Avances en Medición, 6, 85-100
2008

CONSTRUCCIÓN Y VALIDACIÓN DE PRUEBAS DE EXPRESIÓN ESCRITA EN LA UNIVERSIDAD DE COSTA RICA

Tania Elena Moreira–Mora*
Ministerio de Educación Pública de Costa Rica
Universidad de Costa Rica, Costa Rica

Resumen

En este artículo se describen los procesos de construcción y de validación de las pruebas de expresión escrita que han requerido diversos procedimientos técnicos, por tratarse de dos pruebas diferentes: Ensayo y selección única. En la primera se mide la habilidad comunicativa; y en la segunda el uso adecuado del lenguaje, ambos aspectos relacionados con la inteligencia fluida y cristalizada del sistema de tres estratos propuesto por Carroll en 1993. En el caso del ensayo se construyó una escala ordinal, validada por un equipo de jueces, y se realizó una experiencia piloto en el 2006. En la prueba de selección única los ítems son igualmente juzgados para obtener evidencias de validez de contenido y se han aplicado dos pruebas pilotos. En definitiva, esta experiencia ha significado un reto teórico y metodológico para garantizar los estándares técnicos de confiabilidad, objetividad y validez en ambas pruebas.

Palabras clave: Constructo, confiabilidad, validez, ensayo, selección única.

Abstract

This article describes different technical procedures of validity and construction of written expression tests ability; it is because one is an essay test, that measures communicative ability; and the other is a multiple choice format that measures correct usage of language; both of them are related to the fluid and crystallized intelligence of the three strata by John B. Carroll. According to the essay test, it was made a scoring rubric which is validated with some judges and it was made a pilot test in 2006. The multiple-choice item format the items are judged to get evidences of content validity and it were administered two pilot tests. So, experiences in construction and validity items have been a theoretical and methodological challenge to guarantee the technical standards in reliability, objectivity and validity in both tests.

Key words: Construct, reliability, validity, essay test, multiple-choice item.

Introducción

El propósito de este artículo es describir los procedimientos de construcción y de validación de las pruebas de expresión escrita del proyecto de Pruebas Específicas de la Universidad de Costa Rica. Este proyecto nace de una iniciativa por mejorar los indicadores utilizados en la admisión de estudiantes a las carreras universitarias y dentro de esta coyuntura, el Instituto de Investigaciones Psicológicas realizó a finales del 2005 un sondeo a diversas unidades académicas de la Universidad de Costa Rica para determinar el interés en el desarrollo de pruebas específicas para ingreso a carrera, con la intención de añadir poder de predicción al proceso de selección, el cual está limitado actualmente por la Prueba de Aptitud Académica (PAA) y por el promedio de Educación Diversificada, calculado con las puntuaciones de las asignaturas objeto de medición en el programa de Bachillerato en Educación Media. Con este sondeo se destacó la relevancia del cumplimiento de los principios de excelencia y equidad en pruebas de habilidades específicas y no generales. Esta argumentación fue la base del proyecto Construcción de

* Apartado postal: 1437-1100 Tibás, San José, Costa Rica, Centroamérica. E-mail: tmoreira@costarricense.cr

Pruebas Específicas para el Ingreso a Carrera: Fase inicial y experiencia piloto, conformado inicialmente por tres pruebas: Expresión escrita, habilidades cuantitativas e inteligencia fluida.

El principal problema de investigación, en el caso particular de las pruebas de Expresión Escrita, fue determinar cuáles contenidos curriculares y/o competencias lingüísticas son esenciales para un desempeño académico exitoso en una carrera universitaria. En respuesta a este interrogante, se seleccionaron dos habilidades: (a) Comunicativa, y (b) el uso adecuado del lenguaje, bajo la premisa que tales competencias lingüísticas exigen el dominio de ciertas habilidades primarias, por un lado, la organización estructural del texto, la coherencia interna y las estrategias de razonamiento en la prueba de ensayo; y por otro lado, los conocimientos de léxico, morfosintaxis y ortografía en el examen de selección única.

En consonancia con el propósito de esta investigación, se optó por complementar la teoría clásica de los test (TCT) con otros modelos, considerando algunos antecedentes en este campo de investigación, básicamente por tres razones:

1. *La necesidad de emplear nuevas herramientas en el análisis de los resultados*. En este sentido, destacan Zuñiga y Montero (2007), si se cuenta con herramientas útiles para el análisis, se garantiza la calidad técnica de las pruebas y con ello se contribuye a la toma de decisiones adecuadas, según las necesidades de los usuarios e investigadores.

2. *La tendencia actual de complementar la TCT con otros modelos de medida debido a ciertas limitaciones*. Estas críticas se han orientado, específicamente, en cinco dimensiones, señaladas por Gómez e Hidalgo (2003): (a) la definición de pruebas paralelas es muy restrictiva, (b) homocedasticidad de los errores de medida, (c) dependencia muestral de las propiedades psicométricas de las pruebas, (d) dependencia muestral de las características de los ítems, y (e) dependencia de la puntuación observada de la longitud de la prueba.

3. *La búsqueda de instrumentos de medida que se ajusten a un marco que no sea estrictamente estadístico*, sino también al estudio de los procesos cognitivos implicados en la resolución de los ítems (Elosua, 2003).

De esta forma, en este proceso de construcción y validación de las pruebas de expresión escrita se pretende obtener la mayor evidencia empírica tanto con medidas de la TCT como de la *Teoría de Respuesta al Ítem* (TRI) y de la *Teoría de la Generalizabilidad* (Teoría G).

Modelo teórico

El modelo teórico de las pruebas de expresión escrita es el *modelo jerárquico* propuesto por Carroll en 1993, basado en la premisa que la inteligencia tiene una estructura jerárquica. De acuerdo con Lamas (2006), este enfoque tiende a combinar la naturaleza unitaria de la inteligencia con explicaciones factoriales, al considerar la inteligencia como un constructo superordenado y a los factores como entidades subordinadas a la estructura general. Específicamente, en el modelo de Carroll, generado a partir de un modelo factorial exploratorio, las aptitudes o habilidades se clasifican en tres niveles; en el primero se ubican las *habilidades primarias*, en el segundo las *habilidades de generalidad amplia* y en el tercero *la capacidad general de inteligencia* o "*factor g*" de Spearman que se constituye en un rasgo fuente (De Juan-Espinosa, 1997).

Dentro de este sistema o mapa de habilidades, las habilidades del segundo estrato son conocidas como Inteligencia fluida, inteligencia cristalizada, memoria y aprendizaje general, percepción visual, percepción auditiva, capacidad de recuperación, velocidad cognitiva y, finalmente, rapidez de procesamiento y de

decisión. Es de interés particular en esta investigación la inteligencia cristalizada, al englobar aquellas aptitudes relacionadas con el lenguaje y al suponer "*la cristalización de la aptitud fluida en destrezas de comprensión y razonamiento general mediante la exposición a los problemas que se presentan en el ámbito familiar y escolar*" (De Juan-Espinosa, 1997, p. 148). Este es el estrato que representa la inteligencia tradicional, que compete a las habilidades verbales, mecánicas, numéricas y sociales.

La inteligencia fluida, por su parte, es una especie de energía o eficacia neuronal capaz de fluir a través de diversos tipos de actividades, que implica procesos básicos de razonamiento y otras actividades mentales que dependen escasamente del aprendizaje y la aculturación (De Juan-Espinosa, 1997). En suma, la inteligencia fluida es innata e independiente del contexto cultural, la cual permite a los sujetos actuar de forma adecuada en nuevas situaciones; mientras que la inteligencia cristalizada es construida con base en la experiencia y permite refinar procesos ya conocidos.

El factor g, constructo principal, es medido con dos pruebas: Una de ensayo y la otra de selección única. En la primera se mide la capacidad comunicativa y en la segunda el uso adecuado del lenguaje, como se ilustra en la Figura 1

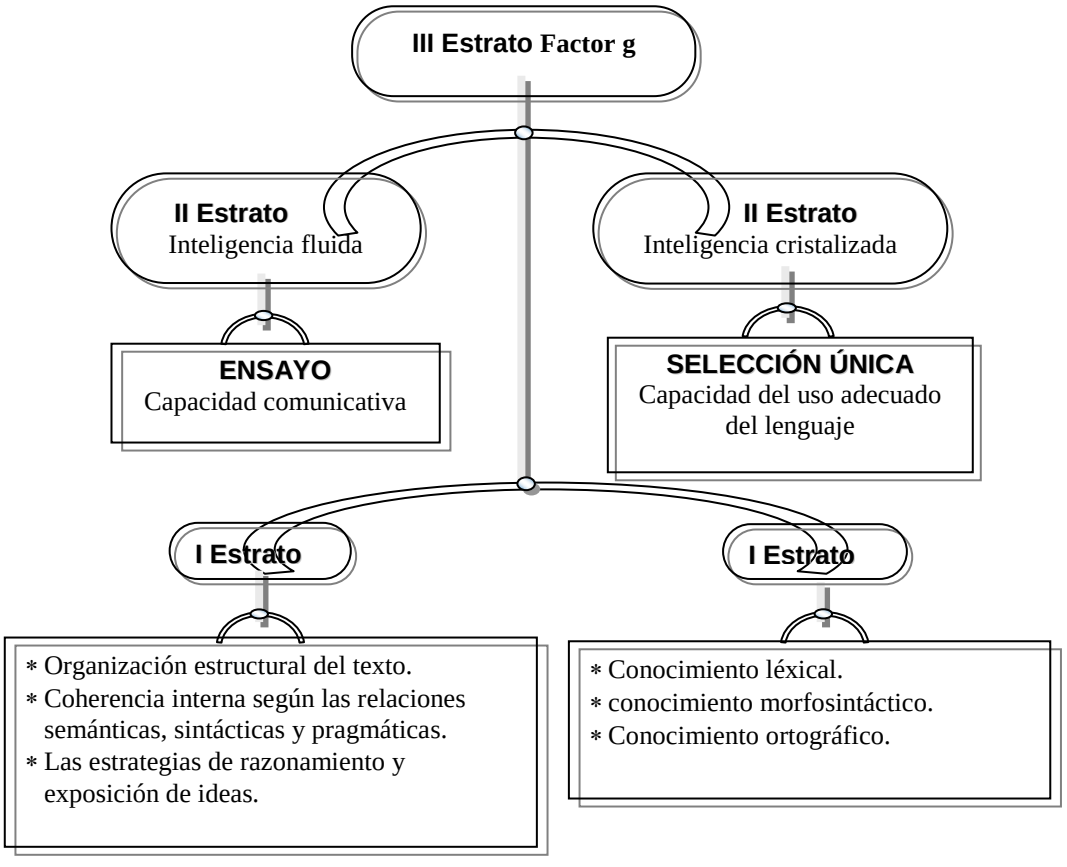


Figura 1: Modelo jerárquico de las pruebas de Expresión Escrita basado en la teoría de los tres estratos de Carroll.

La prueba de ensayo tiene como finalidad medir la competencia comunicativa de los estudiantes mediante tres habilidades básicas: La organización de los elementos del texto, la coherencia interna de acuerdo con las relaciones semánticas, sintácticas y pragmáticas y, finalmente, las estrategias de razonamiento y expositivas.

La organización estructural del texto se conceptualizó como un sistema en que se organizan los elementos o apartados de un texto, caracterizado por la relación de cada parte con el conjunto entero, en una relación de introducción, desarrollo y conclusión. La segunda habilidad primaria se basa en la cohesión interna de las ideas, en virtud de las relaciones semánticas, sintácticas y pragmáticas que mantienen entre sí, y a una relación secuencial, global o temática. Finalmente, en la tercera habilidad se miden las estrategias de razonamiento y expositivas empleadas por el examinado para comunicar las ideas.

El procedimiento para la construcción de la prueba de ensayo implicó, en un primer momento, comprender el proceso cognitivo subyacente en la construcción del ensayo. De acuerdo con Meza, D'Agostino y Cruz (2003), este proceso conlleva: (a) Definir el propósito de la comunicación, (b) revisar información, (c) establecer posición personal, (d) planificar la exposición, (e) concretar el plan, y (f) identificar recursos lingüísticos disponibles.

En cuanto a la prueba de selección única, el objetivo básico es medir el uso adecuado del lenguaje, con base en los conocimientos normativos de ortografía, léxico y morfosintaxis que, de acuerdo con el modelo de Carroll, constituyen el primer estrato. El conocimiento lexical significa que unas personas son más capaces que otras de lograr, mediante la lectura y otras experiencias, un vocabulario amplio y elevado (De Juan-Espinosa, 1997). En concreto, el dominio lexical en esta prueba se mide con el conocimiento de antónimos, de acuerdo con el contexto del enunciado. Los conocimientos ortográficos competen al conjunto de normas que regulan la escritura de una lengua (Real Academia Española, 1999); específicamente, en esta prueba los relativos a las reglas generales de acentuación, empleo de las mayúsculas y de las reglas ortográficas. Finalmente, la morfología identifica las categorías gramaticales y la sintaxis les atribuye funciones. En efecto, como apuntan Piera y Varela (1999), la morfología y la sintaxis tienen un vocabulario compartido, el que identifica a las clases de palabras o categorías gramaticales (sustantivo, adjetivo, verbo…), y éstas se reconocen tanto por su función en la oración como por sus marcas morfológicas. Por ejemplo, en la oración "La casa es amplísima", el vocablo amplísima es un adjetivo con sus marcas morfológicas de grado -ísim- y de género –a, que cumple la función de atributo (sintaxis).

En síntesis, la construcción y validación de las pruebas de expresión escrita requirió tanto de la revisión teórica de los modelos cognitivos, linguísticos y psicométricos, así como de la experticia de los jueces y las evidencias empíricas.

Método

El método empleado es el descriptivo, por ser el más apropiado para lograr los propósitos de esta investigación: La descripción de los procesos de construcción y de validación de las pruebas de expresión escrita, tanto de los procedimientos técnicos como de los resultados de la prueba de ensayo y de selección única.

Participantes

La aplicación piloto de la prueba ensayo fue durante los meses de septiembre y octubre del 2006, con una muestra de 63 examinados de la Facultad de Derecho, Escuela de Computación e Informática, y de

Farmacia de la Universidad de Costa Rica. En la prueba de selección única, la primera aplicación se realizó en el 2006 con una prueba de 63 ítems aplicada a una muestra de 43 examinados.

En el segundo pilotaje se aumentó el tamaño de la prueba (85 ítems) y se aplicó en el mes de mayo del 2007 a una muestra cautiva de 246 estudiantes de primer ingreso de Farmacia, Computación e Informática, Derecho y Estudios Generales.

En la tercera prueba piloto se decidió mantener la misma longitud de la prueba (85 ítems). La principal limitación de estas primeras experiencias fue la aplicación en muestras de una población cautiva, estudiantes que ingresaron a la Universidad de Costa Rica y cursaban el primer año de la carrera o de la Escuela de Estudios Generales. Esta situación se debe principalmente a las limitaciones económicas para lograr la aplicación en la población meta de este proyecto, aspirantes que no han ingresado a la universidad. La principal consecuencia de esta restricción es la subestimación o sobrestimación de los parámetros de dificultad y discriminación de los ítems, por las diferencias en el perfil de los examinados.

En todas las aplicaciones piloto se informó a los estudiantes que estas pruebas eran de una investigación del Instituto de Investigaciones Psicológicas de la Universidad de Costa Rica y que la participación era voluntaria. Asimismo, los resultados se entregaron, en forma impresa, por número de identificación (sin nombre ni apellidos) para garantizar la confidencialidad y se les notificó que esas puntuaciones no tendrían consecuencias académicas.

Instrumentos

En la prueba de ensayo tres expertos en el campo de la Lingüística y la Filología definieron, en términos conceptuales y operativos, las habilidades primarias por medir y, posteriormente, construyeron la rúbrica para calificar esta prueba. La evaluación de esta rúbrica fue hecha por cuatro especialistas, quienes, en un primer momento, realizaron una revisión general de los indicadores y luego incorporaron algunos ajustes en la dimensión operativa, a partir de una aplicación experimental de este instrumento en una muestra de los ensayos.

Después de esta valoración, se establecieron doce indicadores con cuatro valores escalares (de 0 a 3). En caso de la habilidad estructural, se fijaron seis relacionados con: Introducción, desarrollo, conclusión, segmentación lógica de los párrafos, relación de los párrafos con el tema seleccionado y la relación de las ideas del párrafo con su tópico. En la segunda habilidad, (coherencia y unidad temática), los indicadores fueron relación lógica entre ideas del párrafo, desarrollo de las ideas, léxico, uso de conectores o marcadores discursivos, mantenimiento de los referentes textuales. Finalmente, en la habilidad de las estrategias discursivas, el indicador correspondiente fue el uso de argumentos pertinentes y consistentes para sustentar las ideas. Es así como, con el uso de esta rúbrica se pretende, por un lado, minimizar la subjetividad en la calificación de las jueces y, por el otro, lograr una descripción del desempeño del estudiante en las tres habilidades primarias.

Una vez definidas las habilidades por medir y la respectiva rúbrica, el paso siguiente fue la selección de los temas tomando como referencia los tópicos transversales de los programas de estudio de la Educación General Básica de Costa Rica, con la intención de garantizar un dominio básico de los temas para no afectar la medición de la habilidad comunicativa debido a la ausencia de conocimientos sobre los temáticas propuestas. Finalmente, se redactó un instructivo de aplicación de la prueba con dos propósitos: Guiar al estudiante y estandarizar la administración.

En cuanto a la prueba de selección única, uno de los primeros pasos fue responder al interrogante de cuáles eran las habilidades básicas para medir el uso adecuado del lenguaje. Este cuestionamiento implicó,

en primera instancia, la especificación teórica de aquellas habilidades básicas relacionadas con el uso del lenguaje. Una vez definidas estas capacidades, se revisaron los objetivos y contenidos del programa de estudios de la asignatura de Español del nivel de noveno (correspondiente al último nivel de la Educación General Básica en Costa Rica), con la intención de seleccionar los contenidos curriculares concernientes a las normas de acentuación, uso de las mayúsculas y las reglas ortográficas de las letras más frecuentes, calificados de elementales en el uso empírico del lenguaje.

Posteriormente, se construyó la tabla de especificaciones con los objetivos y contenidos curriculares seleccionados, que fueron validados por cuatro expertas del campo de la Lingüística y del Español, quienes además asignaron su respectiva ponderación porcentual. De esta manera, la tabla de especificaciones permite garantizar la cobertura de todos los aspectos relevantes del referente y apreciar el peso en cantidad de ítems correspondientes a cada aspecto (Ravela, 2006). A partir de la validación de los jueces, al componente de la correcta escritura le correspondió un 45%, al de morfología y sintaxis se les asignó otro 45%, y a léxico, específicamente al empleo de los antónimos un 10%.

Es importante destacar que tanto en la aplicación piloto del 2006 como en la primera del 2007, no se utilizó la tabla de especificaciones para la construcción la prueba de selección, puesto que el interés en estas dos primeras experiencias fue probar diferentes formatos de los ítems y obtener evidencias del nivel de dificultad y discriminación de los ítems. Es la segunda prueba piloto del 2007 la que se construye con esta estructura con el propósito de obtener evidencias adicionales de validez de contenido.

Con respecto a los formatos más apropiados para medir los tres componentes se optó, en el caso de los ítems para medir la correcta escritura, por el modelo de dos palabras con errores y cuatro opciones. En los ítems de morfosintaxis se comprobó que el formato más apropiado era la redacción de una o más oraciones con un probable error y cuatro opciones. En el tercer componente de antónimos, se construyó una oración con una palabra subrayada (encabezado) y cuatro palabras (opciones). Finalmente, se establecieron algunos estándares técnicos en este proceso de construcción de los ítems en una guía para orientar tanto a los constructores externos como a los miembros del equipo, quienes también participaron en este proceso. Entre otros estándares, se citan los siguientes:

Medición de los objetivos y áreas de contenido.

Longitud adecuada.

Empleo de un lenguaje sencillo.

Estructuración acorde con el formato seleccionado.

Cumplimiento de las características deseables.

Exclusión de contenidos controversiales y de estereotipos.

Eliminación de opciones absurdas y obvias.

Posición al azar de la clave.

Una vez construidos los ítems se procedía con el juicio, realizado por los miembros del equipo, quienes valoraban el cumplimiento de las normas técnicas establecidas en la guía de construcción para tomar una de las siguientes decisiones: Eliminar el ítem, aceptar con modificaciones y aceptar sin modificaciones. De acuerdo con esta experiencia, la mayoría de los ítems fueron aceptados, pero debían ser modificados para cumplir con los estándares técnicos.

La etapa siguiente consistió en el montaje de la prueba conforme con los criterios preestablecidos para cada una de las pruebas pilotos, cumpliendo con dos normas básicas: Ordenación por nivel de dificultad y distribución conforme con la tabla de especificaciones. Importante acotar la decisión de elaborar dos fórmulas con la finalidad de evitar el fraude durante la aplicación.

Procedimiento

En términos generales, el procedimiento para la construcción y validación de la prueba de ensayo y de selección única se sintetiza en las siguientes etapas: Revisión teórica para construir el marco conceptual, elaboración de la tabla de especificaciones y la rúbrica, construcción de ítems acorde con los estándares técnicos, montaje de las pruebas según las especificaciones, aplicaciones pilotos y análisis de los resultados.

Es necesario destacar que, a partir de las dos últimas etapas, se ha logrado depurar tanto los procesos técnicos como administrativos de ambas pruebas.

Resultados

Prueba de ensayo

Los resultados de la prueba se ensayo fueron analizaron con base en la teoría de la generalizabilidad (teoría G). La razón para implementar este modelo es la necesidad de garantizar la confiabilidad entre los calificadores. En este sentido, Cohen y Swedlik (2001) afirman que la confiabilidad entre jueces es el grado de acuerdo o consistencia que existe entre dos o más jueces y, por lo general, los examinados, sin importar quién realiza la evaluación, esperan que sean evaluados de la misma forma.

La teoría de la generalizabilidad (Teoría G) fue desarrollada por Cronbach, Gleser y Rajaratnam en 1963 y, con frecuencia, se ha utilizado en el desarrollo de la evaluación del desempeño para identificar la magnitud relativa de las múltiples fuentes de error de medición y para hacer proyecciones concernientes a cómo incrementar la confiabilidad de las puntuaciones (Chiu & Wolfe, 2002). Por ello, se ha asumido como una ampliación de la teoría clásica de los test (TCT), puesto que permite la aplicación de procedimientos basados en los modelos de análisis de variancia (ANOVA) y de diseño experimental a los datos (Nunnally & Bernstein, 1995; Martínez, 2005).

El principal objetivo de la teoría G es identificar y estimar la magnitud de las distintas fuentes de variación que pueden intervenir en las diferencias entre las puntuaciones o la variación debida a las puntuaciones del universo y a las múltiples fuentes de error (Martínez, 2005).

En este estudio se empleó el modelo estadístico de efectos aleatorios de dos facetas. Concretamente, los estudiantes constituyeron el objeto de medida (p), quienes fueron calificados por dos jueces (c), con base en una rúbrica constituida por 13 ítems (i). De esta forma, las dos facetas o factores la conforman las calificadoras y los ítems, cuyo diseño se representa como: p x c x i. Entonces cada puntuación observada X_{pci} representa la calificación del estudiante (p) dada por una calificadora (c) en cada uno de los ítems (i).

En este caso, la muestra de examinados fue de 44, por tanto, en el modelo se incluyeron 1144 puntuaciones observadas (44 x 2 x 13). De esta forma, según Martínez (2005), la variancia de las puntuaciones observadas en la población se define como la suma de los componentes de la variancia de los siete efectos principales resumidos en la siguiente fórmula:

$$\sigma 2\,(Xpci) = \sigma 2p + \sigma 2c + \sigma 2i + \sigma 2pc + \sigma 2pi + \sigma 2ci + \sigma 2pci,_e \qquad [1]$$

Con respecto a cada uno de componentes de variancia se asume que:

Los estudiantes difieren en cuanto a su competencia comunicativa (σ^2_p).

Las calificadoras son diferentes, tanto por sus rasgos personalidad como por sus experiencias y conocimientos previos (σ^2_c).

Los ítems muestran variabilidad por su nivel de dificultad (σ^2_i).

La interacción del estudiante con la calificadora (σ^2_{pc}) muestra las inconsistencias en las puntuaciones del estudiante por efecto de las diferencias entre las calificadoras.

La interacción del estudiante con el ítem (σ^2_{pi}) puede reflejar el énfasis o preferencia de los estudiantes por determinados contenidos y por el nivel de dificultad.

La interacción de la calificadora con el ítem (σ^2_{ci}) presenta las inconsistencias de las calificadoras en cada uno de los 13 ítems de la rúbrica.

El componente de la variancia residual ($\sigma^2_{pci,e}$) confunde los efectos de la interacción *p* x *c* x *i* (estudiante x calificadora x ítem) con variaciones aleatorias y otras probables fuentes de variación no incluidas en el diseño. Los resultados de este modelo ANOVA se resumen en la Tabla 1

Tabla 1
Componentes de variancia de la prueba de ensayo (2006)

	Fuentes de variación	Suma de Cuadrados	Gl	Media Cuadrática	Componente de Variancia	Porcentaje
P	Persona	24464,23	43,00	568,94	5,19	9,90
C	Calificadora	13191,41	3,00	4397,14	21,27	40,58
I	Nota _ ítem	2044,53	3,00	681,51	4,09	7,80
Pc	Persona * Calificadora	11583,05	41,00	282,51	21,88	41,73
Pi	Persona * Nota _ ítem	0,00	92,00	0,00	0,00	0,00
Ci	Calificadora * Nota _ ítem	0,00	9,00	0,00	0,00	0,00
Pci	Persona * Calificadora * Nota _ ítem	0,00	33,00	0,00	0,00	0,00
E	Error	0,00	919,00	0,00	0,00	0,00
	Total	51283,23	1143,0	5930,10	52,43	100,00

Variable dependiente: Puntaje. Método: Anova (Tipo uno, suma de cuadrados)

De acuerdo con los resultados, la faceta de las calificadoras explica la mayor variabilidad de las puntuaciones (40.58%). En otros términos, la principal fuente de variación en las puntuaciones de la prueba de ensayo es debido al efecto de las calificadoras (σ^2_c); además, en su interacción con el estudiante (σ^2_{pc}) la variabilidad fue de un 41.73%. Ambos porcentajes son altos, cuando lo deseable es que las puntuaciones varíen, básicamente, por las diferencias en el nivel de competencia de los examinados (σ^2_p), cuyo porcentaje de variancia fue bajo (9.90).

En la teoría G, el grado de generalización de una medida depende del uso de los datos. De acuerdo con Martínez (2005), es frecuente en Psicología y Educación dos formas de usar la medida: Para establecer una ordenación entre individuos o grupos (valor relativo) o para establecer un índice cuantitativo que exprese el nivel absoluto de conocimiento, destreza u otra competencia de un individuo o grupo. El tipo de decisión influye en la definición del error de medida, en la cuantificación de la variancia y en el coeficiente de generalizabilidad. Además, agrega la autora que este coeficiente intenta estimar en qué medida se puede generalizar a partir de las puntuaciones observadas a la media de todas las observaciones o puntuaciones posibles (universo). En este estudio interesa una decisión relativa de los examinados por basarse en el modelo con referencia a normas. Entonces, el cálculo de este coeficiente utiliza los valores correspondientes a cada factor y sus posibles interacciones (ver Tabla 1), según la siguiente fórmula.

$$\sigma_D^2 = \frac{\sigma_{pi}^2}{n_i} + \frac{\sigma_{pc}^2}{n_c} + \frac{\sigma_{pci,e}^2}{n_i n_c} \qquad [2]$$

$$\sigma_D^2 = \frac{0}{13} + \frac{21.88}{2} + \frac{0}{26} = 10.94$$

Donde: n_i es el número de ítems

n_c es el número de calificadoras

Para obtener

$$G = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_D^2} \qquad [3]$$

$$G = \frac{5.19}{5.19 + 10.94} = 0.321$$

Este coeficiente *0,321* evidencia que la variabilidad de las puntuaciones observadas (X_{pci}) de los estudiantes es explicada, además de la variancia de las puntuaciones del universo, por otras fuentes de variabilidad debidas a las condiciones particulares de la medida, como por ejemplo el efecto de los calificadores (40.58%). En suma, con este análisis de la teoría G se demuestra la necesidad de depurar los procesos de capacitación para los y las calificadores(as) con la finalidad de optimizar el uso correcto de la rúbrica (interpretación homogénea) y complementar estos datos con otros análisis de la TCT como el alpha de Cronbach y los coeficientes de correlación intraclase.

Prueba de selección única

En cuanto al análisis de los resultados de la prueba de selección única, desde la perspectiva de la teoría clásica de los test (TCT), se estimaron el nivel de dificultad y discriminación de los ítems, así como el coeficiente de confiabilidad de los resultados.

El índice de dificultad es la proporción de respuesta correcta del ítem del total de examinados que realizaron la prueba. Según Cohen y Swerdlik (2003), el valor de un índice de dificultad puede variar desde el punto de vista teórico de 0 (ninguna respuesta correcta) hasta 1 (todas correctas); si por ejemplo 50 de los 100 examinados contestaron correctamente un determinado ítem, entonces el índice de dificultad sería igual a 50 dividido entre 100 (0.50). Igualmente se puede calcular un índice promedio de la dificultad

de los ítems sumando los índices de dificultad del ítem de todos los reactivos de la prueba y luego dividirlo entre el número total de ítems. La clasificación del índice de dificultad del ítem se realizó en 5 intervalos de clase: *Muy difícil* (de 0.0 a 0.19); *Difícil* (de 0.20 a 0.39); *Intermedio* (de 0.40 a 0.59); *Fácil* (de 0.60 a 0.79); y *Muy fácil* (de 0.80 a 1.0)

En el caso de la prueba piloto 01-2007 la dificultad promedio de la fórmula 1 (F1) fue de 0,68 y de la fórmula 2 (F2) de 0,69 (Ver Tabla 2); ambas ubicadas en un nivel de dificultad fácil. Es necesario apuntar que ambas fórmulas se construyeron con los mismos ítems, pero con un orden aleatorio diferente.

De acuerdo con los datos, los ítems más fáciles tenían un índice de 0,98 (F1) y de 0,99 (F2); mientras que los índices de los más difíciles correspondieron a 0,17 (F1) y a 0,13 (F2).

En cuanto a la discriminación, plantea Lafourcade (1980) que "si una prueba separa convenientemente a los examinados en diferentes niveles de rendimiento, se puede asegurar que es un instrumento que posee un excelente índice de discriminación" (p. 188). Los índices de discriminación de los ítems se pueden clasificar en varios niveles en términos de valores del coeficiente de discriminación, categoría y porcentaje ideal de ítems en esa categoría. Así, un índice de 0 o menor a 0 debería incluir menos del 5% de los ítems; se considera como discriminación Baja los valores entre 0.01 a 0.20, con menos de 15% de ítems en esta categoría; Moderada si toma valores entre 0.21 y 0.40; y Alta si toma valores mayores a 0.41; para estas dos últimas categorías se sugiere un porcentaje ideal mayor al 25% de los ítems (Lafourcade, 1980).

Tabla 2
Datos estadísticos de la fórmula 1 y 2 de la segunda prueba piloto de selección única aplicada en mayo de 2007

Datos estadísticos	**Fórmula 1**	**Fórmula 2**
Casos Válidos	124	122
Confiabilidad	0.843	0.807
Total de ítems	85	85
Dificultad promedio	0.68	0.69
Mínima dificultad	0.98	0.99
Máxima dificultad	0.17	0.13
Discriminación promedio	0.239	0.219
Mínima discriminación	-0.124	-0.255
Máxima discriminación	0.467	0.920
Análisis factorial exploratorio (% de varianza explicada)	10.175	8.891

El índice de discriminación se calcula con el porcentaje de acierto del grupo superior y del grupo inferior, y la diferencia entre los dos porcentajes es el índice de discriminación, que puede encontrarse entre 1,00 y -1,00. En las pruebas estandarizadas con una distribución normal y con muestras grandes se aplica la regla del 27% como punto óptimo, para definir el grupo superior e inferior (Anastasi & Urbina, 1998).

En el caso de los ítems de la prueba de selección única, el índice de discriminación de los ítems más bajo fue de -0.124 (F1) y de -0.255 (F2); mientras que los más altos alcanzaron un índice de 0.467 (F1) y 0.920 (F2). El índice de discriminación promedio de los ítems, calculado como una sumatoria de todos los índices de discriminación dividido entre el total de ítems, fue de 0.239 (F1) y de 0.219 (F2). Según estos

datos, en promedio, el índice de discriminación de ambas fórmulas fue moderado, probablemente porque en esta prueba predominaron ítems con un nivel de dificultad fácil o muy fácil, por tanto, la diferencia entre el porcentaje de acierto del grupo superior e inferior fue baja.

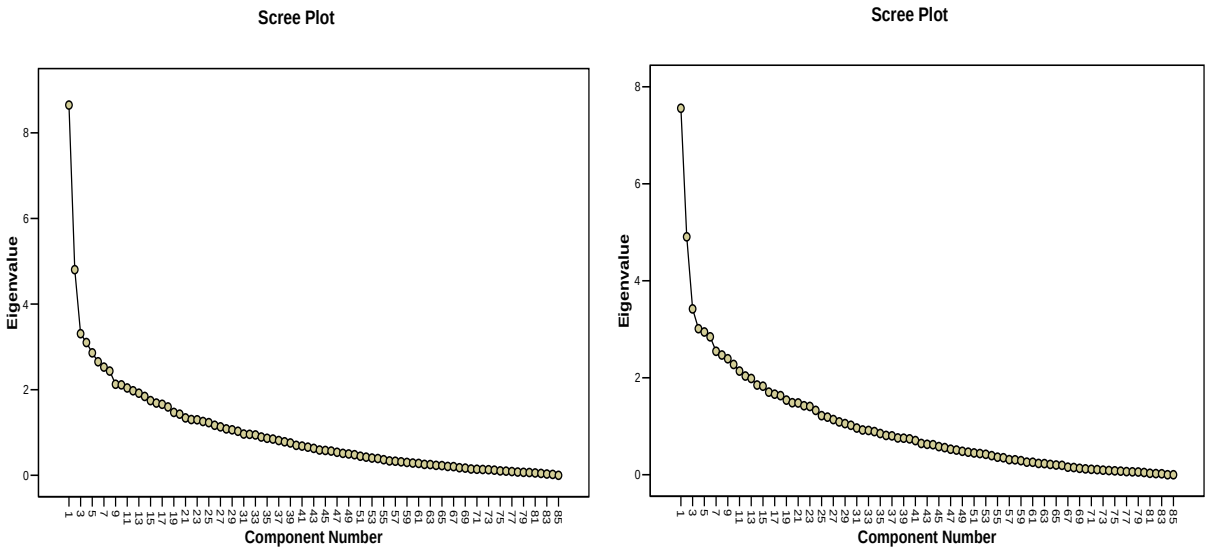
En cuanto al coeficiente de confiabilidad, desde la perspectiva de la TCT, se define como una medida de consistencia interna y de la estabilidad temporal de las puntuaciones. La primera definición recoge el grado de coincidencia existente entre los elementos que la componen y la estabilidad en el tiempo alude a la capacidad del instrumento para arrojar las mismas mediciones cuando se aplica más de una vez a los mismos sujetos (Anastasi & Urbina, 1998; Pardo & Ruiz, 2002). Específicamente en el análisis de la prueba de selección única se empleó el coeficiente del Alpha de Cronbach, asumiendo que la prueba está compuesta por elementos homogéneos que miden la misma característica y que la consistencia interna puede medirse mediante la correlación intraclase existente entre todos sus elementos (Nunnally & Bernstein, 2005; Pardo & Ruiz, 2002). De acuerdo con los resultados, el coeficiente fue de 0.843 (F1) y de 0.807 (F2) a pesar de estar conformados por los mismos ítems. Ambos coeficientes son aceptables para efectos de investigación, pues como lo plantean Nunnaly & Bernstein (2005), el nivel satisfactorio de confiabilidad depende de cómo se use una medida, de tal manera que en las primeras etapas de una investigación de validación predictiva o de constructo, un índice de 0,70 es aceptable (2005, p. 296).

Otros resultados importantes para obtener evidencias asociadas al constructo es el análisis factorial, una de las técnicas más utilizadas para comprobar los supuestos teóricos que sustentan la estructura interna del constructo y sus relaciones con otras variables latentes. Según Martínez (2005), el análisis factorial es un procedimiento de análisis multivariante que intenta explicar mediante un modelo lineal un conjunto de variables observables mediante un número menor de variables hipotéticas, latentes o no observables llamados factores.

El modelo factorial empleado en este análisis fue de tipo exploratorio, con la intencionalidad de obtener evidencias del grado de unidimensionalidad de la prueba, cuya variable hipotética es el uso adecuado del lenguaje, una de las habilidades de expresión escrita medida en esta investigación. Para Martínez (2005), el análisis exploratorio es una técnica muy utilizada para detectar fuentes latentes de variación y covariación en las medidas observadas y suele ser muy útil en los primeros estadios de desarrollo de pruebas. En este análisis se asume que los factores son conceptualmente independientes pero correlacionados, por ejemplo, los conocimientos ortográficos y morfosintácticos son factores independientes, pero se correlacionan en la prueba; por tanto, de las opciones de rotación oblicua se optó por Promax, con un valor Kappa igual a 4 y se escogió el método de extracción análisis de componentes principales.

Una de las principales dificultades es definir el número de factores por ser retenidos en el análisis factorial para evaluar la dimensionalidad de los datos, o en otro sentido, para determinar el grado en que las medidas hipotetizadas de un constructo miden lo mismo. Uno de los criterios más frecuentes ha sido el Scree plot (gráfico de sedimentación), desarrollado por R. B. Catell en 1966, por tratarse de un procedimiento sencillo que separa los factores tempranos importantes de los escombros del error aleatorio, fundamentalmente en un gráfico de la variancia total asociada a cada factor, en orden descendente, contra cada factor, para identificar un punto de transición donde la curva cambia de un descenso fuerte a uno más gradual (Andriola, 2002; Hernández, 1998; Nunnally & Bernstein, 1995). Su principal desventaja es la subjetividad, y según Andriola (2002), presenta problemas cuando las diferencias entre las magnitudes de los autovalores correspondientes a los factores comunes y los factores únicos son muy pequeñas. De acuerdo con el análisis exploratorio de ambas fórmulas (ver Figuras 2a y 2b), se puede observar que sobresalen dos componentes principales en la primera, mientras que en la segunda se retienen tres. Tales

diferencias evidencian que a pesar de ser los mismos ítems en ambas fórmulas, la estructura factorial fue disímil; probablemente ocasionado por las diferencias en el ordenamiento de los ítems.



(a) (b)

Figura 2: Gráficos de sedimentación de la segunda prueba piloto de selección única (01-2007). (a) Fórmula 1, (b) Fórmula 2

Otra conclusión derivada de este análisis es la necesidad de fortalecer el referente de esta prueba con el fin de dar soporte teórico a la organización factorial de las variables latentes hipotetizadas que explicarían las correlaciones de las variables observadas.

En cuanto al otro indicador de unidimensionalidad, según los datos de la Tabla 2, el porcentaje de variancia explicada por el primer componente fue de 10,175 (F1) y de 8.891 (F2). En el caso de la primera fórmula se logra el valor porcentual mínimo aceptable, mas no en la segunda. Al respecto, destaca Andriola (2003), es una cuestión de grado, es decir, cuanta más variancia explique el primer factor, más unidimensionalidad existirá. A partir de ambos indicadores, se puede afirmar que las evidencias de unidimensionalidad de la prueba, en ambas fórmulas, son débiles; sin embargo, como en cualquier proceso de validación de constructo el análisis exploratorio es relevante.

Desde la perspectiva de la Teoría de Respuesta al Ítem (TRI) se analizó la prueba 01-2007 (segundo pilotaje) para complementar los análisis estadísticos de los parámetros de los ítems, por un lado, y para obtener la función de información de la prueba, por el otro. De acuerdo con Montero (2001), el modelo de la TRI es una concepción que, partiendo de ciertas nociones básicas de medición y utilizando como herramientas la Estadística y la Matemática, busca encontrar una descripción teórica para explicar el comportamiento de datos empíricos derivados de la aplicación de un instrumento psicométrico.

Los parámetros estimados por el modelo permiten entonces evaluar la calidad técnica de cada uno de los ítems por separado y de la totalidad del instrumento y, a la vez, examinar el nivel de habilidad de cada sujeto en el constructo de interés. Precisamente Muñiz (1990) apunta que el nombre Teoría de Respuesta a los Ítems proviene de este enfoque centrado en las propiedades de los ítems más que en las de la prueba

global, cuyas principales ventajas son la estimación de la variable latente independientemente del instrumento y la estimación de los parámetros de los ítems invariante a los examinados.

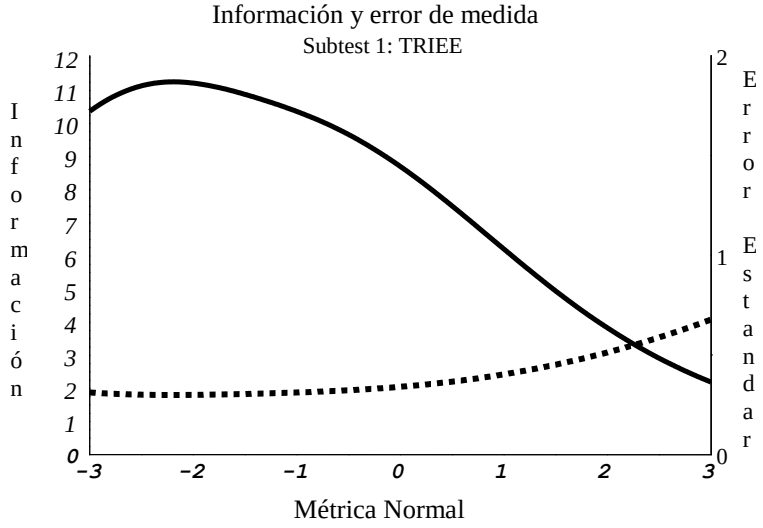


Figura 3: Función de información de la fórmula 1 de la segunda prueba piloto de selección única (01-2007).

El modelo de TRI asume que hay una variable latente θ, no observable directamente y que se desea estimar para cada sujeto a partir de las respuestas dadas en un instrumento de medición. Para ello se puede emplear el modelo Rasch, de un parámetro, que estima únicamente el valor de b (dificultad). En el modelo de dos parámetros se consideran b y a (la discriminación) y en el modelo de tres parámetros debe estimarse también el valor de c o del azar (Cortada de Kohan, 2003; Montero, 2001; Muñiz, 1990; Penfield & Camilli, 2006).

En la interpretación de la función de información (FI) de la prueba es importante considerar la relación recíproca entre la cantidad de información y la variabilidad de las estimaciones de la habilidad. Esto es, a mayor valor de la FI en un nivel θ, la prueba tendrá mayor discriminación para ese nivel y menor error de medida (Martínez, 2005).

Para efectos de ilustración se incluye únicamente la FI de la fórmula 1 de la segunda aplicación piloto (Figura 3). Conforme con este gráfico, se puede afirmar que esta prueba de selección única mide con mayor precisión en niveles bajos de habilidad, próximos al pico de -2, y, por tanto, disminuye el error de medición. A partir de esta evidencia, se deriva una de las principales conclusiones de esta investigación, que para efectos de selección de aspirantes a ingreso de carrera universitaria se debe pensar más bien en una prueba que aporte más información en niveles altos de habilidad.

Conclusiones

En estos dos años de investigación de las pruebas específicas y, particularmente, de las pruebas de expresión escrita, una de las principales conclusiones es que estos procesos de construcción y de validación son elementales para las pruebas con altas consecuencias sociales, como en este caso, la selección de estudiantes que aspiran a ingresar a una determinada carrera universitaria. Igualmente, es una prioridad garantizar los estándares técnicos como la confiabilidad y la unidimensionalidad; así como la obtención de evidencias de validez, en particular predictivas, propósito de esta fase inicial y piloto de las pruebas.

Tal propósito conlleva a un compromiso ético y profesional para desarrollar procedimientos más rigurosos y sistemáticos sustentados en dos pilares: el teórico y el empírico. En el primero se requiere describir y entender el constructo mediante el juicio de expertos y la revisión continua de los procesos cognitivos y modelos teóricos subyacentes en las pruebas. En el campo empírico es clave la evidencia, desde la perspectiva clásica, relacionada con el contenido, el criterio y el constructo mediante el análisis de los resultados y la experimentación. Además, en la construcción de cualquier instrumento la tendencia ha sido, en las etapas iniciales, la obtención de evidencias relativas a la representatividad y relevancia del contenido, mientras que, en las etapas finales, la recolección de evidencias empíricas del constructo medido. En suma, cuantas más evidencias de validez se obtengan de una prueba, más sólido es el diseño y más apropiados sus usos e interpretaciones, especialmente, por las implicaciones educativas.

En estos dos primeros años ha prevalecido la exploración y el ensayo tanto en el formato y contenidos de los ítems de selección única para medir el uso adecuado del lenguaje como en la rúbrica para calificar el ensayo. En relación con los ítems se experimentaron diversos formatos con el fin de seleccionar los más adecuados para medir los tres componentes primarios de la prueba de selección única: Ortografía, morfosintaxis y léxico. En cuanto a la rúbrica, en definitiva, resultaron valiosos la experticia de los jueces, el pilotaje del instrumento para una operacionalización más precisa de los indicadores utilizados en la medición de las tres habilidades primarias: Organización estructural, coherencia interna y estrategias de razonamiento y, finalmente, las evidencias empíricas obtenidas con esta aplicación.

En cuanto a las aplicaciones piloto de ambas pruebas se puede concluir que estas experiencias han contribuido no solo a depurar técnicamente la construcción de las pruebas sino también a la estandarización del proceso administrativo; pues ha implicado, por un lado, la redacción y revisión de los instructivos y, por el otro, la capacitación de los aplicadores para homogeneizar la interpretación de las instrucciones y la administración de las pruebas.

En esta experiencia se han dado avances significativos, mas quedan tareas pendientes como la profundización teórica de los procesos cognitivos subyacentes en la medición de la capacidad comunicativa y del uso adecuado del lenguaje, sustentadas en los supuestos teóricos de los tres estratos de Carroll. No obstante, es importante indagar en otros modelos de la psicología cognitiva que permitan comprender e interpretar estos constructos desde diversas perspectivas. También es necesario ampliar los análisis de la estructura factorial, pensando más en modelos confirmatorios, apoyados tanto en supuestos hipotetizados y en herramientas tecnológicas como el LISREL.

Referencias

Anastasi, A. & Urbina, S. (1998). *Tests psicológicos* (7ª ed.). México: Prentice Hall.

Andriola, W. B. (2002). *Detección del funcionamiento diferencial del ítem (DIF) en test de rendimiento. Aportaciones teóricas y metodológicas*. Tesis de Doctorado para optar al título de Doctor en Investigación, Diagnóstico y Evaluación para la Calidad Educativa, Facultad de Educación, Departamento de Métodos de investigación y Diagnóstico en Educación, Universidad Complutense de Madrid, España. Recuperado el 22 de octubre del 2005 de: http://tede.ibict.br/tde_arquivos/1/TDE-2004-12-13T07:17:46Z-57/Publico/1_WagnerBandeiraAndriola_intro_cap8.pdf

Andriola, W. B. (2003). Descripción de los principales métodos para detectar el funcionamiento diferencial del ítem (DIF) en el área de la evaluación educativa. *Revista de Pedagogía Bordón*, 55(2), 177 – 189.

Chiu, C. W. T & Wolfe, E. W. (2002). A method for analyzing sparse data matrices in the Generalizability Theory framework [Versión electrónica]. *Applied Psychological Measurement*; 26; 321 -338.

Cohen, R. J. & Swerdlik, M. (2001). Pruebas y evaluación psicológicas: Introducción a las pruebas y a la medición (4ª ed.). México: Mc GrawHill.

Cortada de Kohan, N. (2004). Teoría de Respuesta al ítem: Supuestos básicos. [Versión electrónica]. *Evaluar*, *4*, 95-110.

Elosúa, P. (2003). Sobre la validez de los test. [Versión electrónica]. *Psicothema*, 15, 315-321.

Gómez, J. & Hidalgo, M. D. (2003). Desarrollos recientes en Psicometría. [Versión electrónica]. *Avances en Medición, 1*, 17 - 36.

Hernández, O. (1998). *Temas de análisis estadístico multivariado*. San José, Costa Rica: Editorial de la Universidad de Costa Rica.

De Juan-Espinosa, M. (1997). *Geografía de la inteligencia humana*. *Las aptitudes cognitivas*. Madrid, España: Ediciones Pirámide.

Lafourcade, P. (1980). *Evaluación de los aprendizajes.* Buenos Aires, Argentina: Editorial Kapelusz.

Lamas, H. (2006). *Modelos de inteligencia y sus implicancias educativas*. Recuperado el 14 de septiembre del 2007 de htpp://www.monografías.com/trabajos37/modelos-inteligencia/modelos-inteligencia.shtml.

Martínez, R. (2005). Psicometría: Teoría de los tests psicológicos y educativos. Madrid Editorial Síntesis.

Meza, J., D'Agostino, G. & Cruz, A. (2003). Elementos y características del material impreso que favorecen la formación y el aprendizaje a distancia en la UNED. Informe de investigación. San José, Costa Rica: Centro de Mejoramiento de los Procesos Académicos (CEMPA) y Programa de Materiales Impresos de la UNED.

Montero, E. (2001). La teoría de respuesta a los ítemes: una moderna alternativa para el análisis psicométrico de instrumentos de medición. *Revista de Matemática: Teoría y Aplicaciones*. *7*(1-2), 217-228.

Muñiz, J. (1990). *Teoría de Respuesta a los Ítemes*. Madrid: Ediciones Pirámide S.A.

Nunnaly, J. & Bernstein, I. (1995). *Teoría psicométrica* (3ª Ed. México: McGraw-Hill.

Pardo, A. & Ruiz, M. A. (2002). *SPSS 11. Guía para el análisis de datos*. Madrid: McGraw-Hill.

Penfield, R. D. & Camilli, G. (2006). Differential Item Functioning and Item Bias. En C.R. Rao & S. Sinharay (Eds.). *Handbook of Statistics Psychometrics*. Vol. 26 (pp.125-167). Amsterdam: Elsevier.

Piera, C. & Varela, S. (1999). Relaciones entre morfología y sintaxis. En I. Bosque & V. Demonte (Eds.). *Gramática descriptiva de la lengua española* (pp. 4367 – 4422). Madrid: Espasa Calpe S.A.

Ravela, P. (2006). Fichas didácticas para comprender las evaluaciones educativas. Programa de Promoción de la Reforma Educativa en América Latina y el Caribe (OPREAL). Recuperado 24 de setiembre del 2007, de http://www.preal.org/Biblioteca.asp?Id_Carpeta=225&Camino=38%7CGT%20Evaluaci%C3%B3n%20y%20Est%C3%A1ndares/225%7CPublicaciones.

Real Academia Española (1999). *Ortografía de la lengua española.* Madrid: Espasa Calpe S.A.

Zúñiga, M. E. & Montero, E. (2007). Teoría G: un futuro paradigma para el análisis de pruebas. *Revista Actualidades en Psicología,* 21, 117-144.

Manuscrito recibido en Mayo de 2007
Aceptado para publicación en Enero de 2008