



Para citar este artículo, le recomendamos el siguiente formato:

Backhoff, E., Tirado, F. y Larrazolo, N. (2001). Ponderación diferencial de reactivos para mejorar la validez de una prueba de ingreso a la universidad. *Revista Electrónica de Investigación Educativa*, 3 (1). Consultado el día de mes de año en: <http://redie.ens.uabc.mx/vol3no1/contenido-tirado.html>

Revista Electrónica de Investigación Educativa

Vol. 3, No. 1, 2001

Ponderación diferencial de reactivos para mejorar la validez de una prueba de ingreso a la universidad

Differential Weighting of Items to Improve University Admission Test Validity

Eduardo Backhoff Escudero (1)
backhoff@faro.ens.uabc.mx

Felipe Tirado Segura (2)
ftirado@servidor.unam.mx

Norma Larrazolo Reyna (3)
nlar@faro.ens.uabc.mx

(1) y (3) Instituto de Investigación y Desarrollo Educativo
Universidad Autónoma de Baja California

A. P. 453
C. P. 22800
Ensenada, Baja California, México

(2) DEPENDENCIA
Universidad Nacional Autónoma de México, campus Iztacala

Av. de los Barrios s/n
Colonia Los Reyes Iztacala, 54090
Tlalnepantla, Estado de México, México

(Recibido: 29 de enero de 2001; aceptado para su publicación: 21 de febrero de 2001)

Resumen

Este trabajo evalúa diversas formas para incrementar la validez criterial de las pruebas de admisión a la universidad, a través de otorgar distintos pesos a los reactivos que componen las pruebas. Se compararon cuatro métodos para ponderar los reactivos de opción múltiple del Examen de Habilidades y Conocimientos Básicos (EXHCOBA): (1) castigando las respuestas incorrectas con un factor constante; (2) ponderando el castigo de las respuestas incorrectas, considerando los niveles de error; (3) ponderando los aciertos según la dificultad del ítem, basándose en la *Teoría clásica de la medida* y (4) ponderar los aciertos según la dificultad del ítem, basándose en la *Teoría de respuestas al ítem*. Los resultados muestran que ninguno de estos métodos mejoró la validez predictiva del instrumento, aunque sí mejoraron su validez concurrente. Se concluye que es conveniente calificar la prueba sumando simplemente el número de respuestas correctas.

Palabras claves: Ponderación de ítems, calificación ponderada, métodos de evaluación, validez predictiva, exámenes de admisión.

Abstract

This paper deals with different ways to increase university admission test criterion-related-validity, by means of differentially weighting test items. We compared four methods of weighting multiple choice items of the Basic Ability and Knowledge Test (EXHCOBA): (1) punishing incorrect responses by a constant factor, (2) weighting incorrect responses, considering the mistake degree, (3) weighting correct responses, considering the item difficulty, based on *Classic Measurement Theory*, and (4) weighting correct responses, considering the item difficulty, based on *Item Response Theory*. Results showed that none of these methods increased the test predictive validity, but they did improve the test concurrent validity. It was concluded that it's convenient to score the test by simply adding correct responses.

Key words: Test-item, weighted-scores, evaluation-methods, predictive-validity, admission-test.

Introducción

Se dice, por lo general, que un instrumento de medición es válido si mide lo que pretende medir. Sin embargo, estrictamente hablando, uno no valida un instrumento de medición, sino el uso específico que se le da a las puntuaciones o resultados obtenidos (Cronbach, 1971).

Desde hace varias décadas, se han buscado diversos métodos para incrementar la validez de algunas pruebas que se utilizan para la selección y certificación de estudiantes. Uno de ellos consiste en mejorar el procedimiento convencional de calificar la prueba –que se restringe a sumar el número de respuestas correctas- por otro más complejo y más representativo de la ejecución de la persona. Lo anterior se puede lograr otorgando

distintos pesos (ponderar) a los componentes y reactivos que conforman una prueba. Esta ponderación puede realizarse de acuerdo a distintos criterios, tales como: la importancia de los temas evaluados, la dificultad de los reactivos, los tipos de error que cometen los examinados, etcétera.

Tras una revisión bibliográfica de 1982 a la fecha, encontramos una diversidad de estudios sobre “calificación ponderada”, que buscan encontrar la forma de mejorar la validez de los instrumentos utilizados, especialmente la validez predictiva. Podemos clasificar estos trabajos en los que estudian la ponderación de: (1) los resultados de dos o más evaluaciones, (2) las distintas secciones de una prueba y (3) cada uno de los reactivos que conforman el instrumento.

En el primer caso, tenemos estudios sobre la admisión, promoción y certificación de estudiantes. Por ejemplo, Talley y Mohr (1991) estudiaron el efecto de otorgar distintos pesos a las calificaciones de los grados escolares para la admisión a las universidades. Nemeck (1994) y Talley y Mohr (1993) investigaron un sistema ponderado para la promoción escolar o graduación de estudiantes. Siegel (1991) y Bravin (1983) exploraron los pesos diferenciales en las calificaciones de los cursos dependiendo de su nivel de dificultad.

El segundo tipo de estudios investiga las ponderaciones de los componentes de un examen, una prueba, o una evaluación. Por ejemplo, Govindarajulu (1988) analiza cuatro métodos para combinar las calificaciones en un mismo test. Donnelly (1983) investiga las ventajas de ponderar las calificaciones del Medical College Admission Test, y Willis (1993) investiga la ponderación de variables para identificar estudiantes con necesidades especiales.

El tercer tipo de estudios analiza la ponderación de los reactivos de una prueba. Así, Sympson y Haladyna (1988) desarrollan un método de ponderación múltiple para las pruebas referidas a un dominio. Este método pondera cada ítem de acuerdo al promedio percentilar de los examinados que escogieron esa opción. Los resultados muestran que este método de ponderación múltiple arroja la confiabilidad más alta de la prueba, así como la mejor validez relacionada con un dominio. Por su parte, Razel y Eylon (1987) validan distintos modos de calificar la Prueba Coloreada de Matrices Progresivas de Raven. Para ello, comparan la forma convencional de calificar la prueba con tres métodos de calificación de ponderación múltiple: (1) de acuerdo a la teoría de procesamiento cognitivo, (2) según la opinión de expertos (3) en función de las respuestas que seleccionan los estudiantes. Se compara la validez concurrente de estas cuatro formas de calificación, utilizando otras pruebas de inteligencia general. Los resultados muestran que la calificación ponderada es preferible a la convencional, ya que mejora la validez y confiabilidad de la prueba: siendo la ponderación empírica el mejor método. Finalmente, Budescu (1979), utilizando el análisis de regresión múltiple, pondera diferencialmente ítems de opción múltiple para maximizar la validez predictiva de una prueba. Los resultados indican que las reglas de calificación más complejas producen mayor varianza en los ítems y en las calificaciones. Sin embargo, no hay un incremento en la validez predictiva de la prueba.

Aunque no todos los estudios apoyan la idea de que la ponderación mejora la validez criterial de los instrumentos de evaluación, y hay investigadores que, por principio, niegan esta posibilidad (Tristán y Vidal, 2000), hay una tendencia a suponer que mejorando la forma de calificar aumenta la validez predictiva de las pruebas.

Esta hipótesis se apoya en la idea de que nuestros conocimientos y habilidades no se expresan de manera alterna y categórica (aciertos y errores) sino que hay matices, o diferencias de grado, que pueden llegar a ser muy importantes. Nuestro proceder se basa más bien en un cálculo de oportunidades, riesgos, beneficios y costos; de manera tal que no es tan simple, ni pertinente, calificar nuestras respuestas como simplemente “buenas” o “malas”. También se puede argüir que las preguntas en una prueba de conocimientos no son iguales en su dificultad, unas son más fáciles que otras; asimismo, la dificultad de una pregunta no es fija, sino que es relativa a las habilidades y conocimientos de quien está contestando la prueba. Estas consideraciones también abren la discusión sobre la pertinencia de otorgar un peso igual a todas las preguntas.

Sin embargo, no es claro si la ponderación de las puntuaciones mejora la validez de una prueba y, tampoco, cuál es la mejor forma de ponderación. Por lo anterior, el propósito del presente estudio fue comparar cuatro métodos para ponderar reactivos de opción múltiple de un examen de ingreso a la educación superior. Dichos métodos de ponderación son: (1) castigar las respuestas incorrectas con un factor constante, (2) ponderar el castigo de las respuestas incorrectas considerando la “gravedad” del error, (3) ponderar los aciertos de acuerdo al índice de dificultad p del ítem (Teoría clásica de la medida) y (4) ponderar los aciertos de acuerdo al índice de dificultad b del ítem (Teoría de respuestas al ítem).

Metodología

Instrumentos. El Examen de Habilidades y Conocimientos Básicos (EXHCOBA) se utiliza como examen de admisión en varias universidades públicas mexicanas. Su característica principal es que evalúa competencias académicas esenciales de todas las áreas del conocimiento que se aprenden desde la primaria hasta el bachillerato y que se suponen necesarias para tener éxito en los estudios universitarios. Como el EXHCOBA se ha descrito con detalle en otras publicaciones (Backhoff y Tirado, 1992), en este apartado sólo mencionaremos las características más importantes para el diseño de este estudio:

- El examen está conformado por tres secciones: habilidades básicas (60 preguntas), conocimientos básicos (70 preguntas) y conocimientos básicos de especialidad (180 preguntas).
- Todos los estudiantes contestan las primeras dos secciones del examen, que suman un total de 130 preguntas, y sólo contestan 60 preguntas de la tercera sección, de acuerdo al área de conocimientos de la carrera que desean estudiar. En este estudio, por razones prácticas, sólo se analizaron los resultados de las primeras 130 preguntas.
- El examen cuenta con cinco versiones de preguntas equivalentes para cada uno de los 310 ítems del examen. En este estudio se utilizó solamente una de ellas (la versión 2).
- Existen dos formatos del examen: Para lápiz y papel, y para computadora. En este estudio se utilizó la versión computarizada. Una descripción detallada de la interfaz, y

de su equivalencia con el formato de lápiz y papel, se encuentra publicada por Backhoff, Ibarra y Rosas (1995).

VARIABLES CRITERIO. Una decisión metodológica que debimos resolver en este estudio fue la relacionada con el criterio para valorar el grado en que los resultados del EXHCOBA mejoran cuando se utilizan distintos métodos de calificación, ponderada o no ponderada. Siguiendo lo que marca la literatura en los exámenes de selección, establecimos como criterios relacionados el promedio de calificaciones del primer semestre universitario, para determinar la validez predictiva, y el promedio de calificaciones del bachillerato, para la validez concurrente. Así, tanto la validez predictiva como concurrente se obtuvieron correlacionando estas puntuaciones con las calificaciones en el EXHCOBA.

POBLACIÓN. En verano de 1998, se aplicó el EXHCOBA a cerca de 9 000 estudiantes que deseaban ingresar a la Universidad Autónoma de Baja California (UABC). De éstos, se seleccionaron 882 aspirantes que respondieron a la versión 2 del examen. De estos estudiantes, ingresaron a la universidad 653, de los cuales 301 jóvenes ingresaron en el semestre de verano de ese año y el resto ingresó en el semestre de invierno de 1999. Un año después de que los estudiantes respondieron el examen, se solicitó a la Dirección General de Servicios Escolares de la UABC el archivo del cardex donde se almacena el registro de los cursos de todos los estudiantes que ingresan a la universidad. Al igual que el archivo de resultados del examen, este archivo se preparó para realizar los análisis estadísticos correspondientes de aquellos estudiantes que ingresaron a la universidad en el verano de 1998. La preparación consistió en obtener los promedios de calificaciones del bachillerato de todos los estudiantes, así como los promedios de las calificaciones que obtuvieron durante el primer semestre de sus estudios universitarios. En este caso, para calcular los promedios, se asignó una calificación de cero a las materias no acreditadas, cuya calificación era alfanumérica (p.ej. NA, no acreditado; SD, sin derecho) y una calificación de diez a las acreditadas.

PROCEDIMIENTO. Los pasos que se siguieron para administrar y calificar el EXHCOBA fueron los mismos que define la universidad para su proceso de admisión: (1) el estudiante pagó y obtuvo una ficha, en la cual se le indicó la fecha, lugar y hora de su cita; (2) según la carrera a la que aspiró ingresar, el sistema automatizado le presentó uno de los siete tipos de exámenes, correspondientes a las áreas de conocimientos en que se agrupan las carreras de la UABC; (3) el estudiante respondió las preguntas sin ningún tipo de ayuda (calculadoras, diccionarios, etc.), salvo lápiz y papel, y (4) el sistema calificó las respuestas en forma automática y entregó al estudiante un informe por escrito de su ejecución. Durante este proceso, siempre estuvo presente una persona capacitada que resolvió cualquier problema o duda acerca de la versión computarizada del examen.

Conforme los estudiantes terminaban de contestar el examen, los resultados se capturaron en una base de datos. Ésta se preparó posteriormente para realizar los análisis estadísticos correspondientes. La preparación consistió en transformar los resultados crudos de los estudiantes en cuatro distintos formatos, una vez que se depuraron los casos con patrones de respuesta anormales (p.ej. registros vacíos):

1. **Formato categórico.** Donde las respuestas correctas tienen un valor de 1, las incorrectas de -0.25 y las omisiones de 0.
2. **Formato del gradiente de error.** Para calcularlo, el grupo de estudiantes se dividió en tres subgrupos, de acuerdo con el número de aciertos que obtuvieron en el examen, quedando así un tercio superior, un tercio intermedio y un tercio inferior. Para ponderar la calificación, se agregó un puntaje a los aciertos de acuerdo al porcentaje de errores cometido por el tercio superior. Por ejemplo, si 86% de los sujetos del grupo superior optaron por la opción “b” que era incorrecta, a todos los que cometieron este error (del grupo superior, intermedio e inferior) se les sumó 0.86 puntos compensatorios, con la premisa de que esa opción no era tan errónea. Si del grupo superior sólo 3% optó por la opción incorrecta “c”, en ese caso se les agregó 0.03, considerando que esa respuesta resultaba realmente mala o ingenua para el grupo superior. En las opciones erróneas que no seleccionaron los miembros del subgrupo superior, no se otorgó ningún punto compensatorio.
3. **Formato del índice de dificultad (valores p).** En este análisis, partimos de que no todos los aciertos deben tener el mismo peso en la calificación, sino que éste se ponderará por el grado de dificultad, resultando que las preguntas más difíciles agregarán un puntaje mayor a la calificación, que aquellas que resultan ser más fáciles. La graduación del peso se hizo equivalente al valor inverso de p (dificultad de la pregunta, en la Teoría Clásica de la Medida). Así, por ejemplo, si una pregunta fue contestada correctamente por 96% de los estudiantes, entonces se consideró una pregunta muy fácil y el valor del acierto contó 0.04. Pero si la pregunta fue, por el contrario, muy difícil y la contestaron correctamente sólo 2%, entonces se le asignó un valor de 0.98 puntos a cada uno de los participantes que la contestaron acertadamente. En este caso, las respuestas incorrectas tuvieron siempre un valor de 0.
4. **Formato del índice de dificultad (valores b).** Con el mismo razonamiento señalado en el punto anterior, pero ahora con base en la Teoría de Respuestas al Ítem (Muñiz, 1997), y utilizando el programa *Bilog* (Mislevy y Bock, 1982), para el modelo de tres parámetros, se calcularon los índices de dificultad (valores b) para cada uno de los 130 ítems del examen; las respuestas incorrectas tuvieron un valor de 0.

Resultados

Con las cuatro bases de datos elaboradas, se realizaron 11 correlaciones, utilizando el paquete estadístico SPSS (v. 9). La tabla I muestra los resultados obtenidos. Sólo como referencia mostramos las correlaciones entre las calificaciones del bachillerato y las universitarias (ver última columna). Aquí podemos ver que la validez predictiva del promedio de bachillerato es de 0.306; valor relativamente modesto comparado con los informes internacionales.

Tabla I. Correlaciones de las calificaciones del EXHCOBA, utilizando distintos métodos de ponderación con el promedio de calificaciones del bachillerato y de la universidad

Indicadores	Número de aciertos	Ponderación de errores		Ponderación de aciertos		Promedio universitario
		Factor constante	Opciones de respuesta	Valor p	Valor b	
Promedio Bachillerato	0.238	0.244	0.238	0.251	0.246	0.306
Promedio Universidad	0.251	0.243	0.249	0.238	0.239	

Nota: Todas las correlaciones fueron significativas al 0.01 (dos colas).

Ahora bien, nuestro punto de partida para este análisis fue la correlación del número de aciertos en el EXHCOBA con los promedios del bachillerato y del primer semestre universitario (segunda columna). Se podrá observar que la validez predictiva (0.251) es ligeramente mejor que la concurrente (0.238), siendo ambos índices bastante moderados.

Ponderación de errores. En este análisis partimos de dos premisas. Primero, que los errores se dan en un continuo de “gravedad”; es decir, que hay errores más graves que otros. Segundo, que los errores presentan consistencias, y ocurre que quienes más saben cometen menos errores y los errores que cometen son los menos graves.

Se obtuvieron dos tipos de calificación basada en errores: (1) castigando por igual todas las respuestas incorrectas con un cuarto de punto (0.25), que corresponde a la probabilidad de contestar al azar en una prueba con cuatro opciones de respuesta, y (2) ponderando los errores cometidos de acuerdo con la frecuencia obtenida por los mejores estudiantes en cada una de las cuatro opciones de respuesta de las 130 preguntas del EXHCOBA.

En la tercera y cuarta columnas de la tabla I, se presentan los resultados de las correlaciones de estos dos métodos de ponderación de errores con los promedios escolares. Se puede observar que la validez predictiva del examen en ambos casos decrece ligeramente de 0.251 a 0.243 y 0.249. Por el contrario, la validez concurrente del examen mejora ligeramente, de 0.238 a 0.244, cuando se castigan las respuestas incorrectas. Sin embargo, la validez concurrente no cambia cuando se utiliza el método de ponderar las opciones de respuestas incorrectas.

Ponderación de aciertos. En este análisis partimos de que no todos los aciertos deben tener el mismo peso en la calificación, ya que los reactivos difieren según su grado de dificultad. Así, las preguntas más difíciles deben agregar un puntaje mayor a la calificación que aquellas que resultan ser más fáciles.

Se obtuvieron dos tipos de calificación basada en la dificultad de la pregunta. En el primer caso, se utilizó el índice de dificultad calculado con base en la Teoría Clásica de la Medida, valor p . En el segundo, se utilizó la dificultad de la pregunta, calculada con base en la Teoría de Respuestas al Ítem, valor b .

En la quinta y sexta columnas de la tabla I, se presentan los resultados de las correlaciones de estos dos métodos de ponderación de aciertos, con ambos promedios escolares. Se

puede observar que la validez predictiva del examen, en ambos casos, decrece ligeramente de 0.251 a 0.238 y 0.239. Por el contrario, la validez concurrente del examen mejora ligeramente cuando se ponderan las respuestas correctas, con ambos métodos, de 0.238 a 0.251 y 0.246.

En síntesis, podemos decir que los métodos de ponderación de la calificación del EXHCOBA no mejoran la validez predictiva del instrumento, aunque sí mejoran ligeramente su validez concurrente. Asimismo, podemos decir que, prácticamente los cuatro métodos de ponderación producen, en promedio, resultados equivalentes; aunque ligeramente el “mejor” de ellos fue la ponderación de aciertos de acuerdo con la dificultad de los reactivos del examen (valor p).

Discusión

El procedimiento tradicional para calificar una prueba ha suscitado muchas reflexiones en torno a la pertinencia de calificar bajo un proceso binario, en el que todos los aciertos y errores tienen los mismos valores (aciertos = 1, errores y omisiones = 0). En este modelo, no importa la dificultad de la pregunta ni el grado de error de la respuesta equivocada. El modelo asume que dos personas tienen el mismo nivel de conocimientos o habilidades si ambos respondieron correctamente el mismo número de reactivos en una prueba, sin importar en cuáles de ellos acertaron.

Sin embargo, sabemos que nuestros conocimientos y habilidades no se expresan en este binomio de aciertos y errores, sino que nuestras respuestas reflejan diferencias de grado que pueden llegar a ser muy importantes. Por ejemplo, no es lo mismo que alguien piense que Costa Rica se encuentra en Latinoamérica, a que piense que se encuentra en Europa.

Esta consideración parece razonable, y ha motivado a un gran número de investigadores a estudiar mejores formas de calificar las pruebas. Como ya se vio, los resultados son contradictorios y faltan evidencias empíricas que sostengan la postura de ponderar los reactivos de una prueba. Con esta idea, realizamos la presente investigación, y encontramos que no hay indicios para mostrar que se mejora la validez predictiva de un examen de ingreso a la universidad ponderando los errores según su grado de plausibilidad, o los aciertos, según su grado de dificultad; reforzando así la postura de Tristán y Vidal (2000), quienes ofrecen una serie de argumentos para no ponderar los reactivos.

Consideramos que estos resultados se pueden deber a que la prueba utilizada en este estudio no fue configurada con estos propósitos. Es decir, no contempla un gradiente de error en el diseño de las preguntas; por el contrario, la estructura de los reactivos está concebida en aciertos y desaciertos, dado un conjunto de “distractores”. Tampoco es una prueba planificada con pesos relativos a los aciertos en los diferentes reactivos.

Otro argumento que se antoja plausible es que los promedios de calificaciones obtenidos en la universidad y en el bachillerato, no son buenos criterios de validación para los propósitos de este estudio, ya que los criterios de calificación varían de un profesor a otro y de una institución a otra. Lo anterior se evidencia en que los promedios de las calificaciones

escolares no guardan mucha relación entre sí, pues se obtienen correlaciones sólo de 0.306 entre el bachillerato y el primer semestre de licenciatura. Comparando estos resultados con los valores reportados por Trent, Leland y Medsker (1968) de 0.55 entre las calificaciones del bachillerato y la universidad, y de 0.50 para las pruebas de aptitudes académicas y los promedios universitarios, apreciamos que el nuestro es muy modesto.

Si bien lo anterior es plausible, los resultados de la validez concurrente contradicen estos argumentos, ya que, paradójicamente con los mismos procedimientos de ponderación y los mismos criterios de validación, las correlaciones del EXHCOBA con las calificaciones del bachillerato mejoraron ligeramente en la mayoría de los casos.

Poniendo los resultados en una balanza, parece ser que lo más conveniente es seguir aplicando los procedimientos tradicionales, hasta no encontrar mejores evidencias que nos indiquen lo contrario. Sin embargo, dado que los argumentos lógicos apoyan la idea de ponderar la calificación de una prueba, es importante seguir explorando nuevas posibilidades con base en otros procedimientos de ponderación y otros criterios de validación, tales como: el análisis diferencial de los ítems, el cálculo de probabilidades para acertar el ítem, y algunos elementos y criterios que nos proporciona la psicometría moderna.

Referencias

- Backhoff, E., Ibarra, M. A. y Rosas, M. (1995). Sistema Computarizado de Exámenes (SICODEX). *Revista Mexicana de Psicología*, 12 (1), 55-62.
- Backhoff, E. y Tirado, F. (1992). Desarrollo del Examen de Habilidades y Conocimientos Básicos. *Revista de la Educación Superior*, 83, 95-118.
- Bravin, J. (1983). Bright idea: Hard courses should carry more weight than easy courses. *Executive Educator*, 5 (1), 40-30.
- Budescu, D. V. (1979). *Differential weighting of multiple choice items*. Princeton: Educational Testing Service.
- Cronbach, L. J. (1971). Test validation. En R. L. Thorndike (Ed.), *Educational measurement* (2a. ed.). Washington: Consejo Americano en Educación.
- Donnelly, M. B. y otros (1983). Simple Adding versus Differential Weighting of MCAT Subtest Scores. *Journal of Medical Education*, 58 (7), 581-83.
- Govindarajulu, Z. (1988). Alternative methods for combining several test scores. *Educational and Psychological Measurement*, 48 (1), 53-60.
- Mislevy, R. y Bock, R. D. (1982). *Bilog: Maximum likelihood item analysis and test scoring with logistic models*. Mooresville: Scientific Software.

- Muñiz, J. (1997). *Introducción a la Teoría de Respuestas a los Ítems*. Madrid: Pirámide.
- Nemecek, P. M. (1994). Constructing weighted grading systems. *Clearing House*, 67 (6), 325-326.
- Nunnally, J. C. y Bernstein, I. H. (1994). *Psychometric Theory*. New York: Mc Graw-Hill.
- Razel, M., y Eylon, B. S. (1987, abril). Validating alternative modes of scoring for coloured progressive matrices. Trabajo presentado en la *Reunión Anual de la American Educational Research Association*, Washington.
- Siegel, J. y Anderson, C. S. (1991). Considerations in calculating high school GPA and rank in class. *NASSP Bulletin*, 75 (537), 96-109.
- Sympson, J. B. y Haladyna, T. M. (1988, abril). An evaluation of "polyweighting" in domain referenced testing. Trabajo presentado en la *Reunión Anual de la American Educational Research Association*, New Orleans.
- Talley, N. R. y Mohr, J. I. (1993). The case for a national standard of grade weighting. *Journal of College Admission*, 139, 9-13.
- Talley, N. R. y Mohr, J. I. (1991). Weighted averages, computer screening, and college admission in public colleges and universities. *Journal of College Admission*, 132, 9-11.
- Trent, J. W. y Leland, L. M. (1968). *Beyond high school: A psychological study of 10,000 high school graduates*. San Francisco: Jossey-Bass.
- Tristán, A. y Vidal, R. (2000). Análisis de la práctica de asignar pesos a los reactivos y su efecto en el diseño y calificación de pruebas. *Memorias del IV Foro de Evaluación Educativa*. México: Centro de Evaluación de la Educación Superior.
- Willis, J. A. (1993, abril). Chapter 1 Eligibility Factors and Weights: Using probit analysis to determine eligibility criteria. Trabajo presentado en la *Reunión Anual de la American Educational Research Association*, Atlanta.