Estudio de parámetros psicométricos en los sistemas de evaluación de preguntas de respuesta múltiple (*MCQ*): Mejora Genética Animal, un caso de estudio de la ULPGC

J.M. Afonso[*], M.J. Zamorano[*]

[*]Dept. Patología Animal, Producción Animal y Ciencia y Tecnología de los Alimentos, Universidad de Las Palmas de Gran Canaria (ULPGC), Campus Universitario Cardones de Arucas, 35413, Arucas, España

RESUMEN

Un aspecto esencial de los sistemas educativos es cómo son evaluados los conocimientos de los estudiantes. Las preguntas de respuestas multiples, *Multiple-choice questions (MCQ)*, son de las más extendidas en las materias que se imparten en los grados de medicina humana y veterinaria. La validez y calidad de los métodos de evaluación son medidas a través de parámetros psicométricos. Los más extendidos son los índices de dificultad (*IDif*) y discriminación (*IDis*). En la materia de Mejora Genética Animal (*MGA*) de la facultad de Veterinaria de la ULPGC se estimaron dos tipos de puntuación, dentro del marco de un mismo examen; *MCQac* y Verdadero-Falso con puntos negativos (*MCQvf*). Se utilizó una muestra de 442 estudiantes y examenes, donde se analizaron 14.144 preguntas *MCQac* y 21.216 preguntas *MCQvf*, sobre los mismos estudiantes y exámenes, representando un total de 42.432 afirmaciones. Cada pregunta *MCQac* y *MCQvf* estuvo compuesta por 3 y 2 afirmaciones, respectivamente. Cada examen *MCQ* tenía preguntas tipo; problema (*Tp*), texto (*Tt*) e interpretación (*Ti*). Las estimas *IDif* e *IDis* fueron 48,5% y 0,36 en *Tp*, 43% y 0,4 en *Tt* y 42,2% y 0,56 en *Ti*, respectivamente. *MCQac* sobreestimó la nota de los estudiantes suspendidos en un 16% frente a *MCQvf*. Las estimas medias de *IDif* e *IDis* para el sistema de evaluación *MCQac* fueron de 45,2% y 0,4, respectivamente. El sistema *MCQvf* discriminó un 30,15% mejor que el *MCQac* (0,52). Así, la validez y la calidad de *MCQac* y *MCQvf* en *MGA* fueron *Excelentes* en términos de *IDis*, y entre *Acceptable* e *Ideal* en términos de *IDif*.

Keywords: mejora genética, índice de discriminación, índice de dificultad, *MCQ*,

1. INTRODUCCIÓN

La asignatura de Mejora Genética Animal (*MGA*) debe formar a los alumnos del grado de Veterinaria de la Universidad de Las Palmas de Gran Canaria (ULPGC), en el conocimiento y la aplicación de las técnicas genéticas para mejorar la rentabilidad de las explotaciones ganaderas, tanto explorando la variación intrapoblacional como la interpoblacional, incluyendo la biotecnología como herramienta. Para ello, es fundamental el conocimiento causal de las fuerzas que determinan la diversidad [1].

Es extendido entre los estudiantes que cursan la *MGA* en la ULPGC, que se trata de una material árida, debido a varios motivos. Por un lado, porque han de adoptar una actitud mental más abstracta que la utilizada habitualmente en el resto de las asignaturas, ya que muchos de los conceptos que se imparten no son tan tangibles como en otras materias. Por otro lado, porque requiere del conocimiento previo de materias como Bioestadística, Bioquímica y Genética, de cursos previos, cuyos conocimientos son claves para un buen entendimiento y adquisición de las competencias propias de la *MGA*. Finalmente, se trata de una materia que requiere un nivel importante de continuidad, ya que el entendimiento de los nuevos conceptos depende del aprendizaje correcto de los previamente dados.

Sin duda, el profesor juega un papel importante en todo ello. Según la UNESCO, el professor es una herramienta que además de enseñar una determinada disciplina, debe transmitir la forma de autoabastecerse de conocimientos biológicos en el futuro [2]. Esta tendencia de la docencia de las Ciencias Experimentales y Aplicadas de transmitir unos conceptos, herramientas e inculcar en el alumno una vocación autodidacta, como es el caso de la *MGA*, pudiera ser la forma más

ISBN 978-84-09-02374-5

práctica y eficaz de enseñar, es decir, que los alumnos tengan, no sólo una visión real de la disciplina, sino que aprendan a ser usuarios de la misma.

Dentro del marco de cualquier materia, son importantes los conocimientos que se imparten, cómo se imparten y cómo se evalúan, con el fin de que se alcancen las competencias. Para ello, lo habitual es combinar la utilización de diferentes estrategias como son la familiarización con casuística reglada, la resolución de casos prácticos, la estancia en empresas del sector, la asistencia y participación en clase o la realización de exámenes reglados.

La evaluación de los conocimientos mediantes exámenes reglados es ampliamente utilizado en todos los planes de estudio de las diferentes materias, para validar el que los estudiantes hayan alcanzado los hitos académicos establecidos. El sistema de evaluación de preguntas de respuestas multiples, *Multiple-choise questions (MCQs)*, es uno de los más extendidos en el ámbito académico médico [3]. Si bien, los sistemas de corrección que se aplican sobre los datos en bruto (*Raw Data*) del *MCQ* pueden tener un efecto matemático sobre la puntuación final del estudiante [4].

Uno de los objetivos del sistema educativo, es que la puntuación final que obtiene el estudiante represente o covaríe con el esfuerzo realizado por éste. En otros términos, sería conocer la validez y calidad del tipo de examen y sistema de corrección que se aplica, mediante la estimación de parámetros psicométricos [5]. Los parámetros psicométricos más utilizados son los índice de dificultad e índice de discriminicación [6].

En el presente trabajo se estudia el grado de validez y calidad del tipo de examen que afrontan los estudiantes de la facultad de veterinaria de la ULPGC, para supercar la materia de Mejora Genética Animal (*MGA*), a través del estudio comparado de la estimación del efecto matemático que tienen dos sistemas de corrección, dentro del sistema de evaluación de preguntas de respuestas multiples, *Multiple-choise questions (MCQ)*: a) donde solo se consideran como preguntas puntuables aquellas donde se responden solo y a todas las afirmaciones correctas (*MCQac*), y b) donde se considera una corrección de verdadero-falso con puntos negativos (*MCQvf*).

2. METODOLOGÍA

2.1 Materiales

Se utilizaron 442 exámenes de los estudiantes de seis promociones consecutivas de la facultad de veterinaria de la ULPGC (92/93; 93/94; 94/95; 95/96; 96/97; 97/98). Cada examen consistió en 32 preguntas de respuestas multiples (*MCQ*), donde cada una tenía tres afirmaciones a razonar. Por examen, los estudiantes debían razonar 96 afirmaciones, que teniendo en cuenta las seis promociones muestreadas, en este estudio se analizó un total de 14.144 preguntas de respuesta multiple (*MCQ*) y 42.432 afirmaciones. A su vez, el *MCQ* estuvo estructurado en preguntas con tres tipos de formato, para no abordar la valoración del alumno desde una única perspectiva y hacer una valoración más integral del mismo. *Formato tipo problema* (*Tp*), son aquellas en las que se plantea algún problema y los alumnos han de resolverlo y buscar la solución correcta. *Formato tipo texto* (*Tt*), son aquellas preguntas convencionales, en las que los alumnos han de decidir cuáles de las afirmaciones escritas son ciertas. *Formato tipo interpretación* (*Ti*), son aquellas preguntas en las que los alumnos han de interpreter algún escenario y/o completar la afirmación de manera precisa. Además, los alumnos disponían en los exámenes de unas hojas adicionales en las que desarrollar todos los razonamientos que consideraban oportunos, a fin de que la nota del alumno covariase más con sus conocimientos.

2.2 Métodos

Los métodos de corrección validados en el presente estudio fueron dos. *MCQac*, donde de las tres afirmaciones de cada pregunta, podían ser ciertas desde ninguna hasta todas. Una pregunta solo puntuaba en la nota cuando el estudiante contestaba solo a todas las afirmaciones correctas. Así, por ejemplo, si un estudiante contestaba solo una afirmación como correcta, de las dos que tenía una pregunta, además de no puntuar, tampoco se le consideraba como razonamiento correcto la afirmación correctamente respondida. *MCQvf*, donde se hacía una corrección de afirmaciones verdaderas y falsas con puntos negativos. Esto era posible, porque de las 96 afirmaciones totales a razonar por cada estudiante, y distribuidas al azar a lo largo del examen, 48 eran correctas y 48 eran incorrectas. Lo que permitía hacer una corrección simulada de “un examen compuesto por 48 preguntas, cada una con una afirmación verdadera y otra afirmación falsa”, con puntos negativos a lo largo del examen, cuando el estudiante respondía como correcta un afirmación incorrecta.

ISBN 978-84-09-02374-5

2.3 Análisis

La normalidad y homocedasticidad de todos los datos resultados de ambos métodos de corrección fueron testadas, presentando normalidad y homogeneidad de varianzas. Para la comparación de las puntuaciones de los métodos de corrección, se utilizó el siguiente modelo lineal general, mediante el SPSS v15;

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

Donde Y_{ij} es el valor individual del animal, μ es la media de la población, α_i es el efecto del sistema de corrección *i-ésimo*, y ε_{ij} es el error asociado al registro *ij-ésimo*.

El índice de discriminación (*IDis*) es la capacidad de distinguir entre estudiantes con buenos y malos resultados en función de una variable determinada; en este caso, la probabilidad de responder correctamente a las preguntas del test. Además, no está definida de antemano, sino que depende de la curva de resultados de todos los examinados que respondieron al test. La capacidad de discriminación del test viene medida por la media aritmética de los índices de discriminación de cada pregunta. Estos índices se calcularon comparando en cada pregunta las respuestas correctas del grupo de estudiantes que ha obtenido los mejores resultados (formado por el 27% de los mejores exámenes) menos las respuestas correctas del grupo de estudiantes que ha obtenido los peores resultados (formado por el 27% de los peores exámenes), dividido por el número total de estudiantes con buenos y malos resultados que respondieron a cada pregunta y todo multiplicado por dos. Toma valores comprendidos entre +1, cuando las preguntas son contestadas correctamente sólo por los alumnos con buenos resultados, y –1, cuando las preguntas son contestadas correctamente sólo por los alumnos con malos resultados. Como clasificación de la capacidad de discriminación de un tipo de examen se consideró la indicada en la tabla 1 [7].

Tabla 1. Clasificación de calidad de los índices de discriminación (*IDis*).

IDis	**Calidad**	**Recomendaciones**
>0,39	Excelente	Conservar
0,30 – 0,39	Buena	Posibiliad de mejorar
0,20 – 0,29	Regular	Necesidad de revisar
0,00 – 0,19	Pobre	Descartar o revisar en profundidad
< - 0,01	Pésima	Descartar definitivamente

El índice de dificultad (*IDif*), es una dimensión secundaria con una repercusión directa en la discriminación y determina en qué medida, para una población de estudiantes dada, una pregunta de examen es fácil o difícil de contestar correctamente. Se calcula comparando en cada pregunta las respuestas correctas del grupo de estudiantes que ha obtenido los mejores resultados (formado por el 27% de los mejores exámenes) más las respuestas correctas del grupo de estudiantes que ha obtenido los peores resultados (formado por el 27% de los peores exámenes), dividido por el número total de estudiantes con buenos y malos resultados que respondieron a cada pregunta y todo multiplicado por cien. Toma valores comprendidos entre 0 y 100, y es óptimo cuando está entre 50 y 60, aunque se considera como bueno cuando oscila entre 30 y 70 (Tabla 2) [3].

ISBN 978-84-09-02374-5

Tabla 2. Clasificación de calidad de los índices de dificultad (*IDif*).

***IDif* (%)**	**Interpretación**	**Acción**
< 30	Demasiado dificil	Revisar / Descartar
30 – 70	Bueno / Aceptable	Conservar
50 – 60	Excelente / Ideal	Conservar
> 70	Demasiado fácil	Revisar / Descartar

3. RESULTADOS

Los exámenes que se vienen realizando en la actualidad en la asignatura de Mejora Genética Animal de la Facultad de Veterinaria de la Universidad de Las Palmas de Gran Canaria, muestran unos índices medios de discriminación y dificultad de 0,4 y 45,2%, respectivamente (Tabla 3), por lo que discriminan bastante bien entre los grupos de alumnos con buenos y malos resultados, aunque sus niveles de dificultad están a algo menos de 5 puntos de llegar al rango óptimo, siendo en cualquier caso estos últimos entre buenos-aceptables. Los valores están estimados sobre las correcciones múltiples, ya que los exámenes están organizados de ese modo. Sin embargo, de poder hacerse, los índices medios serían más altos si se tuviesen en cuenta los cálculos sobre la corrección verdadero-falso, *MCQvf*. De hecho, una inferencia a través de la comparación de la distancia en nota media entre el 27% de los estudiantes con mejores y peores puntuaciones por sistema de corrección, puso de relive que el sistema *MCQvf* discriminó un 30,15% mejor que el *MCQac* (0,52). Estos resultados están en consonancia con que los coeficientes de variación de los métodos de corrección multiple, *MCQac*, y verdadero-falso, *MCQvf*, fueron del 37% y el 50%, respectivamente.

Tabla 3. Indices de discriminación (*IDis*) y dificultad (*IDif*) por tipo de pregunta en *MCQ*.

Formato de Tipo de pregunta	***IDis***	***IDif* (%)**
Problema	0,36	48,5
Texto	0,40	43
Interpretación	0,56	42,2
Medias ponderada	**0,40**	**45,2**

Como se observa en la tabla 3, aunque todos los formatos de tipos de preguntas del *MCQ* tienen unos índices de discriminación que son entre buenos y excelentes, estos valores mejoraron gradualmente desde las de formato *Tp* a *Tt* y *Ti*. Así, las de formato *Tt* mejoraron a las *Tp* en un 11%, mientras que las *Ti* lo hicieron en un 55%. En las preguntas con formato *Ti*, los estudiantes completan e interpretan las preguntas con sus propias ideas y modo de redactar, lo que aparentemente hace que se manifieste mejor los diferentes esfuerzos y capacidades entre los estudiantes con buenos y malos resultados.

En relación a los *IDif*, el sistema *MCQ* tuvo siempre un grado de dificultad comprendido entre Bueno y Aceptable, tanto en el promedio como por formato de tipo de pregunta. Siendo las preguntas *Tp* las más fáciles, comparativamente (48,5%), las más difíciles las preguntas con formato *Ti* (42,2%) y las con formato *Tt* intermedias (43%).

La nota media de ambos sistemas de corrección fueron de 4,08 y 3,86, para los sistemas *MCQac* y *MCQvf*, respectivamente (un 5,7% superior *MCQac* respecto a *MCQvf*). A pesar de ello, las notas de ambos tipos de sistema de corrección presentaron una correlación alta y positiva, 0,874 ($P<0,01$), indicando que en general los estudiantes que

ISBN 978-84-09-02374-5

obtienen una puntuación alta por el sistema *MCQac* también lo hacen por el sistema *MCQvf*. Y viceversa. Sin embargo, comparando el paralelismo de notas según el sistema de corrección, por intervalos de notas (figura 1), se puede ver que cuando el estudiante estrictamente está suspenso, saca peor puntuación por el sistema *MCQvf* que por el *MCQac*. Mientras que cuando el estudiante está estrictamente aprobado, sucede lo contrario. Es decir, que en la franja inferior al 5, aparentemente el sistema *MCQac* se comporta como un sobreestimador, mientras que en la franja de notas superiores al 5 se comporta como un subestimador.

El beneficio individual del efecto corrector del sistema *MCQvf* frente al *MCQac* es de un 25%, para los estudiantes que obtuvieron puntuaciones en la franja comprendida entre ≥ 4 y 5<, por el sistema *MCQac*.

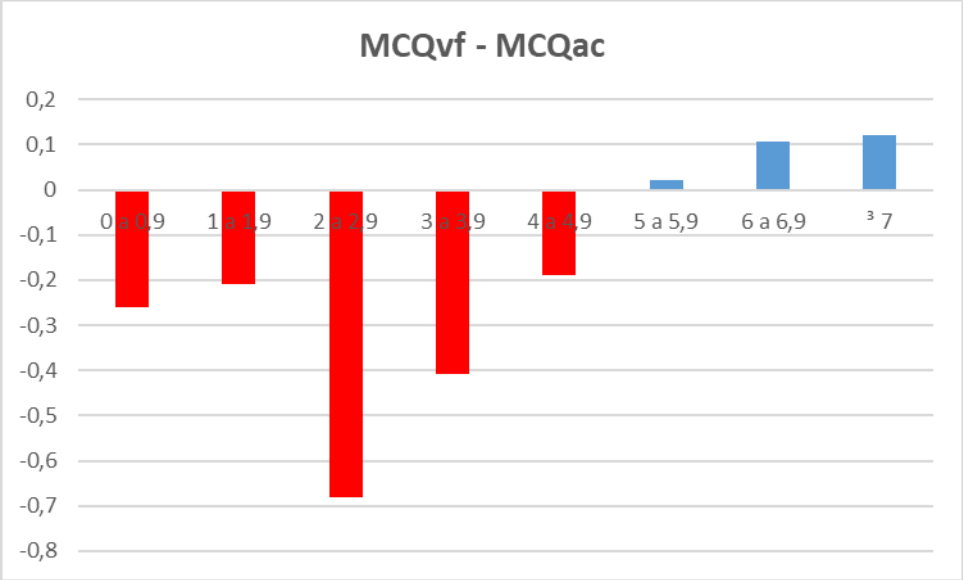

Figura 1. Tendencia de notas de corrección verdadero-falso (*MCQvf*) respecto a corrección de respuestas multiples (*MCQac*), agrupadas por intervalos de notas *MCQac*.

4. DISCUSIÓN

Los sistemas de evaluación basados en preguntas de respuestas multiples, *Multiple-choise questions (MCQs)*, están altamente extendidos en la educación superior [8]. Sin embargo, el sistema de corrección que se implemente sobre el *MCQ* genera diferentes tipos de datos en bruto (*Raw Data*), que se prestan a indicar el grado de validez y calidad del sistema de evaluación, acorde con el método de corrección empleado [4]. Esto hace que el sistema de corrección que se emplee tenga una efecto sobre los parámetros psicométricos del sistema de evaluación [9], entre los que están el índice de discriminación y el índice de dificultad, lo que permite tomar medidas de mejora sobre los métodos de evaluación de los sistemas educativos [6].

La puntuación final del estudiante que se somete a una prueba de evaluación de respuestas multiples (*MCQ*), está fuertemente influenciada por el modo en que el professor procesa (sistema de corrección) los datos en bruto (*Raw Data*) [4]. En el presente estudio se ha utilizado sobre un mismo sistema de evaluación de pregunas de respuestas multiples, *MCQ*, dos sistemas de corrección, denominados *MCQac* y *MCQvf*. En planteamiento del *MCQac* no penaliza las afirmaciones incorrectas contestadas, pero sí que deja de considerar las afirmaciones correctas bien constestadas, cuando en una pregunta hay más de una afirmación correcta y el estudiante no responde a todas ellas. El sistema de corrección *MCQvf*, es una corrección basada en un planteamiento simulado, que ni el estudiante ni el professor puede ver sobre el propio documento del examen *MCQ*. La razón es que de todas las afirmaciones que el estudiante analizada durante el

ISBN 978-84-09-02374-5

ejercicio de calificación, el 50% son correctas y el otro 50% son incorrectas, aunque su distribución en el examen es desconocida. Lo que permite hacer sobre los mismos exámenes, para los mismos estudiantes y en las mismas condiciones ambientales, una corrección adicional consistente en un examen de verdadero-falso con puntos negativos. Es decir, que como cada "*pregunta simulada*" tendría dos afirmaciones, una correcta y otra incorrecta, cuando se contesta la incorrecta se penalizada la nota final. La ventaja del sistema *MCQav* es que no penaliza *a priori* las afirmaciones incorrectas contestadas. Cuando en el sistema de corrección se utiliza la penalización de respuestas incorrectas, puede ser visto como una especie de lotería, ya que la omisión de responder puede ser vista como una "*recompensa*", mientras que la de responder como una "*lotería*" [10].

La validez y calidad de los sistemas de corrección utilizados en este estudio, *MCQav* y *MCQvf*, fueron en ambos casos *Excelentes*, en términos de discriminación [7], y *Bueno-Aceptable*, en términos de dificultad [3]. Es decir, que mostraron un alto grado de discriminación entre estudiantes con buenos y malos resultados, así como un grado de dificultad adecuado [11]. Para cada sistema de evaluación, con su método de corrección, los parámetros psicométricos como *IDis* e *IDif* deben ser estimados, ya que estos no están necesariamente correlacionados [9]. Otros sistemas de corrección han sido aplicados al sistema de evaluación *MCQ*, consistente en elegir la afirmación correcta de cuatro afirmaciones por pregunta sin puntos negativos [6], mostrando *IDis* e *IDif* de 0,28 (una discriminación *Regular*) y 51,63% (una dificultad *Excelente-ideal*), respectivamente. *IDis* e *IDif* para valorar preguntas útiles en los sistemas de evaluación *MCQ* en estudiantes de farmacología [3].

La nota media de ambos sistemas de corrección fue siempre inferior a cinco, siendo el sistema *MCQac* un 5,7% superior respecto a *MCQvf*. Esto fue debido a un efecto de posición de las materias previas necesarias para el correcto aprendizaje de la *MGA*. De hecho, cuando se analizaron las promociones de este estudio, las materias de Bioestadística, Bioquímica y Genética, estaban entre primero, segundo y tercer curso, respectivamente, mientras que la *MGA* lo estaba en quinto. Después de la última modificación del plan de estudios de la facultad de veterinaria, todas las materias se dan en los dos primeros cursos, usando los cuatro semestres de los que se disponen, e impartiéndose la *MGA* en el último de ellos. Desde entonces, con el mismo sistema *MCQ*, la nota media pasó a ser de un cinco.

La singularidad destacable del presente trabajo, es que se emplean dos sistemas de corrección (*MCQac* y *MCQvf*) dentro del marco de un mismo sistema de evaluación (*MCQ*), lo que ha permitido que se refleje mejor el rendimiento del estudiante durante el proceso de aprendizaje. De hecho, de haberse utilizado solo el sistema de corrección *MCQac*, el 25% de los estudiantes con notas comprendidas entre comprendida entre ≥ 4 y 5<, no hubiesen aprobado, algo que si sucede cuando se valora su esfuerzo y rendimiento con la introducción de la corrección *MCQvf*. Igualmente, la disposición de métodos de corrección que permitan corregir el efecto de éstos sobre la puntuación final del estudiante, hacen una prueba más justa, que al final refleja más adecuadamente las competencias adquiridas por el estudiantes, como son saber estimar parámetros genéticos, predicción de la respuesta a la selección, etc…

5. REFERENCIAS

[1] Afonso, J.M., Zamorano, M.J. Proyecto Docente de Mejora Genética Animal. ULPGC (2017).
[2] U.N.E.S.C.O. Nuevo manual de la UNESCO para la enseñanza de las ciencias. Ed., Edhasa, Barcelona (1987).
[3] Kaur, M., Singla, S., and Mahajan, R. Item analysis of in use multiple choice questions in pharmacology. Int J App Basic Med Res 6:170-3 (2016).
[4] Scharf, E.M., and Baldwin, L.P. Assessing multiple choice question (MCQ) tests – a mathematical perspective. SAGE Publications (London, Los Angeles, New Delhi and Singapore) Vol 8(1): 31–4 (2007).
[5] Ganzfried, S. and Yusuf, F. OptimalWeighting for Exam Composition. Educ. Sci., 8, 36 (2018).
[6] Islam, I., and Usmani, A. Psychometric analysis of Anatomy MCQs in Modular examination. Pak J Med Sci.;33(5):1138-1143 (2017).
[7] Escudero. B.F., Reyna, N.L. and Rosas, M. Nivel de dificultad y poder de discriminación del Examen de Habilidades y Conocimientos Básicos (EXHCOBA). Revista electronica de investigación educative. Vol 2, Nº1 (2000).
[8] Lesage, R., Valcke. M., and Sabbe, E. Scoring methods for multiple choice assessment in higher education – Is it still a matter of number right scoring or negative marking?. Studies in Educational Evaluation 39; 188–193 (2013).
[9] Şenela, S., Burcu, E., and Alatlıc, B. Effect of Correction-for-Guessing Formula on Psychometric Characteristics of Test. Procedia - Social and Behavioral Sciences 191; 925 – 929 (2015).

 ISBN 978-84-09-02374-5

[10] Espinosa, M.P., and Gardeazabal, J. Optimal correction for guessing in multiple-choice tests. Journal of Mathematical Psychology 54; 415-425 (2010).
[11] Crocker, L., and Algina, J. Introduction to Classical and Modern Test Theory. New York: Holt, Rinehart and Winston (1986).

ISBN 978-84-09-02374-5