Congreso In-Red 2022
UPV, 6 - 8 de julio de 2022
Doi: https://doi.org/10.4995/INRED2022.2022.15905

Análisis de la calidad de exámenes de alternativa múltiple a través de la teoría de respuesta al item: aplicación en la asignatura de "Evaluación Psicològica"

Assessment of the quality of multiple-choice exams through the Item Response Theory: implementation in the subject of "Psychological Assessment".

Jesús Castro-Calvo[a], Diana Pons-Cañaveras[b], Patricia Beltrán-Martínez[c], Francisco Atienza-González[d], Ascensión Bellver-Pérez[e], Usue De la Barrera-Marzal[f], Amelia Díaz-Martínez[g], Alicia Juan-Hidalgo[h], Laura Lacomba-Trejo[i], Adriana Mira-Pastor[j], Estefanía Mónaco-Gerónimo[k], Inmaculada Montoya-Castilla[l], Konstanze Schoeps[m], Castora Silva-Silva[n] y Maja E. Wrzesien[ñ]

[a]Departamento de Personalidad, Evaluación y Tratamientos Psicológicos, Universitat de València (jesus.castro@uv.es), [b]Departamento de Personalidad, Evaluación y Tratamientos Psicológicos, Universitat de València (diana.pons@uv.es), [c]Departamento de Personalidad, Evaluación y Tratamientos Psicológicos, Universitat de València (patribeltranmartinez@gmail.com), [d]Departamento de Personalidad, Evaluación y Tratamientos Psicológicos, Universitat de València (francisco.l.atienza@uv.es), [e]Departamento de Personalidad, Evaluación y Tratamientos Psicológicos, Universitat de València (bellpeas@uv.es), [f]Departamento de Personalidad, Evaluación y Tratamientos Psicológicos, Universitat de València (usue.barrera@uv.es), [g]Departamento de Personalidad, Evaluación y Tratamientos Psicológicos, Universitat de València (amelia.diaz@uv.es), [h]Departamento de Personalidad, Evaluación y Tratamientos Psicológicos, Universitat de València (alicia.juan@uv.es), [i]Departamento de Personalidad, Evaluación y Tratamientos Psicológicos, Universitat de València (laura.lacomba@uv.es), [j]Departamento de Personalidad, Evaluación y Tratamientos Psicológicos, Universitat de València (adriana.mira@uv.es), [k]Departamento de Personalidad, Evaluación y Tratamientos Psicológicos, Universitat de València (estefania.monaco@uv.es), [l]Departamento de Personalidad, Evaluación y Tratamientos Psicológicos, Universitat de València (inmaculada.montoya@uv.es), [m]Departamento de Personalidad, Evaluación y Tratamientos Psicológicos, Universitat de València (konstanze.schoeps@uv.es), [n]Departamento de Personalidad, Evaluación y Tratamientos Psicológicos, Universitat de València (Castora.Silva@uv.es) y [ñ]Departamento de Personalidad, Evaluación y Tratamientos Psicológicos, Universitat de València (maja.wrzesien@uv.es).

How to cite: Castro-Calvo, J, Pons-Cañaveras, D., Beltrán-Martínez, P., Atienza-González, F., Bellver-Pérez, A., De la Barrera-Marzal, U., Díaz-Martínez, A., Juan-Hidalgo, A., Lacomba-Trejo, L., Mira-Pastor, A., Mónaco-Gerónimo, E., Montoya-Castilla, I., Schoeps, K., Silva-Silva, C. y Wrzesien, M.E. 2022. Análisis de la calidad de exámenes de alternativa múltiple a través de la teoría de respuesta al item: aplicación en la asignatura de "Evaluación Psicològica". En libro de actas: *VIII Congreso de Innovación Educativa y Docencia en Red.* Valencia, 6 - 8 de julio de 2022. https://doi.org/10.4995/INRED2022.2022.15905.

Abstract

Multiple Choice testing (MCT) is one of the most popular approaches to the assessment of knowledge acquisition. Preparing a MCT is complex; however, its metric quality (in terms of difficulty, discrimination capacity, or distractors effectiveness) is not usually assessed. "Item Response Theory" (IRT) is a statistical approach that may assist when evaluating the quality of a MCT. IRT lets to estimate indices that may be used later as a way to improve the quality of the assessment. The aim of this work was to present the results of a teaching innovation project in which the IRT was employed to analyze the MCT used during the last 3 courses in the subject «Psychological Evaluation» (Degree in Speech Therapy). Results

suggested that a significant proportion of the questions analyzed presented certain limitations, such as the excessive prevalence of easy items (between 45%-67.5%) or ineffective distractors (between 17.5%-26.6%). Yet, these issues did not impact the discrimination capacity of the MCT. These results serve as a basis for proposing initiatives to improve evaluations through MCT, which ultimately will result in a fairer and more balanced evaluation of the students of the subject.

Keywords: *Multiple Choice testing (MCT); quality assessment; Item Response Theory (IRT); Psychological Assessment; Degree in Speech Therapy.*

Resumen

Los Exámenes de Alternativa Múltiple (EAM) son la forma más popular de evaluar la adquisición de conocimiento. Elaborar un EAM entraña cierta complejidad; sin embargo, rara vez se analiza en qué grado los EAM cumplen los criterios de calidad exigibles a estas pruebas (p.e., en términos de dificultad, capacidad de discriminación o eficacia de los distractores). La «Teoría de Respuesta al Ítem (TRI)» es una aproximación estadística que puede ayudar a la hora de evaluar la calidad de un EAM a través de la obtención de índices objetivos que, en una fase posterior, se pueden usar para mejorar la evaluación. El objetivo de este trabajo es presentar los resultados de un Proyecto de Innovación Docente en el que se analizó, mediante TRI, los EAM usados durante los últimos tres cursos en la asignatura «Evaluación Psicológica» (Grado en Logopedia). Los resultados sugieren que una proporción importante de preguntas analizadas presentaban limitaciones relacionadas con la excesiva presencia de ítems fáciles (entre el 45%-67.5%) o distractores ineficaces (entre el 17.5%-26.6%), si bien esto no afectó sobre su capacidad de discriminación. Estos resultados sirven de base para proponer iniciativas que permitan mejorar las evaluaciones a través de EAM.

Palabras clave: *Exámenes de Alternativa Múltiple (EAM); evaluación de la calidad; Teoría de Respuesta al Ítem (TRI); Evaluación Psicológica; Grado en Logopedia.*

1. Introducción

1.1. Exámenes de Alternativa Múltiple: ventajas y limitaciones

A pesar de que las primeras referencias a este método de evaluación se remontan a hace más de un siglo, los Exámenes de Alternativa Múltiple (EAM, también conocidos como «exámenes tipo test» o «pruebas objetivas») siguen gozando, a día de hoy, de muy buena salud. Así, los EAM son en la actualidad el método de evaluación más utilizado en la mayoría de contextos y de niveles educativos (Gierl et al., 2017). Un buen ejemplo lo encontramos en el Programa para la Evaluación Internacional de Alumnos de la OCDE (o PISA por sus siglas en inglés): en su edición de 2015, dos tercios de sus preguntas para evaluar conocimientos y habilidades en las áreas de matemáticas, ciencia y lectura se planteaban en formato de EAM (OECD, 2016). En el ámbito de la educación universitaria, se estima que el 80% de las asignaturas usan EAM como forma de evaluar la adquisición de conocimientos entre el alumnado (Birkhead et al., 2018). En muchos casos, los EAM constituirían el único criterio de evaluación (lo que significa que la calificación del alumno/a depende enteramente de su desempeño en un EAM); en otros, los EAM se complementarían con formas alternativas de evaluación (p.e., resolución de casos, preguntas cortas o de desarrollo, etc.). De modo que, con mayor o menor carga, los EAM suelen estar presentes en la guía docente de la mayoría de asignaturas universitarias.

Diferentes aspectos explican la popularidad de los EAM a la hora de evaluar la adquisición de conocimiento en el ámbito de la educación superior. El primero es su versatilidad: siempre y cuando esté bien elaborado (lo que, como veremos posteriormente, no es ni tan sencillo ni tan común), un EAM evaluaría prácticamente cualquier tipo de conocimiento, habilidad o competencia, en prácticamente cualquier rama de conocimiento y a prácticamente cualquier nivel de profundidad (memorización, comprensión de conceptos, elaboración de juicios e inferencias, razonamiento, interpretación de datos, etc.) (Downing, 2006). En segundo lugar, destacaría su eficiencia: los EAM son fáciles de administrar –tanto en formato individual como colectivo– y su corrección es muy rápida y sencilla. Eso supone que, a nivel práctico, el coste de realizar una evaluación a través de un EAM –en términos de tiempo y esfuerzo– sean mucho menores que en el caso de usar formas alternativas de evaluación. La última ventaja de los EAM radica en su objetividad: los EAM permiten una calificación independiente del evaluador, lo que garantiza que todo el estudiantado se enfrente a la prueba en igualdad de condiciones y que su resultado no dependa del criterio del docente (Haladyna, 2018).

Sin embargo, el hecho de que los EAM sean la forma de evaluación más común no significaría –ni mucho menos– que sea la más recomendable, ni tampoco que estén libres de limitaciones. De hecho, cada vez son más los expertos que criticarían su uso (sobre todo, el abuso que de ellos se ha hecho durante mucho tiempo) y que apostarían por formatos de evaluación alternativos (Pereira et al., 2016). Entre sus limitaciones, estos expertos señalan que construir un buen EAM es complejo y requiere más tiempo que implementar cualquier otro método de evaluación (Haladyna, 2018). Así, se han desarrollado multitud de guías que asisten en la elaboración de EAM (p.e., Boland et al., 2010; Coughlin & Featherstone, 2017), lo que demostraría que construir una prueba de este tipo con garantías requiere de un proceso de reflexión y de un conocimiento exhaustivo de la asignatura a evaluar.

Una limitación comúnmente señalada es que muchos EAM se limitan a evaluar el conocimiento superficial. Este sería el caso de los EAM basados en el "reconocimiento": es decir, pruebas que simplemente requieren de la evocación memorística de conceptos y que, por tanto, suelen resultar excesivamente sencillas y apenas suscitarían la activación de niveles de conocimiento profundos o de estrategias cognitivas de orden superior. Otra de las críticas en contra de los EAM tiene que ver con el llamado "*testing effect*", que se define como el incremento en la eficacia a la hora de contestar EAM por el hecho de ganar práctica en su realización (y no tanto por los conocimientos que se tengan sobre el área evaluada) (Marsh et al., 2007). Se ha demostrado que los estudiantes mejoran a la hora de contestar EAM a medida que ganan práctica en su realización: así, un/a estudiante con mucha experiencia en realizar este tipo de evaluación podría obtener una mejor nota que otro con un nivel equivalente de conocimiento pero con menos práctica (Rowland, 2014). Finalmente, también se ha criticado el hecho de que los EAM propicien la creación de "bancos de preguntas": es decir, bases de datos donde se almacenan las preguntas elaboradas a lo largo de los años para la evaluación de los contenidos de una asignatura. Los bancos de preguntas no son, *per se,* una práctica docente negativa; de hecho, es muy común –incluso recomendable– que los/as docentes cuenten con un almacén de preguntas de alternativa múltiple que les hayan sido útiles y que las utilicen posteriormente para la elaboración de sus exámenes (Lane et al., 2016). Utilizado adecuadamente, un banco de preguntas optimiza la tarea de elaborar un EAM y mejora el modo en el que se evalúa los contenidos. El problema es que ciertas deficiencias en el uso de este tipo de recursos pueden pervertir su finalidad: así, cuando disponemos de un banco de preguntas pero no existe una revisión crítica de los ítems que lo integran, sucede que pueden llegar a mantenerse a lo largo del tiempo preguntas que en realidad carecerían de garantías de medida (p.e., que son extremadamente fáciles, difíciles o carecen de capacidad de discriminación).

En resumen, los EAM presentan toda una serie de ventajas que explican que sean la forma más común de evaluación del conocimiento en educación superior; sin embargo, existen limitaciones asociados a su uso que pueden menoscabar su eficacia. En los contextos en los que sucede hoy en día la docencia universitaria (grandes ratios de estudiantes, excesiva carga docente, etc.), no podemos esperar que los EAM vayan a ser sustituidos –ni total ni parcialmente– por formas de evaluación más deseables (como aquellas que permiten evaluar competencias prácticas y/o profesionales). Esto significa que, de momento, nuestro objetivo debería centrarse en mejorar las formas de evaluación comúnmente usadas a día de hoy, entre las que destacaría los EAM. Es en este último aspecto en el que se centra el trabajo aquí presentado, que se deriva de un Proyecto de Innovación Docente (PID-1640371) centrado en mejorar la calidad de los EAM usados en la asignatura "Evaluación Psicológica" a través del análisis de su calidad métrica. En concreto, en este trabajo se exponen los resultados del análisis de la calidad métrica de las preguntas usadas en los EAM de los últimos tres años de la asignatura "Evaluación Psicológica" (Grado en Logopedia), así como las conclusiones y orientaciones que de este análisis se pueden derivar para mejorar la forma en la que se realizan estas evaluaciones.

1.2. Evaluación de la calidad métrica de un EAM a través de la Teoría de Respuesta al Ítem (TRI)

Prácticamente todas las limitaciones de los EAM enumeradas anteriormente resultarían controlables cuando se analiza cuidadosamente su calidad: así, saber en qué medida un EAM cumpliría o no con los criterios de calidad exigibles a este tipo de pruebas permitiría tomar decisiones acerca de su elaboración que, en última instancia, permitirían mejorar la evaluación del alumnado. Teniendo en cuenta que en torno al 80% de las asignaturas universitarias evalúa sus contenidos –completa o parcialmente– a través de EAM (Birkhead et al., 2018), este análisis ya no sólo constituye una "*buena práctica*" docente, sino que debería considerarse una exigencia ética: así, de igual modo que exigimos que cualquier prueba sobre la cual se toman decisiones de calado (p.e., una prueba médica, un test de inteligencia, un cuestionario de psicopatología, etc.) cumpla adecuadamente con el cometido para el cual se ha diseñado, lo mismo debería exigirse de las pruebas usadas para enjuiciar el desempeño del alumnado en una asignatura (Bennett, 2015).

Existen dos niveles de evaluación de la calidad de un EAM: el primero pasa por analizar en qué medida las preguntas de un EAM se adhieren a las directrices propuestas para la elaboración de este tipo de evaluación, mientras que el segundo supone analizar empíricamente la calidad de la prueba a partir de criterios objetivos (es decir, estadísticos). En cuanto al primer nivel de análisis, es de tipo cualitativo y subjetivo ya que implica emitir un juicio acerca del ajuste de los ítems de un EAM a una serie de criterios de calidad estandarizados. Esto permitiría hacer un análisis de calidad del EAM antes de su administración (*pre hoc*), lo que supondría una importante ventaja. Sin embargo, también presenta grandes limitaciones: p.e., (a) que un ítem esté bien construido no significa que luego vaya a funcionar adecuadamente en una población determinada; (b) este análisis no dice nada acerca de la dificultad de las preguntas ni de su capacidad para discriminar entre el estudiantado con mayor o menor nivel de conocimiento; y (c) a pesar de basarse en criterios estandarizados, su valoración tiene un corte subjetivo.

Frente a estas limitaciones, el segundo nivel de análisis de la calidad de un EAM pasa por explorar la calidad métrica a través de criterios estadísticos una vez el examen ha sido ya administrado (evaluación *post hoc*). Una desventaja respecto al nivel anterior es que las mejoras al EAM sólo se podrían introducir en versiones sucesivas del mismo (no en el examen antes de adminsitrarlo). Sin embargo, mientras la evaluación anterior es cualitativa y subjetiva, el segundo nivel de análisis permite una aproximación cuantitativa y objetiva. Se han adoptado cinco modelos distintos –teorías– para evaluar objetivamente la calidad de los EAM: la Teoría Clásica de los Test (TCT), el Análisis Factorial (AF), el Análisis Clúster (AC), los Modelos Dinámicos de Respuesta (MDR) y la Teoría de Respuesta al Ítem (TRI) (Ding & Beichner, 2009). De todas ellas, la que ha tenido un mayor calado ha sido la TRI, ya que se asienta sobre unos axiomas que ajustan bien con los

principios que rigen la medida de los conocimientos, y además provee de una serie de índices que permiten obtener medidas con un impacto directo sobre la *praxis* educativa. La TRI perseguiría "evaluar la habilidad latente de los individuos" (Baker & Kim, 2017). En el caso que nos ocupa, la habilidad latente (o *theta* [ɵ] en su notación estadística) sería el nivel de conocimiento sobre el material que se quiere evaluar. Para modelar dicho nivel de habilidad latente, la TRI dispone de una serie de indicadores que nos permiten evaluar la calidad de los EAM. El primero de ellos sería el índice de dificultad (*b* en su notación estadística) (Ding & Beichner, 2009). Este índice responde a la pregunta de: ¿cómo de difícil es un ítem? Una forma fácil de aproximarse a este índice es estimando el porcentaje de estudiantes que contestan acertadamente a un ítem en un examen: si una pregunta en un EAM es contestada correctamente por más de un 70% de los/as estudiantes se consideraría fácil, de dificultad moderada cuando contestan bien entre el 20-70% de los estudiantes y difícil cuando la aciertan menos del 20% de los/as estudiantes (Abdulghani et al., 2015). En un EAM ideal, la mayoría de las preguntas (en torno al 60%) deberían tener una dificultad media –ya que son las que mejor capacidad de discriminación presentarían–, mientras que el 40% restante debería repartirse a partes iguales entre preguntas fáciles (20%) y difíciles (20%). La dificultad calculada a través de la TRI va más allá del mero conteo del porcentaje de aciertos ya que tiene además en cuenta cómo el nivel de conocimiento latente modula la probabilidad de acertar una pregunta determinada.

Otro índice de interés derivado de la TRI sería la capacidad de discriminación (*a* en su notación estadística). La capacidad de discriminación responde a la pregunta de: ¿en qué medida un ítem permite distinguir entre aquellos alumnos/as con un mayor/menor nivel de conocimientos? Es decir, responde a la pregunta básica que debe guiar cualquier proceso de evaluación educativa (Gajjar et al., 2014). Cuando la mayoría de los/as estudiantes con buen nivel de conocimientos responde acertadamente una pregunta de un EAM y la mayoría de estudiantes con pobre nivel de conocimiento responde mal, decimos que el ítem tendría buena capacidad de discriminación (es decir, permite distinguir a los/as alumnos/as en función de su nivel de adquisición de conocimiento) (Toksöz & Ertunç, 2017); cuando este balance se desequilibra, empeoraría la capacidad de discriminación y cabría plantearse la pertinencia de su uso. En este sentido, posturas estrictas plantean que los EAM no deben incluir absolutamente ninguna pregunta que no reporte una capacidad de discriminación entre buena y excelente (Hingorjo & Jaleel, 2012): no en vano, sería coherente excluir una pregunta cuando no discrimina adecuadamente (esto es, que acertar/fallar no depende del nivel de conocimiento).

El último de los indicadores de calidad derivados de la TRI sería el de la eficacia de los distractores (Ding & Beichner, 2009). Los distractores son aquellas alternativas de respuesta incorrectas que acompañarían a la correcta cuyo objetivo es distraer al evaluado e inducir una duda razonable. La eficacia de un distractor se mide a partir del porcentaje de respuestas que reciben las alternativas incorrectas: cuando una alternativa incorrecta recibe un porcentaje muy bajo de respuestas (<5% de alumnos contestan la alternativa), entonces se considera que el distractor no induce ningún tipo de duda razonable y, por tanto, no cumple su objetivo (Gronlund & Linn, 1990). En esos casos, se recomienda o bien eliminar la pregunta completa o sustituir la alternativa de respuesta por otra más plausible que cumpla su función distractora (Hingorjo & Jaleel, 2012).

Como se desprende de lo comentado hasta el momento, disponemos de una serie de indicadores objetivos, que además están sólidamente fundamentados en una teoría robusta (la TRI) y que nos pueden asistir en la tarea de evaluar la calidad de los ítems que integran los EAM. Ahora bien, ¿qué sucede cuando se analizan los EAM habitualmente utilizados desde el tamiz de esta teoría? Según Haladyna et al. (2002), lo que sucede es que en torno al 50% de los EAM no cumple con los estándares de calidad descritos y, por tanto, presentan pobres capacidades métricas. Eso supone que 5 de cada 10 exámenes no estarían bien escalados en términos de dificultad, poseerían una pobre capacidad de discriminación entre estudiantes con mayor o menor nivel

de conocimiento o no estarían bien construidos respecto a sus distractores. En esta línea, D'Sa & Visbal-Dionaldo (2017) estimaron que sólo el 48% de preguntas de un EAM que usaban para evaluar conocimiento sobre enfermería se ajustaban a los criterios de calidad mencionados. Además, en torno al 50% de preguntas tenían uno o más distractores poco eficaces. Aplicaciones posteriores sugieren que estas cifras podrían estar un tanto sobredimensionadas. Por ejemplo, Rao et al. (2016) analizaron un EAM aplicado a estudiantes de medicina y encontraron que el 30% de los ítems tenían una capacidad de discriminación entre baja y moderada y que sólo el 5% de los distractores eran ineficaces. Estas cifras resuenan con las obtenidas por Toksöz & Ertunç (2017) en un EAM sobre dominios lingüísticos, donde el 28% de los ítems tenían una pobre capacidad de discriminación. Finalmente, en el estudio de Hingorjo & Jaleel (2012) se concluía que el 64% de los ítems sí cumplía con los estándares de calidad exigibles a un EAM; dicho de otro modo, que el 36% de ítems no cumplían adecuadamente su función a la hora de evaluar el conocimiento del alumnado.

2. Objetivos e hipótesis

El objetivo de este trabajo fue analizar la calidad métrica de los EAM usados los últimos tres cursos (2019-2020, 2020-2021 y 2021-2022) en la asignatura "Evaluación Psicológica" (Grado en Logopedia) a través de la TRI. A partir de este análisis, se proponen una serie de estrategias que permitirán mejorar los EAM empleados en la asignatura en cursos sucesivos, lo que en última instancia, redundará en la mejora de la *práxis* evaluativa y en una evaluación más justa y proporcionada del alumnado de la asignatura.

Teniendo en cuenta que los escasos estudios que han analizado la calidad métrica de los EAM identifican porcentajes de preguntas problemáticas de entre el 28%-48%, esperamos que una proporción similar de preguntas de los exámenes que se analizarán presenten algún tipo de limitación. Concretamente, se espera encontrar problemas en el escalamiento de la dificultad de los EAM que posiblemente se acompañe de algnunas limitaciones en cuanto a su capacidad de discriminación y también en cuanto a la eficacia de los distractores.

3. Desarrollo de la innovación

El presente trabajo forma parte de un Proyecto de Innovación Docente (PID-1640371) en el que se utilizaba la TRI para mejorar la evaluación a través de EAM. El primer paso para realizar este proyecto de innovación pasó por recopilar los EAM utilizados en las tres últimas convocatorias (2019-2020, 2020-2021 y 2021-2022) en la asignatura "Evaluación Psicológica", del Grado en Logopedia. Esta asignatura, de carácter semestral e impartida en el 2º año de la titulación, comprende una evaluación final a través de un EAM de 30 (curso 2019-2020) o 40 preguntas (cursos 2020-2021 y 2021-2022) que supondría el 70% de la nota. Es decir, que los resultados en el EAM suponían un porcentaje importante de la nota final del alumnado en la asignatura. Nótese que estos exámenes formaban parte de la evaluación ordinaria de la asignatura, de modo que el presente estudio se trataría de una evaluación retrospectiva y ecológica (es decir, sin intervención de ningún tipo). Como en cualquier otro EAM ordinario, las preguntas que los integraban trataban de evaluar las competencias teóricas contempladas en la guía de la asignatura.

Para el proyecto de innovación, se decidió utilizar únicamente los exámenes correspondientes a la primera convocatoria de los cursos mencionados (esto es, la convocatoria donde se presentan y aprueban una mayor proporción del alumnado). Se desechó la idea de incluir también los EAM de segunda convocatoria, ya que el *n* de alumnos en estos exámenes (entre 10-15 alumnos/as) no habría permitido aplicar la TRI con garantía. Dado que el profesorado está sujeto a una normativa clara en términos de conservación de exámenes durante un periodo concreto, realizar la recopilación de EAM resultó relativamente sencillo. Una vez se dispuso de

los exámenes, se generó una base de datos donde se transcribió la respuesta de cada alumno/a a las preguntas del correspondiente EAM. Las respuestas se codificaron primero indicando la alternativa contestada (nótese que los EAM analizados incluían tres alternativas), y posteriormente se recodificaron en términos de acierto o error (formato dicotómico) para su abordaje estadístico mediante TRI.

Una vez generadas las bases de datos (una para cada EAM analizado), se procedió a calcular los siguientes índices derivados de la TRI: (a) índice de dificultad; (b) capacidad de discriminación; (c) eficiencia de los distractores; y (d) la curva de información del test. Si bien existen fórmulas para el cálculo manual de estos índices, su estimación se realizó a través de dos softwares estadísticos: jMetrik 4.1.1 (Meyer, 2014) y Stata 16.0 (este último únicamente para el cálcula de la curva de información de los EAM).

4. Resultados

Teniendo en cuenta las características de los exámenes analizados y los parámetros de la TRI que interesaba conocer, el método de estimación estadístico escogido fue el 2PL (o "*two-parameter logistic*") (Brown & Abdulnabi, 2017). A través de este método, se obtuvieron los valores de *b* (dificultad) y de *a* (capacidad de discriminación) de los EAM utilizados en la asignatura "Evaluación Psicológica" durante los últimos tres cursos académicos (Tabla 1). En la Tabla 1 se incluiría también el porcentaje de alumnos que respondieron correctamente a cada pregunta, así como la presencia de distractores ineficaces (entendiendo como tal a las alternativas incorrectas de respuesta que fueron escogidas por menos del 5% de los alumnos/as) (Gronlund & Linn, 1990).

El parámetro *b* corresponde a la dificultad de un determinado ítem (o "*cómo de fácil o de difícil es responder acertadamente a un determinado ítem*"). En términos estadísticos, este índice indica el conocimiento latente (ɵ) requerido para tener un 50% de probabilidad de contestar acertadamente a un ítem. Así, cuanto más ɵ requiera contestar acertadamente un ítem, más complicado será. El parámetro *b* oscila entre -∞ y +∞: cuanto mayor sea el valor negativo, más fácil es el ítem; al contrario, valores positivos indican que el ítem es difícil. Si analizamos los EAM desde la perspectiva de su dificultad, los exámenes correspondientes a los cursos 2019-2020 (X_b=-4.24) y 2020-2021 (X_b=-4.22) se considerarían entre fáciles y muy fáciles, mientras que la dificultad aumentaría significativamente en el examen del curso 2021-2022 (X_b=-2.64). Corroborando esta estimación, el 63.3% de preguntas del curso 2019-2020 y el 67.5% de las del 2020-2021 fueron contestadas correctamente por más del 70% de los/as estudiantes (criterio para considerar una pregunta "*muy fácil*"), frente al 45% de las del curso 2021-2022. En el otro extremo, ninguna pregunta de los cursos 2019-2020 y 2020-2021 y sólo un 10% de las del curso 2021-2022 fueron contestadas correctamente por menos del 20% (criterio para considerar una pregunta "*muy difícil*"). Eso supone que el porcentaje de preguntas de un nivel de dificultad media sería del 36.7%, 32.5% y 45% respectivamente.

Tabla 1. Índices de calidad métrica derivados de la TRI

Ítems[a]	***Examen 2019-2020 ($n_{alumnos}$=62)***				***Examen 2020-2021 ($n_{alumnos}$=59)***				***Examen 2021-2022 ($n_{alumnos}$=66)***			
	% respuestas correctas	***b***	***a***	***Distractores ineficaces***	**% respuestas correctas**	***b***	***a***	***Distractores ineficaces***	**% respuestas correctas**	***b***	***a***	***Distractores ineficaces***
Ítem 1	79.03%	-4.38	0.69	Sí (2<5%)	82.76%	-3.54	0.78	Sí (1<5%)	41.54%	-0.28	0.24	No
Ítem 2	77.42%	-2.55	1.83	Sí (2<5%)	51.72%	-2.85	0.48	Sí (1<5%)	35.38%	-5.41	0.22	Sí (1<5%)
Ítem 3	64.52%	-2.03	1.00	No	39.66%	-1.61	0.41	No	66.15%	-2.29	0.74	Sí (1<5%)
Ítem 4	51.61%	-2.50	0.39	Sí (1<5%)	75.86%	-3.25	0.84	Sí (1<5%)	49.23%	-1.17	1.01	No
Ítem 5	48.39%	-1.08	1.62	No	67.24%	-1.30	1.76	No	27.69%	-1.05	0.72	No
Ítem 6	66.13%	-5.20	0.39	Sí (1<5%)	62.07%	-2.08	0.87	Sí (1<5%)	66.15%	-1.72	1.03	Sí (1<5%)
Ítem 7	59.68%	-2.73	0.40	Sí (1<5%)	84.48%	-2.48	1.78	Sí (2<5%)	96.92%	-3.82	1.30	Sí (2<5%)
Ítem 8	77.42%	-2.70	0.97	Sí (1<5%)	70.69%	-1.44	1.21	Sí (1<5%)	61.54%	-2.23	0.81	No
Ítem 9	85.48%	-13.98	0.24	Sí (2<5%)	89.66%	-2.62	1.52	Sí (1<5%)	30.77%	-0.51	1.04	No
Ítem 10	79.03%	-3.30	0.96	Sí (2<5%)	72.41%	-4.67	0.43	Sí (1<5%)	18.46%	-3.04	0.35	No
Ítem 11	74.19%	-5.68	0.48	Sí (1<5%)	87.93%	-4.36	0.72	Sí (1<5%)	76.92%	-4.61	0.58	Sí (1<5%)
Ítem 12	70.97%	-1.73	1.41	Sí (1<5%)	93.10%	-3.48	1.14	Sí (2<5%)	33.85%	0.53	0.94	Sí (1<5%)
Ítem 13	83.87%	-2.24	1.62	Sí (1<5%)	60.34%	-2.79	0.45	Sí (1<5%)	75.38%	-1.85	1.19	Sí (1<5%)
Ítem 14	58.06%	-3.88	0.58	Sí (1<5%)	74.14%	-5.68	0.42	Sí (1<5%)	13.85%	2.78	0.49	No
Ítem 15	88.71%	-2.99	1.15	Sí (1<5%)	94.83%	-3.87	1.25	Sí (2<5%)	67.69%	-3.91	0.39	Sí (1<5%)
Ítem 16	96.77%	-23.60	0.82	Sí (2<5%)	79.31%	-2.88	0.89	Sí (1<5%)	90.77%	-2.97	1.08	Sí (2<5%)
Ítem 17	72.58%	-3.25	0.94	Sí (2<5%)	37.93%	-0.42	0.61	No	83.08%	-2.74	1.04	Sí (1<5%)
Ítem 18	80.65%	-1.94	1.23	No	96.55%	-5.88	0.59	Sí (2<5%)	70.77%	-1.07	1.43	No
Ítem 19	50.00%	-0.31	1.58	No	87.93%	-5.82	0.71	Sí (2<5%)	32.31%	0.62	0.88	No
Ítem 20	54.84%	0.10	2.06	No	86.21%	-2.49	1.18	Sí (1<5%)	72.31%	-5.54	0.24	Sí (1<5%)
Ítem 21	75.81%	-1.48	1.64	Sí (1<5%)	72.41%	-3.40	0.62	Sí (1<5%)	61.54%	-1.64	0.65	Sí (1<5%)
Ítem 22	75.81%	-5.66	0.45	Sí (1<5%)	70.69%	-1.97	0.96	Sí (1<5%)	29.23%	0.27	1.88	No
Ítem 23	83.87%	-1.74	1.76	Sí (1<5%)	96.55%	-23.59	0.82	Sí (2<5%)	70.77%	-3.91	0.26	No
Ítem 24	82.26%	-5.42	0.29	Sí (1<5%)	70.69%	-1.93	1.94	Sí (1<5%)	90.77%	-3.06	1.40	Sí (2<5%)
Ítem 25	70.97%	-2.86	1.00	Sí (1<5%)	79.31%	-2.37	1.39	Sí (1<5%)	95.38%	-4.88	0.92	Sí (1<5%)
Ítem 26	69.35%	-5.88	0.46	Sí (1<5%)	63.79%	-3.34	0.85	Sí (1<5%)	67.69%	-4.26	0.39	Sí (1<5%)
Ítem 27	93.55%	-5.60	0.78	Sí (2<5%)	84.48%	-2.37	1.11	No	98.46%	-5.92	1.21	Sí (2<5%)
Ítem 28	87.10%	-5.70	0.74	Sí (2<5%)	77.59%	-1.85	1.75	Sí (1<5%)	66.15%	-3.31	0.39	Sí (1<5%)
Ítem 29	69.35%	-5.22	0.47	Sí (1<5%)	79.31%	-2.73	0.88	Sí (1<5%)	69.23%	-2.69	0.91	Sí (1<5%)
Ítem 30	38.71%	-1.68	0.48	Sí (1<5%)	94.83%	-4.73	0.96	Sí (2<5%)	84.62%	-5.91	0.40	Sí (1<5%)
Ítem 31					58.62%	-1.79	0.62	No	58.46%	-1.68	0.72	Sí (1<5%)
Ítem 32					75.86%	-4.19	0.59	Sí (1<5%)	90.77%	-10.63	0.28	Sí (2<5%)
Ítem 33					91.38%	-3.64	0.80	Sí (2<5%)	76.92%	-2.15	1.27	Sí (1<5%)
Ítem 34					41.38%	-4.35	0.39	No	76.92%	-2.19	1.17	Sí (1<5%)
Ítem 35					86.21%	-5.92	1.19	Sí (2<5%)	66.15%	-1.97	1.60	Sí (2<5%)
Ítem 36					96.55%	-26.93	0.82	Sí (2<5%)	80.00%	-8.43	0.24	Sí (1<5%)
Ítem 37					60.34%	-2.30	1.24	Sí (1<5%)	13.85%	-0.61	1.13	No
Ítem 38					68.97%	-5.61	0.39	Sí (1<5%)	78.46%	-2.06	1.66	Sí (2<5%)
Ítem 39					46.55%	-0.80	1.55	No	13.85%	1.46	0.46	Sí (1<5%)
Ítem 40					29.31%	-1.49	0.54	No	49.23%	-1.77	0.99	Sí (1<5%)

Nota: [a]= Nótese que aunque los ítems de los tres EAM estén dispuestos en las mismas filas, su contenido difiere entre exámenes (p.e., el ítem 1 del curso 2019-2020 era distinto al del ítem 1 del curso 2020-2021 o 2021-2022). Los índices métricos de los ítems de los tres EAM se incluyen en una misma fila únicamente a fin de ahorrar espacio en cuanto a su presentación (no con fines comparativos).

El parámetro *a* corresponde a la capacidad de discriminación de un ítem (o "*cómo de bueno es un ítem a la hora de discriminar entre personas con mayor y menor nivel de conocimiento*"). El parámetro *a* oscila entre -0.5 y +2. Los ítems con valores negativos se consideran problemáticos, ya que indicarían que las personas con mayor nivel de conocimiento latente tienen mas probabilidad de equivocarse al contestar a una pregunta (lo que atentaría contra la lógica de cualquier escala de medida de conocimientos). Cuanto más positiva sea

por tanto la puntuación de un ítem en capacidad de discriminación, mayor será su potencial a la hora de escalar en función del grado de conocimiento latente. Sin embargo, más no es siempre mejor: un ítem con gran capacidad de discriminación pero una dificultad muy alta será bueno escalando a estudiantes con alto grado de conocimiento, pero no será útil para escalar a aquellos con niveles de conocimiento más modestos. Así, es una condición deseable que los EAM dispongan de ítems con diferente capacidad de discriminación, lo que en interacción con la dificultad, permitirá evaluar correctamente en un mayor rango de conocimientos latentes (ɵ). Si analizamos ahora la capacidad de discriminación de los EAM de "Evaluación Psicológica" (tabla 1), lo que se aprecia en primer lugar es que ningún ítem presentaría una capacidad de discriminación negativa. En los tres exámenes, la capacidad media de discriminación sería muy similar ($X_{a\ [2019\text{-}2020]}$=0.94; $X_{a\ [2020\text{-}2021]}$=0.93; $X_{a\ [2021\text{-}2022]}$=0.84).

El resultado de la interacción entre el nivel de dificultad y la capacidad de discriminación se plasma en un estadístico de gran interés: la "curva de información". La curva de información es la representación gráfica del nivel de conocimiento latente (ɵ) en torno al cual un determinado EAM aporta una mayor cantidad de información. En términos prácticos, esta representación nos permite saber sobre qué perfiles de estudiantes nos aporta más información un EAM: si sobre aquellos que tienen un grado de conocimiento más modesto, sobre los que tienen un conocimiento muy alto o bien en estudiantes con niveles medios de conocimientos. Asumiendo que la distribución de los conocimientos sigue los principios de la distribución normal (es decir, que la mayor proporción de estudiantes se concentrará en niveles medios de conocimiento), una condición deseable para un EAM es que su curva de información se concentre alrededor de niveles de conocimiento medios (representados por valores de ɵ en torno a 0). En la Figura 1 se representan las curvas de información de los tres EAM analizados. Como se aprecia, en los tres casos –sobre todo, para el curso 2019-2020– las curvas se apilan a la izquierda (en torno a niveles negativos de ɵ), lo que significaría que estos tres exámenes dan más información cuando se evalúa a alumnos/as con niveles menores de conocimiento.

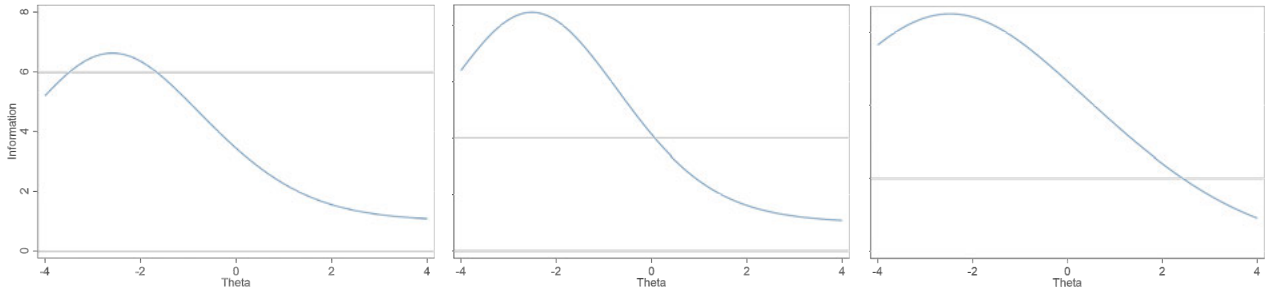

Figura. 1 Curvas de información de los EAM de los cursos 2019-2020 (izquierda), 2020-2021 (centro) y 2021-2022 (derecha)

El último aspecto de los EAM analizados fue la presencia de distractores ineficaces (entendiendo como tal la presencia de alternativas de respuesta escogidas por menos del 5% de los estudiantes). Como se apreciaría en la tabla 1, los tres EAM estarían altamente saturados de distractores ineficaces (un aspecto que también explicaría su facilidad). En el EAM del curso 2019-2020, sólo el 16.6% de los ítems no contenía distractores ineficaces, frente al 56.6% de ítems que tenían un distractor ineficaz y al 26.6% donde ambos distractores fueron ineficaces. Estos porcentajes fueron del 20%, 55% y el 25% respectivamente para el EAM del curso 2020-2021 y del 30%, 52.5% y 17.5% para el del curso 2021-2022.

5. Conclusiones

Este trabajo se planteó con el objetivo de analizar la calidad métrica de los EAM usados durante los últimos tres cursos (2019-2020, 2020-2021 y 2021-2022) en la asignatura "Evaluación Psicológica" (del Grado en Logopedia). Tras analizarlos a través de la óptica de la TRI, la principal conclusión sería que una proporción importante de las preguntas analizadas presentaban limitaciones relacionadas con la excesiva presencia de ítems fáciles o de distractores ineficaces, lo que podría menoscabar en cierto grado la calidad de estos EAM, si bien esto no afectó sobre su capacidad de discriminación.

Un primer aspecto a destacar tendría que ver con la dificultad de los EAM analizados. Idealmente, un EAM debería comprender alrededor de un 60% de preguntas de dificultad media, un 20% de preguntas fáciles y otro tanto de preguntas difíciles (Abdulghani et al., 2015). Esta distribución *a priorística* permitiría asegurar que el nivel de conocimiento latente (ɵ) en el que se agrupa una mayor proporción de estudiantes esté bien representado por una proporción mayor de preguntas, mientras que los extremos –inferior y superior– de la distribución dispongan también de un número apropiado de ítems que permitan el escalamiento en niveles de conocimiento latente más extremos. Esta distribución también permitiría prevenir la aparición de lo que se conoce como "efecto techo" (aquellos EAM en los que estudiantes con poco conocimiento pueden llegar a alcanzar notas muy altas) o "efecto suelo" (EAM tan difíciles que incluso estudiantes con gran ɵ obtienen notas bajas o modestas) (Lane et al., 2016). En los EAM analizados, nos hemos encontrado con unos niveles de dificultad excesivamente bajos, sobre todo en los correspondientes a los cursos 2019-2020 y 2020-2021 (donde el porcentaje de preguntas fáciles o muy fáciles fue del 63.3% y del 67.5% respectivamente). En el extremo contrario, ninguna de las preguntas administradas entraron en la categoría de "difíciles". Las cifras mejoran en el examen correspondiente al curso 2021-2022 (45% de preguntas fáciles, 45% medias y 10% difíciles); sin embargo, todavía queda lejos de la distribución 20-60-20 de la que hablábamos anteriormente. A nivel práctico, eso supone que las notas obtenidas mediante estos EAM pueden estar sobredimensionadas (dicho de otro modo, que el estudiantado ha obtenido notas mayores a lo que realmente merecen atendiendo a su nivel real de conocimiento latente). Este fenómeno principalmente beneficia a aquellos estudiantes con un peor nivel de conocimientos (ya que les permitiría aprobar con facilidad), y al contario, sanciona a los/as estudiantes con un nivel de conocimientos mayor (que, por el principio del "efecto techo", alcanzarían notas similares a las que obtienen otros con un menor nivel de conocimiento que igualmente alcanzan notas altas). Así, uno de los primeras aspectos a la hora de mejorar los EAM aplicados en la asignatura de "Evaluación Psicológica" debería ser aumentar la proporción de preguntas de dificultad media y también de dificultad alta, al tiempo que se reduce notablemente la presencia de preguntas fáciles o muy fáciles.

Como apuntábamos anteriormente, afortunadamente estos problemas en cuanto al escalamiento en términos de dificultad no habrían tenido un impacto muy significativo en la capacidad de discriminación de los EAM. Así, la capacidad de los tres EAM a la hora de distinguir entre alumnos/as con un mayor/menor nivel de ɵ fue positiva y apropiada (X_a entre .84 y .94). Teniendo en cuenta que la capacidad de discriminación es uno de los parámetros más directamente relacionados con la calidad de un EAM (Gajjar et al., 2014), este hecho asegura que, al menos, los exámenes analizados eran válidos para el objetivo para el cual fueron diseñados. Sin embargo, la capacidad de discriminación de un ítem no puede entenderse sin ahondar en la interacción entre este indicador y el nivel de dificultad. El resultado de esta interacción lo ilustrarían las ya mencionadas "curvas de información". Tras observar las "curvas de información" de los tres EAM analizados, se aprecia claramente que los tres dan bastante información cuando se evalúa a alumnos/as con un nivel más bajo de conocimiento. A nivel práctico, el riesgo de EAM así sería que no escalen adecuadamente a estudiantes con nivel medio-alto de conocimiento. Dicho de otro modo: que el EAM sea "apropiado" para examinar a los/as estudiantes con un nivel de conocimiento medio/bajo, pero que no permita distinguir adecuadamente entre

estudiantes con niveles altos de conocimiento (es decir, entre aquellos estudiantes que se muevan en rangos de notas de entre 7 y 10). Esto es lo que sucede en el caso de los EAM correspondientes a los curso 2019-2020 y 2020-2021, que aportan mucha información en niveles de ɵ inferiores a 0 pero muy poca en niveles superiores. El examen del curso 2021-2022 sigue la misma tendencia, pero la información que proporciona en niveles de ɵ mayores a 0 es superior. Teniendo en cuenta que los ítems de los tres exámenes presentaban una buena capacidad de discriminación, la forma apropiada de abordar este problema pasa nuevamente por aumentar la dificultad de los ítems a fin de que la capacidad para diferenciar entre un mayor y menor nivel de conocimiento latente alcance también a estudiantes en niveles de ɵ mayores.

El último aspecto a destacar tendría que ver con la eficacia de los distractores (esto es, con la calidad de las alternativas incorrectas a la hora de inducir una duda razonable) (Ding & Beichner, 2009). En nuestro caso, los tres EAM analizados estarían altamente saturados de distractores ineficaces. Que uno de los distractores resulte ser ineficaz es habitual, y no indicaría grandes problemas en la construcción del EAM; sin embargo, el hecho de que ambos distractores (en preguntas de tres alternativas) sean ineficaces sí indica la necesidad de revisar estas alternativas a fin de proponer distractores más plausibles que cumplan con su objetivo. En concreto, el porcentaje de preguntas con dos distractores ineficaces en el EAM del curso 2019-2020 fue del 26.6%, del 25% en el del curso 2020-2021 y del 17.5% en el del curso 2021-2022. Si bien estas tasas son altas y deben ser revisadas para sucesivas convocatorias, también es cierto que se situarían en un porcentaje notablemente inferior al encontrado en trabajos previos analizando este aspecto (D'Sa & Visbal- Dionaldo, 2017). En cualquier caso, dado que los distractores son uno de los aspectos que más determinan la dificultad de un EAM (precisamente, el principal aspecto a mejorar en los EAM analizados), la mejora de la eficacia de los distractores podría redundar en un beneficio global para la evaluación a través del incremento de la dificultad.

En cualquier caso, este trabajo no está exento de limitaciones. Por ejemplo, una de las críticas habituales a la TRI es que parte de la premisa de que los indicadores estimados son consecuencia de la medida utilizada, sin considerar que existen muchos otros factores más allá del EAM que pueden estar condicionando los parámetros obtenidos: así, un mismo examen puede obtener índices de dificultad altos cuando se aplica en grupos con rendimiento medio inferior y bajos en grupos de alto rendimiento o donde el seguimiento de la docencia ha sido continuo (Fan, 1998). Otra de las limitaciones tendría que ver con el *n* de las muestras analizadas; si bien los tamaños muestrales resultaron apropiados (*n* entre 59 y 66), en algunos casos se ha requerido de múltiples iteraciones para lograr la convergencia de los resultados. Muestras de mayor tamaño posiblemente permitirían asegurar con un mayor rigor la representatividad de los resultados, pero hay que considerar que el tamaño muestral depende del número de alumnos matriculados y presentados al examen (un factor ajeno a cualquier tipo de control). A pesar de las limitaciones enumeradas, creemos que este trabajo demuestra la pertinencia del uso de la TRI para el análisis de la calidad de los EAM e ilustra las posibles mejoras que se pueden introducir en este procedimiento de evaluación.

6. Referencias

Abdulghani, H. M., Ahmad, F., Irshad, M., Khalil, M. S., Al-Shaikh, G. K., Syed, S., Aldrees, A. A., Alrowais, N., & Haque, S. (2015). Faculty development programs improve the quality of Multiple Choice Questions items' writing. *Scientific Reports*, *5*. https://doi.org/10.1038/srep09556

Baker, F. B., & Kim, S. H. (2017). *The Basics of Item Response Theory Using R*. Springer.

Bennett, R. E. (2015). The Changing Nature of Educational Assessment. *Review of Research in Education*, *39*(1), 370–407. https://doi.org/10.3102/0091732X14554179

Birkhead, S., Kelman, G., Zittel, B., & Jatulis, L. (2018). The Prevalence of Multiple-Choice Testing in Registered Nurse Licensure-Qualifying Nursing Education Programs in New York State. *Nursing Education Perspectives*, *39*(3), 139–144. https://doi.org/10.1097/01.NEP.0000000000000280

Boland, R. J., Lester, N. A., & Williams, E. (2010). Writing multiple-choice questions. *Academic Psychiatry*, *34*(4), 310–316. https://doi.org/10.1176/appi.ap.34.4.310

Brown, G. T. L., & Abdulnabi, H. H. A. (2017). Evaluating the Quality of Higher Education Instructor-Constructed Multiple-Choice Tests: Impact on Student Grades. *Frontiers in Education*, *2*, 1. https://doi.org/10.3389/feduc.2017.00024

Coughlin, P. A., & Featherstone, C. R. (2017). How to Write a High Quality Multiple Choice Question (MCQ): A Guide for Clinicians. *European Journal of Vascular and Endovascular Surgery*, *54*(5), 654–658. https://doi.org/10.1016/j.ejvs.2017.07.012

D'Sa, J. L., & Visbal- Dionaldo, M. L. (2017). Analysis of Multiple Choice Questions: Item Difficulty, Discrimination Index and Distractor Efficiency. *International Journal of Nursing Education*, *9*(3), 109–114. https://doi.org/10.5958/0974-9357.2017.00060.5

Ding, L., & Beichner, R. (2009). Approaches to data analysis of multiple-choice questions. *Physical Review s - Physics Education Research*, *5*(2), 1–17. https://doi.org/10.1103/PhysRevSTPER.5.020103

Downing, S. M. (2006). Selected-Response Item Formats in Test Development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 287–301). Routledge.

Fan, X. (1998). Item Response Theory and Classical Test Theory: An Empirical Comparison of their Item/Person Statistics. *Educational and Psychological Measurement*, *58*(3), 357–381. https://doi.org/10.1177/0013164498058003001

Gajjar, S., Sharma, R., Kumar, P., & Rana, M. (2014). Item and test analysis to identify quality multiple choice questions (MCQS) from an assessment of medical students of Ahmedabad, Gujarat. *Indian Journal of Community Medicine*, *39*(1), 17–20. https://doi.org/10.4103/0970-0218.126347

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, Analyzing, and Using Distractors for Multiple-Choice Tests in Education: A Comprehensive Review. In *Review of Educational Research* (Vol. 87, Issue 6). https://doi.org/10.3102/0034654317726529

Gronlund, N. E., & Linn, R. L. (1990). *Measurement and evaluation in teaching*. Macmillan publishing.

Haladyna, T. (2018). Selected-response format: Developing multiple-choice items. In M. E. McDonald (Ed.), *The Nurse Educator's Guide to Assessing Learning Outcomes* (4th editio, pp. 77–132). Jones and Bartlett learning.

Haladyna, T., Downing, S. M., & Rodriguez, C. (2002). Applied Measurement in Education A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education*, *15*(3), 309–333.

Hingorjo, M. R., & Jaleel, F. (2012). Analysis of one-best MCQs: The difficulty index, discrimination index and distractor efficiency. *Journal of the Pakistan Medical Association*, *62*(2), 142–147.

Lane, S., Raymond, M. R., & Haladyna, T. (2016). Handbook of Test Development. In *Journal of Chemical Information and Modeling* (Vol. 53, Issue 9). Routledge.

Marsh, E. J., Roediger, H. L., Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences of multiple-choice testing. *Psychonomic Bulletin and Review*, *14*(2), 194–199. https://doi.org/10.3758/BF03194051

Meyer, J. P. (2014). *Applied Measurement with jMetrik*. Routledge. https://doi.org/10.4324/9780203115190

OECD. (2016). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic and Financial Literacy*. OECD Publishing. https://doi.org/10.1787/9789264255425-en

Pereira, D., Flores, M. A., & Niklasson, L. (2016). Assessment revisited: a review of research in Assessment and Evaluation in Higher Education. *Assessment and Evaluation in Higher Education*, *41*(7), 1008–1032. https://doi.org/10.1080/02602938.2015.1055233

Rao, C., Kishan Prasad, H., Sajitha, K., Permi, H., & Shetty, J. (2016). Item analysis of multiple choice questions: Assessing an assessment tool in medical students. *International Journal of Educational and Psychological Researches*, *2*(4), 201. https://doi.org/10.4103/2395-2296.189670

Rowland, C. A. (2014). The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463. https://doi.org/10.1037/a0037559

Toksöz, S., & Ertunç, A. (2017). Item Analysis of a Multiple-Choice Exam. *Advances in Language and Literary Studies*, *8*(6), 141. https://doi.org/10.7575/aiac.alls.v.8n.6p.141