REDU Revista de docencia Universitaria

Vol. 22(2), julio-diciembre 2024, 233-250
ISSN: 1887-4592

Fecha de recepción: 16/05/2024
Fecha de aceptación: 07/11/2024

¿Innovar en el examen tipo test? La prueba objetiva inversa para mejorar la evaluación sumativa en educación superior

Innovating on the multiple-choice test? Reverse objective testing to improve summative assessment in higher education

Fernando Martínez-Abad
fma@usal.es
Juan Pablo Hernández-Ramos
juanpablo@usal.es
José Carlos Sánchez-Prieto
josecarlos.sp@usal.es
Vanessa Izquierdo-Álvarez
vizquierdo@usal.es
María Teresa del Moral-Marcos
maitedelmoral@usal.es
María Serena Rivetta
serenarivetta@usal.es
Alberto Ortiz-López
aortiz@usal.es
Universidad de Salamanca (España)

Fernando Martínez-Abad
fma@usal.es
Juan Pablo Hernández-Ramos
juanpablo@usal.es
José Carlos Sánchez-Prieto
josecarlos.sp@usal.es
Vanessa Izquierdo-Álvarez
vizquierdo@usal.es
María Teresa del Moral-Marcos
maitedelmoral@usal.es
María Serena Rivetta
serenarivetta@usal.es
Alberto Ortiz-López
aortiz@usal.es
Universidad de Salamanca (Spain)

Resumen

Dadas sus importantes ventajas sobre otras estrategias, la prueba objetiva es un instrumento de evaluación de uso común

Abstract

Given its significant advantages over other strategies, the objective test is a commonly used assessment tool

To cite this article: Martínez-Abad, F., Hernández-Ramos, J.P., Sánchez-Prieto, J.C. Izquierdo-Álvarez, V., Del Moral-Marcos, M.T. Serena Rivetta, M., & Ortiz-López, A. (2024). Curricular internships as a factor for improving employability skills. *REDU. Revista de Docencia Universitaria, 22*(2), 233-250. https://doi.org/10.4995/redu.2024.21752

en la evaluación sumativa universitaria, aunque también se ha asociado a algunas limitaciones como el estudio memorístico, la promoción del conocimiento inmóvil y compartimentalizado, o el desarrollo de procesos de pensamiento de orden inferior. Esta investigación tiene como objetivo presentar y evaluar la eficacia de la prueba objetiva inversa como una modalidad de evaluación sumativa que mejora las prestaciones de las pruebas objetivas convencionales. Se aplica un diseño cuasiexperimental solo postest con grupo control no equivalente en un grupo de 204 estudiantes universitarios de grados de Ciencias de la Educación. Mientras que el grupo control recibe como instrumento de evaluación la prueba objetiva convencional, el grupo experimental completa la misma prueba en la modalidad inversa. Los resultados muestran que los estudiantes del grupo experimental alcanzan niveles de rendimiento académico superiores a los del grupo control, al mismo tiempo que muestran niveles de satisfacción y valoración global de la prueba más elevados. Mientras que ambos grupos perciben razonamientos de orden superior altos y el grupo experimental parece alcanzar niveles sensiblemente superiores, son los estudiantes con mejores calificaciones del grupo experimental los que autoperciben niveles más elevados. Ambos grupos informan niveles similares de ansiedad académica. Se evidencia que la prueba objetiva inversa es una alternativa que permite la mejora de los procesos de evaluación sumativa, promocionando mejores niveles de calidad en la enseñanza universitaria.

Palabras clave: evaluación sumativa, test de opción múltiple, enseñanza superior, pruebas objetivas, innovación educativa, razonamiento de orden superior.

in university summative assessment, although it has also been associated with some limitations such as rote memorization, the promotion of stagnant and compartmentalized knowledge, or the development of lower-order thinking processes. This research aims to present and evaluate the effectiveness of the reverse objective test as a mode of summative assessment that enhances the performance of conventional objective tests. A quasi-experimental design with post-test only and non-equivalent control group is applied to a group of 204 university students majoring in Education Sciences. While the control group receives the conventional objective test as the assessment instrument, the experimental group completes the same test in reverse mode. The results show that students in the experimental group achieve higher levels of academic performance than those in the control group, while also demonstrating higher levels of satisfaction and overall assessment of the test. While both groups perceive high levels of higher-order reasoning, the experimental group appears to achieve significantly higher levels, with students with higher grades in the experimental group perceiving even higher levels. Both groups report similar levels of academic anxiety. It is evident that the reverse objective test is an alternative that allows for the improvement of summative assessment processes, promoting higher levels of quality in university teaching.

Key words**:** summative assessment, Multiple-choice test, Higher education, Objective tests; Educational innovation, Higher-order reasoning.

Introducción y revisión de la literatura

A pesar de la implantación y consolidación del Espacio Europeo de Educación Superior, que supone una transformación radical de los procesos evaluativos en la universidad, la prueba objetiva (informalmente llamada *examen tipo test*) ha mantenido su preponderancia como técnica de evaluación sumativa de los estudiantes. Esta resiliencia de las pruebas objetivas se debe a numerosas ventajas (e.g., Polat, 2020), entre otras: objetividad en la asignación de una calificación numérica al estudiante, simplicidad y eficiencia en su aplicación y corrección, replicabilidad y adaptabilidad, y facilidad para su aleatorización y digitalización. No obstante, también ha recibido cuantiosas críticas, fundamentalmente relacionadas con la promoción del pensamiento memorístico, fragmentado y de orden inferior (Tractenberg et al., 2013), su relación con niveles más elevados de ansiedad académica (Wahyuni et al., 2021), y la influencia de habilidades de adivinación de la opción correcta del estudiante a partir de información contextual de ítems mal construidos (Almalki, 2023; Kissi et al., 2023).

Al respecto, se localiza en la literatura en los últimos años un interés creciente en la mejora de las pruebas objetivas, tanto a partir de su redacción como a través de su formato de presentación, para tratar de promover el pensamiento de orden superior (e.g., Haataja et al., 2023; Scully, 2017; Tractenberg et al., 2013) y para reducir los sesgos de medición y aplicación asociados con factores contextuales y no contextuales del estudiante (Almalki, 2023; Núñez-Peña y Bono, 2021; Singh et al., 2013).

En el marco de la educación superior, este trabajo propone el empleo de la Prueba Objetiva Inversa, una modalidad de aplicación de pruebas objetivas poco conocida, para la promoción del pensamiento de orden superior, la reducción de la ansiedad académica del estudiante, y el control de los sesgos de medición asociados a estas pruebas. Así, retomamos la línea iniciada por estudios previos (Bond et al., 2013; Vanderoost et al., 2018) desarrollados en áreas de Ciencias y Ciencias de la Salud, mostrando que las ventajas de estas pruebas son consistentes también en estudiantes universitarios de Ciencias Sociales. Además, la principal aportación de este estudio es analizar el desarrollo del pensamiento de orden superior con esta modalidad de prueba objetiva en relación con las pruebas convencionales. En un contexto de formación universitaria basada en competencias, en muchos casos con grupos numerosos de estudiantes, poder aplicar modalidades de evaluación eficientes como las pruebas objetivas manteniendo niveles elevados de razonamiento por parte de los estudiantes resulta de vital importancia.

Métodos de aplicación de la prueba objetiva y su eficacia

La aplicación de las pruebas objetivas se ha limitado fundamentalmente a su uso convencional, esto es, al planteamiento en cada ítem de un estímulo (enunciado del ítem o pregunta) y de una serie de opciones de respuesta, de las cuáles el estudiante debe seleccionar la opción que considere correcta. Esta modalidad de aplicación de las pruebas objetivas únicamente da pie a que el estudiante acierte o falle la pregunta, promoviendo una visión reduccionista y estática del conocimiento (Greving y Richter, 2022; Stringer et al., 2021): no permite que el estudiante

demuestre un conocimiento parcial sobre el contenido ni fomenta la reflexión del estudiante sobre el estímulo. Normalmente, los académicos y profesorado entienden que esta es la modalidad única de aplicación de la prueba objetiva, entendemos que por desconocimiento.

De hecho, investigaciones recientes evidencian que las pruebas objetivas convencionales se asocian a varios sesgos importantes de medida. Por un lado, en cuanto a los factores no contextuales y rasgos del estudiante, se observa que los estudiantes con una mayor ansiedad hacia los contenidos de la materia y con una menor confianza en sí mismos tienen una mayor aversión al riesgo, dejando más preguntas sin responder (Núñez-Peña y Bono, 2021), y que las pruebas objetivas convencionales se asocian a mayores niveles de ansiedad académica que las pruebas de elección múltiple (Wahyuni et al., 2021) o las de emparejamiento (Shaha, 1984). También se observan algunos sesgos debidos a factores contextuales o sociodemográficos, como que los hombres se benefician de las pruebas objetivas en relación a las mujeres al tener una menor aversión al riesgo (y, por lo tanto, dejar menos preguntas en blanco) y mostrar una mayor habilidad de adivinación del ítem por la información contextual disponible (Akyol et al., 2022). Esta evidencia se asocia a que las mujeres presentan niveles más elevados de ansiedad académica y menores niveles de autoeficacia en la realización de pruebas objetivas (Singh et al., 2013), mientras que estas pruebas, en su modalidad convencional, benefician a los estudiantes con una mayor seguridad en sí mismos (Stringer et al., 2021). También parece que los estudiantes más capaces (en sentido general) tienden a tener mejores niveles de adivinación por la información contextual, y se benefician de este tipo de pruebas independientemente de su nivel de conocimientos en la materia (Stringer et al., 2021).

Debido a estas limitaciones de las pruebas objetivas convencionales, y teniendo en cuenta sus múltiples beneficios para la evaluación sumativa en contextos saturados de estudiantes, se ha propuesto en la literatura una gran variedad de modalidades de aplicación de las pruebas objetivas, cuya selección permite adaptarse a las necesidades peculiares que pueda tener el contenido, la asignatura, el profesorado o los estudiantes en la prueba de evaluación. La Tabla 1 presenta una enumeración de las principales modalidades planteadas, adaptando la propuesta realizada por Ng y Chan (2009). Como se puede observar, además de la prueba objetiva convencional, de uso generalizado, se han propuesto diferentes modalidades de aplicación que permiten reducir su enfoque memorístico, reduccionista y parcial. Este es el caso por ejemplo de las pruebas objetivas de dos niveles (two-tier multiple-choice), en las que el estudiante debe detectar tanto la opción correcta en el primer ítem (nivel 1 de la pregunta) como la justificación adecuada de porqué esa es la opción correcta (Sıbıç et al., 2020) en el segundo ítem (nivel 2 de la pregunta). A pesar de que esta modalidad ha demostrado su eficacia para desarrollar pensamientos de orden superior (Maulita et al., 2019), su estructura y funcionamiento no contribuye a reducir la ansiedad académica o a mitigar el sesgo de medición por el efecto de adivinación debido a factores contextuales y no contextuales del estudiante.

Tabla 1. Modalidades de aplicación de la prueba objetiva.

Método	Instrucciones	Regla de calificación
Prueba objetiva convencional	Elegir la opción correcta de entre las N opciones	Se obtiene el punto correspondiente al ítem al seleccionar la opción correcta, y cero (o una corrección por azar) al seleccionar la incorrecta.
Prueba objetiva liberal	El sujeto puede elegir más de una opción si tiene dudas sobre la opción correcta	Se obtiene la puntuación máxima al seleccionar la opción correcta únicamente, y una puntuación parcial si se han seleccionado varias opciones (puede haber corrección por azar).
Prueba objetiva inversa	Seleccionar las opciones incorrectas que identifique entre las N opciones	Se obtienen 1/(N-1) puntos por cada opción incorrecta identificada. Puede realizarse corrección por azar (-1 punto) si se selecciona la opción correcta.
Nivel de confianza	El sujeto debe indicar su nivel de confianza sobre la opción elegida como correcta	Se obtiene la puntuación máxima si se selecciona la opción correcta con la máxima confianza, y parcial si la confianza se reduce. La corrección por azar también se puede ponderar en función de la confianza.
Nivel de probabilidad	Indicar la probabilidad de que cada una de las n opciones sea correcta	La puntuación obtenida es igual a la ponderación asignada a la opción correcta.
Orden de preferencia	Las n opciones se ordenan de la que más probablemente sea correcta a la que menos	Se obtiene puntuación máxima si se sitúa la opción correcta en 1ª posición, y una puntuación parcial si se sitúa en posiciones intermedias.
Prueba objetiva con dos estímulos	Se presentan dos estímulos similares. Entre las n opciones, se debe elegir la opción considerada correcta para cada uno de los dos estímulos	Se obtiene el punto correspondiente al ítem solamente cuando se seleccionan ambas opciones correctamente. En cualquier otro caso no puntúa.
Prueba objetiva de dos niveles	Se presentan dos ítems convencionales por pregunta. El primero con el contenido y el segundo para argumentar-razonar la opción seleccionada	Se obtiene la máxima puntuación si se aciertan ambos ítems, y una puntuación parcial si se falla alguno de ellos.

Otra modalidad de prueba objetiva que permite ampliar las posibilidades de la modalidad convencional es de la Prueba Objetiva Inversa (POI), propuesta a lo largo del siglo xx (Arnold y Arnold, 1970; Collet, 1971) aunque muy poco extendida en la literatura científica desde entonces (y con una difusión prácticamente inexistente en España e Iberoamérica). Esta modalidad requiere que el estudiante tenga que reflexionar más en profundidad sobre las opciones de respuesta y permite que éste demuestre su conocimiento parcial, manteniendo la posibilidad de corrección de los efectos de adivinación y mostrando unas propiedades psicométricas incluso superiores a su par convencional (Adair y Jaeger, 2013).

Dado que las POI entienden las pruebas objetivas como pruebas de opción múltiple, permiten que el estudiante tenga que desarrollar un pensamiento y reflexión más amplios, fomentando que se detenga en cada respuesta incorrecta en caso de no conocer la respuesta correcta y permitiéndole demostrar un conocimiento parcial en cada ítem. De este modo, se logra reducir la ansiedad académica y hacia el examen asociada a las pruebas objetivas convencionales (Shaha, 1984; Wahyuni et al., 2021). De hecho, las evidencias recabadas en estudios previos permiten afirmar que las POI incrementan el rendimiento del estudiante, favorecen niveles más elevados

de satisfacción con la prueba de evaluación y reducen la ansiedad académica de los estudiantes hacia la prueba de evaluación y la materia (Bond et al., 2013; Vanderoost et al., 2018; Wu et al., 2018). Todo ello permite limitar el sesgo de género asociado a las pruebas objetivas convencionales (Vanderoost et al., 2018), a la vez que se reduce la aversión al riesgo (Vanderoost et al., 2018) y el efecto de adivinación (Chang *et al*., 2007). Además, Vanderoost et al. (2018) encuentran que los estudiantes tienen preferencia por esta modalidad con respecto a la modalidad convencional.

No obstante, no se ha estudiado la relación entre el desarrollo del pensamiento de orden superior y la modalidad de prueba objetiva, ni se han establecido estudios en el contexto universitario fuera de las áreas de Ciencias y Ciencias de la Salud. Por lo tanto, este estudio trata de confirmar que las ventajas conocidas de esta modalidad son generalizables en el ámbito universitario, promocionando al mismo tiempo niveles elevados en el razonamiento, mayores que con las pruebas objetivas convencionales.

Recomendaciones para el diseño de ítems en las pruebas objetivas inversas

Si bien el empleo de POI parece asociarse a numerosas ventajas, para evitar que el efecto de adivinación tenga un impacto importante y que por tanto se reduzca la fiabilidad de la prueba, es esencial tener en cuenta varias recomendaciones importantes relacionadas con su redacción y formato, sobre todo en lo que a los distractores de los ítems se refiere.

En primer lugar, es importante plantear algunas recomendaciones generales para el diseño de ítems de prueba objetiva (e.g., Gottlieb et al., 2023; Mitra, 2022):

- Planificar la prueba objetiva previamente a su diseño.
- Apoyarse en un equipo de expertos para la redacción de los ítems.
- El enunciado de cada ítem debe plantear un problema definido, único y cerrado, sin necesidad de aportar o conocer información externa al propio enunciado.
- Redacción clara, breve, concisa y precisa, de modo que se evalúe el conocimiento y comprensión de la materia, no las habilidades de los estudiantes con mayor capacidad de adivinación por el contexto. Evitar enunciados complejos, enunciados en sentido negativo, dobles negaciones, información accesoria en el enunciado, opciones de respuesta como 'ninguno de los anteriores' o 'todos los anteriores', o vocabulario excesivamente culto o en desuso (a no ser que este sea vocabulario específico del ámbito profesional de la materia).
- Valorar cuál es el número de opciones de respuesta ideal y redactarlas de modo que todas ellas sean posibles a nivel conceptual, y coherentes con el enunciado plantead, evitando el uso de partículas determinísticas o extremas como 'siempre' o 'nunca'.
- Estudiar las evidencias de validez y fiabilidad de los ítems tras sus primeras aplicaciones: nivel de dificultad variable en los ítems, escalas fiables y consistentes, niveles aceptables de discriminación ítem-total, distractores de calidad, etc.

Se establecen también algunas cuestiones específicas importantes en el caso de las POI y la promoción del pensamiento de orden superior (Scully, 2017; Vanderoost et al., 2018):

- En las POI es especialmente importante que todas las opciones de respuesta sean plausibles. En función del nivel de conocimientos de la persona, ésta será capaz de identificar más o menos distractores como no plausibles. Así, podemos hablar de distractores de mayor y de menor demanda de conocimientos en la materia, siendo importante el planteamiento de distractores con niveles variables de demanda de conocimientos. De este modo es posible alcanzar mayores niveles de fiabilidad que con las pruebas objetivas convencionales (Adair y Jaeger, 2013).
- Para facilitar la evaluación del conocimiento parcial del estudiante y el desarrollo del pensamiento de orden superior, el número de opciones de respuesta en una POI debería ser, como mínimo, de cuatro.
- Inversión de los ítems (*Item Flipping*): Presentar en el enunciado el concepto o fenómeno y las posibles definiciones o ejemplos en las opciones de respuesta se asocia a un nivel bajo de demanda cognitiva. Invertir este formato, presentando un ejemplo de aplicación en el enunciado y los posibles conceptos en los que encaja ese ejemplo en las alternativas de respuesta, se asocia a niveles más elevados de demanda cognitiva, pasando del nivel *conocimiento* de la taxonomía de Bloom al nivel *comprensión*. La Figura 1 presenta cómo se ha empleado esta técnica en la prueba objetiva propuesta en este estudio. El estudiante debe comprender el ejemplo presentado en el enunciado y relacionarlo con un concepto estudiado en la materia.

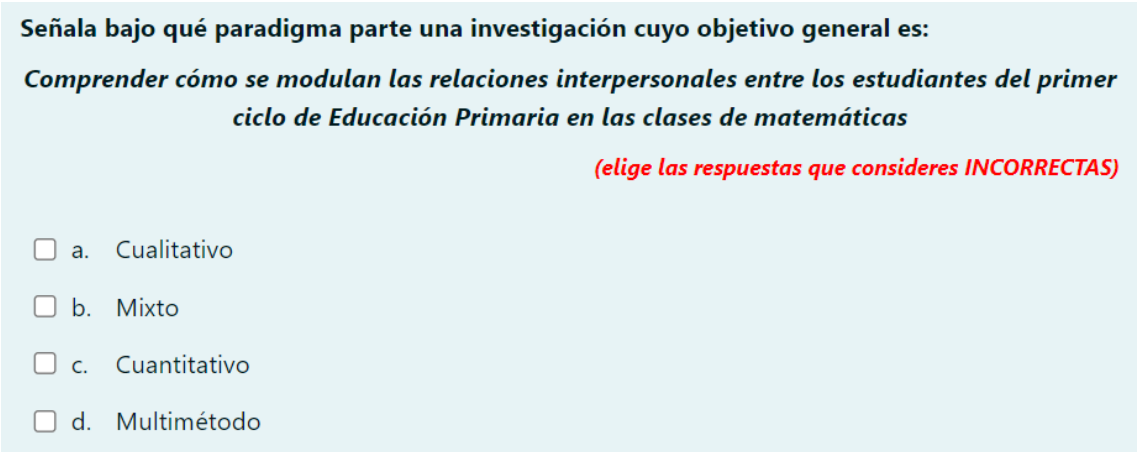

Figura 1. Ejemplo de ítem invertido.

- Planteamiento de enunciados activadores de *múltiples neuronas*. El enunciado debe promover que el estudiante tenga que relacionar adecuadamente diversos conceptos estudiados en la materia. Se debe ser cuidadoso con la aplicación de esta técnica para no complicar en exceso la estructura del enunciado. La Figura 2 presenta un ejemplo de aplicación de esta técnica. Antes de poder responder, el estudiante debe entender qué es una variable dependiente y, a partir de aquí, que el gráfico presenta los resultados de un diseño pretest-postest con grupo control. Además, debe saber interpretar un gráfico de líneas conjuntas en el marco de este tipo de diseños.

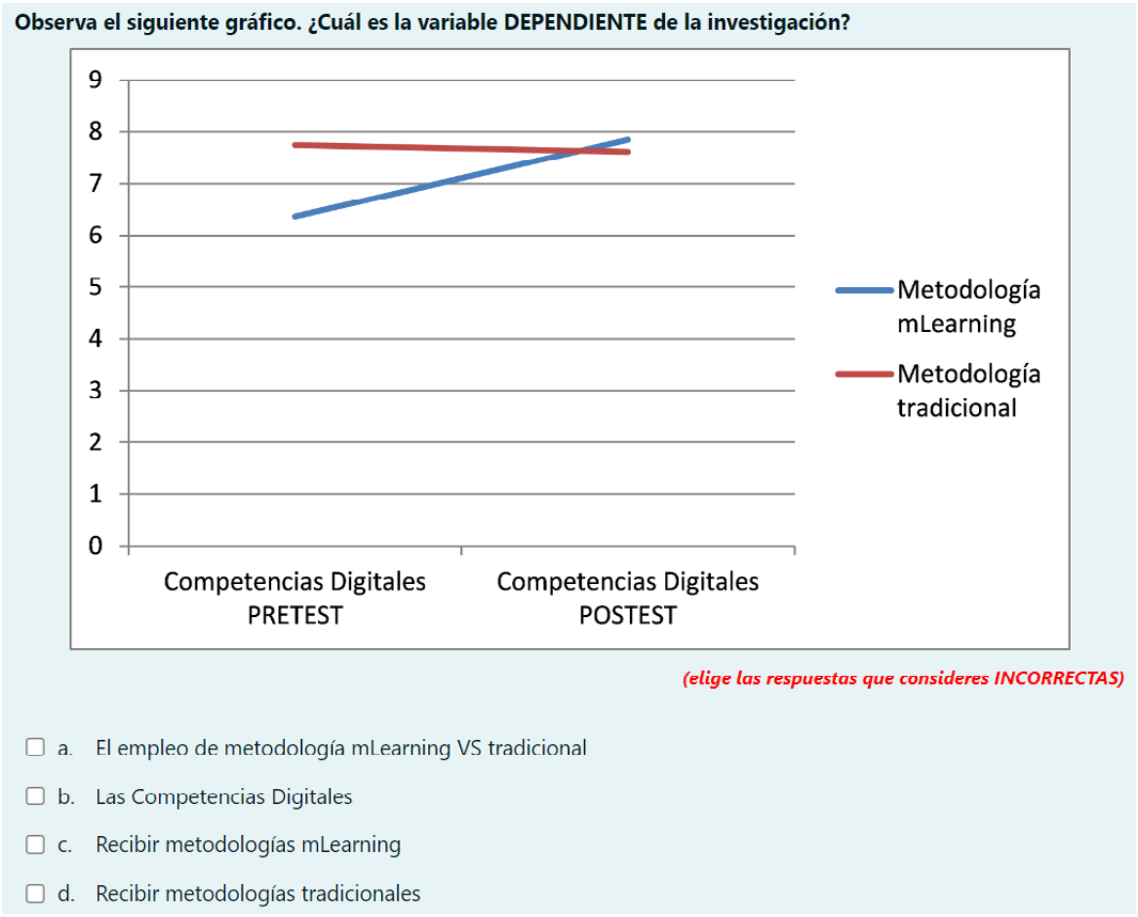

Figura 2. Ejemplo de ítem activador de múltiples neuronas.

Método

El objetivo de este trabajo es analizar la eficacia de la POI para la evaluación sumativa en educación superior, comprobando si su empleo aumenta el rendimiento académico y la satisfacción hacia la propia prueba objetiva, promociona el pensamiento de orden superior, y reduce los niveles de ansiedad académica de los estudiantes.

A partir de la revisión de la literatura, se plantean las siguientes hipótesis:

- H1. El rendimiento académico del grupo experimental es superior al del grupo control.
- H2. El pensamiento de orden superior es más elevado en el grupo experimental.
- H3. Los niveles de satisfacción y valoración del examen son más altos en el grupo experimental.
- H13. Los estudiantes que realizan el examen en modalidad inversa presentan niveles más bajos de ansiedad académica.

Diseño de la investigación

Se aplicó un diseño cuasiexperimental solo postest con grupo control no equivalente (Campbell y Stanley, 1963). Así, en este estudio de carácter cuantitativo se definieron dos grupos de investigación:

- Grupo experimental: Grupo al que se le aplica la modalidad de prueba objetiva inversa, considerada como modalidad innovadora (tratamiento).
- Grupo control: Grupo que recibe la modalidad de prueba objetiva convencional, considerada como modalidad tradicional (placebo).

Participantes

A partir de la población de estudiantes universitarios de grado en titulaciones de Ciencias de la Educación, se obtuvo una muestra no probabilística incidental de n=204 estudiantes de diversos Grados (Grado en Educación Infantil, Grado en Educación Primaria y Grado en Pedagogía) de una universidad pública española. Todos estos Grados comparten una asignatura básica, *Métodos de Investigación en Educación*, incluida en el 1° curso (2° cuatrimestre), y con una guía académica compartida. El estudio se enmarcó en esta asignatura.

Participaron en la muestra todos los estudiantes de primera matrícula que cursaron durante el año 2022 la asignatura y se presentaron al examen teórico (prueba objetiva) de la misma en la 1ª convocatoria.

Variables

En primer lugar, se incluyeron algunas variables de control para comprobar la homogeneidad del Grupo Control y Experimental:

- Sexo.
- Nota de acceso a la universidad.
- Conocimientos previos: Puntuación obtenida en las pruebas de autoevaluación de conocimientos completadas durante el curso.

Por otro lado, las variables dependientes incluidas en el estudio fueron las siguientes:

- Rendimiento académico: Puntuación obtenida por los estudiantes del grupo control y experimental en la prueba objetiva, aplicando la corrección del efecto del azar correspondiente a la modalidad aplicada. Cabe destacar que los ítems presentados a los estudiantes del grupo control y grupo experimental de cada grado fueron los mismos, con la única diferencia de la modalidad de presentación y respuesta de los mismos (modalidad inversa VS modalidad convencional).
- Escala pensamiento de orden superior: Nivel de profundidad en el razonamiento desarrollado en el examen, teniendo en cuenta los 6 niveles de la taxonomía de Bloom. Se aplicó una escala validada adaptada de estudios previos (Olmos Migueláñez et al., 2014).

- Nivel de satisfacción con la actividad examen y valoración global de la modalidad de examen realizada. Se emplearon sendas escalas validadas adaptadas de estudios previos (Martínez Abad y Hernández Ramos, 2017).
- Ansiedad académica ante el examen: Teniendo en cuenta la vinculación de esta asignatura con la estadística, se adaptó una escala validada en estudios previos de ansiedad hacia la estadística (Vigil-Colet et al., 2008).

Los ítems de las 4 escalas, todas de carácter unidimensional, se pueden consultar en la tabla anexa A1.[1] Todos los ítems incluyeron una escala de respuesta de 0 a 10 niveles, correspondiendo el 0 con el menor nivel de acuerdo y el 10 con el mayor. La puntuación factorial de cada factor fue obtenida a partir del promedio de las respuestas a los ítems. En el caso de la escala de pensamiento de orden superior, se obtuvo el promedio de los niveles 2, 3 y 4 de la taxonomía de Bloom (ítems 2-4 de la escala), referidos a la comprensión, aplicación y análisis. Atendiendo a Scully (2017), se asumió el nivel *conocimiento* un nivel de razonamiento bajo, y los niveles *síntesis* y *evaluación-juicio* como niveles que no es posible alcanzar con el empleo de pruebas objetivas.

Procedimiento y análisis de datos

Durante la sesión inicial de la asignatura los estudiantes fueron informados de las dos modalidades de prueba objetiva (POI o convencional) disponibles, aclarando que ambas modalidades constarían de los mismos ítems. Finalmente, se les indicó que deberían seleccionar una de las dos opciones de cara a la realización del examen. Durante el curso, con la intención de que pudieran practicar con la modalidad POI en comparación con las pruebas objetivas convencionales, los estudiantes dispusieron de dos pruebas diferentes de autoevaluación de conocimientos,[2] ambas disponibles en las dos modalidades. Tras la práctica con las pruebas de autoevaluación se pidió a los estudiantes que eligieran la modalidad de examen deseada, quedando configurado el grupo control (modalidad convencional) por n=51 estudiantes y el grupo experimental (modalidad inversa) por n=153 estudiantes.

En cuanto al análisis de datos, se aplicó estadística descriptiva, correlacional e inferencial para dar respuesta a las hipótesis planteadas. El nivel de significación se estableció en el 5%, y se aplicaron contrastes de hipótesis paramétricos al no observarse desviaciones importantes de la normalidad y homocedasticidad. Igualmente, se incluyeron los estadísticos de tamaño del efecto apropiados en cada caso.

[1] https://doi.org/10.5281/zenodo.8220581

[2] La asignatura en la que se aplica este estudio consta de dos bloques de contenido separados y bien definidos (1. Metodología de Investigación; 2. Análisis Estadístico de Datos), por lo que se incluyó una prueba de autoevaluación independiente por cada uno de los dos bloques.

Resultados

Homogeneidad de grupo control y grupo experimental

En primer lugar, cabe destacar que el 75% de los estudiantes, tras la práctica de ambas modalidades de prueba objetiva, decidió realizar en el examen la prueba inversa (Figura 3), no resultando los grupos homogéneos en cuanto a su tamaño. Este resultado da muestras de la preferencia de los estudiantes por este tipo de modalidades de respuesta.

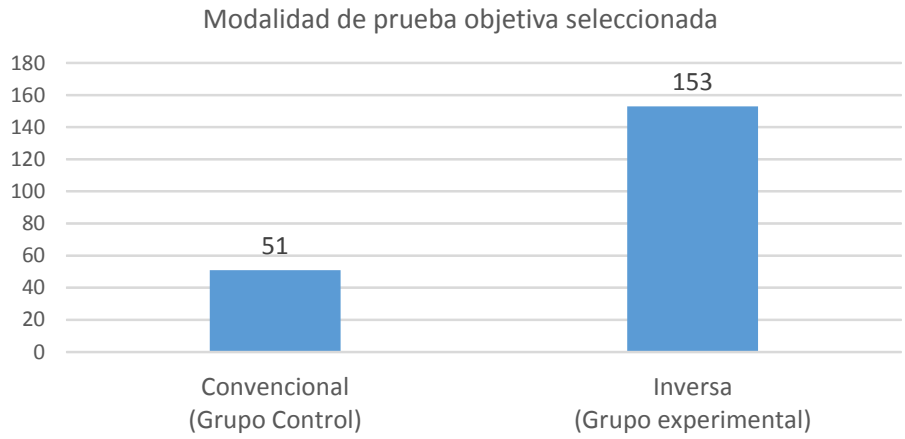

Figura 3. Modalidad de prueba objetiva seleccionada.

La Tabla 2 muestra la homogeneidad de ambos grupos en función de su nivel de conocimientos previos. Los estadísticos descriptivos e inferenciales muestran que, mientras que el grupo experimental accedió al Grado correspondiente con una nota de acceso significativamente superior (con un tamaño del efecto de nivel alto), los nivel de conocimientos previos en la materia demostrados en la prueba de autoevaluación fueron iguales (con un tamaño del efecto bajo), incluso ligeramente superiores en la muestra de estudiantes del grupo control.

Tabla 2. Homogeneidad de Grupos Control y Experimental por conocimientos previos.

Variable	Grupo	Media	D.T.	t	p.	d
Nota de acceso	Control	9.48	2.01	4.68	<.001	.838
	Experimental	10.79	1.39			
Autoevaluación	Control	6.30	2.12	-1.54	.126	.277
	Experimental	5.75	1.95			

En cuanto a la distribución de hombres y mujeres en ambos grupos, la Figura 4 indica que el % de mujeres de la muestra que eligieron la modalidad POI con respecto a la convencional (n_{POI}=138; $n_{conv.}$=41) fue ligeramente superior al de los hombres (n_{POI} =15; $n_{conv.}$=10). No obstante, estas diferencias no resultan significativas en el contraste Chi-Cuadrado aplicado (c^2=3.42; p=.064), con un tamaño del efecto de nivel bajo. Por tanto, se puede afirmar que ambos grupos resultan homogéneos en cuanto a sus niveles de conocimientos previos sobre la materia y en cuanto a la distribución por género en ambos grupos.

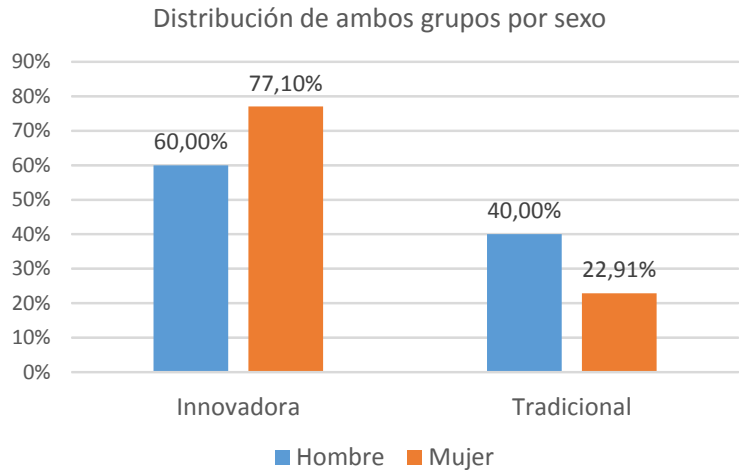

Figura 4. Distribución del sexo en grupo experimental y grupo control.

Análisis descriptivo e inferencial bivariado

La Tabla 3 presenta las diferencias en la puntuación obtenida en el examen en función de la modalidad de prueba objetiva aplicada. Se observan diferencias significativas con tamaños del efecto medios favorables al grupo que completa el examen POI.

Tabla 3. Efecto del tratamiento sobre el rendimiento académico.

Variable	Grupo	Media	D.T.	t	p.	d
Rendimiento académico	Control	5.12	1.94	2.85	.005	.460
	Experimental	6.00	1.91			

La Figura 5 confirma cómo la distribución general de puntuaciones del grupo experimental está en niveles sensiblemente superiores al grupo control.

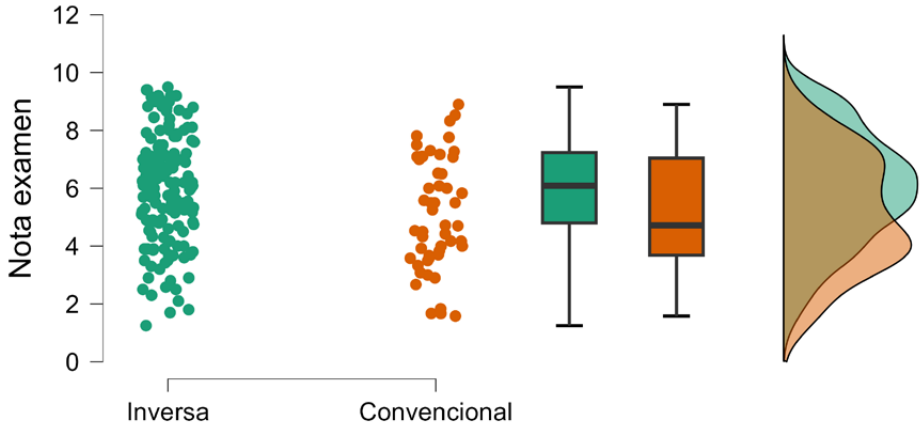

Figura 5. Distribución del rendimiento académico por modalidad de examen.

En cuanto a los efectos de la aplicación de la modalidad inversa sobre el resto de factores, la Tabla 4 muestra valores promedio elevados en ambos grupos, principalmente en la escala de pensamiento de orden superior. Por su parte, mientras que los niveles de razonamiento elevado en la muestra que completa el examen en modalidad inversa

son más elevados, el nivel de significación indica efectos probablemente significativos (inferiores a .1) con tamaños del efecto bajos. Además, tanto la satisfacción con el examen como la valoración global del mismo son significativamente más altos en el grupo experimental, con tamaños del efecto medio y medio-alto respectivamente. Al respecto, se puede destacar también que los niveles de satisfacción con la prueba de las mujeres del grupo experimental (media=6.78), con respecto a los hombres (media=6.12) resultaron significativamente superiores con tamaños del efecto medios (t=2.14; p.=.03; d=.583). Finalmente, los niveles de ansiedad académica parecen iguales en ambos grupos.

Tabla 4. Efecto causal del tratamiento sobre los factores dependientes.

Variable	Grupo	Media	D.T.	t	p.	d
Pensamiento de orden superior	Control	7.24	1.33	1.70	.091	.279
	Experimental	7.57	1.11			
Satisfacción examen	Control	6.17	1.52	2.74	.007	.445
	Experimental	6.73	1.17			
Valoración examen	Control	6.67	1.84	4.05	<.001	.665
	Experimental	7.65	1.31			
Ansiedad académica	Control	5.51	2.03	0.52	.604	.085
	Experimental	5.67	1.91			

Así, mientras que se verifican las hipótesis de investigación 1 y 3, la hipótesis 2 se valida parcialmente y la 4 no se cumple.

Análisis correlacional por grupo

La Tabla 5 muestra la correlación entre el rendimiento académico y el resto de factores en cada grupo. En primer lugar, los estudiantes con mejor calificación perciben mayores niveles de pensamiento de orden superior durante el examen, aunque únicamente en el grupo experimental, con efectos medios-bajos. Este mismo efecto se observa en los niveles de ansiedad académica, que se reducen a nivel general en las personas con calificaciones más altas, aunque de manera más clara en el grupo experimental. Por otro lado, la satisfacción con el examen es más elevada en los estudiantes con mayor rendimiento en ambos grupos. Por último, la valoración global del examen realizado no depende de la puntuación obtenida.

Tabla 5. Correlación entre el rendimiento académico y los factores dependientes.

	Grupo control		Grupo experimental	
	R_{xy}	p.	R_{xy}	p.
Pensamiento de orden superior	.052	.720	.181	.034
Satisfacción examen	.241	.088	.236	.004
Valoración examen	-.020	.888	-.018	.836
Ansiedad académica	-.158	.269	-.240	.005

Estos resultados muestran matizan los resultados obtenidos anteriormente en torno a las hipótesis 2 y 4:

- En cuanto a la H2, se observa que la modalidad de prueba objetiva media en la relación entre el pensamiento de orden superior desarrollado en el examen y el rendimiento: esta interacción evidencia que los niveles de pensamiento de orden superior puestos en juego en los estudiantes de modalidad inversa de rendimientos elevados (grupo experimental) son más superiores a los de los estudiantes de rendimiento inferior. Este efecto no se observa en los estudiantes de modalidad convencional.
- En el caso de la H4 se observan efectos similares. Los estudiantes de modalidad inversa con buenas calificaciones muestran niveles de ansiedad académica claramente inferiores a aquellos que obtienen peores calificaciones. Aunque esta relación también se observa en el grupo control, es de magnitud menor.

Discusión y conclusiones

A nivel general, teniendo en cuenta los resultados obtenidos podemos afirmar que la POI tiene ventajas importantes sobre la modalidad convencional, demostrando su eficacia.

En primer lugar, en cuanto a la primera hipótesis planteada, referida al rendimiento académico, los estudiantes que aplicaron la prueba objetiva de modalidad inversa obtuvieron resultados significativamente superiores al grupo control, de modalidad convencional. Este resultado, congruente con estudios previos (Bond et al., 2013; Vanderoost et al., 2018), está relacionado con que la POI permite al estudiante que demuestre un conocimiento parcial, sin perder las propiedades psicométricas de la prueba (Adair y Jaeger, 2013). En este sentido, debemos tener en cuenta que la calificación del estudiante en una prueba objetiva tiene relación directa con la estrategia que aplica a la hora de responderla (Almalki, 2023), por lo que es esencial dar instrucciones claras y precisas sobre las alternativas que tiene el estudiante a la hora de responder a la prueba y la corrección por azar aplicada. Esta cuestión es crítica en el caso de la modalidad inversa, ya que se trata de una modalidad de elección múltiple (Wahyuni et al., 2021), con numerosas posibilidades de respuesta por ítem.

La H2 se refería a los niveles de pensamiento de orden superior desarrollados con las pruebas inversas. Siguiendo las recomendaciones de Scully (2017), en este estudio se diseñaron ítems de alta demanda cognitiva, lo que dio lugar a valores promedio elevados en el desarrollo de los niveles comprensión, aplicación y análisis de la taxonomía de Bloom. En esta línea, una de las principales aportaciones de este estudio es que el empleo de POI podría estar relacionado con el desarrollo de razonamientos de orden superior por parte de los estudiantes. Los valores promedio de razonamiento resultaron superiores en el grupo experimental que en el control, y el contraste de hipótesis obtuvo resultados cercanos al nivel de significación establecido. Teniendo en cuenta que Stringer et al. (2021) indican que los estudiantes con mayor nivel académico y mayor seguridad en sí mismos tienen a emplear un pensamiento de orden inferior y a buscar patrones en las pruebas objetivas convencionales, es relevante el hallazgo aquí obtenido en relación a los niveles de razonamiento percibidos en función del rendimiento: los estudiantes que realizan las POI con buenas calificaciones perciben niveles de razonamiento más

elevados que los que tienen peores notas. Estos resultados sugieren que el empleo de la modalidad de respuesta inversa refuerza el empleo del pensamiento de orden superior por parte de los estudiantes, y que precisamente los estudiantes que son capaces de poner en marcha este tipo de razonamientos elevados son los que tienden a alcanzar mejores calificaciones. Los resultados parecen confirmar que existe una asociación significativa entre la demanda cognitiva del ítem y la calificación del estudiante, confirmando la relación observada por Haataja et al. (2023), y además añaden que esta relación es más intensa en las pruebas inversas. Por tanto, este estudio confirma que la modalidad de respuesta de la prueba objetiva de hecho influye sobre la demanda cognitiva del ítem, siendo recomendable el empleo de POI si se pretende asociar el nivel de razonamiento cognitivo alcanzado por el estudiante con la calificación obtenida en la prueba.

En relación a la H3, tal y como cabía esperar en base a estudios previos (Bond et al., 2013; Vanderoost et al., 2018), se observan mayores niveles de satisfacción y de valoración global de la prueba objetiva cuando se aplica en modalidad inversa. Los resultados obtenidos en este estudio añaden que, mientras que los niveles de satisfacción generales con la prueba son medios-altos, las mujeres se muestran significativamente más satisfechas con la POI que los hombres. Esta evidencia sugiere que el uso de POI puede ayudar a limitar el sesgo de género asociado a las pruebas objetivas convencionales (Vanderoost et al., 2018), ya que mejora los niveles de confianza de las mujeres, reduciendo su aversión al riesgo.

Los resultados más sorprendentes teniendo en cuenta los antecedentes (Bond et al., 2013; Núñez-Peña y Bono, 2021; Wahyuni et al., 2021) son los obtenidos en relación a la H4. El presente estudio observa niveles de ansiedad académica hacia el examen similares en ambos grupos, independientemente de la modalidad de prueba objetiva aplicada. Además, la asociación entre la ansiedad y el rendimiento es inversa en ambos grupos pero de mayor intensidad en el grupo experimental. Los resultados sugieren que la calificación obtenida por los estudiantes está en el caso de la prueba inversa más asociada a los niveles de ansiedad. Esta discrepancia con los estudios previos puede estar debida a que la variable incluida en este estudio fue la ansiedad ante el examen, no la ansiedad matemática o científica.

Teniendo en cuenta que las pruebas objetivas seguirán siendo una herramienta esencial para la evaluación sumativa en la enseñanza universitaria, podemos concluir que la manera de redactar y estructurar el ítem, junto con la modalidad de respuesta facilitada, son cuestiones que influyen de manera determinante tanto en la calificación obtenida por el estudiante como en sus niveles de satisfacción y las estrategias aplicadas en su resolución. Así, resulta fundamental que el profesorado universitario tenga en cuenta estas cuestiones, y que las planifique cuidadosamente durante el proceso de diseño y desarrollo de sus pruebas objetivas. En este sentido, la POI ha demostrado ser una herramienta eficaz que permite mantener las amplias ventajas de las pruebas objetivas para la evaluación sumativa (objetividad, simplicidad, eficiencia, replicabilidad, adaptabilidad o facilidad de aleatorización y digitalización), poniendo solución a algunas de las desventajas más notables, fundamentalmente posibilitar el conocimiento parcial del estudiante y fomentar el pensamiento de orden superior. Estas ventajas, unidas a la posibilidad de digitalización y automatización que ofrecen las pruebas objetivas, dan cuenta del gran potencial que atesoran las POI en la evaluación sumativa de los estudiantes universitarios.

Llama la atención el limitado uso docente de modalidades alternativas de prueba objetiva a la convencional teniendo en cuenta los numerosos trabajos que evidencian sus ventajas. En un contexto como el universitario en el que la innovación docente tiene cada vez una mayor relevancia, resulta fundamental dar a conocer este tipo de técnicas, reduciendo las resistencias al cambio y promoviendo el desarrollo de procesos de evaluación sumativa objetivos pero de calidad. A pesar de estar denostadas en algunos foros educativos, evidencias como las aquí presentadas muestran que las pruebas objetivas bien empleadas atesoran un gran potencial para la evaluación educativa manteniendo estándares de eficiencia elevados.

A pesar de las claras evidencias obtenidas, se observan algunas limitaciones importantes en este estudio que deberían ser abordadas en futuras investigaciones. La limitación fundamental tiene que ver con la muestra obtenida, de carácter incidental y restringida al ámbito de las Ciencias de la Educación. Se obtiene una muestra de estudiantes de grado de una sola universidad pública española. Si bien el tamaño de la muestra es importante, su restricción geográfica y temática supone un sesgo importante en la generalización de los resultados a la población de estudiantes universitarios. Es importante en futuros estudios ampliar la muestra a otros ámbitos de conocimiento, regiones, materias y niveles universitarios, de modo que sea posible confirmar si las tendencias aquí observadas siguen estables. Por otro lado, la definición en el diseño de grupos de investigación no aleatorizados, en los que los estudiantes deciden formar parte del grupo control o experimental, puede estar asociado a sesgos de comparación entre ambos grupos que afecten a las variables dependientes. Consideramos que en futuros estudios es de suma importancia tratar de controlar mejor estos sesgos, tomando medidas de control más fuertes para asegurar la equivalencia de ambos grupos de investigación, e incluso incorporando medidas pretest que confirmen los efectos causales aquí apuntados.

Dadas sus características, las pruebas objetivas son y seguirán siendo herramientas de uso generalizado para evaluar el dominio de los estudiantes universitarios sobre una materia. Por tanto, la mejora en los procesos de diseño e implementación de las mismas son fundamentales para mejorar la calidad de la enseñanza universitaria. Este estudio aporta evidencias empíricas en esa línea, presentando una alternativa a las pruebas objetivas convencionales viable y que promueve la calidad educativa.

Referencias bibliográficas

Adair, D. y Jaeger, M. (2013). A probabilistic scoring method in multiple-choice testing incorporating partial knowledge. En R. White (Ed.), *Curriculum Development, Innovation and Reform* (pp. 109-124). Nova Science Pub. Inc.

Akyol, P., Key, J. y Krishna, K. (2022). Hit or miss? Test taking behavior in multiple choice exams. *Annals of Economics and Statistics*, *147*, 3-50. https://doi.org/10.2307/48684785

Almalki, M.S. (2023). Multiple Choice Test-Taking Strategies, Test Anxiety, and EFL Students" Achievement. *World Journal of English Language*, *13*(2), 248-259. https://doi.org/10.5430/wjel.v13n2p248

Arnold, J.C. y Arnold, P.L. (1970). On Scoring Multiple Choice Exams Allowing for Partial Knowledge. *The Journal of Experimental Education*, *39*(1), 8-13. https://doi.org/10.1080/00220973.1970.11011223

Bond, A.E., Bodger, O., Skibinski, D.O.F., Jones, D.H., Restall, C.J., Dudley, E. y Van Keulen, G. (2013). Negatively-Marked MCQ Assessments That Reward Partial Knowledge Do Not Introduce Gender Bias Yet Increase Student Performance and Satisfaction and Reduce Anxiety. *PLoS ONE*, *8*(2). https://doi.org/10.1371/journal.pone.0055956

Campbell, D.T. y Stanley, J. (1963). *Experimental and quasi-experimental designs for research*. Wadsworth Publishing.

Chang, S.-H., Lin, P.-C. y Lin, Z.-C. (2007). Measures of partial knowledge and unexpected responses in multiple-choice tests. *Educational Technology and Society*, *10*(4), 95-109.

Collet, L.S. (1971). Elimination Scoring: An Empirical Evaluation. *Journal of Educational Measurement*, *8*(3), 209-214. https://doi.org/10.1111/j.1745-3984.1971.tb00927.x

Gottlieb, M., Bailitz, J., Fix, M., Shappell, E. y Wagner, M.J. (2023). Educator's blueprint: A how-to guide for developing high-quality multiple-choice questions. *AEM Education and Training*, *7*(1). https://doi.org/10.1002/aet2.10836

Greving, S. y Richter, T. (2022). Practicing retrieval in university teaching: Short-answer questions are beneficial, whereas multiple-choice questions are not. *Journal of Cognitive Psychology*, *34*(5), 657-674. https://doi.org/10.1080/20445911.2022.2085281

Haataja, E.S.H., Tolvanen, A., Vilppu, H., Kallio, M., Peltonen, J. y Metsäpelto, R.-L. (2023). Measuring higher-order cognitive skills with multiple choice questions –potentials and pitfalls of Finnish teacher education entrance. *Teaching and Teacher Education*, *122*. https://doi.org/10.1016/j.tate.2022.103943

Kissi, P., Baidoo-Anu, D., Anane, E. y Annan-Brew, R.K. (2023). Teachers' test construction competencies in examination-oriented educational system: Exploring teachers' multiple-choice test construction competence. *Frontiers in Education*, *8*. https://doi.org/10.3389/feduc.2023.1154592

Martínez Abad, F. y Hernández Ramos, J.P. (2017). Flipped Classroom con píldoras audiovisuales en prácticas de análisis de datos para la docencia universitaria: Percepción de los estudiantes sobre su eficacia. En S. Pérez Aldeguer, G. Castellano Pérez y A. Pina Calafi (Eds.), *Propuestas de Innovación Educativa en la Sociedad de la Información* (pp. 92-105). Adaya Press. https://doi.org/10.58909/ad17267178

Maulita, S.R., Sukarmin y Marzuki, A. (2019). The Content Validity: Two-Tier Multiple Choices Instrument to Measure Higher-Order Thinking Skills. *Journal of Physics: Conference Series*, *1155*(1), 012042. https://doi.org/10.1088/1742-6596/1155/1/012042

Mitra, A.K. (2022). The Art of Designing a Quality Multiple Choice Question in Chemistry. *Resonance*, *27*(6), 1017-1031. https://doi.org/10.1007/s12045-022-1394-2

Ng, A.W. Y., y Chan, A.H.S. (2009). *The testing methods and gender differences in multiple-choice assessment*. *1174*, 236-243. https://doi.org/10.1063/1.3256252

Núñez-Peña, M.I. y Bono, R. (2021). Math anxiety and perfectionistic concerns in multiple-choice assessment. *Assessment and Evaluation in Higher Education*, *46*(6), 865-878. https://doi.org/10.1080/02602938.2020.1836120

Olmos Migueláñez, S., Martínez Abad, F., Torrecilla Sánchez, E.M. y Mena Marcos, J.J. (2014). Análisis psicométrico de una escala de percepción sobre la utilidad de Moodle en la universidad. *RELIEVE - Revista Electrónica de Investigación y Evaluación Educativa*, *20*(2), preprint 1. https://doi.org/10.7203/relieve.20.2.4221

Polat, M. (2020). Analysis of Multiple-Choice versus Open-Ended Questions in Language Tests According to Different Cognitive Domain Levels. *Novitas-ROYAL (Research on Youth and Language)*, *14*(2), 76-96.

Scully, D. (2017). Constructing multiple-choice items to measure higher-order thinking. *Practical Assessment, Research and Evaluation*, *22*(4), 1-13.

Shaha, S.H. (1984). Matching-Tests: Reduced Anxiety and Increased Test Effectiveness. *Educational and Psychological Measurement*, *44*(4), 869-881. https://doi.org/10.1177/0013164484444009

Singh, A., Bhadauria, V., Jain, A. y Gurung, A. (2013). Role of gender, self-efficacy, anxiety and testing formats in learning spreadsheets. *Computers in Human Behavior*, *29*(3), 739-746. https://doi.org/10.1016/j.chb.2012.11.009

Sıbıç, O., Akçay, B. y Arik, M. (2020). Review of Two-tier Tests in the Studies: Creating a New Pathway for Development of Two-tier Tests. *International Journal of Contemporary Educational Research*, *7*(2), Article 2. https://doi.org/10.33200/ijcer.747981

Stringer, J.K., Santen, S.A., Lee, E., Rawls, M., Bailey, J., Richards, A., Perera, R.A. y Biskobing, D. (2021). Examining Bloom's Taxonomy in Multiple Choice Questions: Students' Approach to Questions. *Medical Science Educator*, *31*(4), 1311-1317. https://doi.org/10.1007/s40670-021-01305-y

Tractenberg, R.E., Gushta, M.M., Mulroney, S.E. y Weissinger, P.A. (2013). Multiple choice questions can be designed or revised to challenge learners' critical thinking. *Advances in Health Sciences Education*, *18*(5), 945-961. https://doi.org/10.1007/s10459-012-9434-4

Vanderoost, J., Janssen, R., Eggermont, J., Callens, R. y De Laet, T. (2018). Elimination testing with adapted scoring reduces guessing and anxiety in multiplechoice assessments, but does not increase grade average in comparison with negative marking. *PLoS ONE*, *13*(10). https://doi.org/10.1371/journal.pone.0203931

Vigil-Colet, A., Lorenzo-Seva, U. y Condon, L. (2008). Development and validation of the Statistical Anxiety Scale. *Psicothema*, *20*(1), 174-180. https://doi.org/10.1037/t62688-000

Wahyuni, L.D., Citraini, R., Hutomo, B.A. y Rakhman, G.G.F. (2021). Anxiety and Test Form: The Differences of Test Anxiety Levels in Terms of Test Form. *Jurnal Pengukuran Psikologi Dan Pendidikan Indonesia*, *10*(2), 184-192. https://doi.org/10.15408/jp3i.v10i2.17974

Wu, Q., De Laet, T. y Janssen, R. (2018). Elimination Scoring Versus Correction for Guessing: A Simulation Study. En M. Wiberg, S. Culpepper, R. Janssen, J. González y D. Molenaar (Eds.), *Quantitative Psychology* (pp. 183-193). Springer International Publishing. http://dx.doi.org/10.1007/978-3-319-77249-3_16