada) que también sirve como evidencia para la admisión del próximo nivel académico. Si el test

es puramente de certificación el diseño óptimo requiere una función de información que se concentre en el punto de corte. Sin embargo, los tests que jueguen un doble papel de certificación-admisión podrían requerir una función de información concentrada desde el punto de corte en adelante. Menciono todo esto porque en el primer caso adaptar el test no tiene sentido pero sí se justifica en el segundo caso.

Finalmente, con relación a encuestas educativas, me parece que tienen gran aplicabilidad, si nos fijamos en la demanda que existe por este tipo de tests en todas partes del mundo.

En resumen, la psicometría relacionada con tests de certificaciones y las encuestas educativas me parece que son completamente aplicables al entorno español. Los tests de admisiones adaptativos quizá tengan aplicabilidad en la selección de aspirantes pero son mucho menos aplicables en el ámbito educativo.

En fin, me parece que en EE.UU. se tomó la delantera en psicometría por razones históricas, pero dentro de poco, y como parte de la globalización y ampliación de perspectivas que resultan del intercambio de ideas que ahora se acelera por el Internet, no percibiremos gran diferencia entre la aplicabilidad o contenido de las investigaciones psicométricas en España y EE.UU.

Agradecimientos

Esta entrevista se elaboró en el *Turbull Hall* del *Educational Testing Service* (Princeton, N.J., EE.UU.) durante el mes de febrero del 2000. El excelente trato del Dr. Isaac Bejar, su perfecto español (es de procedencia cubana) y el magnífico entorno (entre árboles, ardillas y nieve) hicieron que el trabajo se desarrollara en un ambiente muy cordial y agradable. Quede aquí reflejado mi agradecimiento a Isaac por haber hecho todo esto posible.

Nota

La estancia en el *Educational Testing Service* fue subvencionada en parte por la Dirección General de Universidades e Investigación de la Junta de Andalucía.

Referencias

Aguado, D., Santa Cruz, C., Dorronsoro, J.R. y Rubio, V.J. (2000). Algoritmo mixto mínima entropía-máxima información para la selección de ítems en un test adaptativo informatizado. *Psicothema,* vol. 12(Supl. nº2), 12-14.

Almond, R.G. y Mislevy, R.J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement, 23(3),* 223-237.

Bennett, R.E. (1998). *Reinventing Assessment: speculations on the Future of Large-Scale Educational Assessment.* Princeton, NJ: Educational Testing Service.

Bennett, R. (1999). ETS Scientist Awards Announced. *News Alert, Vol.1, No.7.*

Bejar, I.I. (1983). *Achievement testing: recent advances.* Beverly Hills, CA: Sage.

Bejar, I.I. (1991). A methodology for scoring open-ended architetectural design problems. *Journal of Applied Psychology, 76,* 522-532.

Bejar, I.I. (1993). A generative approach to psychological and educational measurement. En N. Frederiksen, R.J. Mislevy e I.I. Bejar (Eds.). *Test theory for a new generation of tests*. pp. 323-358. Hillsdale, NJ: LEA.

Bejar, I.I. (1995). From adaptive testing to automated scoring of architectural simulations. En E.L. Mancall y P.G. Bashook (Eds.). *Assessing clinical reasoning: the oral examination and alternative methods.* Evanston, IL: American Board of Medical Specialties.

Bejar, I.I. y Bennett, R. E. (1999). La puntuación de las respuestas como un parámetro del diseño de exámenes: implicaciones en la validez. En J. Olea, V. Ponsoda y G. Prieto, (Eds.). *Tests informatizados*. Madrid: Pirámide.

Bejar, I.I. y Yocom, P. (1991). A generative aprproach to the modeling of isomorphic hidden figure items. *Applied Psychological Measurement, 15(2),* 129-137.

College Board (2000). *ACCUPLACER™.* [On Line]. Obtenido 10/02/2000 en URL: http://www.collegeboard.org/accuplacer/html/accupla1.html

LaDuca, A., Templeton, B., Holzman, G. B., Staples, W. I. (1986). ítem-modelling procedure for constructing content-equivalent multiple choice questions. *Medical education, 20,* 53-56.

Linn, R. L. (1990). Has ítem response theory increased the validity of achievement test scores? *Applied Measurement in Education, 3*(2), 115-141.

Lord, F.M. (1971a). The self-scoring flexible test. *Journal of Educational Measurement, Vol. 8 (3),* 147-151.

Lord, F.M. (1971b). Robbins-Monro procedures for Tailored Testing. *Educational and Psychological Measurement, Vol.31,* 3-31.

Lord, F.M. (1971c). The Theoretical Study of the Measurement Effectiveness of Flexilevel Tests. *Educational and Psychological Measurement, Vol.31,* 805-813.

Olea, J., Ponsoda, V. y Prieto, G. (Eds.). (1999). *Tests informatizados.* Madrid: Pirámide.

Reckase, M.D. (1989). Adaptative Testing: The evolution of a good idea. *Educational Measurement: Issues and Practice 8,* 11-15.

Revuelta, J. (2000). Estimación de habilidad mediante isomorfos. Efectos en la fiabilidad de las puntuaciones. *Psicothema,* vol. 12(2), 303-307.

Revuelta, J. y Ponsoda, V. (1998). Un test adaptativo informatizado de análisis lógico basado en la generación automática de ítems. *Psicothema,* vol. 10(3), 753-760.

Simon, H. A., & Hayes, J. R. (1976). The understanding process: Problem isomorphs. *Cognitive Psychology, 8*(2), 165-190.

Ward, W.C. (1984). Using microcomputers to administer tests. *Educational Measurement: Issues and Practice 3,* 16-20.

Weiss, D.J. (1985). Adaptative Testing by Computer. *Journal of Consulting and Clinical Psychology 53,* 774-789.

Aceptado el 27 de diciembre de 2000