Vol. XXV (1), 2007

REVISTA DE PSICOLOGÍA

Sandra Castañeda
Anne Marie Costalat-Founeau
Daniel González
Jorge Haddad
Dirk Hermans
Carlos Iberico
Pierina Traverso
Debora Vansteenwegen
César Varela
Bram Vervliet

DEPARTAMENTO DE PSICOLOGÍA

FONDO EDITORIAL
PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ. 90 AÑOS

REVISTA DE PSICOLOGÍA
Vol. XXV (1), 2007

CONTENIDO

ARTÍCULOS

Revista de Psicología Vol. XXV (1), 2007 (ISSN 0254-9247)

Proceso de respuesta a examen de egreso en contabilidad: validación de constructo[1]

Daniel González Lomelí[2]
Universidad de Sonora

Sandra Castañeda Figueiras[3]
Universidad Nacional Autónoma de México

César W. Varela Romero[4]
Universidad de Sonora

Se investiga la validación de constructo de seis dimensiones que están en la base del examen de egreso de una licenciatura en contabilidad, con el fin de entender el proceso de respuesta subyacente. A partir de las respuestas a 17 ítems objetivos aplicados de manera grupal a 313 participantes seleccionados según muestreo intencional, se realizó un análisis factorial confirmatorio MRMM, en el que se representaron tres operaciones cognitivas demandadas para resolver los ítems y los métodos, así como los tres campos de conocimiento del contenido que se evalúa en ellos. El modelo resultante muestra bondad de ajuste, validez convergente entre dos constructos y sus variables manifiestas, además de validez divergente solo entre el conocimiento técnico y el combinado. El proceso de respuesta fue explicado, simultáneamente, por la operación comprender, influida por el conocimiento teórico. Se discuten las implicancias para comprender el proceso de respuesta en este tipo de examen a partir de la validación de constructo realizada.
Palabras clave: validez de constructo, proceso de respuesta a ítems objetivos, multirasgo-multimétodo, examen de egreso, contabilidad.

The process of responding to graduate examinations in Accountancy colleges: Construct validation

The construct validity of six hypothesized dimensions was explored for an Accountancy major graduation examination, in order to understand the underlying process of responding. Using answers from 313 examinees to an intentional sample of 17 objective items administered on large scale, a confirmatory factorial analysis was carried out, with a Multi-Trait Multi-Method matrix (MTMM): Traits represented three cognitive operations required to solve the items and the methods represented three knowledge fields of the content. Each item was linked to a cognitive operation and to a knowledge field. The resulting model showed convergent validity in two constructs and its manifest variables and divergent only between technical and combined knowledge. Thus, the answering process was explained, simultaneously by the Understanding operation influenced by Theoretical knowledge. Implications to understanding the answering process in this type of examination through construct validation are discussed.
Keywords: construct validation, objective exam responding process, multi-trait multimethod, Accountancy license exam.

La concepción contemporánea de validez la establece como un concepto unitario, esencialmente intrínseco, de los factores que determinan la respuesta al examen (Embretson, 1999; Messick 1994). En pocas palabras, la concibe como validez de constructo. Al reconocerse que lo relevante y representativo del proceso de examen es lo que sucede en la mente del sustentante cuando se enfrenta a las demandas incluidas en los ítems y no a su contenido superficial, los estudiosos de la validez de constructo se ven retados a establecer cuáles propiedades de los ítems permiten medir lo que se intenta medir y a establecer la utilidad de la evidencia generada por la medición para derivar inferencias válidas. De esta manera, la validez de constructo se constituye en la base empírica de la interpretación del puntaje derivado de la examinación, en la medida en que el significado del constructo provee una base racional para hipotetizar resultados potenciales.

Las teorías que subyacen a los exámenes, cualquiera que sea su aproximación, comprometen modelos para elaborar inferencias acerca de lo que los sustentantes conocen y son capaces de hacer en un

[1] Se agradece al Proyecto CONACyT 40608-H por el financiamiento de esta investigación.

[2] Doctor en Psicología por la UNAM. Profesor titular en la Universidad de Sonora. Su investigación incluye estudios sobre modelamiento de habilidades y estrategias de aprendizaje en escenarios educativos. Dirección postal: Blvd. Luis Encinas y Rosales s/n, Col. Centro. Hermosillo, Sonora, México. Correo electrónico: dgonzalez@psicom.uson.mx

[3] Doctora en Psicología Experimental. Profesora titular del postgrado de la UNAM. Ha recibido una Medalla al Mérito Universitario y Cátedra Especial, en la UNAM. Premio Nacional 2004 a la Enseñanza de la Psicología. Trabaja en evaluación y fomento del desarrollo cognitivo y el aprendizaje complejo. Dirección postal: Av. Universidad 3004, edificio D, último piso, cubículo 6, México D.F. 04510, México. Correo electrónico: sandra@servidor.unam.mx

[4] Doctor en Ciencias Sociales por la Universidad Autónoma de Sinaloa. Profesor de la Escuela de Psicología y Ciencias de la Comunicación de la Universidad de Sonora. Ha realizado investigación sobre la construcción y la validación de instrumentos en estrategias de aprendizaje y autorregulación, y el efecto de estas en el rendimiento académico. Dirección postal: Blvd. Luis Encinas y Rosales s/n, Col. Centro. Hermosillo, Sonora, México. Correo electrónico: cvarela@psicom.uson.mx

dominio de conocimiento particular. Estos marcos de trabajo generan universos de discurso dependientes de los tipos de aserciones que se elaboran acerca del proceso de responder y de las maneras con las que se colectan los datos que las apoyen. Así, la teoría subyacente a un examen constituye la maquinaria de inferencias, los razonamientos acerca de lo que conocemos teóricamente y de lo que observamos en los datos —siempre en presencia de la incertidumbre—, dado que la naturaleza de la información con la que se trabaja en la examinación es típicamente incompleta y susceptible de tener más de una explicación. Validar constructos supone, en este contexto, establecer el peso y la cobertura de la evidencia (puntaje) de lo que se está midiendo.

Pero ¿cómo establecer lo que se puede considerar como evidencia? Para Mislevy (1993), los datos generados en los exámenes son pistas que adquieren significado solamente con relación a la red de conjeturas establecidas; es decir, los datos solo pueden ser evidencia cuando se establece su relevancia para una o más hipótesis, o sea, cuando muestran incrementar o reducir lo que la hipótesis plantea. De aquí que establecer las inferencias a ser hechas requiera haber validado, previamente, las evidencias (puntajes) generadas por los constructos hipotetizados en la examinación.

Sin embargo, generar tales evidencias requiere superar una importante limitación de los estudios tradicionales de validación de constructo (Embretson, 1983): el que los reactivos de examen estén construidos más por especificaciones acerca de formatos y *syllabus,* que por fundamentos teóricos, con amplia base empírica, que permitan representar los constructos en términos de las demandas cognitivas que el sustentante debe satisfacer para resolver los ítems. En este sentido, la revolución cognitiva ha generado evidencia sólida acerca de mecanismos cognitivos que han mostrado ser responsables del éxito en una amplia gama de actividades humanas. Si en el proceso de examen el constructo es representado como la capacidad para ejecutar las clases de tareas que especifican las diferencias entre la ejecución exitosa de la

no exitosa (Wiley, 2002), entonces la representación de los constructos utilizados en la examinación debería tomar ventaja de la base empírica provista por la investigación de mecanismos cognitivos que han mostrado ser responsables del éxito deseado (Hornke & Habon, 1986). Tal base constituye, hoy día, una fuente de consulta obligada en y para la identificación de los mecanismos que dan cuenta del proceso de responder a la examinación (constructos) y también para los estudios que los validan.

En este contexto, y a manera de ilustración, Castañeda (1993, 1998, 2002), asumiendo que la medición es un proceso inferencial, desarrolló un marco de trabajo que permite identificar componentes del proceso de responder a examinación variada. El marco utiliza análisis funcional de desempeños críticos (componentes de macroestructura) y cognitivo de tareas (componentes de microestructura) para descomponer, recursivamente, los elementos que componen lo que va a ser medido. El procedimiento analítico comienza identificando los desempeños críticos de la ejecución deseada en el dominio que constituirán la macroestructura de la medición. Los productos de esta etapa inicial especifican un número reducido de desempeños críticos de gran importancia (dimensiones a ser evaluadas), que abarcan a otros más elementales (subdimensiones). Se asume que la ejecución demostrada constituye una muestra representativa de lo deseado y que la anidación dimensiones-subdimensiones posibilita la interpretación de los resultados en un conjunto significativo y comprensible, más que en la mera descripción de un conjunto atomizado de datos.

Toda vez establecida la macroestructura, el marco de trabajo utiliza el análisis cognitivo de tareas (ACT) para identificar los microcomponentes en los que los desempeños críticos serán medidos, por ejemplo, las operaciones cognitivas a ser demandas, los tipos de conocimiento a ser evaluados y los contextos en los que serán medidos, todos expresados en gradientes de complejidad creciente.

Este procedimiento apoya al interesado a identificar, en un primer paso, conocimientos, habilidades, disposiciones, tareas y resultados de ejecución esperada, asociados con el dominio de conocimiento que se desea medir (análisis del dominio), en una secuencia sistemática y progresiva de mayor nivel de detalle y precisión. Toda vez identificados, el procedimiento auxilia al interesado a modelar la interrelación entre conocimientos, tareas y niveles de demanda incluidos, lo que facilita construir la estrategia que da solución a una tarea específica, por ejemplo, la que un ítem particular requiera. Así, el interesado está en capacidad de modelar el dominio a ser medido, en términos de los rasgos de tareas que eliciten la ejecución esperada, en los niveles de demanda requeridos. De esta manera, el procedimiento permite identificar mecanismos del proceso de responder a la examinación (constructos), construir las medidas que generarán las evidencias requeridas y, gracias a estas últimas, validar los constructos hipotetizados en la medición. Atiende a la precaución que Messick señaló para la medición: "una aproximación centrada en el constructo … debería empezar por preguntar ¿cuál complejo de conocimientos, habilidades y otros atributos debería ser medido?, porque, presumiblemente, está ligado a objetivos instruccionales explícitos o implícitos o a otros valores de la sociedad" (1994, p. 16).

El marco también apoya al interesado a establecer, explícitamente, las fuentes de demanda a ser incluidas en los ítems; asume que hacerlas explícitas hará más transparente la manera en la que estas afectan la ejecución del sustentante. En otras palabras, hace factible construir o mejorar ítems, y dar al diseñador de exámenes mayor seguridad de que los datos recabados soporten la medición de lo que los estudiantes conocen y pueden hacer. Este aspecto completa lo sugerido por Messick (1994): "¿qué conductas o ejecuciones deberían revelar esos constructos y qué tareas o situaciones deberían elicitar estas conductas?".

En el establecimiento de las fuentes de demanda a ser incluidas en los ítems, el marco apoya al especialista a identificar propiedades del

contexto en el que el sustentante producirá su respuesta, por ejemplo, las de los tipos de conocimientos que serán evaluados y las del uso que se pide que se les dé. De igual manera, apoya la identificación de las propiedades que la evidencia debe mostrar para establecer que se domina o no lo que está siendo medido. En lo general, el marco asume que una tarea o un ítem particular circunscribe circunstancias específicas que le dan al examinando la oportunidad para actuar en formas que producen la evidencia acerca de lo que sabe o puede hacer (Mislevy, Wilson, Ercikan & Chudowsky, 2003). En lo operacional, permite que, para cada tarea o ítem, se asignen puntajes a sus constituyentes, vistos estos como fuentes de contenido incluidas en los ítems. Así se obtienen valores que evidencian la situación en la que el examinando ejecutará; es decir, se generan datos susceptibles de análisis cuantitativos (Castañeda, González, López, García-Jurado & Pineda, 2003) que caracterizan, apriorística y empíricamente, fuentes de facilidad-dificultad asociadas a los ítems que pueden generar varianza irrelevante de constructo.

La ventaja de hacer explícitas las fuentes de contenido incluidas en los ítems se hace más evidente cuando se toma en cuenta que los puntajes obtenidos en la medición reflejan una compleja relación entre el atributo que se mide y el error de medición (Messick, 1989). Estimar la cantidad de error contenida en los puntajes es un asunto prioritario, dado que el error de medición es el principal responsable de su falta de precisión. Entonces, el especialista debe estar atento a las múltiples fuentes de error, porque este se puede generar desde la misma construcción del reactivo. Por ejemplo, demandar habilidades lingüísticas complejas en ítems de resolución de problemas trigonométricos en los que una pobre habilidad lingüística podría ser la responsable de un puntaje bajo en un examinando que poseyera alta habilidad trigonométrica. Errores como el ejemplificado sistemáticamente afectan los puntajes y los procedimientos usuales para estimar la confiabilidad de los puntajes no son sensibles para ellos: solo lo son para la presencia de errores aleatorios. Así y ante tal limitación, generar evidencia en favor de fuentes incluidas en los ítems que generen puntajes representativos del atributo

medido constituye una importante y deseable línea de generación de evidencia en favor de lo que se intenta medir: el constructo.

A partir de lo planteado, el problema abordado en este trabajo centra su preocupación en qué ítems del examen de egreso para contaduría son válidos según los puntajes que generan. Cualquier examen debe asegurar que tanto los constructos subyacentes como la evidencia recabada reflejen, válidamente, lo que se desea medir; de otra manera, no se podrán prevenir explicaciones perniciosas para el sustentante y terceros interesados.

El principal objetivo del presente estudio es validar evidencias (puntajes) que hipotéticamente representan demandas de dos fuentes de contenido incluidas en los ítems de un banco de examinación de egreso en Contaduría, la operación cognitiva requerida para contestar al reactivo y el campo de conocimiento en el que se evalúa la información que el reactivo presenta. La literatura internacional respecto de la generación de ítems (Bejar, 2002; Embretson, 2002; Irvine, 2002) marca la necesidad de comprender a profundidad las fuentes de contenido relacionadas con los ítems, pues hacerlo aporta fundamentos más defendibles para validar los constructos y entender el proceso de responder. Esta fue la meta del estudio descrito aquí.

Tomando en cuenta la necesidad de avanzar la red de evidencias, el trabajo que presentamos utiliza la aproximación multirasgo-multimétodo. En el diseño MRMM dos o más rasgos o características son medidos, cada uno con dos o más métodos. Los rasgos pueden ser habilidades, actitudes, operaciones cognitivas, conductas o características de personalidad, mientras que los métodos hacen referencia a medidas o situaciones variadas, como son el campo de conocimiento que se evalúa, los formatos de reactivo utilizados, el patrón en el que se presenta el contenido del reactivo, entre otros. Al reconocer que los resultados de cualquier medición reflejan tanto el rasgo medido como el método utilizado para medirlo, Campbell y Fiske (1959) señalan que es fundamental recoger

la varianza que explican tanto los rasgos como los métodos en cualquier situación estudiada. Con lo anterior no solo se obtiene una mayor proporción de varianza explicada para cada observación o reactivo de investigación, sino que también se estiman los indicadores de validez convergente y divergente del constructo. La validez convergente hace referencia a relaciones altas y significativas entre variables observadas y las variables latentes correspondientes, mientras que la validez discriminante hace referencia a las correlaciones menores y tal vez no significativas entre algunas variables observadas y uno o más factores que no corresponden —según la teoría— con estas variables observadas (Corral-Verdugo, 1995; Hair, Anderson, Tatham & Black, 1999). Se apoya la validez de constructo cuando la validez convergente y la discriminante son altas y los efectos del método resultan ser menores (Marsh & Grayson, 1995).

Aunque entre los investigadores ha existido una tendencia predominante a buscar relaciones únicas entre constructos e indicadores, en la naturaleza la mayoría de las variables se relacionan significativamente con más de un factor a la vez (Corral-Verdugo, 2002). En el caso particular presentado en este trabajo, se empleó un sistema de análisis que es capaz de probar, cuando menos, la existencia de dos tipos de factores explicando el uso de cada operación cognitiva y cada campo de conocimiento, con el fin de obtener validez de constructo.

Metodología

Participantes

De una población de 462 contadores, hombres y mujeres, egresados de instituciones de educación superior que sustentaron grupal y voluntariamente, bajo procedimiento estandarizado, un examen de egreso de la Licenciatura en Contaduría en 2003, se seleccionaron 313, egresados de universidades públicas y privadas, cuyas respuestas aparecen en ambas fuentes analizadas. La edad promedio fue de 25 años; 58.4%

del sexo masculino, 41.6% del femenino, 70.80% solteros y 78.57% obtuvieron un promedio general de licenciatura entre 8 y 9.5 de calificación. El 85% egresó de instituciones públicas.

Instrumentos

1. *Escala de Valoración de Fuentes de Contenido de Reactivos Objetivos* (Castañeda, 1998; Castañeda, González, López, García-Jurado & Pineda, 2003): instrumento de lápiz y papel que caracteriza y asigna valores de dificultad apriorística a las fuentes de contenido incluidas en los ítems que se utilizan para medir. Fue construida con base en lo que la literatura internacional señala acerca de mecanismos responsables del proceso de responder a examinación (Castañeda & López, 1998; Mislevy, Wilson, Ercikan & Chudowsky op. cit; Pollit & Ahmed, 1999). La escala caracteriza al ítem con base en diversas fuentes de contenido, por ejemplo, las operaciones o procesos cognitivos requeridos para resolver el ítem, los patrones en los que la pregunta y la respuesta requieren interactuar para resolverlo e, incluso, la dificultad del lenguaje, así como la claridad y exactitud en los términos teóricos o técnicos que se utilizan. La escala fue validada por jueces expertos independientes ($Q = 12$, $g.l. = 13$, $p = 0.528$).

En este estudio, solo se utilizaron las fuentes que mostraron predecir una buena proporción de varianza de la dificultad apriorística del ítem y que mostraron efectos razonables sobre su dificultad empírica, definida esta como su ajuste a la tendencia latente (Castañeda, Ortega & García Jurado, 2005). Las fuentes de contenido utilizadas para caracterizar los ítems fueron las siguientes:

- *Operación cognitiva demandada para resolver el reactivo:* definida como el procesamiento cognitivo subyacente a la ejecución requerida para resolverlo. Incluye tres tipos de demanda: a) *de comprensión*: capacidad para identificar, clasificar, ordenar temporalmente y/o jerarquizar información conceptual presentada en el reactivo;

b) *de aplicación*: capacidad de utilizar, en tareas profesionales iniciales y rutinarias, conceptos, principios, procedimientos, técnicas e instrumentación acordes con el nivel científico en el que se les reconoce, y c) *de resolver problemas*: capacidad de evaluar e integrar conceptos, principios, métodos, técnicas, procedimientos, estructuras de tareas y/o planes de acción en función de los principios de adecuación y/o valores profesionales requeridos para resolver situaciones problemáticas, así como para identificar y corregir errores importantes en soluciones preestablecidas.

- *Campo de conocimiento evaluado en el reactivo*: la cualidad de la información que se evalúa en el reactivo e incluye tres tipos: a) *solo teórico*: *factual* (fechas, personajes, lugares y fórmulas), *conceptual* (definiciones de conceptos y reglas) y *procedimental* (definiciones de procedimientos, técnicas e instrumentación); b) *solo técnico* (destrezas técnicas dirigidas a la acción profesional), c) *combinado* (conocimiento teórico y destreza técnica integrados).

Tres expertos independientes, previamente entrenados, clasificaron en 2001 el total de ítems del banco del examen general de egreso en Psicología Clínica; de entre ellos fueron seleccionados los 38 ítems que se utilizaron en este estudio.

2. Un banco intencional de 17 reactivos objetivos que satisficieron el requisito de calibración logística de un parámetro y el de discriminación (PtBs ≥.16). La calibración se realizó mediante el calibrador Rascal (*Assessment Systems Corporation*, 1992) y el índice de discriminación se obtuvo a partir de los resultados del calibrador *BigSteps* (Linacre & Wright, 1994). El criterio de elegibilidad de cada reactivo requirió que este hubiera sido elegido por tres jueces independientes, con base en su ajuste a la tendencia latente y a su índice de discriminación. El tipo de reactivo es objetivo, de opción múltiple, con cuatro opciones de respuesta, de las cuales solo una es correcta y el resto son tres distractores verosímiles.

Procedimiento

La recogida de datos se realizó mediante aplicación estandarizada de un examen de egreso, por aplicadores entrenados en la promoción 2001 del examen.

Calibración de reactivos

La calibración se realizó con datos de 220 reactivos, proporcionados por un centro de evaluación especializado en este dominio[1]. De entre estos, se seleccionaron los 17 que, además de satisfacer los parámetros, mostraron en su mayoría atender a ambas fuentes investigadas. Su distribución se muestra en la Tabla 1.

Tabla 1
Distribución de los ítems por fuentes de contenido

		No. ítems
Campo de conocimiento	Teórico	5
	Técnico	8
	Combinado	4
	Total	17
Operación cognitiva	Comprender	10
	Aplicar	7
	Resorlver	0
	Total	17

A partir de la distribución de ítems (Tabla 1), se calcularon índices de consistencia interna Alfa de Cronbach, por tratarse de datos obtenidos una sola vez (Tabla 2).

[1] Los ítems son propiedad intelectual del centro de evaluación.

Tabla 2
Índice de consistencia interna para las fuentes de contenido

		M	*DE*	α
Campo de conocimiento	Teórico	2.00	1.55	0.67
	Técnico	3.03	2.07	0.68
	Combinado	0.83	0.87	0.18
Operación cognitiva	Comprender	3.96	2.43	0.68
	Aplicar	1.9	1.21	0.35
	Resolver			

Especificación del modelo

El modelo a prueba se representa gráficamente en la Figura 1 y está constituido tanto por la operación cognitiva ("los rasgos") como por los campos de conocimiento ("los métodos"): se teoriza que la ejecución de los sustentantes sobre cada una de las operaciones cognitivas solicitadas se liga a una dimensión del constructo *operación cognitiva*, lo que conforma el componente "rasgos" del modelo, y a una dimensión del constructo *campo de conocimiento,* el componente "métodos". Dada esta adaptación de la estrategia de Campbell y Fiske (1959), podemos hablar de un modelo de multioperación cognitiva-multicampo de conocimiento (MOCMCC).

1. El modelo de MOCMCC está constituido por los constructos *comprender*, *aplicar* y *resolver* como operaciones cognitivas y por los constructos *solo teórico*, *solo técnico* y *combinado*, como campos de conocimiento.

2. Finalmente, cada una de las variables observadas está ligada a dos variables latentes —a un rasgo o tipo de operación cognitiva y a un método o tipo de campo de conocimiento—, como se ejemplifica a continuación.

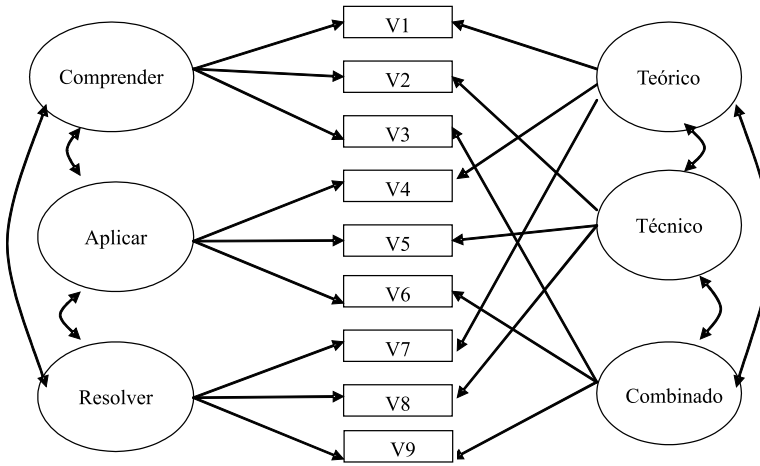

Figura 1. Modelo teórico Multioperación Cognitiva - Multicampo de Conocimiento (MOCMCC).

En la Figura 1, se presenta el modelo hipotético de relaciones estructurales a través del cual se desean validar los constructos *tipos de operación cognitiva* y *tipos de campo de conocimiento*. Las variables latentes son seis constructos, mientras que las variables manifiestas o indicadores son las puntuaciones obtenidas en los ítems que conforman esa dimensión.

Construcción de la matriz

A partir de haber especificado el modelo teórico, se clasificaron los reactivos con base en su doble condición de medir alguna de las tres operaciones cognitivas en alguno de los tres métodos posibles. Los resultados se muestran en la Tabla 3.

Tabla 3
Frecuencia de ítems por campo de conocimiento y operación cognitiva

		Campo de conocimiento			
		Teórico	Técnico	Combinado	Total
Operación cognitiva	Comprender	5	4	1	10
	Aplicar	0	4	3	7
	Resolver	0	0	0	0
	Total	5	8	4	17

Se construyó la matriz MOCMCC y posteriormente se realizó un Análisis Factorial Confirmatorio (AFC); la prueba del modelo incluyó la medición de bondad de ajuste entre el modelo inclusivo y el modelo restringido (o modelo propuesto). El modelo inclusivo refiere una interrelación total de factores y variables observadas y, a pesar de que se acepte que ese tipo de relaciones existe —aunque sea en forma mínima— en la realidad, en ciencia se buscan, sobre la base del principio de parsimonia, modelos simples que expliquen lo más posible.

Para contrastar ambos modelos, se utilizó el estadístico χ^2, que compara el grado de diferencias entre dos modelos. Aquí una χ^2 alta y significativa refiere que los dos modelos sean diferentes, por lo cual debemos buscar una χ^2 no significativa, es decir, cuya probabilidad asociada sea mayor a 0.05, de manera que nos muestre que el modelo restringido no es diferente del modelo inclusivo en términos de poder explicativo. Otros índices de ajuste utilizados fueron el Índice Bentler-Bonet de Ajuste Normado (IBBAN), el Índice Bentler-Bonet de Ajuste No Normado (IBBANN) y el Índice de Ajuste Comparativo (IAC), incluidos dentro del programa EQS (Bentler, 1993). Estos índices producen resultados que van de 0 al 1.0, y se acepta .90 como índice de ajuste adecuado.

Posteriormente, se estimaron las correlaciones entre las variables medidas y los factores, y las covarianzas de las variables latentes entre sí, así como de los errores correspondientes a cada factor. Se buscó que

las relaciones entre las variables observadas y las variables latentes correspondientes fueran altas y significativas, con el fin de que la teoría y la validez de constructo convergente de las medidas fueran confirmadas. Además, se buscó validez de constructo divergente o discriminante pretendiendo mostrar que las correlaciones entre algunas variables observadas y uno o más factores que no corresponden —según la teoría— con estas variables observadas fueran menores y tal vez no significativas (Corral-Verdugo, 2002).

Resultados

Los datos sociodemográficos de la muestra utilizada revelan que la proporción del sexo masculino (58%) fue ligeramente mayor que la del sexo femenino (42%), que la mayoría no tiene pareja —es decir, son solteros, divorciados o viudos— (71%), que están en un rango de edad de 22 a 25 años (56%) y que su promedio en la licenciatura se encuentra en el rango de 8 a 9.5 de calificación (79%).

Configuración dimensional identificada

La Figura 2 representa gráficamente los resultados obtenidos del AFC empleando la estrategia MRMM con las respuestas al banco de ítems utilizado. El modelo multi-operación cognitiva y multi-campo de conocimiento quedó construido por una *operación cognitiva comprender* y un *campo de conocimiento*, el *técnico*, conformados a partir de las variables observadas. Los pesos factoriales significativos entre cada factor y sus indicadores establecen la validez convergente de cada constructo (Gorsuch, 1983). Así se puede establecer que solo la *operación cognitiva comprender* y el *campo de conocimiento técnico* poseen validez de constructo convergente. Al buscar validez de constructo discriminante, los constructos de operación cognitiva presentaron una covarianza significativa, lo que indica que las ejecuciones de los sustentantes no fueron capaces de discriminar entre estos constructos. Algo similar

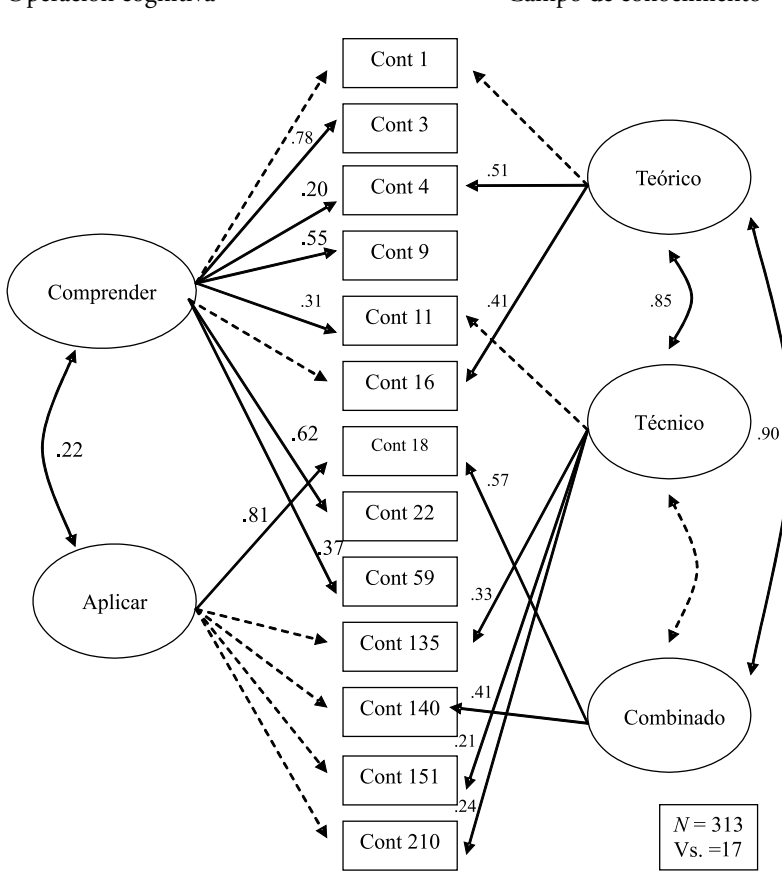

Figura 2. Modelo Multioperación Cognitiva-Multicampo de Conocimiento de un banco de reactivos intencional de un examen de egreso de la Licenciatura en Contaduría. La $\chi^2 = 30$ (*g. l.* = 52), $p = .99$; IBBAN = .92, IBBANN = .99, IAC = .99 y RMSEA = .000. Los pesos factoriales son significativos a $p < .05$. Las líneas discontinuas representan relaciones no significativas.

se presentó entre los constructos *campo de conocimiento teórico*, *técnico* y *combinado*. No se presentó covariación significativa entre el *técnico* y el *combinado*. De tal modo, se puede decir que el modelo no tiene validez divergente.

Los indicadores de bondad de ajuste muestran que los datos respaldan al modelo. La χ^2 resultante fue de 30, con 52 *g. l.*, asociada a una p = .99; el IBBAN fue igual a 0.92; el IBBANN fue de .99; el IAC, considerado el índice más preciso para la medida de ajuste, fue igual a .99; y el RMSEA fue igual a 0.000, pero no se logró computar su intervalo de confianza. Esto significa que este modelo teórico no es significativamente diferente —en cuanto a poder de explicación— del modelo saturado y que es teóricamente plausible (MacCallum & Austin, 2000).

Discusión

En términos del modelo probado, se mostró que, mediante AFC, es posible obtener factores de una operación cognitiva hipotetizada, la de *comprender*, y de un campo de conocimiento, el *técnico*. Los factores identificados explicaron el proceso de responder en este banco intencional de examinación de egreso en Contaduría solo cuando los sustentantes enfrentan las demandas de *comprender conocimiento teórico* (11% de varianza explicada entre sustentantes) y de *comprender conocimiento técnico* (8%), particularmente en sustentantes que comprenden y aplican conceptos y principios de la disciplina para responder un examen de egreso en el área. Aun cuando los puntajes mostraron que los sustentantes ejecutaron significativamente mejor en *comprender* que en *aplicar* (t = 6.52, *g. l.*= 312, p = 0.000), la interpretación de sus puntajes debe tomar en cuenta que los puntajes de estos constructos no lograron validez discriminante entre ellos. Asimismo, la interpretación de los puntajes de los conocimientos *teórico* y *combinado* debe tomar en cuenta que no se configuraron en factores, ni mostraron validez discriminante. Lo mismo sucedió con la operación *resolver problemas*:

al no configurarse como factor la interpretación de que los sustentantes ejecutaron mejor en *comprender* que en *resolver problemas* (t = -13.66, *g. l.* = 312, p = 0.000), deba ser hecha cautelosamente.

Con referencia a los campos de conocimiento, la evidencia mostró diferencias significativas en favor del campo *técnico* sobre la ejecución en el *teórico* (t = -17.38, *g. l.* = 312, p = 0.000) y en la del *combinado* (t = 2.97, *g. l.* = 312, p = 0.000). Sin embargo, los puntajes obtenidos en los campos *teórico* y *combinado* no validaron el constructo hipotetizado, en tanto que sí lo hicieron para el conocimiento *teórico* en la muestra intencional de reactivos provista por el centro de evaluación. De aquí que el diseñador de exámenes pueda ganar confianza en la interpretación de los puntajes derivados de la medición de conocimientos *técnicos*, pero tomar con precaución los del conocimiento *teórico* y los del *combinado*. El índice de consistencia interna (α = 0.18) en el método *combinado* muestra insuficiente homogeneidad entre los reactivos que lo componen.

Así, el modelo probado mostró fortalezas que deben ser tomadas en cuenta para interpretar los puntajes, pero también mostró debilidades que deben ser tomadas en cuenta para que la interpretación de los puntajes sea relevante.

Seis indicadores de la operación *comprender* se relacionaron significativamente con el factor hipotetizado, y solo uno de ellos mostró estar influido significativamente por el conocimiento *teórico*, lo que podría estar mostrando que este factor es más de rasgo que de método. La proporción de varianza explicada por el rasgo *comprender*, en interacción con el método conocimiento *técnico* (19%), constituye evidencia débil en favor del rasgo y en apoyo de la elección del mejor método para medirlo. El índice de consistencia interna del campo *teórico* muestra insuficiente homogeneidad entre los ítems que lo componen (α = .18), lo que no permite confiar en las medidas que lo constituyen. Requiere resolver este problema psicométrico, aumentar la representatividad del constructo y mejorar las medidas a fin de ganar validez en las inferencias a ser hechas.

La relación entre los rasgos *comprender* y *aplicar* (.22), aunque significativa, es menor a los pesos factoriales de las variables medidas y del constructo que las agrupa; los métodos no presentan validez discriminante. El conocimiento *teórico* covaría significativamente con el *técnico* al .85 y con el *combinado* al .90; de aquí que el *teórico* esté midiendo lo mismo que los otros dos. La única validez discriminante encontrada fue entre el conocimiento *técnico* y el *combinado*.

Las operaciones cognitivas *aplicar* y *resolver problemas* no se configuraron como factores, a diferencia de lo que teóricamente se esperaba. Esto puede deberse a que sus medidas mostraron covariar muy bajo o negativamente con el constructo correspondiente, y a que en *resolver problemas* existe una sub-representación de constructo (Messick, 1995) que pone en riesgo la validez de su medición. Una situación psicométricamente similar fue encontrada en las dimensiones *campo de conocimiento teórico* y *combinado*, que no pudieron ser conformadas como factores.

En resumen, la estrategia utilizada permitió recoger varianza que explicó la ejecución de los sustentantes en función de uno de los rasgos medidos influido, a su vez, por uno de los métodos en los que fue medido. También permitió modelar las relaciones estructurales entre los constructos hipotetizados, de manera integral y con carácter confirmatorio. Así, ha sido factible entender cómo la interacción entre las diversas fuentes de contenido incluidas en los ítems investigados explica o no los resultados en una examinación a gran escala en una disciplina particular. El modelo obtenido, además de verificar la validez de constructo de las medidas analizadas, permitió corroborar la pertinencia de estudiar el proceso de responder bajo la perspectiva AFC con matriz multirasgo-multimétodo.

Y, en la medida en la que la validez siempre es aproximada como hipótesis —del significado interpretativo deseado a partir de los datos generados por la medición—, se hace necesario enfatizar aquí la impor-

tancia de generar una cadena de evidencias que ligue la interpretación de los puntajes a las redes de teoría e hipótesis que han sido utilizadas para que los datos soporten o refuten la racionalidad de la interpretación de los puntajes de los sustentantes.

En el contexto empírico de construir exámenes objetivos a gran escala, las agencias evaluadoras podrán tomar ventaja de evidencias como las generadas aquí, que le den al constructor de exámenes confianza en sus medidas, particularmente en países cuya tradición en este tipo de medición y para efectos de certificación de conocimientos es reciente, como es el caso de México. Si el éxito en las tareas de un examen es una muestra representativa del éxito en las tareas deseadas en el dominio de conocimiento, las inferencias a ser hechas podrán sostenerse cuando los puntajes constituyan evidencia empírica sólida para los constructos hipotetizados.

Debido a que la medición es un proceso inferencial, estudios que validen diversas fuentes de contenido constituyen líneas de generación de evidencia importantes en favor del atributo a ser medido. Señalan, también, nuevas líneas de investigación para una mejor caracterización del fenómeno de responder exámenes objetivos aplicados a gran escala. Es a través de evidencias sólidas que el especialista tendrá seguridad en que el dato obtenido apoya las inferencias a ser hechas.

Cabe enfatizar que los resultados de este estudio solo pueden ser generalizados bajo el modelo teórico puesto a prueba y en las condiciones en las que fueron recolectados los puntajes. Las evidencias empíricas obtenidas al analizar la estructura de los constructos solo describen la utilidad del enfoque utilizado respecto de estos constructos y con la muestra intencional de reactivos utilizada. Si bien los hallazgos sugieren mejora en las medidas de las operaciones *aplicar* y *resolver problemas*, y en los *campos de conocimiento teórico* y *combinado*, estos no pueden generalizarse a la totalidad de la población que sustentó el examen de egreso a partir del cual se extrajeron los datos utilizados en este estudio,

debido a que las muestras de ítems y de sustentantes fueron intencionales. Otra limitante del estudio fue que el banco de ítems no incluyó todos los ítems del banco extenso. Será necesario desarrollar estudios tomando en cuenta mayor número de ítems por cada factor.

Entender el proceso de responder a examinación de egreso requiere, indiscutiblemente, validar fuentes de contenido incluidas en los ítems, en los que la complejidad creciente entre las operaciones cognitivas demandas —desde comprender y aplicar fundamentos disciplinares hasta resolver problemas de la vida profesional— y los tipos de conocimiento en los que es evaluada la operación reflejen, como sería de esperarse en un examen de egreso, la compleja interacción entre categorías heterogéneas de conocimientos, habilidades y valores que caracterizan al aprendizaje derivado de la formación profesional inicial. La examinación objetiva, a gran y pequeña escala, necesita en el futuro inmediato representar mejor los constructos subyacentes a la examinación, construir o seleccionar medidas válidas a esos constructos y validar las evidencias que el arreglo de medición genera para estar en capacidad de elaborar las inferencias a ser hechas.

Referencias

Assessment Systems Corporation. (1992). RASCAL, *Rasch Analysis Program*, *version 3.5*. St. Paul, MN: Autor.

Bejar, I. (2002). Generative testing: From conception to implementation. En S. H. Irvine & C. P. Kyllonen (Eds.), *Item generation for test development* (pp. 199-218). Hillsdale, NJ: LEA.

Bentler, P. M. (1993). *EQS: Structural Equations Program Manual.* Los Ángeles: BMPD Statistical Software.

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validations by multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Castañeda, S. (1993). *Procesos cognitivos y educación médica*. México: UNAM.

Castañeda, S. (1998). Evaluación de resultados de aprendizaje en escenarios educativos. *Revista Sonorense de Psicología, 12*(2), 57-67.

Castañeda, S. (2002). A cognitive model for learning outcomes assessment. *International Journal of Continuing Engineering Education and Life-long Learning, 12*(1-4), 106.

Castañeda, S. (2003). Construyendo puentes entre la teoría y la práctica. *Pensamiento Educativo, 32*, 155-176.

Castañeda, S., González, D., López, O., García-Jurado, R. & Pineda, L. (2003). *Escala de valoración de fuentes de contenido en ítems objetivos*. Documento de trabajo del proyecto de investigación CONACYT 40608-H.

Castañeda, S. & López, M. (1998). Elaboración de un instrumento para la medición de conocimientos y las habilidades en estudiantes de psicología. *Revista Intercontinental de Psicología y Educación, 1*, 9-15.

Castañeda, S., Ortega, I. & García-Jurado, R. (2005, julio). *Exploring aprioristic and empirical difficulties in sources of content of large-scale objective exam items*. Documento presentado en el IX Congreso Europeo de Psicología, Granada, España.

Corral-Verdugo, V. (1995). Modelos de variables latentes para la investigación conductual. *Acta Comportamentalia, 3*, 171-190.

Corral-Verdugo, V. (2002). Structural equation modeling. En R. Bechtel & A. Churchman (Eds.), *Handbook of environmental psychology* (pp. 256-270). Nueva York: John Wiley.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*(1), 179-197.

Embretson, S. E. (1999). Cognitive psychology applied to testing. En F. T. Durso, R. S. Nickerson, R. W. Schvanevelt, S. T. Dumais, D. S. Lindsay & M. T. H. Chi (Eds.), *Handbook of applied cognition*. Nueva York: John Wiley.

Embretson, S. E. (2002). Generating abstract reasoning items with cognitive theory. En S. H. Irvine & P. C. Kyllonen (Eds.), *Item*

generation for test development (pp. 219-250). Hillsdale, NJ: Lawrence Erlbaum.

Gorsuch, R. L. (1983). *Factor analysis.* Hillsdale, NJ: Erlbaum.

Hair, J., Anderson, R., Tatham, R. & Black, W. (1999). *Análisis multivariante.* Madrid: Prentice Hall Iberia.

Hornke, L. F. & Habon, M. W. (1986). Rule-base item bank construction and evaluation within the linear logistic framework. *Applied Psychological Measurement, 10*, 369-380.

Irvine, S. H. (2002). Item generation for test development: An introduction. En S. H. Irvine & C. P. Kyllonen (Eds.), *Item generation for test development* (pp. 199-218). Hillsdale, NJ: LEA.

Linacre, J. M. & Wright, B. D. (1994). *A user's guide to BigSteps*. Chicago: MESA.

MacCallum, R. C. & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology, 51*, 1-226.

Marsh, H. W. & Grayson, D. (1995). Latent variable models of multitrait-multimethod data. En R. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 177-198). California: SAGE.

Messick, S. (1989). Validity. En R. L. Linn (Ed.), *Educational measurement* (3a. ed.). Nueva York: American Council of Education & McMillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 32*(2), 13-23.

Messick, S. (1995). Validity of psychological assessment. Validation on inferences from person's responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

Mislevy, R. (1993). Foundations of a new test theory. En N. Frederiksen, R. Mislevy & I. Bejar (Eds.), *Test theory for a new generation of test.* Hillsdale, NJ: Lawrence Erlbaum.

Mislevy, R. J., Wilson, M., Ercikan, K. & Chudowsky, N. (2003). Psychometric principles in student assessment. En T. Kellaghan & D. L. Stufflebeam (Eds.), *International handbook of educational evaluation*. Ámsterdam: Kluwer Academic Press.

Pollit, A. & Ahmed, A. (1999, mayo). *A new model of the question answering process*. Documento presentado en la XXV Conferencia Anual de la International Association for Educational Assessment, Bled, Eslovenia.

Wiley, D. E. (2002). Validity of constructs versus construct validity of scores. En H. I. Braun, D. N. Jackson & D. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 207-227). Hillsdale, NJ: Lawrence Erlbaum.

Recibido 5 de enero, 2006
Aceptado 25 de abril, 2007