¿QUÉ SABEMOS DE LA MEDIDA DE LAS COMPETENCIAS? CARACTERÍSTICAS Y PROBLEMAS PSICOMÉTRICOS EN LA EVALUACIÓN DE COMPETENCIAS

What do we know about the assessment of educational performance? Psychometric features and problems in performance assessments

MARÍA CASTRO MORERA
Universidad Complutense de Madrid

La evaluación de competencias se puede definir como un procedimiento en el que se requiere que el estudiante complete tareas o procesos en los que se demuestre su habilidad para aplicar conocimiento y destrezas o aplicar conocimientos en situaciones simuladas similares a la vida real. La evaluación de competencias se utiliza en educación como una medida del logro académico, teniendo una trascendencia importante en la vida de los estudiantes. Parece entonces fundamental entender cómo se obtienen estas puntuaciones, qué información proporcionan, cuáles son los requisitos técnicos que deben cumplir para considerar estos instrumentos como válidos y fiables y por tanto cuáles son sus principales limitaciones y por dónde debe seguir la investigación en este ámbito. Este artículo es una síntesis de las evidencias que hay en la literatura científica sobre las características psicométricas de las pruebas utilizadas en la evaluación de competencias, cuáles son sus fortalezas y debilidades. La investigación muestra que es poca la evidencia que apoya la solidez técnica de este tipo de medidas, de hecho, la nota común predominante es que hace falta más investigación psicométrica. El principal problema de las pruebas de competencias es que las puntuaciones no suelen ser comparables dada la constatada variabilidad en las calificaciones otorgadas por jueces humanos que aplican el mismo criterio de evaluación. Además características psicométricas básicas como la fiabilidad de las puntuaciones y la validez de las inferencias no encuentran en la literatura suficiente base empírica.

Palabras clave: *Test de desempeño, Evaluación de competencias, Evaluación educativa, Ítems de crédito parcial, Generalizabilidad, Teoría de Respuesta al Ítem Multidimiensional, Fiabilidad interjueces.*

Introducción

La idea de evaluar competencias no es nueva. Ya Aristóteles afirmaba «La excelencia moral es resultado del hábito. Nos volvemos justos realizando actos de justicia; templados, realizando actos de templanza; valientes, realizando actos

de valentía». Esta sencilla definición del constructo de competencia en justicia, por ejemplo, muestra que la medida de la misma pasa por el desempeño, por la demostración de la competencia definida. Sin embargo, una rápida búsqueda en Google del término *performance assessment* produce 13.300.000 resultados, que no es más que una muestra de su rabiosa actualidad.

Madaus y O'Dwyer (1999) establecen los orígenes de las pruebas de desempeño durante la dinastía Han en China. Hasta ese momento, los puestos de funcionario y de ministro eran hereditarios. La modernización de la administración supuso la introducción de los criterios de promoción y carrera profesional. Entonces se instauraron exámenes públicos competitivos como medida de selección objetiva para ocupar un cargo en la administración pública, midiendo competencias en las áreas de leyes, milicia, agricultura y geografía. La dinastía Han perdura en el poder hasta la primera década del siglo XX, y lógicamente se produce una evolución y cambios importantes en la evaluación del desempeño de los funcionarios públicos a lo largo de todo este periodo, incorporando ejercicios que incluyen desde la memorización de pasajes clásicos, la descripción poética o la demostración de la capacidad de razonamiento a través de la discusión de conflictos clásicos.

En la Edad Media se utilizaron similares evaluaciones en los gremios, las pruebas exigían la realización del periodo de aprendiz y la realización de una pieza maestra. En las incipientes universidades europeas se utilizaban discusiones orales en latín de contenido teológico para la valoración de los estudiantes. En estos ejercicios también se incluían los estándares de desempeño: «*The student had to show the ability to remember relevant and aceptable knowledge, the ability to present it in eloquent form, and a tacit conformity to orthodoxy, educational and social*» (Hoskin, 1979: 138).

Un salto en el tiempo, no exento de ejemplos de evaluación de competencias, nos sitúa en Europa en el ámbito de la selección de personal. Hace más de 60 años aparecen los *Assessment Center* hoy conocidos como Centros de Desarrollo y de Evaluación que utilizan muestras de trabajo y ejercicios simulados para evaluar competencias difíciles de medir con pruebas convencionales. En 1942 el British War Office Selection Board inició procesos de selección donde se evaluaba de forma clara la vocación para el liderazgo de los oficiales británicos. La generalización de estas técnicas ha sido rápida y extensa tanto geográficamente como en el tipo de puestos a los que se aplican. Actualmente, en este tipo de pruebas se usan simulaciones, muestras de trabajo y proponen estímulos que demandan respuestas centradas en la actuación del sujeto (Thorton y Rupp, 2006).

En nuestro entorno cotidiano estamos rodeados de múltiples pruebas que miden competencias, piénsese por ejemplo en el examen práctico de conducir, en los exámenes de cualquier instrumento musical en los conservatorios de música o en las pruebas de acceso para algunas facultades universitarias como las de Bellas Artes. En todas ellas se demanda al estudiante que demuestre el nivel de dominio efectivo del manejo del coche, del virtuosismo del instrumento o de la técnica de dibujo necesaria para poder cursar con éxito unos estudios.

En los contextos profesionales, donde surge la necesidad de la comprobación de la competencia del profesional, se desarrollan sistemáticamente procedimientos y técnicas para la evaluación de competencias. En distintos estados de Estados Unidos se pide una certificación periódica a médicos (United States Medical Licensure Examination, USMLE, 2009) o abogados (Multistate Performance Test, del National Conference of Bar Examiners and American Bar Association, 2005) por la cual acreditan la actualización de sus niveles de competencia profesional para poder seguir ejerciendo. En grandes empresas, como IBM, AT&T o General Electric, se pide a sus empleados que demuestren su competencia profesional a través la presentación de los casos

en los que han trabajado para poder avanzar en su carrera profesional. Johnson, Penny y Gordon (2009) ofrecen descripciones detalladas de evaluaciones de desempeño utilizadas en procesos de acreditación profesional.

Así, la primera nota definitoria de la medida de una competencia pasa por la comprobación del grado de ejecución y dominio de esa competencia. La competencia se demuestra. ¿Alguno de ustedes se fiaría de un conductor que sólo ha sido examinado de libros sobre mecánica? ¿Qué opinión les merecería que la acreditación de piloto sólo exigiera superar un simulador de vuelo? Parece claro pues que la medida de la competencia pasa por una prueba que incluya una demostración del dominio, sea esa competencia la justicia, la música, el pilotaje de un avión o la cirugía. La segunda nota característica necesaria para la medida de las competencias es la necesidad de definición del constructo. La tercera nota característica es que la demostración de las competencias supone el dominio de contenidos propios del constructo evaluado. Así, la competencia es una definición altamente específica, tan específica que permite definir una ejecución, como hemos visto. En el fondo, la competencia define un determinado «*saber hacer*» que es preciso definir.

La evaluación de competencias en educación está claramente de moda. No hay evaluación que se precie si no se centra en la valoración del grado que los estudiantes exhiben sus competencias educativas. Y esta afirmación es independiente de la finalidad de la evaluación, tanto si es una evaluación diagnóstica a gran escala como si es una evaluación para el desarrollo de un plan de mejora de un centro educativo como si se trata de definir el currículum de estudios universitarios. En general, el profesorado y los equipos directivos de centros no universitarios muestran una actitud favorable hacia este tipo de evaluaciones, pues consideran que producen cambios positivos en la instrucción y en las prácticas de evaluación de clase (Lane, Parke y Stone, 2002; Parke, Lane y Stone, 2006).

La evaluación de competencias es un indicador del logro académico que tiene una trascendencia importante en la vida de los examinados. Parece fundamental entender cómo se obtienen estas puntuaciones, qué información proporcionan, cuáles son los requisitos técnicos que deben cumplir para considerar estos instrumentos como válidos y fiables y por tanto cuáles son sus principales limitaciones y por dónde debe seguir la investigación en este ámbito. Las esperanzas abiertas con este tipo de evaluaciones deben estar refrendadas en la superioridad empírica de este tipo de pruebas para medir los logros y los aprendizajes escolares.

Este artículo es una síntesis de las evidencias que hay en la literatura científica sobre las características psicométricas de las pruebas utilizadas en la evaluación de competencias, cuáles son sus fortalezas y debilidades.

La medida de competencias en educación

Los ejemplos de evaluaciones descritos previamente se agrupan genéricamente bajo el término inglés de *performance assessment* que se traduce frecuentemente como «evaluación del desempeño». Este término procede de la evaluación educativa y de las certificaciones profesionales. Aunque ha sido generalizado en el ámbito educativo como «evaluación de competencias».

Aparece con fuerza la demanda y la necesidad de evaluar habilidades cognitivas de alto nivel (como son planificación, estructuración de tareas, obtención de información, construcción de respuesta, explicación de procesos, integración de conocimientos e informaciones), pues se consideran resultados valiosos del sistema educativo (Linn, 1993; Hambelton, 2000; Ryan, 2006). Si se asume además que la evaluación influye claramente en los contenidos de la enseñanza (Stiggins, 1987; Frederiksen y Collins, 1989; Wiggins, 1989), parece entonces

que estas habilidades complejas deben incluirse en las evaluaciones de los niveles de logro de los sistemas educativos. Así, la evaluación de competencias llega al sistema educativo (tanto no universitario como universitario), influenciada por la necesidad de mejora de las medidas de logro académico unido al desarrollo e implantación de evaluaciones internacionales a gran escala (de las que el proyecto PISA de la OCDE es su máximo exponente).

Desde mediados de los ochenta, las prácticas de evaluación han ido cambiando gradualmente, desde el uso exclusivo de pruebas de opción múltiple al empleo de formatos mixtos de cuestiones y reactivos que pueden incluir portafolios, tareas de desempeño, redacción de ensayos, cartas, respuestas cortas, resolución secuencial de problemas, elaboración de proyectos, presentaciones orales y otras muchas aproximaciones. La estrategia de medida pasa de la selección de una respuesta correcta a la construcción de una respuesta para un amplio conjunto de problemas o situaciones. Se demanda que el examinado produzca algo en un periodo de tiempo y se evalúan los procesos o los productos con relación a unos criterios de rendimiento establecidos (estándares). Este cambio está motivado parcialmente por la creencia de que las pruebas de opción múltiple no son la manera adecuada de medir habilidades complejas (como la resolución de problemas o el pensamiento crítico) ni para medir la evaluación de destrezas para el aprendizaje a lo largo de la vida (como pensamiento independiente, persistencia, flexibilidad).

Sin embargo, las definiciones de una realidad tan heterogénea son muy variadas y en ocasiones confusas. El análisis de las distintas descripciones y definiciones de la evaluación de competencias muestra que la mayoría de ellas son independientes del objeto de evaluación, y aunque comparten algunas características comunes, la mayoría de ellas se definen bien por el formato de la respuesta, bien por la observación de la discrepancia entre la respuesta y el criterio de interés (Palm, 2008). Aunque su principal modo de definición viene por la oposición a los reactivos de opción múltiple, vinculando los distintos procedimientos de medida a la mejor manifestación de altas habilidades y procesos cognitivos, como muestra la definición clásica de la *Office of Technology Assessment* del Congreso de los Estados Unidos de América (OTA, 1992: 19):

> «Esta forma de enfrentarse con la medición de competencias se comprende mejor considerándola como un continuo de formatos que va desde la más simple respuesta construida por el estudiante hasta amplios conjuntos de trabajos recogidos a lo largo de un periodo de tiempo […]. Las cuestiones de respuesta construida requieren de los estudiantes elaborar una respuesta y no meramente seleccionarla entre una serie de posibilidades, como se hace en los ítems de elección múltiple […]. Ejemplos de respuestas construidas serían, entre otros, completar textos rellenando espacios en blanco, resolver problemas matemáticos o escribir respuestas cortas.»

En el ámbito de la evaluación educativa la definición de los constructos de evaluación supone una *aproximación* al futuro desempeño que deberían poder realizar los estudiantes. La demostración del conocimiento aplicado a situaciones de la vida real se realiza mayoritariamente a través de estímulos que simulan la vida real. De ahí que la definición que aparece en los *Standards for Educational and Psychological Test* (*American Educational Research Association* —AERA—, *American Psychological Association* —APA— y *National Council on Measuremente in Education* —NCME—, 1999: 179) destaque que «las evaluaciones de desempeño *emulan* el contexto o las condiciones en las que se aplican los conocimientos y destrezas que se intentan evaluar». Son, como hemos dicho, aproximaciones al desempeño real de los examinados.

En esta misma línea, Johnson, Penny y Gordon (2009) señalan que en las evaluaciones de

competencias los examinados demuestran sus conocimientos y habilidades a través de la implicación en procesos o a través de la construcción de un producto. En la misma línea (Stiggins, 1987; Ruiz-Primo y Shavelson, 1996; Shavelson, Solano-Florez y Ruiz-Primo, 1998) definen las evaluaciones de competencias como un sistema que incluye a) un propósito de la evaluación; b) un conjunto de tareas que emulan el desempeño; c) una respuesta del examinado que define su desempeño y d) un conjunto sistemático de métodos para codificar y ordenar los distintos niveles de competencia.

Sin embargo, Berk (1986) en una propuesta ya clásica de definición de evaluación de competencias señala que estas pruebas deben incluir una observación directa de la conducta de interés como estrategia opuesta a las respuestas escritas de lápiz y papel. Johnson, Penny y Gordon (2009) indican que los test de desempeño se pueden clasificar por lo que evalúan, es decir, por el resultado de la tarea (*products*), o por los procesos que sigue el examinado para llegar a una solución (*performances*). Aunque en la mayoría de las ocasiones se combinan procesos y productos.

Así, en el ámbito educativo la evaluación de competencias se operacionaliza fundamentalmente como una simulación a través de tareas que probablemente se demandan al estudiante en la vida no académica. Dista pues de la definición inicial por la cual la evaluación de la competencia supone una demostración. Cabe pues preguntarse si con esta operacionalización se está realizando una evaluación de competencias propiamente dicha.

Características psicométricas de las pruebas de competencia en educación

Las evaluaciones educativas quieren informar de lo que los estudiantes saben y pueden hacer. Los atributos medidos en el ámbito educativo constituyen representaciones mentales, conocimientos y procesos que no son directamente observables. Así, el desempeño tampoco es directamente observable. Una evaluación es el resultado de una inferencia de lo que saben a partir de las respuestas emitidas a un conjunto de estímulos que incluye el test. A todos los profesores nos gustaría tener una «regla» que se aplicara directamente al cerebro de nuestros estudiantes y que nos indicara sin lugar a dudas lo que ese estudiante sabe y puede hacer. Sin embargo, esa «regla», definida de forma unívoca y directa, no existe.

Más formalmente, se puede decir que la evaluación es un proceso de razonamiento desde la evidencia (Mislevy, Steinberg y Almond, 2002; Mislevy, Wilson, Ercikan y Chudowsky, 2003; Mislevy 2006;) que concluye en la emisión de un juicio de valor sobre el nivel de desempeño del evaluado. El proceso de acumulación de evidencias se justifica por el carácter indirecto de la medida en el ámbito educativo. Podríamos pensar en la evaluación como un proceso de *estimación* que se basa en las inferencias realizadas a partir de las respuestas emitidas por un examinado a una *muestra de estímulos* determinados. Esta muestra de estímulos procede de un universo de conocimientos y desempeños que un individuo puede conocer y demostrar. Y a los que lógicamente no puede responder en su totalidad. Lo que el estudiante conoce y puede hacer no coincide fielmente con lo que demuestra en una situación de evaluación. Necesariamente en la selección de estímulos perdemos información. Así, cuando emitimos una valoración sobre un examinado, esta estimación indudablemente depende de la cantidad de información que podamos extraer del conjunto de estímulos propuestos (entre otras cosas).

Para que las estimaciones se conviertan en evidencia precisamos que los datos o resultados de la evaluación se conviertan en evidencia de dominio, y para ello es necesario un modelo de representación del conocimiento y de desarrollo de las competencias. Se necesita además de un

conjunto de tareas o situaciones que permitan observar el desempeño de forma coherente con el modelo de representación definido. Y por último se precisa de métodos de interpretación para la realización de las inferencias. Estos tres elementos (modelo, tareas o estímulos de evaluación y pauta de interpretación) se denominan «triángulo de la evaluación» (National Research Council, 2001) y representa el proceso de construcción de una regla de medida en una competencia. En la misma línea, los *Standards for educational and psycholgical test* (AERA *et al.,* 1999) señalan que para la construcción de medidas de competencias se precisa una definición del constructo, pues guía la adecuada representación del dominio, la selección de las tareas, los criterios para establecer las puntuaciones y la detección de la posible varianza irrelevante, junto con una definición del propósito de la evaluación y de las inferencias que se vayan a realizar con las evaluaciones.

Además, las pruebas de competencias deben cumplir con los criterios psicométricos exigibles a todo procedimiento de evaluación (Kane, 2004). Como se ha visto, la evaluación de competencias exige que los examinados construyan sus respuestas a muy variados tipos de problemas y situaciones. Aunque su uso tiende a incrementarse cada vez más, la mayoría de la metodología disponible para evaluar la calidad psicométrica de las pruebas se ha desarrollado para pruebas de lápiz y papel. La aplicabilidad de estos métodos a medidas procedentes de evaluaciones de competencias es limitada, cuestionable o incluso no está probada.

Se van a revisar a continuación algunos de los rasgos psicométricos más relevantes para la evaluación de competencias educativas, sin embargo, esta revisión no es ni exhaustiva ni técnica, simplemente destaca algunos de los principales problemas y algunos de los más relevantes desarrollos psicométricos para la evaluación de competencias. Esta medida debe tener algunas características (Viswesvaran, 2001; SIOP, 2004; Martínez Arias, 2010): a) la relevancia del constructo a medir, b) la fiabilidad o consistencia de las mediciones, c) la capacidad de discriminación entre evaluados, ordenándose en función de las diferencias reales entre examinados en el constructo, d) la estandarización de la medida asegurando así la comparabilidad de las puntuaciones entre individuos y en el tiempo y e) la validez de las medidas. Entraremos a continuación en los cuatro últimos puntos mencionados. Excede con creces el marco de este trabajo entrar en la consideración de cuáles son las competencias escolares relevantes que han de evaluarse.

Modelos de medida, discriminación y comparabilidad de las puntuaciones

Como ya hemos dicho, el proceso de medida pasa por establecer los vínculos entre las respuestas observables a estímulos de medida y el constructo latente que se quiere medir. Este constructo latente está definido a través de un modelo estadístico que describe cómo las respuestas observadas están relacionadas con esas inferencias sobre el desempeño. El modelo estadístico nos permite estimar y asignar valores numéricos a la competencia que queremos medir, y se denomina modelo de medida. Los modelos de medida aportan una vía sistemática por la que se asocian valores numéricos (estimaciones de competencia) basados en respuestas observables.

Los modelos estadísticos a utilizar con las pruebas de competencias deben tener en cuenta las características específicas de este tipo de evaluaciones. Por un lado, la evaluación de competencias está asociada muy a menudo con pautas de corrección complejas y generalmente suponen una valoración graduada del desempeño (valoraciones que van más allá de la corrección o incorrección de la respuesta, generando repuestas politómicas o de crédito parcial). Los modelos basados en la Teoría de Respuesta al Ítem (TRI) se desarrollan inicialmente para ítems de opción múltiple (con respuestas

dicotómicas), y contrasta con la variedad de formatos de los estímulos incluidos en las pruebas de competencias. Se han desarrollado modelos TRI politómicos para ítems de respuesta construida y se han estudiado sus características (Master, 1982; Hemker, Sijtsma, Molenaar y Junker, 1997; Wu, Adams y Wilson, 1998 que son los diseñadores del paquete CONQUEST). Sin embargo, aún se necesita más investigación que estudie la adecuación de estos modelos a algunos formatos de respuestas.

Y por otro lado, la unidimensionalidad es un rasgo hasta ahora casi imprescindible en los modelos de medida tanto dicotómicos como politómicos de TRI. El supuesto de unidimensionalidad establece la necesidad de que sea una única habilidad la que se mida en cada tarea de evaluación o en cada test. Así las puntuaciones obtenidas están vinculadas a una escala que refleja la evolución del dominio en un único constructo. No obstante, la multidimensionalidad suele ser un rasgo propio de la demanda de evaluación en las pruebas de competencias, pues se requieren múltiples habilidades para completar con éxito un ejercicio o tarea (Gibbons, Bock, Hedeker, Weiss, Segawa, Bhaumik, Kupfer, Frank; Grochocinsky y Stover, 2007; Reckase, 2009). Recientes avances en modelos de medida, los modelos MIRT (Multidimensional Item Response Theory) incorporan la consideración de la multidimensionalidad a los modelos de la Teoría de Respuesta al Ítem. Los modelos MIRT representan una metodología para representar la posición de los examinados en un hipotético espacio cognitivo multidimensional (Reckase, 2009). Sin embargo, su uso no está generalizado debido a la novedad de los modelos y a la dificultad de interpretación de sus índices.

En cualquier caso, los modelos estadísticos permiten estimar y asignar valores para el constructo latente de competencia, que es variable en función de las respuestas dadas por un examinado. El proceso de construcción de la escala supone la definición de unidades de medida junto con la identificación del modelo de medida (pues no están determinadas únicamente por el modelo de medida). El propósito principal de la escala de medida es la ordenación de los sujetos en función del nivel de desempeño demostrado. De ahí que sea fundamental, que ante distintas pruebas las puntuaciones que se asignen sean comparables. Los procedimientos de equiparación, tanto horizontal como vertical, son absolutamente imprescindibles para garantizar esta comparabilidad de las puntuaciones, y generalmente exigen compartir estímulos comunes. El principal problema de las pruebas de competencias es que las tareas de evaluación no suelen ser comparables, pues pueden medir distintos constructos subyacentes. Además, son fácilmente recordables, por lo que no se pueden reutilizar como se hace con los ítems de opción múltiple y dadas las condiciones temporales de aplicación no se pueden incluir muchas tareas en la misma prueba.

Fiabilidad

La fiabilidad de las pruebas es comúnmente entendida como la consistencia o replicabilidad de las puntuaciones en la medida de un determinado atributo. La medida de la fiabilidad en pruebas de competencias necesita un enfoque más amplio que el establecido por la Teoría Clásica de los Test (TCT), pues el número de fuentes de error que afectan a la consistencia de las puntuaciones es mayor en pruebas de competencias que en pruebas más convencionales. La Teoría de la Generalizabilidad (TG), sistematizada por Cronbach, Gleser, Nanda y Rajaratnam (1972), es una extensión de la TCT que utiliza el Análisis de los Componentes de Varianza para estimar simultáneamente los efectos de las distintas fuentes de variabilidad o error (facetas) sobre las puntuaciones (Martínez Arias, 2010).

En el caso de las pruebas de competencias, las puntuaciones se ven afectadas fundamentalmente

por las tareas que incluye la prueba y los evaluadores humanos que juzgan el desempeño. Aunque también se pueden incluir las ocasiones de medida, la administración o el formato del test. La aportación de la TG se centra en la posibilidad de mejorar la fiabilidad realizando estudios para determinar el número de tareas y evaluadores necesarios al descomponer la varianza de error en diferentes fuentes (para un tratamiento más extenso de la TG recomendamos Martínez Arias, 1995; Brennan, 2000; Martínez Arias, Hernández Lloreda y Hernández Lloreda, 2006).

La tarea se convierte en una fuente de error debido al reducido número de estímulos que se pueden incluir en una prueba para evaluar una competencia. Lane y Stone (2006) encuentran una baja consistencia entre las tareas y las interacciones con los sujetos. Miller y Linn (2000) apuntan dos soluciones para reducir los errores asociados con la heterogeneidad de las tareas de evaluación: a) incrementar el número de tareas o estímulos y b) definir de manera precisa el constructo a medir y la tarea de evaluación.

Además, la evaluación de competencias precisa de correctores o jueces, que si bien son entrenados en la aplicación de la pauta de corrección o rúbrica, también es cierto que cada evaluador aplica la pauta de una manera diferente; existiendo discrepancias entre las puntuaciones ofrecidas por distintos jueces (fiabilidad inter jueces). Incluso un mismo evaluador puede ser inconsistente al evaluar el mismo ejercicio en dos o más ocasiones distintas (fiabilidad intra jueces). El nivel de coincidencia, primera aproximación a la fiabilidad interjueces, es muy heterogéneo. Por ejemplo Lane y Stone (2006) encuentran correlaciones que oscilan entre 0,33 y 0,91 en la evaluación de escritura. La fiabilidad interjueces de las certificaciones médicas (van der Vleuten y Swanson, 1990) oscila entre 0,50 y 0,93. La fiabilidad inter jueces también está afectada por la competencia evaluada, encontrándose mayor consistencia en ciencias y matemáticas que en escritura (Shavelson, Baxter y Gao, 1993).

En general, parece haber acuerdo en que la variabilidad de las tareas contribuye más al error de medida que el evaluador dada la gran heterogeneidad posible y existente entre tareas (Lane y Stone, 2006; Shavelson *et al.,* 1993). Además, es posible observar interdependencias en las pautas de corrección y se constata una baja capacidad de generalización en las puntuaciones de una tarea o ejercicio, pues desempeñarse bien en un grupo de tareas no significa mostrar un alto nivel de desempeño en otras.

Evidencias de validez

Cuando se dice que los estudiantes deben ser «competentes» en ciencias o en comprensión lectora en un idioma, la definición de lo que se considera «competente» tiene una gran relevancia. Según las teorías de medida, este nivel de desempeño es considerado como un constructo no observado o latente que explica las diferencias individuales en los niveles de logro de un conjunto de medidas observables. El establecimiento de la relevancia de las medidas observadas para respaldar las inferencias referidas a este constructo de «competencia en ciencias» es el objetivo de la validación de los test.

La definición de la validez de las puntuaciones de los test de desempeño es la establecida en los *Standards for Educational and Psychological Test* (AERA *et al.,* 1999) que entiende la validez como el grado en que la evidencia y la teoría apoyan las interpretaciones que se van a hacer de las puntuaciones obtenidas en las pruebas de medida; y es similar a la de otros tipos de tests estandarizados, con la validez de constructo como concepto unificador (Messick, 1980, 1995, 1996).

Nos vamos a referir aquí a tres tipos de validez: de contenido, sustantiva y externa, por considerar que son las más relevantes en el ámbito de las pruebas de competencias.

Las pruebas de competencias suelen incluir menor cantidad de estímulos que las tradicionales

pruebas de lápiz y papel. Esta menor cantidad de estímulos supone una mayor selección de estímulos y conlleva una menor cantidad de información recogida por la prueba. La *validez de contenido* se centra en la relevancia y representatividad del contenido de la evaluación. Esta característica representa dos grandes amenazas a la validez de contenido comparada con los test convencionales. Messick (1989, 1996) señala el potencial riesgo de la *infrarrepresentación del constructo* debido a la menor cantidad de ítems en las pruebas de desempeño y el riesgo de *varianza irrelevante* debido múltiples fuentes como la elección del tema por parte de los examinados, a la tendencia de los evaluadores humanos a fijarse en aspectos irrelevantes de la respuesta, o la motivación de las respuestas, sobre todo si la prueba no tiene consecuencias para los examinados.

A las pruebas de competencias se les atribuye generalmente la capacidad de medir habilidades complejas y de alto nivel cognitivo. Messick (1996: 9) destaca «la necesidad de obtener evidencias empíricas de los procesos puestos en juego por los examinados cuando realizan la tarea». Sin embargo, son pocas las investigaciones que avalen esta *validez sustantiva* (o de los procesos de la respuesta). Martínez Arias (2010) realiza una revisión de estas investigaciones, mostrando que los resultados obtenidos son poco consistentes.

La *validez externa* se define como el análisis de las relaciones de las puntuaciones de un test con variables externas (AERA *et al.,* 1999). Esta parte de los estudios de validez es usada para probar si las relaciones de las evaluaciones son consistentes con la teoría en la que se basa el constructo medido. Así, los constructos que se miden en las evaluaciones deben racionalmente mostrarse en un patrón de correlaciones con variables externas.

Es frecuente que las pruebas de competencias no estén basadas en teorías muy bien definidas (Miller y Linn, 2000), aunque hay diversas hipótesis acerca de estos test consistentes con la literatura. En primer lugar, el desempeño debe estar relacionado con los efectos instructivos, ofreciendo mejores medidas para la rendición de cuentas (Wiggins, 1989). Además, todas las medidas de rendimiento deben ser sensibles a los efectos de la instrucción. En segundo lugar, la varianza del constructo debe ser más grande que la del método (Campbell y Fiske, 1959). Esto tiene una particular relevancia dado que puede haber una variedad de procedimientos y tareas de medida para medir el mismo constructo.

El procedimiento habitual para obtener evidencias de validez externa consiste en el estudio de correlaciones según las expectativas teóricas o hipótesis del constructo. Martínez Arias (2010) señala que en el ámbito educativo hay pocos trabajos sobre este tipo de evidencias, encontrándose relaciones débiles pues suele haber una mayor proporción de varianza entre las puntuaciones debida al contexto que al constructo. En el ámbito educativo se ha realizado recientemente un trabajo de estudio de la validez concurrente entre pruebas convencionales de medición del rendimiento y pruebas de competencias. Los resultados del estudio de las pruebas ENLACE en México realizado bajo el auspicio de la OCDE (Zúñiga-Molina y Gaviria, 2010) son realmente interesantes.

El programa ENLACE es una evaluación nacional del rendimiento académico en México que se aplica censalmente desde 3º de Educación Primaria hasta 3º de Educación Secundaria en las materias de español y matemáticas. Las pruebas ENLACE se aplican anualmente a todos los estudiantes mexicanos y están compuestas por ítems de opción múltiple. Son pues unas pruebas «convencionales» de evaluación de rendimiento aplicadas a gran escala.

En el estudio de Zúñiga-Molina y Gaviria (2010) se muestra una característica particularmente relevante, la validez concurrente de las pruebas ENLACE con pruebas de evaluación de competencias, similares a las pruebas PISA.

El programa SEP-ISA aplicado a estudiantes mexicanos está construido en colaboración con el *Australian Council on Educational Research* (ACER), y utiliza ítems del mismo banco utilizado para la elaboración de las pruebas PISA así como los ítems liberados de las pruebas PISA. La correlación empírica entre los resultados de ENLACE y de SEP-ISA aplicados a 11.717 estudiantes de 15 años es de 0,780. Si en lugar de variables empíricas se utilizan estimaciones que eliminan el efecto de atenuación, las correlaciones superan los valores de 0,80.

Obviamente, estos resultados son sorprendentes en dos sentidos. En primer lugar, las altas correlaciones entre ambos tipos de pruebas, consideradas tan diferentes, indican que tienen un comportamiento psicométrico muy similar, siendo un indicador de que ambas pruebas están midiendo mismo constructo. En segundo lugar, la escasa literatura existente está referida las investigaciones en los denominados centros de evaluación y muestra correlaciones inferiores a 0,40 (Arthur, Day, NcNelly y Edens, 2003; Salgado y Moscoso, 2008). Se pone así de manifiesto una gran diferencia en la magnitud de las evidencias de validez en el ámbito educativo comparado con el psicológico.

Estos datos plantean serias dudas sobre la diferenciación de las pruebas de competencias frente a otros procedimientos de medida que además son claramente más económicos; aunque hace falta evidencia adicional en este sentido.

Conclusiones

La evaluación de competencias se puede definir como un procedimiento en el que se requiere que el estudiante complete tareas o procesos en los que se demuestre su habilidad para aplicar conocimiento y destrezas o aplicar conocimientos en situaciones simuladas similares a la vida real (Wiggins, 1993; Messick, 1996; Nitko, 1996; Payne, 1997). La evaluación de competencias abarca una amplia variedad de formatos de pruebas (respuesta construida, ensayos, demostraciones, presentaciones orales, portfolios, observación directa…). Se pide a los examinados que produzcan, creen o desempeñen algo durante un periodo de tiempo y tanto los procesos, como los productos o ambos se evalúan considerando un estándar de desempeño (Oosterhof, 1994; Messick, 1996). Se cree además que este tipo de evaluaciones es más apropiado para medir habilidades complejas de pensamiento o en la evaluación de la resolución de problemas (Haertel y Linn, 1996; Sax, 1997).

En las pruebas de competencias, la estructura de las respuestas está definida por el examinado, de ahí que las respuestas construidas puedan ser evaluadas en distintos niveles de calidad, más allá de la mera corrección o incorrección de las respuestas. Se entiende que así los estudiantes podrán demostrar habilidades que no podrían fácilmente evaluarse con pruebas convencionales como las que incluyen ítems de opción múltiple.

Desde el punto de vista de la medida, estos nuevos formatos de test plantean serios retos con respecto a los deseables e imprescindibles requisitos psicométricos. La nota común más predominante es que hace falta más investigación, pues es poca la evidencia que apoya la solidez técnica de este tipo de medidas. Lo que sí sabemos hasta ahora se podría resumir en los siguientes puntos:

a) Habitualmente, el número de ítems utilizados en las pruebas de competencias es menor que los utilizados en pruebas de opción múltiple. Esto puede suponer riesgos, pues es posible elaborar pruebas basadas en una muestra inadecuada del constructo de referencia (Dunbar, Koretz y Hoover, 1991) y obtener puntuaciones muy inestables. Es muy grande la dificultad para representar adecuadamente el dominio por el número limitado

de tareas que se pueden incluir en un test de competencias. La cantidad de información recogida seguramente es más rica pero es menor que la que se puede recoger por procedimientos denominados convencionales.

b) Es fácil encontrar constructos de medida y tareas de evaluación multidimensionales y con comportamiento inestable en función del contexto de aplicación de la evaluación (Mislevy, 1992; Bond, Moss y Carr, 1996; Haertel y Linn, 1996). La variabilidad de las tareas por sí mismas puede ser una razón importante, pero los efectos del contexto, de la práctica y de la varianza debida a definiciones pobres del constructo contribuye también a la inestabilidad de la dimensionalidad de las pruebas entre contextos de aplicación. Esta circunstancia afecta a la capacidad para realizar adecuadas comparaciones de las puntuaciones procedentes de distintas evaluaciones. Se precisa de modelos complejos como los MIRT para la estimación de las puntuaciones o para la equiparación o calibración de los parámetros de los ítems.

c) Las pruebas de competencias habitualmente producen puntuaciones politómicas, basadas en evaluaciones de jueces que no dejan de ser subjetivas, incluso cuando se desarrollen guías de puntuación muy claras, precisas y costosas y se entrene a los jueces. La (in)consistencia de las puntuaciones intra e inter jueces constituyen una adicional fuente de error (Kolen y Brennan, 1995). Sin considerar que son evaluaciones tanto económica como temporalmente costosas, lo que dificulta el uso formativo de sus resultados.

d) Se precisa más investigación sobre el comportamiento de las pruebas que incluyen ítems de muy distinta naturaleza, pues la evidencia acumulada hasta ahora sugiere que los resultados procedentes de distintos tipos de ítems pueden no ser comparables (Miller y Linn, 2000; Patz, 2006).

e) La fiabilidad de las pruebas de competencias es en general baja. Además, se precisa realizar estudios basados en la Teoría de la Generalizabilidad para obtener índices adecuados de fiabilidad de las pruebas, necesitando más estudios sobre el número de tareas a incluir y la fiabilidad intra e inter jueces.

f) La fiabilidad de las pruebas de competencias está limitada por la gran cantidad de varianza de interacción entre examinado y tarea. Para controlar esta limitación se precisa emplear un gran número de estímulos, siendo difícil dadas las habituales restricciones de tiempo con las que se cuentan en las pruebas de desempeño académico.

g) Es necesaria una precisa definición del constructo a medir para asegurar la validez de contenido. Es preciso además desarrollar estudios que demuestren la supuesta validez sustantiva de las pruebas de competencias, y comprobar las presumibles consecuencias positivas de estas pruebas sobre la enseñanza y el aprendizaje.

h) No hay muchas evidencias sobre la validez externa de las pruebas de competencias. Sin embargo, los sólidos y contundentes resultados del estudio de Zúñiga-Molina y Gaviria (2010) plantean serias dudas sobre la utilidad y la diferencia real de las pruebas de competencias frente a otras que además son claramente más económicas.

Cabe preguntarse si estamos claramente en condiciones de sustituir las pruebas de evaluación del logro académico más convencionales, cuyas características psicométricas son bien conocidas, por unas pruebas de competencias que albergan muchos deseos y aspiraciones pero de las que aún no sabemos exactamente cuánta confianza podemos depositar en esas medidas. Es preciso acumular evidencia de que los distintos formatos

de ítem propuestos desde las pruebas de competencias son técnicamente tan solventes como las pruebas de opción múltiple. Además, parece razonable considerar el objetivo y amplitud de la evaluación en la elección del tipo de pruebas. Así, por ejemplo, en función de la necesidad de comparabilidad de las puntuaciones sería adecuado ligar las pruebas convencionales a las evaluaciones a gran escala, y vincular las pruebas de competencias a evaluaciones en las que se pueda valorar la riqueza de las respuestas como un elemento de mejora.

Referencias bibliográficas

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION, AMERICAN PSYCHOLOGICAL ASSOCIATION Y NATIONAL COUNCIL ON MEASUREMENTE IN EDUCATION (1999). *Standards for Educational and Psychological Testing*. Washington, DC: AERA.

ARTHUR, W.; DAY, E. A.; NCNELLY, T. L. y EDENS, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimension. *Personnel Psychology*, 56, 125-154.

BERK, R. A. (1986). Preface. En R. A. BERK (ed.), *Performance assessment: Methods & applications*. Baltimore, Maryland, John Hopkins University Press, ix-xiv.

BOND, L.; MOSS, P. y CARR, P. (1996). Fairness In large-scale performance assessment. En G. PHILLIPS (ed.), *Technical issues in large-scale performance assessment*. Washington, DC: National Center for Education Statistics, 117-140.

BRENNAN, R. (2000). Performance assessment from the perspective of the Genralizability theory. *Applied Psychological Measurement*, 24, 339-353.

CAMPBELL, D. T. y FISKE, D. W. (1959). Convergent and discriminant validation by the multitraitmultimethod matrix. *Psychological Bulletin,* 56, 81-105.

CRONBACH, L. J.; GLESER, G. C.; NANDA, H. y RAJARATNAM, N. (1972). *The dependability of behavioural measurements: theory of generalizability for scores and profiles*. New York: Willey.

DUNBAR, S.; KORETZ, D. y HOOVER, H. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education*, 4 (4), 289-303.

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.; GROCHOCINSKY, V. J. y STOVER, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement*, 31, 4-19.

HAERTEL, E. y LINN, R. (1996). Comparability. En G. PHILLIPS (ed.), *Technical issues in large-scale performance assessment*. Washington, DC: National Center for Educational Statistics, 59-78.

HAMBELTON, R. K. (2000). Advances in assessment performance assessment methodology. *Applied Psychological Measurement*, 24, 291-293.

HEMKER, B. T.; SIJTSMA, K.; MOLENAAR, I. W. y JUNKER, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika*, 62, 331-347.

HOSKIN, K. (1979). The examination, disciplinary power and rational schooling. En *History of Education*. London: Taylor y Francis, vol. 8, 135-146.

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: Interdiciplinary Research and Perspectives*, 2, 13-170.

LANE, S. y STONE, C. A. (2006). Performance assessment. En R. BRENNAN (ed.), *Educational Measurement* (4th Edition). Westport, CT: American Council on Education and Praeger, 387-431.

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.

LINN, R. L. (1993). Linking results of distinct assessment. *Applied Measurement in Education*, 6, 83-102.

MADAUS, G. y O'DWYER, L. (1999). A short history of performance assessment. *Phi Delta Kappan*, 80 (9), 688-695.

MARTÍNEZ ARIAS, R. (1995). *Psicometría*: *Teoría de los test psicológicos y educativos*. Madrid: Síntesis.

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

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

MASTER, G. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 42 (2), 149-174.

MESSICK, S. (1980). Test validity and the ethics of assessment. *American Psychologist,* 35*,* 1012-1027.

MESSICK, S. (1989). Validity. En R. LINN (ed.), *Educational Measurement* (3rd ed). Washington, DC: American Council on Education, 13-103.

MESSICK, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist,* 50*,* 741-749.

MESSICK, S. (1996). Validity of performance assessments. En G. PHILLIPS (ed.), *Technical issues in large-scale performance assessment*. Washington, DC: National Center for Education Statistics, 1-18.

MILLER, D. M. y LINN, R. L. (2000). Validation of performance assessments. *Applied Psychological Measurement*, 24, 367-378.

MISLEVY, R. J. (1992). Scaling procedures. En E. G. JOHNSON y N. L. ALLEN (eds.), *The NAEP 1990 technical report* (Report No. 21-TR-20). Washington DC: National Center for Education Statistics, 199-213.

MISLEVY, R. J. (2006). Cognitive psychology and educational assessment. En R. BRENNAN (ed.), *Educational Measurement* (4th Edition). Westport, CT: American Council on Education and Praeger, 257-305.

MISLEVY, R. J.; STEINBERG, l. y ALMOND, R. (2002). Design and analysis in task-based language assessment. *Language Testing*, 19(4), 477-496.

MISLEVY, R. J.; WILSON, M.; ERCIKAN, K. y CHUDOWSKY, N. (2003). Psychometric principles in student assessment. En T. KELLAHAN y D. STUFFLEBEAM (eds.), *International handbook of educational evaluation*. Boston: Kluwer Academic, 489-532.

NATIONAL RESEARCH COUNCIL (2001). *Knowing what students know: the science and design of educational assessment*. Washington, DC: National Academy Press.

NITKO, A. J. (1996). *Educational assessment of students* (2nd ed.). Englewood Cliffs NJ: Prentice-Hall.

OFFICE OF TECHNOLOGY ASSESSMENT, U.S. Congress (1992). *Testing in american schools: Asking the right questions* (OTA-SET-519). Washington, DC: U.S. Government Printing Office.

OOSTERHOF, A. (1994). *Classroom applications of educational measurement* (2nd ed.). New York: Macmillan.

PALM, T. (2008). Performance Assessment and authentic assessment: a conceptual analysis of the literature. *Practical Assessment, Reseach and Evaluation*, 13 (4). Disponible en http://pareonline.net/pdf/v13n4.pdf.

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 assessment: psychometric and practical considerations. *Measurement* 4 (4), 199-239.

PAYNE, D. A. (1997). *Applied educational measurement*. Belmont CA: Wadsworth.

RECKASE, M. D. (2009). *Multidimensional item response theory*. New York: Springer.

RUIZ-PRIMO, A. y SHAVELSON, R. J. (1996). Rhetoric and reality in science performance assessment: an update. *Journal of Research in Science Teaching*, 33 (10), 1045-1063.

RYAN, T. (2006). Performance assessment: critics, criticim, and controversy. *Internacional Journal of Testing*, 6(1), 97-104.

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.

SAX, G. (1997). *Principles of educational and psychological measurement and evaluation* (4th ed.). Belmont CA: Wadsworth.

SHAVELSON, R. J.; BAXTER, G. P. y GAO, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, 30, 215-232.

Shavelson, R. J.; Solano-Flórez, G. y Ruiz-Primo, A. (1998). Toward a science performance assessment technology. *Evaluation and Program Planning*, 21 (2), 129-144.

Society For Industrial And Organizational Psychology, INC. (2004). Principles for the validation and use of personnel selection procedures (4ª Edición). Disponible en www.siop.org.

Stiggins, R. (1987). Design and development of performance assessment. *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 standarized patines: State of the art. *Teaching and learning in Medicine*, 2, 58-76.

Viswesvaran, C. (2001). Assessment of individual job performance: a review of the past century and a look ahead. En N. Anderson; D. S. Ones; H. K. Sinangil y C. Viswesvaran (ed.). *Handbook of industrial and organizational psychology*. 1. London: Sage, 110-126.

Wiggins, G. (1989). A true test. Toward more authentic and equitable assessment. *Phi Delta Kappan*, 70, 703-713.

Wiggins, G. (1993). Assessment: Authenticity, context, and validity. *Phi Delta Kappan,* 75(3), 200-214.

Wu, M. L.; Adams, R. J. y Wilson, M. R. (1998). *ACER ConQuest: Generalized item response modeling software [Computer program]*.Melbourne: ACER Press.

Zúniga Molina, L. y Gaviria, J. L. (2010). Challenges and Opportunities for the Further Development of the ENLACE Assessment for Evaluation and Teacher Incentives in Mexico. *OECD Working Paper for the Cooperation Agreement between the OECD and the Government of Mexico.*

Abstract

What do we know about the assessment of educational performance? Psychometric features and problems in performance assessments

Performance assessments demand students to complete complex tasks showing their ability to apply knowledge and skills to real life in simulated situations. In education, performance assessment is used as an academic achievement measure, having high impact over student's life. Thus, it is capital to have a good understanding about how these scores are obtained, what information is given, which are the technical requirements to be consider as reliable and valid, which are their main limitations and which way should have to be followed in the future research. This article is a synthesis of scientific research literature that evidences psychometric features of performance tests, their strengths and weaknesses. Research evidences are limited and their main conclusions claim for the development of more research about educational performance in test psychometric features. The main problem of this kind of test is the score comparability, due to inter rater variability, among others. Moreover, basic psychometric features as test reliability and validity don't find enough empirical bases on research literature.

Key words: *Performance Assessment, Educational Assessment, Partial Credit Items, Generalizability, Multidimensional Item Response Theory, Inter-rater Reliability.*

Résumé

Qu'est-ce que nous savons sur la mesure des compétences ? Caractéristiques et problèmes psychométriques dans l'évaluation des compétences

L'évaluation des compétences peut être définie comme une procédure qui demande à l'étudiant d'effectuer des taches ou des processus qui prouvent sa capacité pour appliquer des savoirs et des habilités, ou pour appliquer des savoirs dans des situations simulées semblables a la vie réelle. L'évaluation des compétences s'utilise en éducation comme une mesure de la réussite académique, ayant une importante transcendance dans la vie des étudiants. Il semble donc essentiel de comprendre, comment ses scores sont obtenus, quelle information ils apportent, quelles sont les exigences techniques qui doivent accomplir pour tenir compte de ces instruments comme valables et fiables, et donc, quelles sont ses principales limitations et où doit continuer la recherche dans ce domaine. Cet article est une synthèse des évidences trouvées dans la littérature scientifique sur les caractéristiques psychométriques des preuves utilisées pour l'évaluation des compétences, quelles sont ses forces et ses faiblesses. L'étude montre qu'il existe peu de preuves qui appuient la validité technique de ce type de mesures, en fait, la note dominante commune est que plus de recherche psychométrique est nécessaire. Le principal problème des tests de compétences est que les ponctuations ne sont habituellement pas comparables en raison de la constaté variabilité dans les qualifications attribuées par les juges humains qui appliquent un même critère d'évaluation. En plus, des caractéristiques psychométriques basiques comme la fiabilité des ponctuations et la validité des inférences, ne trouvent pas une suffisante base empirique dans la littérature.

Mots clés : *Test de Performance, Évaluation des compétences, Évaluation pédagogique, items à crédit partiel, Généralisabilité, Théorie de la Réponse à l'Item Multidimensionnel, fiabilité inter-juges.*

Perfil profesional de la autora

María Castro Morera

Profesora titular en la Universidad Complutense de Madrid, actualmente desempeña el cargo de Inspectora de Servicios en la misma universidad. Sus líneas de investigación se centran en la medición y evaluación educativas, con énfasis en modelos multinivel, valor añadido en educación y técnicas sobre estudios de jueces. Participa en diversos proyectos competitivos, tanto en evaluaciones de sistemas educativos, como en otros relativos a factores asociados de calidad educativa y explicación del rendimiento y logro académico.
Correo electrónico de contacto: maria.castro@edu.ucm.es