

LA EVALUACIÓN DEL DESEMPEÑO PERFORMANCE ASSESSMENT

Rosario Martínez Arias
Universidad Complutense de Madrid

Las prácticas de evaluación han evolucionado desde el uso casi exclusivo de tests formados por ítems de elección múltiple a la combinación de formatos múltiples, incluyendo tareas de desempeño. El objetivo del artículo es proporcionar una visión del concepto, diseño, uso y características psicométricas de los tests de desempeño. Comienza con el concepto y la justificación de su uso. En la sección 2 se presentan los principales usos actuales de este tipo de tests. La sección 3 describe algunos aspectos relativos al diseño y puntuación. La sección 4 muestra algunas cuestiones relativas a las características psicométricas. La sección 5 concluye con una valoración de los tests de desempeño, presentando sus principales fuerzas y debilidades, así como las necesidades de futuras investigaciones. Se necesita un esfuerzo continuado en modelos y métodos de medida que permitan mejorar la generalizabilidad y las evidencias de validez de los tests de desempeño.

Palabras clave: Test de desempeño, Centros de evaluación, Guías de puntuación, Generalizabilidad, Evidencias de validez.

Assessment practices have gradually shifted from almost exclusively objectively score multiple-choice test items to the use of a mixture of formats, including performance assessments. The purpose of this article is to provide an overview on the concept, design, use and psychometric characteristics of performance assessment. The article is divided in five sections. It begins with the concept and rationale for the use of performance assessment. Section 2 presents the main uses of these tests. Section 3 describes some questions related to the design and scoring. Section 4 shows some issues related to psychometric characteristics. Section 5 concludes with an evaluation of performance assessment tests indicating the main strengths and weaknesses and needs of future research. Continued work is needed on measurement models and methods that can improve the generalizability and the evidences of validity of performance assessments. The computers can contribute to practical issues.

Key words: Performance assessment, Assessment centers, Scoring rubrics, Generalizability, Evidences of validity.

EL CONCEPTO DE TEST DE DESEMPEÑO

Muchas personas, incluso profesionales de la psicología, consideran el test estandarizado como sinónimo de test de elección múltiple o de respuesta construida única. Esta consideración está justificada ya que estos formatos han dominado el campo de los tests de inteligencia, aptitudes y rendimiento académico durante muchos años y por buenas razones, relacionadas sobre todo con la cobertura de contenido y las facilidades para la corrección y puntuación. No obstante, bajo la etiqueta de test es-

tandarizado caben otros formatos que cumplen con todos los requisitos exigibles a un test y que pueden mostrar adecuadas propiedades psicométricas. Entre ellos se encuentran los que aquí denominamos "tests de desempeño"¹ (*performance assessment*), de uso cada vez más frecuente en la evaluación psicológica y educativa.

En la Tabla 1 se presenta una clasificación de los distintos tipos de formato, que pueden adoptar los tests estandarizados y que se pueden graduar a lo largo de varios continuos (Gronlund, 2006).

Dentro de los formatos anteriores, suelen considerarse evaluación del desempeño los ensayos, proyectos, simulaciones y muestras de trabajo. Puede observarse que estos formatos se encuentran más próximos a los extremos caracterizados como de mayor autenticidad, complejidad cognitiva, cobertura en profundidad y respuesta estructurada por el propio sujeto. También se caracterizan por un mayor costo.

Dada la diversidad de formatos que pueden adoptar, se presenta a continuación una definición integradora que permita recoger su diversidad: "los tests de desempeño son procedimientos estandarizados de evaluación en los que se demanda de los sujetos que lleven a cabo tareas o procesos en los que demuestren su capacidad para aplicar conocimientos y destrezas a acciones en si-

Correspondencia: Rosario Martínez Arias. Departamento de Metodología de las Ciencias del Comportamiento. Universidad Complutense de Madrid. E-mail: rmnez.arias@psi.ucm.es

¹ Se ha traducido la expresión inglesa "performance assessment" por "evaluación del desempeño" o tests de desempeño. El término procede de la evaluación educativa y de las certificaciones profesionales, no obstante, en psicología también se utilizan estos tests desde hace mucho tiempo, especialmente en el ámbito de la selección de personal. Aunque no se utiliza el término "performance assessment", las tareas de simulación y muestras de trabajo utilizadas en los centros de evaluación (*assessment centers*) muestran todas las características de este tipo de tests: demandan respuestas que ponen el acento en la actuación del sujeto y que requieren métodos sistemáticos para su valoración. Con la llegada de las nuevas tecnologías su uso se está extendiendo a otros ámbitos como la psicología clínica y la neuropsicología.

tuaciones simuladas o de la vida real”.

Estos tests pueden ser tan diversos como escribir un ensayo, interpretar una composición musical, hacer una presentación oral, diagnosticar a un paciente estandarizado, planificar las actividades del día o proponer una solución a un problema empresarial. En todos los casos el sujeto debe producir algo durante un período de tiempo y se evalúan los procesos o productos con relación a criterios de rendimiento establecidos.

La definición es integradora en el sentido de que recoge los dos grandes grupos en los que suelen dividirse las definiciones: las que ponen el acento en el formato de la respuesta y las que lo ponen en la semejanza entre la respuesta demandada y el criterio de interés (Palm, 2008). En este grupo, la mayor parte ponen el acento en la actuación del examinado (Stiggins, 1987).

Algunas van más allá del formato de respuesta, insistiendo en sus características de *autenticidad* y de *simulación* de la situación criterio. Así, Fitzpatrick y Morrison (1971) los definen como “aquellos tests en los que se simula una situación criterio con más fidelidad y globalidad que en los usuales tests de papel y lápiz” (p.268). Kane, Crooks y Cohen (1999) destacan que “suponen una muestra de la actuación del sujeto en algún dominio, interpretando las puntuaciones resultantes en términos del rendimiento típico o esperado en dicho dominio.....siendo su característica definitoria la estrecha semejanza entre el tipo de actuación observada y la de interés” (p.7). En esta misma línea se encuentra la definición de los *Standards for Educational and Psychological Tests* (American Educational Research Association (AERA), American Psychological Association (APA) y National Council on Measurement in Education (NCME), 1999), que consideran que “las evaluaciones del desempeño *emulan* el contexto o las condiciones en las que se aplican los conocimientos y destrezas que se intenta evaluar” (p.137).

Esta insistencia en la emulación del rendimiento de interés lleva a confusión con la denominada *evaluación au-*

téntica (Wiggins,1989). Ésta comparte muchas características con los tests de desempeño y es una de sus formas, pero implica otros aspectos que van más allá de los exigidos a estos tests.

Con frecuencia se destaca también la complejidad cognitiva, ya que exigen de los sujetos la utilización de estrategias de orden superior, como planificar, estructurar la tarea, obtener información, construir las respuestas y explicar el proceso, integrando conocimientos e información (Ryan, 2006).

Los tests de desempeño suelen clasificarse por lo que evalúan y en este sentido suele hablarse de *productos* (*products*) o resultados de la tarea y *desempeños* (*performances*), que son los procesos que sigue el examinado para llegar a la solución. Ejemplos típicos del primer tipo son los ensayos escritos, informes de laboratorio, actuaciones artísticas, etc. Entre los segundos se encuentran las presentaciones orales y las demostraciones. La mayor parte de las veces suponen una combinación de procesos y productos.

USOS DE LOS TESTS DE DESEMPEÑO

Los tests de desempeño no representan algo nuevo; Madaus y O’Dwyer (1999) establecen sus orígenes en el 210 aC durante la Dinastía Han en China. Evaluaciones similares se utilizaron en los gremios durante la Edad Media y en las Universidades para la valoración de los estudiantes. En la psicología del trabajo tienen una larga tradición en el ejército y desde hace más de 60 años se emplean en los denominados *Centros de Evaluación* (*Assessment Centers*), hoy conocidos como *Centros de Evaluación y Desarrollo* (Thorton y Rupp, 2006), en los que se emplean muestras de trabajo y ejercicios simulados para evaluar a los sujetos en competencias difíciles de medir con tests convencionales. Su uso se remonta a 1942 en la *War Office Selection Boards* del Reino Unido para la selección de altos mandos y pronto se extendieron a los Estados Unidos y otros países, especialmente

TABLA 1
CONTINUOS DE FORMATOS DE TESTS

Muestras de Trabajo	Simulaciones	Proyectos	Ensayos	Respuesta corta	Elección Múltiple	Verdadero/ Falso
Más auténtica	←—————→					Menos auténtica
Cognitivamente. más compleja	←—————→					Cognitivamente menos compleja
Cobertura en profundidad	←—————→					Cobertura en contenido
Respuesta estructurada por el sujeto	←—————→					Respuesta estructurada por el test
Mayor costo	←—————→					Menor costo

de lengua alemana. Desde los años cincuenta se emplean para la selección de puestos directivos, aunque en la actualidad su uso se está generalizando a muchos tipos de puestos (Thorton y Rupp, 2006).

En la evaluación educativa, las fuertes críticas de los años sesenta y setenta al formato de elección múltiple llevaron a la inclusión de tareas de desempeño en la evaluación. Durante los años noventa se pasa del uso exclusivo de formatos de elección múltiple a formatos mixtos que incluyen tareas de desempeño, como ensayos escritos, secuencias de solución de problemas, presentaciones orales e incluso *portafolios* de los estudiantes (Hambleton, 2000). Las razones del cambio son diversas, pero básicamente tienen que ver con las limitaciones de los tests de elección múltiple para lograr algunos objetivos educativos: 1) evaluar habilidades de nivel cognitivo superior; 2) evaluar destrezas para el aprendizaje a lo largo de la vida (pensamiento independiente, persistencia, flexibilidad,...); 3) evaluación de las estrategias de resolución de problemas y dificultades; 4) alineación de destrezas y habilidades con las competencias importantes para la vida, junto con contextos realistas y 5) integrar la evaluación con la instrucción de acuerdo con las teorías del aprendizaje y la psicología cognitiva. Estos objetivos están incluidos en las reformas educativas en las que se pone el acento en la enseñanza de habilidades cognitivas superiores (Linn, 1993a) y en la unión entre evaluación e instrucción, por considerar la evaluación como un instrumento valioso para la mejora de la instrucción y del aprendizaje (Frederiksen y Collins, 1989; Stiggins, 1987). Intentan superar las denunciadas reducciones del currículo generadas por los tests de elección múltiple, en la creencia de que la evaluación determina lo que los profesores enseñan y lo que los estudiantes aprenden (Wiggins, 1989).

Los avances experimentados por la psicología cognitiva fueron un importante detonante para la inclusión de tareas de desempeño en las evaluaciones. En 1998 el *Board on Testing and Assessment* del *National Research Council* (NRC) formó un comité de 18 expertos presidido por Pellegrino y Glaser para establecer un puente entre los avances de la psicología cognitiva y los métodos de medición educativa. El producto final fue un excelente libro: *Knowing What Students Know: The Science and Design of Educational Assessment* (NRC, 2001). En el texto se destacan las limitaciones de los tests tradiciona-

les para captar los conocimientos y las destrezas complejas exigidas por los nuevos estándares de rendimiento y la escasa validez de las inferencias derivadas de sus puntuaciones. El comité desarrolló un marco teórico para la evaluación, *el triángulo de la evaluación*, basado en la idea de que la evaluación es un *proceso de razonamiento desde la evidencia* (Mislevy, 2006; Mislevy, Steinberg y Almond, 2002; Mislevy, Wilson, Ercikan y Chudowsky, 2003), que se apoya en tres pilares: a) un modelo de representación del conocimiento y desarrollo de las competencias, b) tareas o situaciones que permitan observar el desempeño de los estudiantes y c) métodos de interpretación para extraer inferencias.

Por otra parte, la llegada de los ordenadores abrió la posibilidad de usar nuevos formatos de ítem y de respuesta, facilitando tanto la administración como la puntuación de estas tareas (Drasgow, Luecht y Bennett, 2006; Zenisky y Sireci, 2002).

En la actualidad los tests de desempeño están presentes en la mayor parte de las evaluaciones a gran escala, generalmente acompañados de ítems de formato estructurado. En Estados Unidos comenzaron a incluirse en el *National Assessment of Educational Progress* (NAEP) durante los años 90 y hoy muchos estados evalúan el desempeño en sus programas anuales de tests. También se incluyen en todas las evaluaciones internacionales a gran escala, como *Trends in International Mathematics and Science Study*, TIMSS (Arora, Foy, Martin y Mullis, 2009) y en el programa PISA (OECD, 2007). En España se han incorporado a las pruebas de diagnóstico desarrolladas por el Instituto de Evaluación.

En el ámbito del trabajo, las evaluaciones del desempeño tienen una fuerte representación en las certificaciones profesionales, especialmente para el ejercicio de la medicina y la abogacía. Como ejemplo de las primeras se encuentran los tests de *United States Medical Licensure Examination* (USMLE, 2009). Un ejemplo de las segundas es el *Multistate Performance Test*, utilizado en 30 estados de Estados Unidos (National Conference of Bar Examiners & American Bar Association, 2005).²

Una gran parte de las tareas de estos tests son similares a las que se utilizan en los *centros de evaluación* para la selección de personal. En estos sistemas el tipo de tareas adopta múltiples formas, aunque las más comunes son las siguientes: tests en la bandeja, *role-play* en interacciones, análisis de casos escritos de la organización,

² Descripciones detalladas de las evaluaciones del desempeño utilizadas en diversas acreditaciones profesionales pueden encontrarse en Johnson, Penny y Gordon (2009).

presentaciones orales, liderazgo en discusiones de grupo, búsqueda de hechos relevantes a partir de presentaciones orales, juegos de empresa y combinaciones de varias tareas o ejercicios. Una descripción de las tareas de los centros de evaluación puede consultarse en Thornton y Rupp (2006).

Las conductas de los sujetos se evalúan en *dimensiones* relevantes para los puestos de trabajo. Su número y tipo difiere según el objetivo del centro de evaluación (Thornton y Rupp, 2006). Algunas son comunes a la mayor parte de los centros y similares a las de las certificaciones: solución de problemas, comunicación oral, liderazgo, gestión de conflictos, búsqueda de información, planificación y organización, adaptabilidad cultural, generación de soluciones, usos de los recursos,... (Arthur, Day, McNelly y Edens, 2003; Brummel, Ruth y Spain, 2009).

DESARROLLO, ADMINISTRACIÓN Y PUNTUACIÓN DE LOS TESTS DE DESEMPEÑO

Los tests de desempeño deben asegurar que los ejercicios o tareas estén estandarizados, sean válidos, fiables, equitativos y legalmente defendibles. Para conseguirlo deben seguir en su desarrollo los estándares y guías para la construcción y uso de los tests como los *Standards for educational and psychological tests* (AERA et al., 1999). En el caso de los ejercicios de los centros de evaluación, deben cumplir además con algunas guías específicas como los *Principles for the validation and use of personnel selection procedures* (Society for Industrial and Organizational Psychology, 2003) y con las *Guidelines and Ethical Considerations for Assessment Center Operations* (International Task Force on Assessment Center Guidelines (2000).

El proceso de desarrollo comienza con la *definición del marco*, que supone la descripción del constructo o de las tareas, el propósito de la evaluación y las inferencias que se harán con las puntuaciones. El marco conceptual guía el desarrollo de las *especificaciones*, que reflejan el contenido, los procesos, las características psicométricas de las tareas y otra información pertinente para la evaluación. Pueden seguirse dos aproximaciones, centrada en el constructo o en la tarea, aunque se recomienda la primera (Messick, 1994). El constructo guía la adecuada representación del dominio, la selección de las tareas, los criterios para establecer las puntuaciones y la detección de posible varianza irrelevante. Patz (2006) presenta una buena descripción del desarrollo de una evaluación de ciencias. En los centros de evaluación, el

marco de definición de los constructos o competencias se deriva de un riguroso análisis del puesto de trabajo (Thornton y Rupp, 2006).

Para la adecuada estandarización es necesario determinar las condiciones de la administración que permitan la comparabilidad de las puntuaciones (AERA et al., 1999). Se elaboran *Guías* en las que se establecen los tiempos, ítems o tareas de ensayo, equipamiento y materiales, así como instrucciones para la aplicación (Cohen y Wollack, 2006).

La clave del éxito de estos tests y uno de los aspectos más controvertidos es la correcta *asignación de puntuaciones* a las tareas realizadas. Para ello se elaboran las *Guías para la especificación de puntuaciones (scoring rubrics)* en las que se establecen los criterios de valoración de las respuestas junto con un procedimiento para puntuarlas (Clauser, 2000). Deben ser claras, completas e ilustradas con ejemplos de respuestas tipo (Welch, 2006). Su objetivo es obtener puntuaciones consistentes e invariantes a través de evaluadores, tareas, localizaciones, ocasiones y otras condiciones. Combinadas con un adecuado entrenamiento de los evaluadores permiten alcanzar niveles adecuados de fiabilidad.

Hay dos tipos de guías, las *holísticas* o *globales* y las *analíticas*. En las globales los evaluadores emiten un único juicio sobre la calidad del proceso o producto, asignando una puntuación basada en descripciones de *anclaje* de los distintos niveles. En las analíticas, las descripciones del desempeño se separan en partes (aspectos, criterios evaluativos, dimensiones, dominios,...). Además de los epígrafes de las guías, se incluyen respuestas ejemplares para operacionalizar cada uno de los criterios evaluativos, denominados "*anclajes*" o puntos de referencia.

Una guía analítica especifica rasgos o aspectos detallados de las respuestas y el número de puntos que se deben atribuir a cada uno, permitiendo la ponderación. Los distintos rasgos suelen puntuarse por medio de escalas tipo Likert con varios grados. En los centros de evaluación se utiliza un procedimiento similar al que se aplica en las guías analíticas, conocido como *Behaviorally Anchored Rating Scales (BARS)*, que incluyen descripciones ejemplares ("*anchored*") de conductas y permiten valorar cada dimensión en escalas que suelen tener cinco puntos.

Una variación del sistema de puntuaciones analítico es el de las *listas de conductas (checklists)* en las que cada aspecto se valora como Sí o No, según que la actuación esté o no presente. Es el procedimiento habitual en las

acreditaciones médicas y legales y en ocasiones en los centros de evaluación en lugar de las BARS.

Cuando las tareas se basan en las teorías cognitivas del aprendizaje dentro de un dominio las puntuaciones pueden reflejar criterios de progresión en el aprendizaje (Wilson, 2005).

La elección de una u otra forma depende en gran medida del constructo, el propósito de la evaluación, si lo evaluado es un proceso o producto y de las inferencias que se derivarán de las puntuaciones. El número de categorías o de puntos de la escala depende de la facilidad de diferenciación y discriminación. Lane y Stone (2006) indican que tendrá el número suficiente de categorías para diferenciar entre niveles de rendimiento y que no sean tantas que la diferenciación se haga difícil.

Uno de los aspectos más investigados son los méritos relativos de los dos sistemas de puntuaciones, analizados a partir de la fiabilidad entre jueces. Por el momento no hay respuestas claras, no pudiéndose hablar de un procedimiento superior en todas las situaciones. Parece que las guías holísticas se ven más afectadas por las fuentes de sesgo de los evaluadores que las analíticas y que las listas de conducta mejoran el acuerdo entre jueces. Johnson et al., (2009) y Arter y McTighe (2001) recomiendan las holísticas para las tareas relativamente simples, como las incluidas en las evaluaciones a gran escala. Las analíticas son más adecuadas para ejecuciones complejas con múltiples elementos, como es el caso de las licencias y certificaciones y de los centros de evaluación (Welch, 2006).

Los grandes costos derivados de la puntuación de estos tests han motivado el desarrollo de algunos sistemas informatizados para la corrección (Bennett, 2004; Livingston, 2009; Williamson, Mislevy y Bejar, 2006). Para su implementación se identifican un gran número de respuestas tipo evaluadas por calificadores expertos, que representan el rango total de las puntuaciones de la escala y posteriormente se utilizan algoritmos para la obtención de las puntuaciones que emulan a los evaluadores humanos (Williamson et al., 2006).

Otro aspecto esencial es la formación de los evaluadores con los que se intenta llegar a adecuados grados de acuerdo, corrigiendo sus sesgos. Los sesgos más frecuentes se presentan resumidos en la Tabla 2, adaptada de Johnson et al. (2009).

Un procedimiento frecuente de entrenamiento supone la inclusión de protocolos previamente corregidos por expertos, que permiten detectar evaluadores con sesgos y la monitorización con evaluadores experimentados.

CARACTERÍSTICAS PSICOMÉTRICAS DE LOS TESTS DE DESEMPEÑO

Los tests de desempeño deben cumplir con los criterios psicométricos exigibles a todo procedimiento de evaluación (Kane, 2004), para lo que se utilizan los diferentes modelos de la teoría de los tests. Ciertas características específicas exigen el uso de modelos más avanzados que la Teoría Clásica de los Tests (TCT), como la Teoría de la Generalizabilidad (TG) y la Teoría de la Respuesta al Ítem (TRI). Las nuevas concepciones de la validez también llevan a algunas diferencias con respecto a los planteamientos tradicionales (véanse los artículos de Muñiz (2010) sobre las teorías de los tests y de Prieto y Delgado (2010) sobre fiabilidad y validez, en este mismo número).

A continuación se revisan brevemente algunos aspectos psicométricos de los tests de desempeño: el tratamiento de los errores de medida y la consistencia de las puntuaciones (fiabilidad), los procedimientos para obtener estimaciones de la habilidad y las evidencias de validez. Esta clasificación convencional es difícil en estos tests, ya

TABLA 2
SESGOS MÁS FRECUENTES DE LOS EVALUADORES

Tipo de sesgo	Tendencia del evaluador a...
Apariencia	Puntuar fijándose en aspectos que considera importantes
Tendencia central	Asignar puntuaciones en torno al punto medio
Conflicto de estándares	Sus estándares personales no están de acuerdo con los de las guías
Fatiga	Estar afectado por el cansancio
Efecto halo	Atribuir puntuaciones altas por algún aspecto valioso para el evaluador
Escritura del sujeto	Asignar puntuaciones basadas en características del escrito
Arrastre de ítems	Valorar un ítem en función de lo que ha hecho en otros
Lenguaje	Valorar basándose en el lenguaje utilizado por el evaluado
Longitud	Valorar más las respuestas más largas
Indulgencia/severidad	Tendencia a puntuaciones altas/bajas
Repetición	Valorar menos porque ha visto el tópico repetidamente
Prejuicios	Puntuación baja debido a algún aspecto de la respuesta que no le gusta al evaluador
Efectos de otros tests ya corregidos	Puntuar menos una respuesta de lo que dice la guía porque va precedida de respuestas excelentes de otros examinados
Aspecto particular	Poner el acento en un aspecto y darle demasiado peso

que la generalizabilidad de las puntuaciones se trata con frecuencia como uno de los aspectos de la validez (Brennan, 2000a; Kane, Crooks y Cohen, 1999; Miller y Linn, 2000; Messick, 1996).

La fiabilidad y consistencia de las puntuaciones

En ocasiones, la fiabilidad puede tratarse con la TCT (Johnson et al., 2009), pero generalmente se necesita utilizar la TG. Este modelo, sistematizado por Cronbach, Gleser, Nanda y Rajaratnam (1972) tuvo escaso eco en la construcción de tests hasta la llegada de los tests de desempeño en los que su uso se ha generalizado. La TG es una extensión de la TCT que utiliza modelos de Análisis de la Varianza (componentes de la varianza) que permiten estimar simultáneamente efectos de diferentes fuentes de variabilidad o error (*facetas*) sobre las puntuaciones. Las facetas consideradas con mayor frecuencia son las tareas y los evaluadores, aunque en algunos estudios se incluyen ocasiones, administración y formato del test. Permite analizar los efectos principales de cada faceta, así como sus interacciones con el sujeto y entre ellas. La TG contempla dos tipos de estudios, los G o de Generalizabilidad y los D o de Decisión. En los primeros se estima la contribución relativa de cada faceta y de sus interacciones sobre la varianza error y estas estimaciones permiten optimizar el procedimiento de medida, determinando el número óptimo de tareas, evaluadores, etc., en cada aplicación para la reducción del error. En los estudios D se calculan los *coeficientes de generalizabilidad* bajo las condiciones de medida concretas utilizadas; éstos pueden ser de dos tipos, según que las decisiones sean *absolutas* o *relativas*. La posibilidad de descomponer la varianza error en diferentes fuentes es lo que hace imprescindible a la TG en los tests de desempeño. La fiabilidad mejora cuando se realizan estudios para determinar el número de tareas y de evaluadores necesarios.

Razones de espacio nos impiden extendernos más en la descripción de la TG. Un completo tratamiento puede encontrarse en Brennan (2000b). Una exposición resumida se presenta en Martínez Arias, Hernández Lloreda y Hernández Lloreda (2006) y la descripción de sus principales características en el artículo de Prieto y Delgado (2010), en este monográfico.

Las fuentes de error más investigadas son las relativas a *tarea* y *evaluador*. Se ha encontrado que los efectos de las tareas son los más críticos, debido al reducido número que se puede incluir para cada habilidad o competencia, encontrándose baja consistencia entre tareas e interacciones con los sujetos (Lane y Stone, 2006).

El efecto de los evaluadores es importante, tanto como

efecto principal como en interacción con tareas y con sujetos. En evaluaciones de la escritura se han encontrado correlaciones entre jueces muy heterogéneas que van de .33 a .91 (Lane y Stone, 2006) y algo más altas en las certificaciones médicas, con valores entre .50 y .93 (van der Vleuten y Swanson, 1990). En cuanto al tipo de competencia evaluada, se encuentra mayor consistencia en ciencias y matemáticas y menor en escritura (Shavelson, Baxter y Gao, 1993).

En general, puede decirse que la variabilidad de las tareas contribuye más al error que el evaluador en la mayor parte de las materias (Lane y Stone, 2006; Shavelson et al., 1993).

Algunos modelos de la TRI desarrollados en el marco del modelo de Rasch (Adams, Wilson y Wang, 1997) permiten incorporar los efectos del evaluador en la estimación de las puntuaciones.

Aunque tareas y evaluadores son las fuentes de variabilidad más estudiadas, también se han investigado los efectos de otras facetas: ocasiones (tiempos de la medida), formato de la evaluación y comité de calificadoros. Una faceta importante es la ocasión (Cronbach, Linn, Brennan y Haertel, 1997; Fitzpatrick, Ercikan, Yen y Ferrara, 1998), especialmente en evaluaciones periódicas en las que se examinan cambios y puntúan evaluadores diferentes.

La estimación de la competencia o habilidad

Para obtener estimaciones de la competencia o habilidad de los sujetos suelen utilizarse como marco los modelos de la TRI para respuestas politómicas ordenadas (Abad, Ponsoda y Revuelta, 2006). Recientes avances en modelos TRI multidimensionales (*Multidimensional Item Response Theory*, MIRT) permiten tratar con la complejidad de estas evaluaciones, en las que es difícil lograr la unidimensionalidad asumida (Gibbons et al., 2007; Reckase, 2009).

Un problema frecuente es la combinación de diferentes formatos de respuesta en el mismo test. El uso de modelos de TRI con software especializado para modelos politómicos permite obtener estimadores únicos de las habilidades o rasgos en estas condiciones.

En relación con la estimación de las puntuaciones surge el problema de la *equiparación*, cuando se utilizan diferentes conjuntos de ítems, en la misma evaluación o en tiempos distintos para evaluar cambios. Las características de estos tests plantean problemas especiales para la aplicación de las técnicas de equiparación estricta (Kolen y Brennan, 2004), debiendo tratarse con frecuencia mediante formas más débiles como la calibración, pre-

dicción o moderación (Linn, 1993b). Los principales problemas se deben a la frecuente multidimensionalidad, la dificultad de encontrar ítems de anclaje comunes, ser ítems politómicos y la dependencia entre ítems (Muraki, Hombro y Lee, 2000), así como los efectos del evaluador (Kolen y Brennan, 2004). Para su tratamiento se utilizan con frecuencia los modelos multigrupo de la TRI (Bock, Muraki y Pfeifferberger, 1988). Reckase (2009) propone algunos procedimientos en el contexto de los modelos multidimensionales. Un tratamiento reciente de estos problemas puede encontrarse en Dorans, Pommerich y Holland (2007).

Las evidencias de validez de los tests de desempeño

La definición de validez de las puntuaciones de los tests de desempeño es la establecida en los *Standards for Educational and Psychological Tests* (AERA et al., 1999), similar a la de otros tipos de tests estandarizados, con la validez de constructo como concepto unificador. En el artículo de Prieto y Delgado (2010), en este monográfico, se presenta la definición y los tipos de evidencias. En los tests de desempeño se mencionan a veces otros aspectos como la autenticidad, significación para los evaluados (Linn, Baker y Dunbar, 1991) y validez sistémica (Fredericksen y Collins, 1989). Messick (1996) considera estos aspectos dentro de la representación del constructo (autenticidad) y de los aspectos sustantivos y consecuenciales de la validez (significación y validez sistémica).

A continuación se revisan brevemente las evidencias de validez, con algunas consideraciones sobre el sesgo y la equidad, que también podrían tratarse dentro de los aspectos de varianza irrelevante para el constructo o de las consecuencias.

Evidencias de validez de contenido

En los tests de desempeño se presentan con mayor frecuencia que en los convencionales las dos grandes amenazas a la validez de contenido señaladas por Messick (1989, 1996): *infrarrepresentación del constructo* y *varianza irrelevante*. La primera suele deberse al reducido número de ítems que se incluyen. La segunda tiene múltiples fuentes: la elección del tema por los sujetos, la tendencia de los evaluadores a fijarse en aspectos irrelevantes o sesgos (Messick, 1994, 1996; véase Tabla 2), los procedimientos de corrección automatizados (Lane y Stone, 2006) y la motivación de los sujetos, especialmente en situaciones de tests sin consecuencias (DeMars, 2000; O'Neil, Subgure y Baker, 1996).

Evidencias de validez desde los procesos de la respuesta o sustantiva

Messick (1996) destaca "la necesidad de obtener evidencias empíricas de los procesos puestos en juego por los examinados cuando realizan la tarea" (p.9). Dadas las expectativas puestas en estos tests para evaluar procesos cognitivos superiores es preciso justificar que efectivamente lo hacen (Hambleton, 1996; Linn et al., 1991). Por el momento son escasas las investigaciones y los resultados son poco consistentes (Ayala, Shavelson, Shue y Schultz, 2002). Algunos desarrollos inspirados en los modelos del *Latent Trait Logistic Model* (Fischer, 1973), como los de Embretson (1998) y Gorin y Embretson (2006) son prometedores en este sentido. Adams, Wilson y Wang (1997) desarrollaron una versión multidimensional, adecuada para este tipo de tests.

Un marco teórico interesante es el del *triángulo de la evaluación*, mencionado en la sección segunda de este artículo. En las evaluaciones de los aprendizajes educativos, las teorías del desarrollo y progresión del aprendizaje también permiten sustentar este tipo de validez (Briggs, Alonzo, Schwab y Wilson, 2006; Wilson, 2005).

Estructural

Según AERA et al. (1999) "el análisis de la estructura interna de un test puede indicar el grado en que las relaciones entre los ítems y los componentes del test se adecuan al constructo en el que se basan las interpretaciones de las puntuaciones" (p.13).

La evaluación de la dimensionalidad suele establecerse por medio de técnicas de análisis factorial. Hay pocos trabajos publicados sobre la estructura factorial de los tests de desempeño en educación. Las razones son variadas: 1) la complejidad de los estímulos lleva a la recomendación de análisis de contenido y análisis sustantivo (Ackerman, Gierl y Walker, 2003); 2) los esquemas de puntuación pueden tener impacto en la dimensionalidad, llevando en ocasiones a la multidimensionalidad y 3) diferentes puntos de la escala de valoración pueden reflejar diferentes combinaciones de destrezas (Reckase, 1997). Los avances en los modelos multidimensionales de la teoría de la respuesta al ítem (Gibbons et al., 2007; Reckase, 2009) pueden aportar evidencias estructurales. En el ámbito de los centros de evaluación se ha tratado más este aspecto, encontrando resultados contradictorios. Rupp et al. (2006) encuentran evidencias de dimensiones claras, pero otros autores las cuestionan (Lance, 2008).

Externa

Según AERA et al. (1999) "el análisis de las relaciones de las puntuaciones del test con variables externas es otra fuente importante de evidencia de validez" (p.13). Para obtener estas evidencias se examinan los patrones de correlaciones empíricas según las expectativas teóricas o hipótesis del constructo. Messick (1996) pone el acento en la importancia de las evidencias de validez *convergente* y *discriminante* mediante las matrices multimétodo-multirrasgo (MMMR). La "evidencia convergente significa que la medida está coherentemente relacionada con otras medidas del mismo constructo, así como con otras variables con las que debe estar relacionada sobre bases teóricas. La *evidencia discriminante* significa que la medida no debe estar relacionada con otros constructos" (Messick, 1996, p.12).

Debe justificarse que la varianza debida al constructo supera considerablemente a la varianza del método o de las tareas. En el ámbito educativo hay pocos trabajos sobre estas evidencias, pero es un tema muy investigado en los centros de evaluación, encontrándose en general bajas evidencias, ya que suele ser mayor la proporción de varianza ligada al contexto que al constructo (Lance, 2008). No obstante, Rupp, Thorton y Gibbons (2008) atribuyen estos resultados a deficiencias metodológicas en el diseño de las matrices multimétodo-multirrasgo.

En relación con las evidencias referidas a criterios externos, la mayor parte de la investigación se ha realizado en el ámbito de los centros de evaluación. En un meta-análisis, Arthur et al. (2003) encontraron correlaciones entre .25 y .39 según el tipo de competencias. Salgado y Moscoso (2008) revisan la fiabilidad y la validez operativa (corregidos los sesgos de falta de fiabilidad del criterio y restricción del rango) de diversos instrumentos de selección, encontrando para las simulaciones de los centros de evaluación coeficientes de fiabilidad de .70 y validez de .37, siendo esta última inferior a la de otros procedimientos (tests de aptitudes generales y razonamiento, de conocimientos del puesto y entrevista conductual estructurada). Estos datos plantean dudas sobre la utilidad de estos procedimientos frente a otros que son además más económicos.

Las consecuencias del uso de los tests

Este aspecto de la validez de constructo tiene que ver con las consecuencias deseadas y no deseadas del uso de los tests y su impacto sobre la interpretación de las puntuaciones (Messick, 1996). Estas evidencias se han estudiado en el contexto educativo donde son uno de los argumentos más frecuentes para el uso de estos tests. En-

tre las consecuencias positivas para los examinados se incluyen la motivación, el aprendizaje y la aplicación de lo que han aprendido.

Hasta el momento las investigaciones son escasas. Stecher et al. (2000) en una encuesta al profesorado encontraron que 2/3 de los profesores de 4º y 7º grado respondieron que los estándares del estado y los tests de desempeño les influyeron en sus estrategias docentes. El impacto del Maryland State Performance Assessment Program fue examinado por Lane y colaboradores (Lane, Parke y Stone, 2002; Parke, Lane y Stone, 2006) encontrando que tanto los equipos directivos como el profesorado consideraban que la evaluación había llevado a cambios positivos en la instrucción y en las prácticas de evaluación en clase. No obstante, este resultado puede deberse a las consecuencias de la evaluación para las escuelas (rendición de cuentas).

Sesgo y equidad

Generalmente se entiende por sesgo "la validez diferencial de una interpretación de la puntuación de un test para cualquier subgrupo definible de sujetos que responden al test" (Cole y Moss, 1989, p.205). Para evitarlo se recomienda utilizar las técnicas de detección del *funcionamiento diferencial de los ítems*, que permiten detectar tareas o ítems que potencialmente pueden contribuir al sesgo. Para una descripción más detallada, véase el artículo de Gómez Benito, Hidalgo Guilera (2010), en este monográfico.

Se han realizado escasas investigaciones sobre el funcionamiento diferencial de los ítems de los tests de desempeño (Lane y Stone, 2006). La mayor parte de las investigaciones se limitan al análisis de las diferencias entre grupos. En los ensayos escritos se encuentran diferencias entre varones y mujeres, favorables a éstas (Ryan y DeMark, 2002) y diferencias étnicas (Engelhard, Gordon, Walker y Gabrielson, 1994). En estudios más adecuados para el análisis del sesgo, se encontraron diferencias no paralelas entre formatos de desempeño y de elección múltiple (Livingston y Rupp, 2004). Cuando hombres y mujeres muestran resultados similares en los de elección múltiple, las mujeres son superiores en los de respuesta construida; cuando son similares en los de respuesta construida, los varones son superiores a las mujeres en los de elección múltiple.

Se considera que los tests de desempeño pueden mostrar más factores irrelevantes, que favorecen el funcionamiento diferencial (Penfield y Lamm, 2000). Su detección es más difícil, debido a las dificultades ya mencionadas a propósito de la equiparación.

CONCLUSIONES

Los tests de desempeño hoy forman parte del repertorio de las técnicas de evaluación y su uso es creciente. Han generado muchas expectativas por su validez aparente y sus potenciales ventajas: mayor autenticidad mediante la emulación de situaciones reales, posibilidad de medir habilidades y competencias difíciles de evaluar con otros formatos, medición de los procesos además de los productos, su valor educativo y formativo y la detección de los progresos de aprendizaje. Todo ello hace que se consideren imprescindibles en las evaluaciones, normalmente combinados con tests o tareas de formatos tradicionales. Las innovaciones derivadas del uso de nuevas tecnologías ayudan a su aplicación, posibilitando evaluar nuevas competencias y dimensiones.

No obstante, a pesar de sus innegables ventajas y de lo extendido de su uso, presentan todavía numerosos retos a la investigación psicométrica. Sus principales limitaciones son las siguientes:

- ✓ Dificultades para la representación adecuada del dominio por el número limitado de tareas que se pueden incluir.
- ✓ Problemas de generalizabilidad, debidos sobre todo a la varianza debida a las tareas y a la interacción de éstas con sujetos y evaluadores.
- ✓ Inconsistencias y sesgos de los evaluadores, que obligan al desarrollo de Guías de Puntuaciones muy claras, elaboradas y costosas. También requieren costosos procesos de entrenamiento de los evaluadores, para obtener puntuaciones consistentes entre calificadores y ocasiones.
- ✓ Los costos de corrección son altos, requiriendo a veces demasiado tiempo, lo que dificulta el uso formativo de los resultados.
- ✓ La complejidad de las tareas lleva a menudo a estructuras multidimensionales, que dificultan el empleo de los modelos de TRI unidimensionales para la estimación, equiparación o calibración y funcionamiento diferencial.
- ✓ Se requiere más investigación sobre el funcionamiento diferencial de las tareas de desempeño en relación con otros formatos y la influencia de factores motivacionales.
- ✓ Aunque su validez aparente está clara, las diferentes evidencias de validez psicométrica deben investigarse más. Dentro de este bloque son muy importantes ciertos aspectos irrelevantes en los que se fijan los evaluadores, para eliminar su influencia. Las evidencias sustantivas referidas a los procesos deben conti-

nuar investigándose, por medio del uso de modelos de medida que permitan su evaluación, así como la progresión de los aprendizajes. También parece necesario el examen de las evidencias de relaciones con otras variables.

Los desarrollos actuales de modelos psicométricos tanto dentro de la TRI, como en otros marcos (Mislevy, 2006), que permiten la introducción de componentes ligados a los procesos representan un importante avance. Los modelos de la TRI multidimensionales y jerárquicos también permitirán tratar con algunas de las limitaciones anteriores. Se necesita más investigación sobre la combinación adecuada de tareas de formato de elección múltiple y respuestas cortas con tareas de desempeño para optimizar la información.

La introducción de nuevas tecnologías puede mejorar muchas limitaciones. La presentación y respuesta por ordenador mediante tests adaptativos permite reducir considerablemente el tiempo del test, mejorando la representación del dominio. Permiten además el uso de tareas dinámicas, como las que se usan en el diagnóstico de pacientes, mejoran la autenticidad, que puede verse incrementada por la inclusión de múltiples recursos (gráficos, vídeo, audio, materiales de referencia,...). También mejoran el registro de los procesos mediante seguimiento, poniendo de relieve evidencias de validez sustantiva. Las respuestas emitidas a través del ordenador pueden corregir algunos aspectos irrelevantes relacionados con la escritura y forma de exposición. Por otra parte, debe continuarse en el desarrollo de los sistemas automatizados de corrección que reducirán considerablemente los costos.

Finalmente, podríamos preguntarnos si los tests de desempeño deben sustituir a los formatos tradicionales como los de elección múltiple y la respuesta es negativa, ya que hay muchos aspectos de las evaluaciones que pueden evaluarse adecuadamente con estos formatos más económicos en tiempo y dinero. Lo ideal es la combinación adecuada de los distintos tipos.

En este artículo se ha presentado una visión general y limitada de los tests de desempeño. Las personas interesadas en el tema pueden encontrar un tratamiento extenso en las referencias citadas de Johnson et al., (2009) sobre aplicaciones en educación y en acreditaciones; en el libro de Thorton y Rupp (2006) se trata con bastante profundidad el tema de los centros de evaluación. Por último, ejemplos de tareas de tests de desempeño típicas de la evaluación educativa se encuentran entre los ítems hechos públicos del estudio PISA (<http://www.pisa.oecd.org>). Información sobre tests de desempeño en las certificaciones y acreditaciones pueden encontrarse en las páginas web

de la American Board of Pediatric Dentistry (http://www.abdp.org/pamphlets/oral_handbook.pdf), en la National Board of Medical Examiners (http://www.usmle.org/Examinations/step2/step2ck_content.html) y en la ya mencionada de la National Conference of far Examiners (<http://www.ncbex.org/multistate-tests/mbe>).

REFERENCIAS

- Abad, F.J., Ponsoda, V. y Revuelta, J. (2006). *Modelos politómicos de respuesta al ítem*. Madrid: La Muralla.
- Ackerman, T.A., Gierl, M.J. y Walker, C.M. (2003). An NCME instructional module on using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, 22, 37-53.
- Adams, R. J., Wilson, M. R., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21, 1–23.
- Arora, A., Foy, P., Martin, M.O. y Mullis, I.V.S. (Eds.) (2009). *TIMSS Advanced 2008: technical report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- American Educational Research Association, American Psychological Association & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.
- Arter, J. y McTighe, J. (2001). *Scoring rubrics in the classroom*. Thousand Oaks, CA: Corwin Press.
- Arthur, W., Day, E.A., McNelly, T.L. y Edens, P.S. (2003). A meta-analysis of the criterion-related validity of assessment center dimension. *Personnel Psychology*, 56, 125-154.
- Ayala, C.C., Shavelson, R.J., Yue, Y., y Schultz, S.E. (2002). Reasoning dimensions underlying science achievement: the case of performance assessment. *Educational Assessment*, 8, 101-121.
- Bennett, R.E. (2004). *Moving the field forward: Some thoughts on validity and automated scoring* (ETS RM 04-01). Princeton, NJ: Educational Testing Service.
- Bock, R.D., Muraki, E., & Pfeifferberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement*, 25, 275-285.
- Brennan, R. L. (2000a). *Generalizability theory*. New York: Springer-Verlag.
- Brennan, R. (2000b). Performance assessment from the perspective of the generalizability theory. *Applied Psychological Measurement*, 24, 339-353.
- Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment*, 11, 33–63.
- Brummel, B.J., Rupp, D.E. & Spain, S.M. (2009). Constructing parallel simulation exercises for assessment centers and other forms of behavioral assessment. *Personnel Psychology*, 62, 137-170.
- Clauser, B. (2000). Recurrent issues and recent advances in scoring performance assessments. *Applied Psychological Measurement*, 24(4), 310–324.
- Cohen, A. y Wollack, J. (2006). Test administration, security, scoring and reporting. En R. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 355-386). Wesport, CT: American Council on Education/Praeger.
- Cole, N.S. y Moss, P.A. (1989). Bias in test use. En R.L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 201-220).
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioural measurements: Theory of generalizability for scores and profiles*. New York: Wiley.
- Cronbach, L., Linn, R., Brennan, R., & Haertel, E. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement* 57, 373–399.
- DeMars, C. (2000). Test stakes and item format interactions. *Applied Measurement in Education*, 13, 55- 77.
- Dorans, N.J., Pommerich, M. y Holland, P.W. (Eds.). (2007). *Linking and aligning scores and scales*. New York: Springer.
- Drasgow, F., Luecht, R.M. y Bennett, R.E. (2006). Technology and testing. En R.L. Brennan (Ed.), *Educational Measurement*. Pp.471-515.
- Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, 3, 300–396.
- Engelhard, G., Gordon, B., Walker, E.V y Gabrielson, S. (1994). Writing tasks and gender: Influences on writing quality of black and white students. *Journal of Educational Research*, 87, 197-209.
- Fischer, G. H. (1973). Linear logistic test model as an instrument in educational research. *Acta Psychologica*, 37, 359–374.
- Fitzpatrick, R., Ercikan, K., Yen, W.M. y Ferrara, S. (1998). The consistency between raters scoring in different test years. *Applied Measurement in Education*, 11, 95-208.
- Fitzpatrick, R., & Morrison, E. (1971). Performance and product evaluation. In R. Thorndike (Ed.), *Educational measurement* (pp. 237–270). Washington, DC: American Council of Education.

- Frederiksen, J.R. y Collins, A. (1989). A systems approach to educational testing. *Educational Researcher*, 18, 27-32.
- Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D., Kupfer, D. J., Frank, E., Grochocinski, V. J., & Stover, A. (2007a). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement*, 31, 4-19.
- Gómez-Benito, J., Hidalgo, M.D. y Guilera, G. (2010). El sesgo de los instrumentos de medida. *Tests justos. Papeles del Psicólogo*, 31 (1), 75-84.
- Gorin, J.S. y Embretson, S.E. (2006). Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement*, 30, 394-411.
- Hambleton, R.K. (1996). Advances in assessment models, methods, and practices. In D.C. Berliner and R.C. Calfee (Eds.), *Handbook of Educational Psychology* (pp. 899-925). New York: Macmillan.
- Hambleton, R.K. (2000). Advances in performance assessment methodology. *Applied Psychological Measurement*, 24, 291-293.
- International Taskforce on Assessment Center Guidelines (2000). Guidelines and ethical considerations for assessment center operations. *Public Personnel Management*, 29, 315-331.
- Johnson, R.L., Penny, J.A. y Gordon, B. (2009). *Assessing performance: designing, scoring, and validating performance tasks*. New York: Guilford Press.
- Kane, M.T. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives*, 2, 13-170
- Kane, M. T., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, 18 (2), 5-17.
- Kolen, M.J. y Brennan, R.L. (2004). *Test equating, linking, and scaling: Methods and practices* (2nd Ed.). New York: Springer.
- Lance, C.E. (2008). Why assessment centers do not work the way they are supposed. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 1, 84-97.
- Lane, S., Parke, C.S. y Stone, C.A. (2002). The impact of a state performance-based assessment and accountability program on mathematics instruction and student learning: Evidence from survey data and school performance. *Educational Assessment*, 8, 279-315.
- Lane, S. y Stone, C.A. (2006). Performance assessment. En Brennan (Ed), *Educational Measurement*, (4th ed., pp. 387-431). Westport, CT: American Council on Education and Praeger.
- Linn, R.L. (1993a). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis*, 15, 1-16.
- Linn, R. L. (1993b). Linking results of distinct assessments. *Applied Measurement in Education*, 6, 83-102.
- Linn, R.L., Baker, E. L. y Dunbar, S.B. (1991). Complex, performance-based assessment: expectations and validation criteria. *Educational Researcher*, 20 (8), 15-21.
- Livingston, S.A. (2009). *Constructed-response test questions: Why we use them; how to score them*. (R & D Connections, nº 11). Princeton, NJ: Educational Testing Service.
- Livingston, S.A. y Rupp, S.L. (2004). *Performance of men and women on multiple-choice and constructed-response tests for beginning teachers*. (ETS Research Report No.RR.04-48). Princeton, NJ: Educational Testing Service.
- Martínez Arias, R., Hernández Lloreda, MV y Hernández Lloreda, MJ. (2006). *Psicometría*. Madrid: Alianza.
- Madaus, G., & O'Dwyer, L. (1999). A short history of performance assessment. *Phi Delta Kappan*, 80(9), 688-695.
- Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement*, (3rd ed., pp. 13-103). Washington, DC: American Council on Education.
- Messick, S. (1994). The interplay of evidence and consequence in the validation of performance assessments. *Educational Researcher*, 23(2), 13-23.
- Messick, S. (1996). Validity of performance assessments. In G. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 1-18). Washington, DC: National Center for Education Statistics.
- Miller, D.M. y Linn, R.L. (2000). Validation of performance assessments. *Applied Psychological Measurement*, 24, 367-378.
- Mislevy, R.J. (2006). Cognitive psychology and educational assessment. En R.L. Brennan (Ed.), *Educational Measurement*, pp. 257-305.
- Mislevy, R., Steinberg, L., & Almond, R. (2002). Design and analysis in task-based language assessment. *Language Testing*, 19(4), 477-496.
- Mislevy, R., Wilson, M., Ercikan, K., & Chudowsky, N. (2003). Psychometric principles in student assessment. In T. Kellaghan & D. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 489-532). Boston: Kluwer Academic.
- National Research Council (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.
- Muñiz, J. (2010). Las teorías de los tests: Teoría Clásica

- y Teoría de la Respuesta a los Ítems. *Papeles del Psicólogo*, 31,
- Muraki, E., Hombo, C.M. y Lee, Y.W. (2000). Equating and linking of performance assessments. *Applied Psychological Measurement*, 24, 325-33.
- National Conference of Bar Examiners (NCBE) y American Bar Association (ABA). (2005). *Bar admission requirements*. Disponible en <http://www.ncnex.org/tests.htm>. ECD (2007). *PISA 2006 Science Competencies for Tomorrow's World*. París: OECD.
- O'Neil, H.F., Subgure, E. y Baker, E.L. (1996). Effects of motivational interventions on the National Assessment of Educational Progress Mathematics Performance. *Educational Assessment*, 3, 135-157.
- Palm, T. (2008). Performance assessment and authentic assessment: A conceptual analysis of the literature. *Practical Assessment, Research & Evaluation*, 13, nº 4. Disponible en <http://pareonline.net/getvn.asp?v=13&n=4>
- Parke, C.S., Lane, S. y Stone, C.A. (2006). Impact of a state performance assessment program in reading and writing. *Educational Research and Evaluation*, 12, 239-269
- Patz, R.J. (2006). Building NCLB science assessments: Psychometric and practical considerations. *Measurement*, 4, 199-239.
- Penfield, R.D. y Lamm, T.C.M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practices*, 19, 5-15.
- Prieto, G. y Delgado, A.R. (2010). Fiabilidad y validez. *Papeles del Psicólogo*, 31,
- Reckase, M.D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, 21, 25-36.
- Reckase, M.D. (2009). *Multidimensional item response theory*. New York: Springer.
- Rupp, D. E., Gibbons, A.M., Baldwin, A. M., Snyder, L. A., Spain, S. M., Woo, S. E., et al. (2006). An initial validation of developmental assessment centers as accurate assessments and effective training interventions. *Psychologist-Manager Journal*, 9, 171-200.
- Thorton, G.C. y Gibbons, A.M. (2008). The construct validity of the assessment center method and usefulness of dimensions as focal constructs. *Industrial and Organizational Psychology*, 1, 116-120.
- Ryan, T. (2006). Performance assessment: Critics, criticism, and controversy. *International Journal of Testing*, 6(1), 97-104.
- Ryan, J. M., y DeMark, S. (2002). Variation in achievement scores related to gender, item format and content area tested. En G. Tindal & T. M. Haladyna (Eds.), *Large scale assessment programs for all students: validity, technical adequacy, and implementations issues*, (pp. 67-88). Mahwah, NJ: Lawrence Erlbaum Associates.
- Salgado, J.F. y Moscoso, S. (2008). Selección de personal en la empresa y las AAPP: de la visión tradicional a la visión estratégica. *Papeles del Psicólogo*, 29, 16-24. <http://www.cop.es/papeles>
- Shavelson, R.J., Baxter, G.P. y Gao, X. (1993). Sampling Variability of performance assessments. *Journal of Educational Measurement*, 30, 215-232.
- Stecher, B., Klein, S., Solano-Flores, G., McCaffrey, D. M. Robyn, A., Shavelson, R. y col. (2000). The effects of content, format and inquiry level on science performance assessment scores. *Applied Measurement in Education*, 13, 139-160.
- Society for Industrial and Organizational Psychology (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.
- Stiggins, R. (1987). Design and development of performance assessments. *Educational Measurement: Issues and Practices*, 6(3), 33-42.
- Thorton, G.C. y Rupp, D.E. (2006). *Assessment centers in human resource management*. Mahwah, NJ: Erlbaum.
- United States Medical Licensure Examination (2009). *Examinations*. Disponible en <http://www.usmle.org/examinations/index.html>
- Van der Vleuten, C. y Swanson, D. (1990). Assessment of clinical skills with standardized patients: State of the art. *Teaching and learning in Medicine*, 2, 58-76.
- Welch, C. (2006). Item and prompt development in performance testing. In S. Downing y T. Haladyna (Eds.), *Handbook of test development* (pp. 303-327). Mahwah, NJ: Erlbaum.
- Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan*, 70, 703-713.
- Williamson, D.M., Mislevy, R.J. y Bejar, I.I. (Eds.) (2006). *Automated scoring of complex tasks in computer based testing*. Mahwah, NJ: Erlbaum
- Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Zenisky, A.L. y Sireci, S.G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education*, 15, 337-362.