Nuevas tecnologías y medición educativa*

por Vicente PONSODA
Universidad Autónoma de Madrid

Introducción

Las Teorías de los Tests proporcionan procedimientos para obtener las puntuaciones que reflejan adecuadamente el nivel de conocimiento de los estudiantes y determinar la calidad métrica de las evaluaciones y de los instrumentos de evaluación (o tests). La Evaluación Educativa ha sido en el pasado y sigue siendo el motor de buena parte de la investigación en Teoría de los Tests. Una buena evaluación plantea desafíos métricos impresionantes. Hay que conseguir instrumentos que permitan comparar las puntuaciones obtenidas por distintos centros, países, culturas, idiomas..., hay que equiparar las puntuaciones obtenidas con distintos instrumentos, teniendo control de que todos ellos midan realmente lo que se pretende medir, sin sesgos que puedan contaminar las puntuaciones. En el caso de personas con alguna discapacidad hay que introducir las acomodaciones oportunas para que su rendimiento en el test sea el que corresponda a su nivel de conocimiento. A veces, hemos de elaborar instrumentos que sean informativos para el proceso de enseñanza-aprendizaje, y otras veces que permitan evaluar si la clase, el centro, el distrito... cumple o no sus objetivos estratégicos. Debe informar de la calidad métrica cuando los ítems son de respuesta seleccionada o de respuesta abierta, cuando evaluamos con un test de opción múltiple o con un portafolio.

Las principales teorías son dos: la Teoría Clásica de los Tests (TCT) y la Teoría de la Respuesta al Ítem (TRI). En español, pueden consultarse los libros de Abad, Olea, Ponsoda y García (2011), Martínez Arias, Hernández Lloreda y Hernández Lloreda (2006) y Muñiz (1997, 2000), entre otros, para aprender más sobre ellas. Hay otras, como la teoría de la Generalizabilidad, pero su importancia para la evaluación educativa es menor.

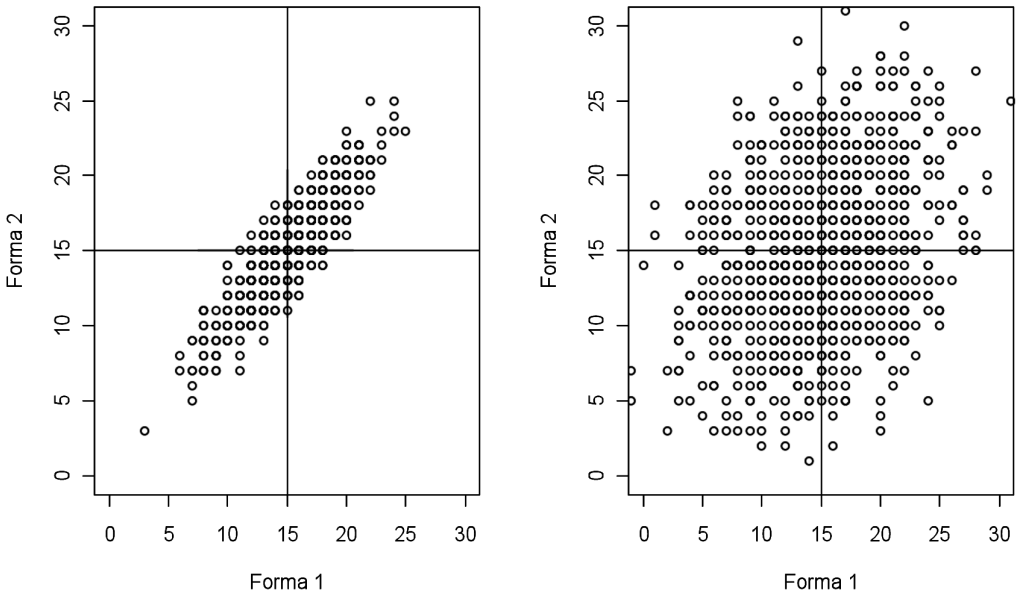
La TCT parte de unos pocos supuestos y de la definición de formas paralelas, y permite, junto con la aplicación de las técnicas de la reducción de dimensionalidad,

* El autor agradece al Ministerio de Ciencia e Innovación (proyecto PSI2008-0-1685) y al Instituto de Ingeniería del Conocimiento (cátedra de Patrocinio MAP) la financiación recibida.

como el análisis factorial, guiar el proceso de construcción de los tests, determinar cuantas medidas tiene sentido extraer de una prueba, establecer sus propiedades métricas básicas en relación a la precisión y validez, fijar puntos de corte, e interpretar las puntuaciones mediante los baremos. Son muchos los tests que se siguen desarrollando bajo la TCT. Sus procedimientos se pueden entender y aplicar fácilmente. La puntuación de la persona en la dimensión es la suma de puntos que ha obtenido en los ítems que evalúan la dimensión. La dificultad de la pregunta viene a ser la puntuación media en el ítem de los que lo han respondido. Su capacidad de discriminación se puede indicar de varias maneras, pero una muy común es mediante la correlación de Pearson entre las puntuaciones en el ítem y las puntuaciones en el test. Dos indicadores muy conocidos de la precisión son el coeficiente de fiabilidad y el coeficiente alfa de Cronbach. En ambos casos, valores próximos a 1 (0) indican buena (mala) precisión.

En la Universidad Autónoma de Madrid es muy común utilizar exámenes de opción múltiple como parte de la evaluación del estudiante de los primeros cursos de grado. En un trabajo reciente (Sierra, Alonso, Chavero, García, Ponsoda y Pulido, 2010) hemos obtenido las propiedades métricas de los exámenes y de las preguntas de todos los exámenes de opción múltiple corregidos automáticamente durante el segundo semestre de 2010, en total más de 300. El principal resultado es que hemos encontrado exámenes de más de 100 ítems y de menos de 10, con el consiguiente impacto en los indicadores de precisión. La peor fiabilidad encontrada ha sido 0.3 y la mejor, 0.9. ¿Qué consecuencias tiene que evaluemos con uno o con otro examen? En el GRÀFIC0 1 se muestra el resultado de simular las respuestas de 1000 estudiantes a un test de opción múltiple de 30 preguntas. En realidad, lo que muestra la simulación son las puntuaciones que cada estudiante simulado daría a dos tests formalmente idénticos (es decir, formas paralelas) que tuviesen la precisión indicada. Se ha fijado como punto de corte el valor 15. El coeficiente de fiabilidad del test que aparece a la izquierda es 0.89 y el de la derecha es 0.32. La gráfica de la izquierda muestra que habríamos obtenido con las dos formas la misma decisión en relación con el punto de corte en el 86% de los estudiantes (un 50% de los estudiantes han obtenido puntuaciones en ambas formas por encima de 15 y un 36% en ambas por debajo o iguales a 15). Un 14% de los estudiantes han obtenido en una forma puntuaciones por encima y en la otra por debajo o iguales a 15. En el caso del test de menor precisión (gráfica derecha), llegaríamos a la misma decisión con ambas formas en el 61% de los estudiantes y a distintas en el restante 39% de los casos. Vemos entonces que la consistencia en las decisiones depende claramente de la fiabilidad o precisión del test. Si evaluamos con poca precisión tendremos poca certeza de obtener parecidas puntuaciones para los mismos estudiantes de haberles evaluado con otro test de propiedades similares. En el caso del test de fiabilidad 0.32 es inaceptablemente alta la probabilidad de que el estudiante dé una puntuación apreciablemente distinta ante otro test similar, quedando entonces la duda de si la decisión tomada con la puntuación observada en el test administrado corresponde o no a su nivel de conocimiento.

GRÁFICO 1: *Consecuencias de la diferente precisión de dos tests*



La TCT tiene algunas desventajas. Una importante es que carece de una propiedad (invarianza) que debiera poseer todo buen procedimiento de medida. En un examen, si los estudiantes han estudiado mucho responderán muchos ítems correctamente y éstos tendrán entonces medias altas (es decir, el indicador de la dificultad de los ítems nos dirá que son fáciles). Lo contrario ocurrirá si los estudiantes saben poco. La TCT no permite separar lo que los alumnos saben, de las propiedades métricas de los ítems y del test en su conjunto. La TRI surgió como una teoría alternativa, que resuelve este y otros problemas. Por poseer la propiedad de la invarianza, la puntuación que asignemos al estudiante no depende de los ítems concretos administrados. Si administramos ítems fáciles, el estudiante tendrá más aciertos (y, entonces, una mayor puntuación según la TCT), que si le administramos ítems más difíciles; pero será básicamente la misma puntuación la que le asignaríamos con la TRI en ambos casos. Justamente en esta propiedad se basan los tests adaptativos informatizados, como veremos más adelante.

La TRI no solo tiene ventajas. Es más difícil de aplicar y utiliza procedimientos estadísticos más sofisticados. A las respuestas de la muestra a un ítem se ajusta un modelo estadístico, de manera similar a como se hace en regresión. Las estimaciones de los parámetros del modelo ajustado proporcionan la descripción métrica del ítem. Aplicado el test, a partir de las respuestas del estudiante a cada ítem y de los valores de sus parámetros, aplicando un estimador apropiado, como el de máxima verosimilitud o bayesiano, se obtiene la puntuación del estudiante. Es posible obtener la función de información del test, que nos indica la precisión con la que se mide cada nivel de rasgo. Tras las evaluaciones PISA está el modelo TRI *logit multinomial de coeficientes mixtos* (Adams, Wilson y Wang, 1997), que no es especialmente simple. Como en

los más sencillos, cada modelo TRI obtiene la probabilidad de dar cada una de las posibles respuestas al ítem en función de sus parámetros y del parámetro o parámetros (en los modelos multidimensionales) que caracterizan al estudiante. Al final del proceso, tras aplicar la TRI, volvemos a obtener como en la TCT las características de los ítems y de los estudiantes; pero por otros procedimientos y con otras propiedades.

Desde una perspectiva aplicada, una ventaja importante de la TRI es que varios de sus modelos (por ejemplo, los modelos logísticos de 1, 2 y 3 parámetros) utilizan la misma escala para describir la dificultad de los ítems y el nivel de las personas, lo que posibilita el *mapeo de ítems*. Por ejemplo, se utiliza esta técnica para facilitar la interpretación de las puntuaciones de la escala de Matemáticas de la evaluación NAEP de grado 4 de 2007 (Institute of Education Sciences). La escala va a de 0 a 500. En el enlace suministrado en la bibliografía se muestra la escala y sus distintos niveles (*Básico*, entre 214 y 248 puntos; *Competente*, entre 249 y 281; y *Avanzado* por encima de 282 puntos). Además, se muestra un ítem apropiado para varias puntuaciones de la escala. Por ejemplo, cuando la puntuación es 287, el ítem que se muestra es un problema sencillo de probabilidad. La respuesta correcta es característica de esta puntuación pero atípica en puntuaciones inferiores. Por tanto, a unas cuantas puntuaciones se adjuntan los ítems o tareas y las respuestas esperadas del estudiante que tenga dichas puntuaciones. Esta técnica da por tanto una detallada información de lo que el estudiante que alcanza cada puntuación probablemente hará (los ítems y tareas asociadas a esa puntuación e inferiores) y no hará (a las superiores).

El objetivo principal de este trabajo es mostrar algunos avances y desarrollos recientes producidos en las teorías de los tests como resultado de la llegada de las nuevas tecnologías, en general, y en particular de la informatización. Entre los asuntos a considerar están los ítems innovadores, los tests de desempeño, el ensamblaje automático de tests, la generación automática de ítems, y los tests adaptivos informatizados. El impacto de la informatización en la práctica y teoría de los tests ha sido considerable. Los *Standards for Educational and Psychological Measurement* (AERA, APA y NCME, 1999) reúnen las principales recomendaciones sobre las buenas prácticas en relación a la construcción y administración de tests. Cuando se escriben estas líneas, están en una fase muy avanzada de una profunda revisión. Una de las cinco razones aducidas para la revisión es precisamente tratar con más detalle el papel de las nuevas tecnologías en los tests (Wise, Drasgow, Hansen, Sackett y Tippings, 2010).

Impacto de la informatización

La evaluación mediante tests consta de distintas fases. Se ha de determinar en primer lugar qué es exactamente lo que queremos medir; habremos de elegir o construir el test o procedimiento de evaluación; se habrá de administrar; se habrá de analizar su calidad métrica y, en su caso, depurarlo; se habrán de obtener las puntuaciones de los estudiantes en las variables que interese; se habrá de interpretar e informar de las puntaciones obtenidas a los interesados. Todas las fases anteriores se han visto afectadas en mayor o menor medida por la informatización.

La informatización permite registrar respuestas que no eran antes accesibles. Se puede registrar el tiempo que se tarda en responder cada ítem; es posible dar retroalimentación en tiempo real y ver su efecto; la información puede presentarse en otros formatos, como video y audio; es posible registrar otras medidas, como los movimientos del cursor… En definitiva, resulta posible medir de otra manera las variables tradicionales y, además, medir variables que no resultaba posible medir antes, como las directamente relacionadas con el tiempo invertido en responder el ítem (Olea, Abad, y Barrada, 2010).

Las posibilidades comentadas están en la base de los nuevos *ítems innovadores* y las nuevas estrategias de *evaluación del desempeño*. En la construcción de tests los ordenadores han cambiado sustancialmente la manera de proceder. Existen programas de *ensamblaje automático* que seleccionan de un banco de ítems los que deben componer el test de acuerdo con los objetivos prefijados. En los últimos tiempos ha habido un considerable desarrollo de la *generación automática de ítems* (GAI). En la GAI no existe un banco de ítems preparados de antemano para su aplicación. El ítem a administrar en cada momento se genera de manera automática. Los ordenadores y la TRI han posibilitado los tests que se adaptan al rendimiento que va manifestando el evaluado. De todos ellos, los *tests adaptativos informatizados* han recibido especial atención en los últimos años. De todos estos temas hablaremos en los apartados que siguen.

El análisis psicométrico es un tipo especial de análisis de datos. Como muestran los restantes artículos de este monográfico, estos análisis se han visto profundamente afectados por las nuevas tecnologías y la informatización. Disponemos en este momento de gran variedad de paquetes y programas, de libre distribución y comercializados, que hacen posible el análisis clásico y moderno de los tests, y las tareas específicas, como el estudio del funcionamiento diferencial, la dimensionalidad, la calibración del banco, el control de supuestos… Especial mención merece el considerable desarrollo actual y previsible desarrollo futuro de los paquetes en R (Elosua, 2011). En el momento de escribir estas líneas, en la página principal de R, dentro del bloque *Psychometrics*, se ofrecen 7 conjuntos de programas (*paquetes*, en la terminología de R) para hacer tareas relacionadas con la TCT, 21 con la TRI y 23 con los modelos de ecuaciones estructurales y el análisis factorial.

También la investigación se ha visto afectada por la informatización. En los estudios teóricos es muy común utilizar la simulación estadística o método de Monte Carlo (Revuelta y Ponsoda, 2003). Por ejemplo, si se quiere saber cómo la longitud del test afecta a la precisión de las puntuaciones, la simulación (y solo la simulación) permite conocer el verdadero nivel de cada estudiante, lo que posibilita la comparación de los niveles verdaderos con los niveles estimados obtenidos en tests de diferente longitud.

Por último, la simulación resulta muy recomendable en docencia. Por ejemplo, el autor explica a sus estudiantes el significado del coeficiente de fiabilidad y otros conceptos métricos con simulaciones como las mostradas en el GRÁFICO 1.

Ítems innovadores

Scalise (2010) propone una taxonomía de tareas de evaluación educativa cruzando dos dimensiones: las restricciones y la complejidad. Propone en el eje de la dimensión *restricción* 7 niveles, desde el 1 al 7, que van desde *respuesta completamente seleccionada* hasta *completamente construida*. El eje de la dimensión *complejidad* tiene 4 niveles y va desde menos (A) hasta más (D) complejo. Un ejemplo de ítem del tipo 1A (menor nivel en ambas dimensiones) consiste en proporcionar al estudiante una afirmación y ha de indicar si la considera correcta o falsa. Un ejemplo de ítem del tipo 6A muestra una tabla con la población de truchas en diferentes años y el correspondiente gráfico de líneas. El estudiante ha de modificar dicho gráfico para que informe adecuadamente de los datos que contiene la tabla. Un ítem que requiere una respuesta más compleja, su tipo es 6B, sería el siguiente: se muestran las alturas de los 31 niños de una clase. Se pide al estudiante que coloque las 31 alturas de modo que tras una inspección rápida de los datos alguien pueda hacerse una idea de la altura de la clase. Estos y otros ejemplos de los 28 tipos de ítems pueden descargarse de la web *(http://pages.uoregon.edu/kscalise/)*. Los ejemplos muestran muy bien a) que es posible utilizar en evaluación educativa ítems muy distintos a los que empleamos habitualmente, b) las nuevas medidas que se pueden registrar cuando se utiliza el ordenador como medio de presentación y recogida de respuestas, y c) las posibilidades de medida de variables nuevas que estos ítems abren.

La evaluación del desempeño

La *evaluación del desempeño* viene teniendo un considerable desarrollo en los últimos años (Martínez Arias, 2010). Es cada vez más factible medir el desempeño, y hacerlo con los adecuados controles métricos, en tareas como escribir un ensayo, realizar un proyecto, crear un portafolio, resolver un problema..., con fines de evaluación o acreditación. Son varios los retos métricos de la evaluación del desempeño, algunos compartidos con la evaluación más tradicional, pero otros más específicos. Un problema específico es cómo conseguir disminuir el esfuerzo que requiere la corrección de estos ítems y la obtención de correctores fiables. Las agencias de tests están optando por el desarrollo de los procedimientos automáticos de corrección. La agencia *Educational Testing Service* ha desarrollado varios, para la corrección de distintos tipos de repuestas, como textos y problemas de álgebra. El último *SpeechRater* evalúa la eficiencia en la comunicación en inglés (Livingston, 2009).

SpeechRater evalúa el nivel de inglés hablado en no nativos. Dos son sus principales módulos: El primero estima la puntuación que un evaluador asignaría tras considerar la fluencia, pronunciación, diversidad del vocabulario y nivel gramatical. El segundo módulo combina las puntuaciones obtenidas en los distintos ítems y proporciona la puntuación total y el correspondiente intervalo de confianza. Los autores (Higgings, Xi, Zechner y Williamson, 2011) reconocen que queda bastante por hacer: no mide aspectos importantes de la eficiencia hablada, como la entonación, y tampoco aspectos de más alto nivel (como la complejidad de las estructuras lingüísticas, la coherencia temática y el desarrollo que se hace del contenido). Además, el ni-

vel de acuerdo con los evaluadores humanos es inferior al que tienen estos entre sí.

Los nuevos ítems innovadores y los procedimientos de corrección automática son asuntos específicamente considerados en los nuevos *Standards*. Se pretende específicamente resolver el conflicto de intereses existente entre los propietarios de los algoritmos que puntúan y corrigen automáticamente los ítems y los usuarios de los tests que demandan información concreta sobre su funcionamiento para poder evaluarlos (Wise *et al.*, 2010).

Un problema compartido entre los ítems de respuesta seleccionada y respuesta construida es el problema del *sesgo*. Al evaluar nivel de conocimiento, por ejemplo, hemos de hacer las cosas de modo que sea eso, y preferiblemente solo eso, lo que evaluemos. En ítems de respuesta seleccionada se sabe que existen diferencias en rendimiento en las pruebas de acceso a la universidad, que favorecen por lo general a los varones, y que estas diferencias no se mantienen en el desempeño real de unos y otros. No parece, por tanto, que la explicación de la diferencia sea un mayor nivel en lo que mide el test; es decir, se trataría más bien de un sesgo de esos tests de acceso a la universidad. Posibles explicaciones de esas diferencias son el diferente comportamiento de mujeres y varones ante el riesgo, lo que haría que unos y otros se comportasen de forma diferente cuando se penalizan las opciones incorrectas, o diferencias en la forma de responder cuando se dispone de poco tiempo *(FairTest)*. Pues bien, el estudio del sesgo es algo conocido, que se debe tener en cuenta en el proceso de elaboración del test. Se recomienda que los ítems sean administrados a una muestra piloto y que sean eliminados o modificados los que muestren comportamiento diferencial (Gómez- Benito, Hidalgo y Guilera, 2010). Si los ítems de respuesta seleccionada tienen el problema de posibles sesgos, lo tienen también y quizás el problema es mayor en los ítems de respuesta construida. González-Espada (2009) mostró a sus estudiantes de un examen de Física una imagen en la que aparecía el inspector Gadget elevándose desde el suelo, como si fuera un helicóptero, con solo un propulsor. Los estudiantes tenían que decir si era posible volar como se indicaba en la imagen. Se encontró que el 70% de los varones dieron con la respuesta correcta, frente al 50% de las mujeres. El autor propone como posible explicación el previsible mayor contacto con helicópteros y aviones de juguete que han podido tener los varones cuando eran niños. Sea o no tal hipótesis la causa de la diferencia observada, lo que parece cierto es que los ítems abiertos, precisamente por suponer tareas más próximas a la vida real, tienen más riesgo de resultar afectados por experiencias previas u otros factores no anticipados por el constructor del ítem y de padecer problemas de sesgo.

Ensamblaje automático

La composición de un test a partir del banco de ítems es un problema complejo, pues requiere tener en cuenta simultáneamente objetivos y restricciones que a veces son parcialmente contradictorios. Se viene denominando *ensamblaje automático de tests* al conjunto de estudios que se ocupa específicamente de estos problemas (Van der Linden, 2005). Diao y Van der Linden (2011) muestran cómo aplicar el

software de libre distribución *R* y el paquete *lp_solve* versión 5.5 para resolver algunos problemas concretos de ensamblaje.

Supongamos, por ejemplo, que tenemos un banco de 165 ítems y que cada uno mide una de *C* categorías de contenido. Queremos construir dos tests de 55 ítems, sin ítems repetidos, tales que se acerquen lo más posible a una función de información del test objetivo prefijada y que cada forma muestre al menos n_c ítems de la categoría *c* (*c*: 1, 2... C).

El problema entonces consiste en encontrar qué 55 ítems de los 165 han de ir a una forma y qué otros 55 a la segunda, de manera que ambas satisfagan lo mejor posible las restricciones y objetivos propuestos. La estrategia consiste en convertir los objetivos y restricciones en un conjunto de ecuaciones e inecuaciones que el programa resuelve. Por ejemplo, siendo

$$x_{if} = \begin{cases} 1 & \text{si el item i (i:1,2... 165) es asignado a la forma f} \\ 0 & \text{en otro caso} \end{cases}$$

el programa ha de encontrar la solución (los dos vectores anteriores, uno por forma) que minimiza la diferencia entre las funciones de información de cada forma y la función de información objetivo y al mismo tiempo satisface un conjunto de restricciones. Por ejemplo, la restricción de que los ítems de ambas formas sean diferentes se indica con la inecuación $\sum_{f=1}^{2} x_{if} \leq 1$ (para todo *i*) y la de que cada forma tenga 55 ítems con la ecuación $\sum_{i=1}^{165} x_{if} = 55$ (para ambas formas). Diao y Van der Linden (2011) muestran las otras 6 ecuaciones necesarias para que el problema de programación entera recoja todas las restricciones planteadas y esté bien especificado. Estas estrategias se pueden aplicar para obtener tests de otras características. En el citado artículo se muestran otros dos ejemplos más complejos.

Generación automática de ítems

La *generación automática de ítems* (GAI) surge como respuesta a la necesidad de reducir el coste de generación de cada ítem individual (Bejar, 1993). Los dos principales requerimientos de la GAI son que podamos describir bien las características de cada clase de ítem, de forma que un programa de ordenador pueda crearlos, y que conozcamos suficientemente bien los determinantes de su dificultad, para que podamos considerar de la misma dificultad todos los ítems desarrollados a partir de una clase. A la clase se le suele llamar *modelo* y a los ítems que resultan del modelo, *variantes* (Drasgow, Luecht y Bennet, 2006).

Drasgow *et al.* (2006) distinguen entre GAI derivada de una teoría fuerte y la derivada de una débil. En el primer caso, se conocen los verdaderos determinantes de la dificultad de los ítems y, a partir de ellos, se predice la dificultad del ítem generado automáticamente. Por ejemplo, Embretson (1998) aplicó la GAI a los ítems tipo Raven. Un ítem típico muestra, en una tabla de 3x3 celdillas, 8 figuras que han sido construidas con una cierta lógica por filas y columnas. La tarea consiste en descubrirla y seleccionar entre las posibles opciones del ítem la figura que debe ir

en la celdilla vacía. Embretson (1998) ha mostrado que a partir de dos variables de proceso y tres perceptuales se puede predecir adecuadamente la dificultad de estos ítems. Tal resultado permite generar de forma automática ítems con una dificultad predicha conocida, a partir de los particulares valores de las variables proceso y perceptuales utilizadas en la generación de cada ítem particular. Conviene advertir que tal procedimiento de generar ítems genera medidas algo más imprecisas. Embretson (1998) ha mostrado que la correlación entre los parámetros predichos y reales (entre 0.7 y 0.8) es inferior que la de los parámetros estimados y reales (0.9) que se obtiene con ítems tipo Raven cuando no se aplica la GAI.

Muchas veces no tenemos una teoría de cómo se resuelven los ítems, en ese caso se puede también aplicar la GAI. Se parte de un ítem *padre* del que conocemos sus propiedades métricas y el objetivo es generar *variantes* o *isomorfos* de ese padre alterando las características que por experiencia, intuición, teoría... pensamos que no afectan a sus propiedades métricas. La calibración de estos ítems es un asunto especialmente difícil, pues no tenemos detrás la teoría fuerte que permita predecir, por ejemplo, la dificultad. Los procedimientos que se pueden aplicar proporcionan estimaciones de los parámetros menos precisas también en este caso (más detalles en Revuelta, 2000).

Tests adaptativos informatizados

La combinación de los ordenadores y la TRI ha hecho posible la aparición y el desarrollo de los *tests adaptativos informatizados* (TAIs). Su principal característica es que los ítems a administrar se adaptan al nivel de competencia que va manifestando el evaluado en los ítems previos. Cuando el rendimiento manifestado por el estudiante es alto, el TAI administrará ítems más difíciles que cuando es bajo. De lo anterior se desprende que el TAI seleccionará ítems distintos a los distintos evaluados. Gracias a la propiedad de invarianza de la TRI, las estimaciones del nivel de rasgo obtenidas en los distintos tests son comparables. La idea básica del TAI es precisamente aprovechar esta propiedad y presentar preferentemente los ítems que resultan altamente informativos para estimar el nivel de cada estudiante.

Una vez calibrado el banco de ítems, el proceso de aplicación de un TAI a un evaluando puede resumirse en los siguientes pasos. Se inicia con una determinada estrategia de arranque, que permite establecer el nivel de rasgo inicial del evaluando (por ejemplo, cero, en la escala típica). Después de que el evaluando responde a cada ítem, se realiza una estimación de su nivel de rasgo mediante los procedimientos de estimación habituales en TRI. Conocida la estimación provisional del nivel de rasgo, el siguiente paso es seleccionar del banco el ítem a administrar. El procedimiento estándar consiste en elegir el ítem que resulta más informativo para el último nivel provisional estimado; es decir, el ítem que produciría un menor error estándar. El TAI entra entonces en un proceso iterativo –administración de ítem, nuevo nivel de conocimiento provisional y nueva selección del ítem– hasta que se satisface el criterio de parada, que suele ser bien que se haya administrado un

cierto número de ítems, o bien que se haya alcanzado una determinada precisión.

Un buen test no solo debe ser muy preciso, debe también muestrear los contenidos adecuadamente y debe procurar que los evaluados reciban, en la medida de lo posible, ítems distintos. En esos casos, un algoritmo adecuado de selección deberá incluir restricciones en la tasa de exposición de los ítems (por ejemplo, que cada ítem no sea administrado en más del 20% de los tests) y otras restricciones, para garantizar un adecuado muestreo de contenidos. En los últimos años ha habido mucha investigación sobre el problema de control de la exposición en los TAIs, que consiste en lo siguiente. Un punto débil de los TAIs es que una parte importante del banco de ítems, en ocasiones hasta un 80% de los ítems disponibles (Hornke, 2000), no se administra nunca. Para hacer un buen test, se elabora un banco de, por ejemplo, 500 ítems, muy cuidado, se estudian todos ellos minuciosamente, se eliminan los ítems defectuosos... y, después, el nuevo TAI es tan eficaz (al presentar solo los ítems buenísimos) que ¡400 de los 500 ítems del banco no se administran nunca! Para resolver este problema surgen los procedimientos de control de la exposición, que hacen posible la administración de muchos más ítems del banco y reducen la tasa de exposición de los que se administrarían en todos o en casi todos los tests. Los procedimientos de control consiguen estos objetivos minimizando la pérdida de precisión (Revuelta y Ponsoda, 1998). Los procedimientos siguen distintas estrategias para evitar que el ítem más informativo sea el ítem seleccionado. Por ejemplo, el método proporcional (Barrada, Ponsoda, Olea y Abad, 2010) en vez de elegir el ítem más informativo, obtiene para cada ítem el cociente entre la información del ítem y la información del banco y elige cada ítem al azar, siendo dichas proporciones las probabilidades de ser seleccionado.

Los TAIs, dada su condición adaptativa, tienen importantes ventajas adicionales a las de cualquier test informatizado (Olea et al., 2010): La primera es que mejoran la seguridad del test, ya que gran parte de los ítems que se presentan a los evaluados son diferentes. La segunda ventaja es que reducen el tiempo de aplicación (a veces a menos de la mitad), ya que consiguen niveles similares de precisión que los tests convencionales con un número apreciablemente menor de ítems. En caso de que el TAI tenga el mismo número de ítems que un test convencional, realizan estimaciones más precisas. Por último, la tercera ventaja es que resultan especialmente indicados cuando el test ha de aplicarse a personas con niveles muy heterogéneos en su nivel del rasgo. En estos casos, los test no adaptativos suelen contener muchos ítems de dificultad media y miden con poca precisión a las personas con niveles altos o bajos.

Los TAIs tienen también sus desventajas. Una es que su eficiencia depende del tamaño y calidad del banco, y tener un gran y buen banco supone un coste considerable. Una segunda desventaja es la imposibilidad de revisar las respuestas. En la mayoría de los TAIs no se permite al evaluando que cambie la respuesta a un ítem ya respondido, algo que no plantea problemas en los tests convencionales. Una tercera desventaja es que, a las complicacio-

nes técnicas de la TRI, hay que añadir las relativas al equipo y programas informáticos con los que se administra el test.

El crecimiento de los TAIs ha sido exponencial (Wainer, 2000) desde 1990 al 2000. Más de 20 programas de evaluación que aplican TAIs aparecen listados en la web (http://www.psych.umn.edu/psylabs/catcentral/). Olea *et al.* (2010) exponen varios desarrollados dentro y fuera de España. El principal campo de aplicación es la evaluación educativa, la certificación y acreditación. Fetzer, Dainis, Lambert y Meade (2008) afirman que el número de TAIs que se administran al año en el mundo está entre 4 y 6 millones. Más información sobre TAIs puede encontrarse en Olea, Ponsoda y Prieto (1999), Olea y Ponsoda (2003) y Renom (1993).

Otros tipos de tests informatizados adaptativos

Existen TAIs con características diferentes de las expuestas, como los TAIs basados en testlests y los TAIs multietapa (Drasgow et al., 2006). En los últimos tiempos están ganando relevancia los TAIs basados en *testlests*. La mayoría de los modelos TRI requieren el supuesto de independencia local; es decir, el resultado en un ítem debe depender solo del nivel de conocimiento de quien responde y de las características del ítem, y no de si se ha acertado o no algún otro ítem. En determinadas situaciones se ha comprobado que tal supuesto no se cumple. Por ejemplo, cuando presentamos un fragmento de texto y hacemos varias preguntas para evaluar su comprensión, suele aparecer la dependencia local. Si el fragmento ha sido entendido, lo más probable es que se acierten todas o casi todas las preguntas; mientras que si no lo ha sido, lo normal es que se fallen. Wainer y Kiely (1987) introdujeron el término *testlet* para referirse de forma genérica a los elementos que forman el test, que pueden ser un ítem tradicional o el tipo de ítem comentado antes, formado por un fragmento de texto y varios ítems sobre el mismo fragmento. El *testlet* es en ambos casos el elemento que forma el test, al ser el elemento que se puntúa y a partir del que se obtiene, juntando las puntuaciones en los distintos *testlets,* la puntuación en el test de cada evaluado.

Se han propuesto tests adaptativos cuyo elemento básico es el *testlet*. Funcionan como los TAIs descritos anteriormente, pero a partir de un banco de testlests. En cada momento el TAI elige el más apropiado y a continuación lo presenta. Por lo general, presenta todos los ítems del *testlet*, pero se ha explorado si resulta más eficiente presentar también adaptativamente dichos ítems (Keng, 2008), siendo los primeros resultados prometedores. En este caso, el TAI elige el mejor *testlet* y a continuación, en vez de presentar todos sus ítems, presenta adaptativamente un cierto número de ellos.

Resulta a veces difícil en un TAI tener la seguridad de que cumple ciertas exigencias que se quiere que el test satisfaga, como un adecuado muestreo de los contenidos a evaluar. En los últimos tiempos han recibido renovado interés los tests multietapa, que en adaptabilidad están a medio camino entre el test tradicional y los TAIs. En estos tests se establecen unas pocas posibles rutas a seguir por el estudiante, de modo que es posible asegurarse

que todas ellas cumplen los requisitos que queremos satisfacer. Estos tests aúnan las ventajas de los TAIs y el juicio de los expertos que seleccionan los testlets que integran cada ruta (Hendrickson, 2007). Los elementos de un test multietapa basado en testlets son los testlets (o *módulos*), las *etapas* y los *paneles*. En la primera etapa suele haber un único módulo. El estudiante responde y en función de su resultado se le muestra el módulo más adecuado de la segunda etapa; a continuación, de nuevo en función del resultado, se le muestra el más adecuado de la tercera etapa; y así sucesivamente. Al conjunto de etapas, módulos y reglas de bifurcación que se ofrecen al estudiante se llama panel. Lo normal es tener preparados varios paneles, para evitar que los estudiantes se enfrenten a muchos módulos idénticos, y asignar a cada estudiante un panel al azar.

Conclusiones

Las nuevas tecnologías, y en especial la informatización, están modificando prácticamente todas las fases de la evaluación mediante tests: el ensamblaje del test, su administración, la calibración, la corrección de las respuestas, el análisis de los resultados... incluso la investigación, con la progresiva mayor relevancia de la simulación.

En las páginas precedentes hemos detallado algunos de estos impactos, pero no todos. Resulta evidente que mejorando nuestras evaluaciones incrementamos nuestra calidad docente. En la introducción, se ha expuesto que, como no podía ser de otra manera, la precisión de los exámenes universitarios de opción múltiple es mejorable. Dado que muchos de esos exámenes utilizan sólo texto en sus ítems, se nos ocurrió dar un primer paso de lo que podría ser un sistema automático de verificación del cumplimiento de las recomendaciones dadas por Haladyna, Downing y Rodríguez (2002) para la correcta redacción de ítems. García, Ponsoda y Sierra (2011) elaboraron 39 criterios que recaban información relacionada con algunas de las anteriores recomendaciones. Por ejemplo, el primer criterio (número de palabras que tiene el ítem) se considera un posible indicador de las recomendaciones 13 (minimizar la cantidad de material a leer en el ítem) y 16 (reducir al mínimo la información verbal no necesaria). Otros criterios reflejan lo similares que son el enunciado y las opciones, entre las opciones entre sí... García et al. (2011) han elaborado un programa de ordenador que lee el examen y obtiene para cada ítem sus valores en los 39 criterios. El objetivo final es disponer de un sistema que automáticamente pueda avisar al profesor, antes de aplicar el examen, de sus fallos de redacción. En la siguiente fase de este proyecto estamos aplicando las técnicas de *análisis semántico latente*, pues se ha comprobado que responden a los ítems de opción múltiple de forma bastante similar a como lo hacen los propios estudiantes (Lifchitz, Jhean-Larose y Denhière, 2009). Estas técnicas podrían sugerir al profesor, antes de la aplicación del examen, que revise alguna de las respuestas que ha especificado como correcta.

La lista de impactos en medición debido a las nuevas tecnologías no acaba aquí. Un asunto de considerable relevancia es el de la invarianza métrica, o estudio de los posibles cambios en la dimen-

sión o dimensiones a medir como resultado del cambio del medio u otras condiciones de administración del test. Por ejemplo, se han hecho muchos estudios para comprobar si cambian o no las dimensiones que el test mide cuando se presenta en modo informatizado o en modo lápiz y papel. Con el incremento de la administración de tests por internet, se está estudiando si cambian las dimensiones cuando el test se administra en condiciones controladas (con vigilantes) y en condiciones no controladas.

La informatización y las nuevas tecnologías han posibilitado nuevos modos de evaluación, como las *simulaciones* y los *juegos serios* (Winkley, 2010). Las *simulaciones* se proponen cuando la interacción del estudiante con la realidad resulta desaconsejable por ser peligrosa (aprender a pilotar un avión, por ejemplo), demasiado lenta o cara. Los *juegos serios* son juegos pensados para el aprendizaje. El juego serio *FloodSim* pretende hacer al jugador consciente de las muchas variables que intervienen en las inundaciones, del gasto que se genera y de la importancia del problema donde el riesgo de inundaciones es alto. El jugador se enfrenta a una simulación y ha de decidir cuánto gastar en construir defensas, dónde construir las nuevas casas, cómo informar a los ciudadanos de los riesgos... El jugador toma las decisiones, comprueba sus consecuencias y puede cambiarlas. En las simulaciones se generan varias medidas que son las puntuaciones del estudiante en las variables de interés. Sin embargo, muchas de estas medidas no pueden ser tratadas con los modelos de las teorías de los tests disponibles.

Pese a que la tecnología viene impulsando la medición educativa en la dirección correcta, conviene recordar que existe un riesgo con los nuevos ítems y tests. La innovación no debe ser un objetivo en sí misma y pierde su sentido si no consigue mejorar la evaluación. Existe el riesgo de que tras algunos nuevos ítems y tests haya solo un mero cambio de apariencia para hacer el test más atractivo (Winckley, 2010). Bartram (2011) revisa algunos de las recientes innovaciones en los tests y afirma que la tecnología, por avanzada que sea, no podrá nunca hacer innecesarios los estudios de validez.

Dirección para la correspondencia: Vicente Ponsoda. Departamento de Psicología Social y Metodología. Facultad de Psicología. C. Iván Pavlov, 6. 28049 - Madrid

Fecha de recepción de la versión definitiva de este artículo: 2.VII.2011

Bibliografía

ABAD, F. J.; OLEA, J.; PONSODA, V. y GARCÍA, C. (2011) *Medición en ciencias sociales y de la salud* (Madrid, Síntesis).

ADAMS, R. J.; WILSON, M. y WANG, W. C. (1997) The Multidimensional Random Coefficients Multinomial Logit Model, *Applied Psychological Measurement*, 21, pp. 1-23.

AERA, APA y NCME (1999) *Standards for Education and Psychological Testing* (Washington, AERA).

BARRADA, J. R.; PONSODA, V.; OLEA, J. y ABAD, F. J. (2010) A Method for the Comparison of Item Selection Rules in Computerized Adaptive Testing, *Applied Psychological Measurement*, 34:6, pp. 438-452.

BARTRAM, D. (2011) *Testing Goes Global*, Universidad Autónoma de Madrid. Seminario 3 de la cátedra MAP.

BEJAR, I. I. (1993) A Generative Approach to Psychological and Educational Measurement, en FREDERIKSEN, N.; MISLEVY, R. J. y BEJAR. I. I. (eds.) *Test Theory for a New Ge-*

neration of Tests (Hillsdale, LEA), pp. 323-359.

DIAO, Q. y VAN DER LINDEN, W. J. (2011) Automated Test Assembly Using lp_Slve Version 5.5 in R, *Computer Software Review,* 25:5, pp. 398-409.

DRASGOW, F.; LUECHT, R. M. y BENNET, R. A. (2006) Technology and Testing, en BRENNAN, R. L. (ed.) *Educational Measurement* (Westport, Praeger).

ELOSUA, P. (2011) *Introducción al entorno R* (Bilbao, Universidad del País Vasco).

EMBRETSON, S. E. (1998) A Cognitive Design System Approach to Generating Valid Tests: Application to Abstract Reasoning, *Psychological methods,* 3, pp. 380-396.

FAIRTEST. Ver http://fairtest.org/gender-bias-college-admissions-tests.

FETZER, M.; DAINIS, A.; LAMBERT, S. y MEADE, A. (2008) *Computer Adaptive Testing in an Employment Context,* Previsor, White Paper, April.

GARCIA, C.; PONSODA, V. y SIERRA, A. (2011) Prediction of Item Psychometric Indices from Item Characteristics Automatically Extracted from the Stem and Option Text, *International Journal of Continuing Engineering Education and Life-Long Learning,* 21:2/3, pp. 210-221.

GÓMEZ-BENITO, J.; HIDALGO, M. D. y GUILERA, G. (2010) Los sesgos de los instrumentos de medición. Tests justos, *Papeles del Psicólogo,* 31:1, pp. 75-84.

GONZALEZ-ESPADA, W. J. (2009) Detecting Gender Bias Thorough Test Item Analysis, *The Physics Teacher,* 47, pp. 175-179.

HALADYNA, T. M.; DOWING, S. M. y RODRIGUEZ, M. C. (2002) A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment, *Applied Measurement in Education,* 15:3, pp. 309-334.

HIGGINGS, D.; XI, X.; ZECHNER, K. y WILLIAMSON, D. (2011) A Three-Stage Approach to the Automated Scoring of Spontaneous Spoken Responses, *Computer Speech and Language,* 25, pp. 282-306.

HENDRICKSON, A. (2007) An NCME Instructional Module on Multistage Testing, *Educational Measurement: Issues and Practice,* 26:2, pp. 44-52.

HORNKE, L. F. (2000) Item Response Times in Computerized Adaptive Testing, *Psicologica,* 21, pp. 175-189.

INSTITUTE OF EDUCATION SCIENCES. Ver http://nces.ed.gov/nationsreportcard/itemmaps/

KENG, L. (2008) *A Comparison of the Performance of Testlet-Based Computer AdaptiveTests and Multistage Tests,* Doctoral dissertation, University of Texas.

LIFCHITZ, A.; JHEAN-LAROSE, S. y DENHIERE, G. (2009) Effect of Tuned Parameters on an LSA Multiple Choice Questions Answering Model, *Behavior Research Methods*, 41:4, pp. 1201-1209.

LIVINGSTON, S. A. (2009) Constructed-Response Test Questions: Why We Use Them; How We Score Them, *R&D Connections,* 11, September.

MARTÍNEZ ARIAS, R. (2010) La evaluación del desempeño, *Papeles del Psicólogo,* 31:1, pp. 85-96.

MARTÍNEZ ARIAS, M. R.; HERNÁNDEZ LLOREDA, M. J. y HERNÁNDEZ LLOREDA, M. V. (2006) *Psicometría* (Madrid, Alianza).

MUÑIZ, J. (1997) *Introducción a la teoría de respuesta a los ítems* (Madrid, Pirámide).

MUÑIZ, J. (2000) *Teoría clásica de los tests* (Madrid, Pirámide).

OLEA, J.; ABAD, F. J. y BARRADA, J. R. (2010) Tests informatizados y otros nuevos tipos de tests, *Papeles del Psicólogo,* 31:1, pp. 97-107.

OLEA, J. y PONSODA, V. (2003) *Tests adaptativos informatizados* (Madrid, UNED).

OLEA, J.; PONSODA, V. y PRIETO, G. (1999) *Tests informatizados. Fundamentos y aplicaciones* (Madrid, Pirámide).

RENOM, J. (1993) *Tests adaptativos computerizados* (Barcelona, PPU).

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

REVUELTA, J. y PONSODA, V. (1998) A Comparison of Item Ex-

posure Control Methods in Computerized Adaptive Testing, *Journal of Educational Measurement,* 35, pp. 311-327.

REVUELTA, J. y PONSODA, V. (2003) *Simulación de modelos estadísticos en ciencias sociales* (Madrid, La Muralla).

SCALISE, K. (2010) *Innovative Item Types: New Results on Intermediate Constraint Questions and Tasks for Computer-Based Testing Using NUI Objects,* NCME Annual Meeting, Denver.

SIERRA, A.; ALONSO, E.; CHAVERO, J.; GARCIA, V.; PONSODA, V. y PULIDO, E. (2010) *Análisis cuantitativo de la calidad de las pruebas de opción múltiple en la UAM* (Madrid, Universidad Autónoma de Madrid).

VAN DER LINDEN, W. J. (2005) *Linear Models for Optimal Assembly* (New York, Springer).

WAINER, H. (2000) CATs: Whither and Whence, *Psicologica,* 21, pp. 121-133.

WAINER, H. y KIELY, G. (1987) Item clusters and computerized adaptive testing: A case for testlets, *Journal of Educational Measurement,* 24, pp. 185-202.

WINKLEY, J. (2010) *E_assessment and Innovation,* Becta, Technical Report.

WISE, L.; DRASGOW, F.; HANSEN, J-I. C.; SACKETT, P-R. y TIPPINGS, N. T. (2010) *Revision of the Standards for Education and Psychological Testing*. SIOP Annual Meeting, Atlanta.

Resumen:
Nuevas tecnologías y medición educativa

El artículo revisa algunos cambios que la informatización ha producido en la medición educativa actual. Todas las fases de la evaluación mediante tests se han visto en mayor o menor medida afectadas por las nuevas tecnologías. Los cambios expresamente comentados son los ítems innovadores, los tests de desempeño, el ensamblaje automático, la generación automática de ítems y varios tipos de tests adaptativos informatizados. La informatización ha introducido cambios también en el tratamiento psicométrico de los datos, con la aparición de programas de ordenador para tareas como la calibración, la determinación del sesgo de los ítems y tests, el estudio de los supuestos y del ajuste de los modelos... y ha cambiado también la manera de investigar, siendo más y más frecuente la aplicación del método de Monte Carlo. Se concluye que los cambios están impulsando la medición educativa en la dirección correcta. No obstante, existe el riesgo de olvidar que estas innovaciones pierden su verdadero sentido si no consiguen mejorar la calidad métrica de la evaluación.

Descriptores: ítems innovadores, tests de desempeño, ensamblaje automático, tests adaptativos informatizados.

Summary:
New technologies and educational measurement

The effects the new technologies have had on current educational measurement are reviewed. They relate to most of the tasks involved in an educational measurement process. The specific topics considered are: innovative items, performance assessment, automatic test assembly, automatic item generation and different types of computerized adaptive testing. Technology has also changed the psychometric analysis of item and test data, as there are now available computer programs for most psychometric tasks, as item calibration, differential item and test functioning, the study of model fit and mo-

del assumptions... Technology has also affected the way to do research, as the use of Monte Carlo methods is becoming more and more common. It is concluded that in general technology is pushing educational measurement in the right direction. However, it should be kept in mind that these innovations may lose their true meaning if they do not really improve the metric quality of the assessment.

Key Words: innovative items, performance test, automatic test assembly, computerized adaptive testing.