Congreso In-Red 2024
UPV, 11 - 12 de julio de 2024
Doi: https://doi.org/10.4995/INRED2024.2024.17847

Comportamiento del tipo de pregunta/ítem en un cuestionario Moodle: un análisis estadístico-psicométrico de los resultados obtenidos en asignaturas de Química de Educación Superior

Question types behaviour in Moodle quizzes: analysis of statistical-psychometric data obtained in Chemistry subjects of Higher Education

Isabel López-Tocón[a], Fuensanta Sánchez-Rojas[b] y Catalina Bosch-Ojeda[c]

[a]Departamento de Química Física, Facultad de Ciencias, Universidad de Málaga, tocon@uma.es, , [b]Departamento de Química Analítica, Facultad de Ciencias, Universidad de Málaga, fsanchezr@uma.es, y [c]Departamento de Química Analítica, Facultad de Ciencias, Universidad de Málaga, cbosch@uma.es, .

How to cite: López-Tocón, I.; Sánchez-Rojas, F. y Bosch-Ojeda, C. (2024). Comportamiento del tipo de pregunta/ítem en un cuestionario Moodle: un análisis estadístico-psicométrico de los resultados obtenidos en asignaturas de Química de Educación Superior. En libro de actas: *X Congreso de Innovación Educativa y Docencia en Red.* Valencia, 11 - 12 de julio de 2024. Doi: https://doi.org/10.4995/INRED2024.2024.17847

Abstract

Four types of items in Moodle quizzes, such as: true/false, numerical, matching and multiple-choice, were investigated by analyzing the statistical-psychometric data calculated on the basis of a classical theory of tests. To this end, four quizzes were performed with the same type of item, in addition to a fifth quiz in a random mode where the entire set of items is integrated. The experience was carried out in two different populations of students as an integrated part of the blended-learning methodology of the subjects of General Physical Chemistry in the Chemistry Degree and Analytical Chemistry in the Chemical Engineering Degree of the University of Malaga, during the academic year 2022-2023. Several statistical and psychometric parameters were analyzed such as, employed time, average grade, facility index, discrimination index and internal consistency coefficient, among others. In conclusion, the multiple-choice and matching items show the highest quality to be used in online knowledge tests as a virtual exam, while the true/false and numerical items would be the least suitable for an evaluative test of the subject. However, the combination of all items in a random quiz also has an evaluative nature, with a quality similar to that of multiple-choice items

Keywords: *Moodle quizzes; Online Assessment; Statistical-Psychometric analysis*

Resumen

Se ha investigado el comportamiento de cuatro tipos de ítems en un cuestionario Moodle como son: verdadero/falso, numérica, emparejamiento y multi-respuesta mediante el análisis estadístico-psicométrico de los resultados, utilizando la teoría clásica de los tests. Para ello, se han realizado cuatro cuestionarios con el mismo tipo de ítem además, de un quinto cuestionario en modalidad aleatorio donde se integra todo el conjunto de ítems. La

experiencia se ha realizado en dos poblaciones de estudiantes diferentes como parte integrada en la metodología blended-learning de las asignaturas de Química Física General en el Grado de Química y de Química Analítica en el Grado de Ingeniería Química de la Universidad de Málaga, durante el curso 2022-2023. Los principales parámetros estadísticos y psicométricos que se han analizado son: el tiempo de realización, la nota media, el índice de facilidad, el índice de discriminación y el coeficiente de consistencia interna. Los resultados muestran que los ítems multi-respuesta y emparejamiento son los que presentan una mayor calidad como para ser utilizados en pruebas de conocimiento online a modo de examen virtual mientras que, los ítems verdadero/falso y numérico serían los menos indicados. No obstante, la combinación de todos estos tipos de ítems en un cuestionario aleatorio presenta también carácter evaluativo

Palabras clave: *Cuestionarios Moodle; Tipos de ítems; Análisis estadístico-psicométrico*

1. Introducción

La evaluación en el ámbito educativo es una actividad relevante y de gran importancia en el proceso de enseñanza-aprendizaje de una asignatura, tanto para estudiantes como docentes, ya que va a permitir determinar si se han alcanzado los objetivos educativos marcados. Para el estudiante, se trata de una prueba donde poner en evidencia el aprendizaje realizado durante el curso, y por tanto, los conocimientos adquiridos sobre la materia con la finalidad de superar la asignatura, mientras que para el docente, va a suponer una etapa para la reflexión y el análisis de la metodología educativa, con un efecto de retroalimentación para identificar fortalezas y debilidades de la experiencia educativa realizada y de esta forma, tomar decisiones o realizar cambios si es necesario, orientadas al perfeccionamiento de la docencia.

En el caso de seguir una metodología docente tradicional, la evaluación de una asignatura consiste en un único examen escrito que se realiza al finalizar el programa docente de la materia. Sin embargo, hoy en día, la incorporación de las tecnologías de información y comunicación (TIC) en el campo de la educación (Blanco y Ginovart, 2012; Ferrao, 2010; Crews y Curtis, 2011) y la aparición de nuevas plataformas educativas como Moodle, Teams, Kahoo!, Mentimeter, etc, donde se insertan actividades online que pueden ser puntuables de forma automática como son, por ejemplo, los cuestionarios o pruebas de nivel (López-Tocón, 2021), han favorecido la aplicación de una nueva metodología educativa tipo blended-learning en la mayoría de las asignaturas, donde la evaluación de las capacidades y habilidades de los estudiantes suele estar distribuida entre el examen escrito y los cuestionarios virtuales.

El empleo de cuestionarios virtuales como parte del proceso de evaluación de una asignatura implica la necesidad de elaborar pruebas que contengan preguntas/ítems de calidad, de forma que se obtengan resultados con fuerte evidencia de validez y elevada confiabilidad (Kane, 2001; Downing, 2003). Determinar la validez de un cuestionario no es tarea fácil ya que, dependiendo del objetivo de la prueba y sus propósitos se puede hablar de diferentes tipos de validez, bien en relación a un criterio, de contenido y de constructo. Para un proceso educativo y ocupacional se busca, principalmente, la validez de contenido (McGartland, 2003) y de constructo (Messick, 1995) ya que, los cuestionarios son utilizados como pruebas de rendimiento académico donde el principal interés es determinar el grado en que las preguntas representan el contenido disciplinar que se quiere medir y el proceso cognitivo que se espera que se ponga

en juego a la hora de resolver la pregunta (Paz, 1996). La validez de contenido viene determinada por el juicio de expertos en la materia que, generalmente, recae en el criterio de los profesores implicados en la docencia de la asignatura, mientras que la validez de constructo se refiere a la propia estructura interna de la prueba y puede determinarse de forma cuantitativa con el análisis estadístico-psicométrico de los resultados obtenidos haciendo uso de la teoría clásica de los tests (Muñiz, 2017).

También, la calidad o validez de un cuestionario está directamente relacionada con la elaboración y el diseño de la prueba (Muñiz y Fonseca, 2019), que puede ser muy amplia y variada ya que se pueden modificar diferentes aspectos de la prueba como son, el tiempo de realización, el tipo de modalidad: aleatorio o predeterminado, el tipo de presentación: secuencial o directa, los tipos de preguntas o ítems; verdadero/falso, numérica, muti-respuesta, etc; el número de intentos de la prueba, el acceso a la prueba: restringido en el aula o libre, entre otros. Dependiendo de estas variables se pueden adaptar los cuestionarios (Parshall, 2010) a las distintas necesidades del proceso de enseñanza-aprendizaje, presentando diferentes características como son los de tipo formativo, donde el objetivo es aprender de forma que, si se identifica que ignora algo el estudiante, lo aprende al resolver el examen, o bien, los de tipo evaluativo o sumativo, donde el objetivo es valorar si supera los créditos correspondientes al tema/asignatura.

2. Objetivos

Con este trabajo se pretenden alcanzar varios objetivos, como son:

a) Estudiar el comportamiento de un determinado tipo de pregunta/ítem en cuestionarios realizados en asignaturas de química de educación superior, mediante el análisis estadístico-psicométrico de los resultados obtenidos en base a la teoría clásica de los tests.

b) Diseñar y analizar cuatro tipo de cuestionarios, dentro de la plataforma Moodle, que contienen un mismo tipo de ítem y a su vez, son diferentes entre sí. Los ítems que se van a estudiar son los más frecuentemente utilizados y son: verdadero/falso, numérica, emparejamiento y multi-respuesta, también conocida como opción múltiple.

c) Analizar el efecto de utilizar todos estos tipos de ítems en conjunto mediante la realización de un quinto cuestionario en modo aleatorio.

d) Establecer qué tipo de ítem y qué modalidad de cuestionario tiene suficiente calidad como para ser utilizado en pruebas de nivel con carácter evaluativo.

3. Desarrollo de la innovación

3.1. Entorno docente

Esta experiencia educativa se ha realizado como parte integrante de la metodología blended-learning que se desarrolla en las asignaturas de “Química Física General” en el Grado en Química, y en “Química Analítica” en el Grado de Ingeniería Química de la Universidad de Málaga, durante el curso académico 2022-2023.

La asignatura de Química Física General se imparte en el primer semestre del primer curso del Grado en Química a un grupo de aproximadamente 80 estudiantes. Es una asignatura íntegramente teórica de 6 créditos (60 horas en el aula) donde se combinan las clases presenciales y las actividades electrónicas

desarrolladas en el Campus Virtual. El programa docente consta de 11 temas agrupados en cinco bloques temáticos como son, Materia, Disoluciones, Termodinámica, Electroquímica y Cinética.

La asignatura de Química Analítica se imparte en el segundo semestre del tercer curso del Grado en Ingeniería Química a un grupo de aproximadamente 40 estudiantes. Es una asignatura teórico-práctica de 6 créditos (45 horas de teoría y 15 horas de prácticas de laboratorio). La metodología docente combina clases presenciales, actividades virtuales integradas en el Campus Virtual y experiencias en el laboratorio. El programa docente consta de 15 temas agrupados en tres bloques temáticos como son, Proceso analítico e interpretación de los resultados, Equilibrios en disolución: volumetrías y Técnicas instrumentales.

Ambas asignaturas están integradas en el Campus Virtual de la Universidad de Málaga donde se encuentra implementada la plataforma educativa Moodle.

3.2. Diseño de los cuestionarios Moodle

Para ambas asignaturas se ha seguido un mismo procedimiento en la elaboración de los cuestionarios. Se ha creado un banco de preguntas con los diferentes ítems que se van a utilizar en las pruebas de nivel. Se trata de 15 preguntas/ítems del tipo verdadero/falso, y otras tantas del tipo numérica, emparejamiento y multi-respuesta. De esta forma, el banco de preguntas contiene un total de 60 ítems. Hay que resaltar que los ítems de tipo numérico son ejercicios numéricos simples donde únicamente hay que aplicar alguna expresión matemática relacionada con algún concepto químico, y que no requiere razonamientos o cálculos numéricos complejos. En el caso de los ítems muti-respuestas (opción múltiple) se trata de preguntas con una única respuesta correcta y tres distractores, o respuestas incorrectas, ya que se ha demostrado que este modelo de pregunta es suficiente como para obtener resultados con fuerte evidencia de validez y elevada confiabilidad en los cuestionarios (Haladyna, 2004; McCombrie, 2004). Tanto es así, que este modelo de preguntas es mayormente utilizado en procesos selectivos concurso-oposición de diferentes materias, que pueden ser de gran relevancia, como son las que entran a formar parte del campo de la medicina (Jurado-Núñez, 2013).

Las 15 preguntas/ítems que forman el cuestionario engloban todo el contenido del programa docente de las asignaturas y son las mismas para todos los estudiantes del mismo curso. En el caso de la asignatura de Química Física General con un programa de 11 temas, se ha seleccionado una pregunta por cada tema y se han añadido 2 preguntas más de los temas de Termodinámica y Electroquímica hasta completar las 15 preguntas. Se han seleccionado estos temas porque se trabaja con conceptos y conocimientos más complejos que no han sido estudiados en secundaria y que es importante su correcta asimilación para seguir estudiando otras materias de Química Física en cursos superiores. Para la asignatura de Química Analítica con un programa de 15 temas simplemente se ha seleccionado una pregunta/ítem por tema.

Además de estos cuatro cuestionarios, que contienen el mismo tipo de ítem, se ha realizado un quinto cuestionario en modalidad aleatoria. En este caso, la prueba contiene 20 preguntas, cinco preguntas más que en los anteriores cuestionarios por tratarse de la última prueba online que realizan los estudiantes, y porque además, las preguntas son ya conocidas por los estudiantes al haber realizado los anteriores cuestionarios. Esta modalidad presenta la peculiaridad que utiliza todos los ítems del banco de preguntas y la propia plataforma selecciona de forma aleatoria los 20 ítems del cuestionario cada vez que un estudiante accede a la prueba, por lo que se trata de una prueba casi individual. El objetivo es conocer el efecto del uso de los distintos tipos de ítems de forma aleatoria en los resultados estadísticos, y compararlos con las pruebas realizadas en modalidad no aleatoria y con un mismo tipo de ítem.

Otros aspectos relacionados con el diseño de los cuestionarios son: el tiempo de ejecución es de una hora, la disposición de los ítems así como de las respuestas en los ítems multi-respuesta es aleatorio para cada estudiante, la navegación durante la prueba es secuencial, es decir, tienen que contestar a las preguntas en el orden en el que aparecen sin posibilidad de volver atrás, solo hay un intento para realizar la prueba y no existe retroalimentación en las respuestas. Al acabar el cuestionario se indica únicamente la nota obtenida.

Estas pruebas de nivel se realizan al finalizar el periodo docente de cada una de las asignaturas y, por tanto, sirven como repaso de los contenidos además de preparación para el examen escrito. Se realiza un cuestionario al día de forma consecutiva empezando por verdadero/falso, seguidos de los tipos numérico, emparejamiento, opción múltiple y por último, el aleatorio. Los estudiantes son informados previamente de las características del cuestionario y del tipo de ítems que los contiene. A la hora de ejecutar estas pruebas, que son puntuables para la nota final de la asignatura, los estudiantes se encuentran familiarizados con el entorno Moodle al haber realizado, con anterioridad, otras actividades docentes formativas de tipo online con características similares.

3.3. Parámetros estadístico-psicométricos

La calidad de los cuestionarios se ha analizado en base a dos aspectos: a la prueba en su totalidad y a los ítems que la componen. En relación a la prueba como un todo se evalúa su validez y su confiabilidad (Brown, 2000). La validez está relacionada con el grado que una prueba mide aquello que debe medir y es útil para el propósito que se construyó. En este caso, al tratarse de pruebas de rendimiento académico, la validez se considera a juicio de los docentes encargados de la asignatura y a su propio criterio para determinar los conocimientos más relevantes de la materia que deben asimilar los estudiantes para proseguir en cursos superiores, como en el caso de la asignatura de Química Física General de primer curso, o bien, para favorecer su inminente inserción en el mercado laboral como es el caso de la asignatura Química Analítica de tercer curso. Por otro lado, la confiabilidad hace referencia a la consistencia y la precisión en las medidas arrojadas por una prueba. En este caso, se puede determinar cuantitativamente mediante el coeficiente de consistencia interna (CCI) aplicando variados métodos de toma de muestras de diferentes ítems y se revisan los efectos de éstos sobre la confiabilidad (Aiken, 1996), como es el caso del parámetro Alfa de Cronbach que es una media ponderada de las correlaciones entre los ítems que forman parte de la escala, en este caso, el cuestionario. Cuanto más se aproxime a su valor máximo de 1 (ó 100%) mayor es la fiabilidad de la escala. En general, se considera que valores de CCI o del alfa superiores a 0,6 (60%) ó 0,7 (70%) son suficientes para garantizar la fiabilidad de la escala. Un valor inferior indicaría que los ítems en su conjunto no son homogéneos o bien no son buenos para discriminar la diferente habilidad de los estudiantes (Cabrera, 2013; Valero, 2022).

Otros parámetros que se han analizado son: el número de participantes, el tiempo medio de realización de los cuestionarios, la nota media de las calificaciones, así como la dispersión o desviación estándar (DS), la asimetría y planaridad de la distribución de las notas, también llamado sesgo y curtosis, respectivamente, y el error estándar (ES). Este último parámetro es una medida de la incertidumbre en la calificación de cualquier estudiante, es decir, si ese mismo estudiante resolviera otra prueba de características similares en la misma materia, se esperaría que su calificación estuviera dentro de más-menos ese error estándar de la calificación obtenida. Se estima un valor adecuado de 8-10% (Valero, 2022). También se indica la tasa de error que estima el porcentaje de la DS que se debe a efectos aleatorios en lugar de diferencias significativas en las capacidades entre los estudiantes. Este parámetro está directamente relacionado con el CCI, ya que a mayor CCI menor será la tasa de error, y viceversa.

En relación a los ítems que componen la prueba se han evaluado dos parámetros como son: el índice de facilidad (IF) y el índice de discriminación (ID). El IF de un ítem indica la proporción de sujetos que resuelven correctamente el ítem. Un rango entre el 20-85% se considera aceptable (Guilford, 1975), aunque cada investigador o usuario puede fijar el rango de aceptación de acuerdo a los datos obtenidos y al objetivo de la prueba (Lord y Novick, 1968). No obstante, una limitación de este parámetro es que depende de la población que responde el ítem y del dominio que tienen sobre la materia. Así, un elevado dominio de la materia resultará más fácil responder al ítem y viceversa. El ID indica si el ítem tiene alto poder para diferenciar la capacidad entre estudiantes y es la correlación entre las calificaciones ponderadas obtenidas por los estudiantes en dicho ítem y las obtenidas en el test. Indica que tan efectiva es la pregunta/ítem para discernir a los estudiantes más capaces de los menos capaces. Se considera un valor adecuado superior al 20% ó 30% (Guilford, 1975; Cano, 2004).

Los parámetros psicométricos como son el IF, ID y el CCI se han calculado en base a la teoría clásica de los tests (Muñiz, 2017) que, junto con el resto de parámetros estadísticos, vienen reportados en el apartado de estadísticas dentro de la plataforma Moodle. El análisis de estos parámetros en los distintos cuestionarios nos va a permitir conocer cómo se comporta el tipo de ítem en estas pruebas de nivel online y además, detectar qué tipo de ítem es más idóneo para emplearlo en pruebas evaluativas.

4. Resultados y discusión

4.1. Asignatura de Química Física General (Grado en Química)

En la Tabla 1, se recogen los valores promedios de los cuestionarios realizados. En general, la participación ha sido elevada, del 75%, aproximadamente, en cualquiera de las pruebas. La calificación media obtenida es mayor para el cuestionario verdadero/falso con un valor de 7.68 mientras que la más baja corresponde al de tipo numérico con un 4.91. Los de tipo emparejamiento, opción múltiple y aleatorio presentan valores intermedios de 6.91, 5.22 y 5.69, respectivamente. El ES que mide la incertidumbre en la calificación de cualquier estudiante se encuentra en un rango bajo de 7-11%, como es aconsejable en estas pruebas (Valero, 2022).

El tiempo medio empleado en la prueba de verdadero/falso es de 29 min., el valor más bajo con respecto al resto de pruebas que están sobre 40 min. Este comportamiento se puede apreciar en la Fig. 1 donde se representa el tiempo empleado y la calificación obtenida por los participantes.

Tabla 1. *Valores estadísticos obtenidos para los cuestionarios de Química Física General*

Cuestionario Tipo de ítem	Participación/ Nº matrículas	Nota media	Tiempo (min.)	DS (%)	Sesgo	Curtosis	CCI (%)	ES / Tasa error (%)
Verdadero/Falso	66 / 84	7.68	28.33	12.98	-1.00	1.12	49.66	9.21 / 70.95
Numérica	61 / 84	4.91	40.40	12.11	-1.30	2.43	60.34	7.63 / 62.98
Emparejamiento	59 / 84	6.91	40.49	12.87	-0.76	0.23	72.55	6.74 / 52.40
Opción Múltiple	61 / 84	5.22	37.17	13.94	-0.71	0.78	39.89	10.81 / 77.53
Aleatoria	61 / 84	5.69	34.27	16.73	-0.61	0.15	54.89	11.23 / 67.16

DS: Desviación estándar; CCI: Coeficiente de Consistencia Interna (Alfa de Cronbach); ES: Error Estándar

La mayoría de los estudiantes emplean más de 20 min. en resolver cualquiera de las pruebas. No obstante, una pequeña proporción emplean un tiempo inferior que puede ser atribuido a diferentes causas como problemas informáticos de cierre inesperado de la aplicación, o bien a estudiantes repetidores que conocen las respuestas y contestan rápidamente obteniendo elevada calificación, o simplemente, a

estudiantes poco motivados que contestan al azar, por probar cómo es la prueba sin interés en la calificación que se pueda alcanzar.

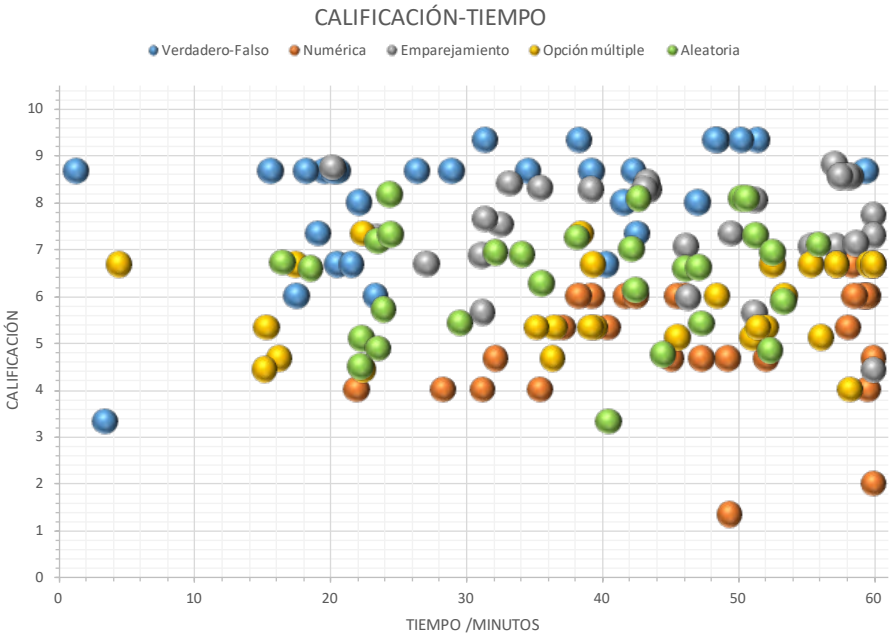

Fig. 1. *Calificación-tiempo de todos los participantes en los cuestionarios de la asignatura de Química Física General*

Las calificaciones obtenidas en cualquiera de los cuestionarios presentan una baja dispersión con respecto al promedio, con una DS en el rango de 12-17% que se considera adecuado para este tipo de pruebas (Valero, 2022). Respecto a los parámetros que caracterizan la distribución de las calificaciones como son el sesgo y la curtosis se obtienen valores dentro de lo establecido. El sesgo es una medida de la asimetría de la distribución y es recomendable un valor entre -1.0 y 1.0. En este caso, se obtienen valores en un rango de -0.6 a -1.0 para todos los cuestionarios excepto para los de tipo verdadero/falso y numérico con un valor de -1.0 y -1.3, respectivamente, indicando la presencia de pocos valores bajos de calificaciones estando éstas concentradas en valores altos de la escala. Por otro lado, la curtosis es una medida de la forma y de cuanto de aplanada es la distribución, es decir, indica la deformación vertical de la misma. Su valor deseable está entre 0-1. También son las pruebas verdadero/falso y numérica las que se salen de este rango con unos valores de 1.12 y 2.43, respectivamente. Por tanto, estos dos tipos de pruebas con ítems verdadero/falso y numérico, no ofrecen muy buena calidad ya que no brindan una muy buena discriminación entre los estudiantes a los que le va mejor que al promedio.

Sin embargo, el análisis del CCI que mide la confiabilidad de las pruebas indica que solamente las pruebas numérica, emparejamiento y aleatoria, presentan valores dentro de lo recomendable, sobre un 60% o superior, mientras que las de tipo verdadero/falso y opción múltiple son las que tienen un valor ligeramente inferior, sobre un 40 y 50%, respectivamente. Esto puede ser debido a que algunos ítems no están funcionando correctamente y tienen una calidad diferente del resto de ítems, provocando que la prueba en su conjunto no sea homogénea, bien porque no se contesta al ítem como ocurre en los ítems de tipo numérico, bien porque hay un alto grado de respuestas debidos al azar. De esta forma, las diferencias entre las calificaciones finales de los estudiantes estarían asociadas al azar. Tanto es así, que la tasa de

error (Tabla 1) que estima el porcentaje de la DS que se debe a efectos aleatorios son las más elevadas en los cuestionarios verdadero/falso y opción múltiple, sobre 70-80% (70.95 y 77.53), algo inferior para los de tipo numérico y aleatorio, sobre 60-70% (62.98 y 67.16), y el más bajo corresponde al de tipo emparejamiento con un valor de 52.40%. Por tanto, según los resultados estadístico-psicométricos analizados, el cuestionario con mejor calidad evaluativa para esta asignatura sería el de emparejamiento. Esta conclusión se encuentra respaldada con el análisis del IF y ID de cada uno de los ítems (Fig. 2).

En la Fig. 2 se puede observar que la mayoría de los ítems verdadero/falso presentan un IF elevado entre 80-100% y por tanto un ID bajo, inferior al 20%. Los ítems de tipo numérico presentan un comportamiento parecido, con la diferencia que hay muchas preguntas que no son respondidas por los estudiantes, tanto es así, que solo resuelven la mitad de los ítems, y por tanto, hay ítems con IF elevado y otros con IF nulo, lo que hace que el ID sea bajo en la mayoría de los ítems. En definitiva, estos dos tipos de ítems, verdadero/falso y numérico, no son adecuados para realizar una prueba compuesta únicamente por estos ítems. Sin embargo, los ítems de emparejamiento funcionan mejor dando una distribución en el IF y el ID adecuado a estas pruebas online. El IF se distribuye ente un 50-90% y el ID en un rango de 20-50% indicando que la dificultad en las preguntas es moderada y que éstas son adecuadas para distinguir las distintas capacidades entre los estudiantes. En el caso de los ítems opción múltiple, a pesar de presentar un IF moderado e incluso con dificultad elevada ya que hay ítems con valores de IF inferior al 20%, presentan un ID bajo lo que indica que no establece diferencias significativas entre los estudiantes mejores calificados y el promedio y que las calificaciones pueden deberse al azar como en el caso de los ítems verdadero/falso.

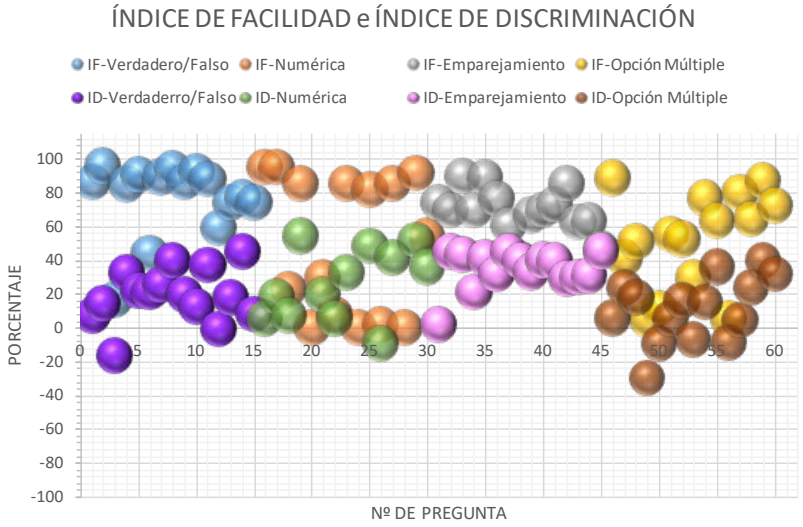

Fig. 2. *Índice de Facilidad (IF) e Índice de Discriminación (ID) de los diferentes ítems que forman parte de los cuatro cuestionarios con el mismo tipo de ítem en la asignatura de Química Física General*

No obstante, el comportamiento de estos dos tipos de ítems, emparejamiento y opción múltiple, son los que mejor funcionan en el cuestionario tipo aleatorio (Fig. 3) ya que presentan valores del ID elevados, próximo al 100%, para la mayoría de los ítems e IF entre un 20-80% que suelen aportar importantes diferencias ente el nivel de conocimiento, habilidad y preparación entre los estudiantes.

En resumen, la experiencia docente realizada en la asignatura de Química Física General indica que los cuestionarios elaborados únicamente con ítems de emparejamiento son los que tienen calidad para una

evaluación online. Los de opción múltiple también serían adecuados, aunque conllevan una contribución al azar que puede ser significativa. La combinación de estos dos tipos de ítems en una prueba de modalidad aleatoria funcionaría de forma satisfactoria. Sin embargo, los ítems verdadero/falso y numérico no serían adecuados para realizar pruebas evaluativas en ninguna de las dos modalidades, único o aleatorio. No obstante, su participación en modalidad aleatoria, junto con los ítems de emparejamiento y opción múltiple, puede dar lugar a una prueba de conocimiento con elevada confiabilidad, siendo conveniente no abusar de esos tipos de ítems ya que, harían disminuir la calidad del mismo.

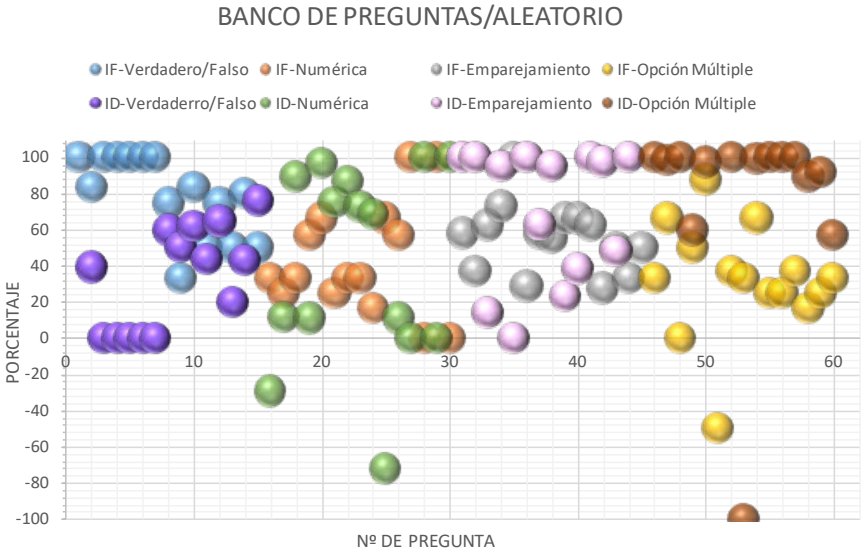

Fig. 3. *Índice de Facilidad (IF) e Índice de Discriminación (ID) de los ítems que forman el banco de preguntas en el cuestionario aleatorio de la asignatura de Química Física General*

4.2. Asignatura de Química Analítica (Grado en Ingeniería Química)

En la Tabla 2, se recogen los valores promedios de los cuestionarios realizados. En general, la participación no ha sido excesivamente elevada, sobre el 61%, en cualquiera de las pruebas. La calificación media obtenida es mayor para el cuestionario verdadero/falso con un valor de 8.78, ligeramente inferior para el de tipo opción múltiple y aleatorio con valores de 7.73 y 7.49, respectivamente, mientras que la más baja corresponde al de tipo numérico con un 5.03 y a la de emparejamiento con un 5.70. La incertidumbre en estas calificaciones, estimadas por el ES, indica que se encuentra en un rango bajo de 8-10%. Este comportamiento es similar al analizado previamente en la asignatura de Química Física General, donde se obtiene una mayor puntuación media para los cuestionarios verdadero/falso y una menor puntuación para los de tipo numérico.

También se ha observado un comportamiento similar en el tiempo medio empleado en la realización de las pruebas, siendo el más bajo en la prueba de verdadero/falso, de 25 min., con respecto al resto de pruebas que están sobre 40 min. excepto la de emparejamiento donde se llega casi a consumir todo el tiempo que dura la prueba, 60 min. Este comportamiento se puede apreciar en la Fig. 4 donde se representa el tiempo empleado y la calificación obtenida para todos los participantes. La mayoría de los estudiantes emplean más de 20 min. en resolver cualquiera de las pruebas, aunque una pequeña proporción emplean un tiempo inferior sobre todo en las pruebas de verdadero/falso donde se obtiene elevada calificación empleando poco tiempo. Esto puede ser atribuido a una participación colaborativa

entre varios estudiantes donde se pasan las respuestas de las preguntas. El resto de casos en las otras pruebas, numérica y opción múltiple, puede ser debido a estudiantes con poco interés en este tipo de actividades.

Tabla 2. *Valores estadísticos obtenidos para los cuestionarios de Química Analítica*

Cuestionario Tipo de ítem	Participación/ Nº matrículas	Nota media	Tiempo (min.)	DS (%)	Sesgo	Curtosis	CCI (%)	ES / Tasa error (%)
Verdadero/Falso	24/39	8.78	25.20	8.49	-0.06	-1.17	17.70	7.71 / 90.72
Numérica	24/39	5.03	48.30	12.58	-0.07	-1.30	29.52	10.56 / 83.95
Emparejamiento	23/39	5.70	54.00	11.61	-0.23	0.13	53.12	7.95 / 68.47
Opción Múltiple	25/39	7.73	38.17	18.16	-1.45	2.89	73.53	9.34 / 51.45
Aleatoria	19/39	7.49	39.35	14.19	-0.80	0.83	69.28	7.86 / 55.42

DS: Desviación estándar; CCI: Coeficiente de Consistencia Interna (Alfa de Cronbach). ES: Error Estándar

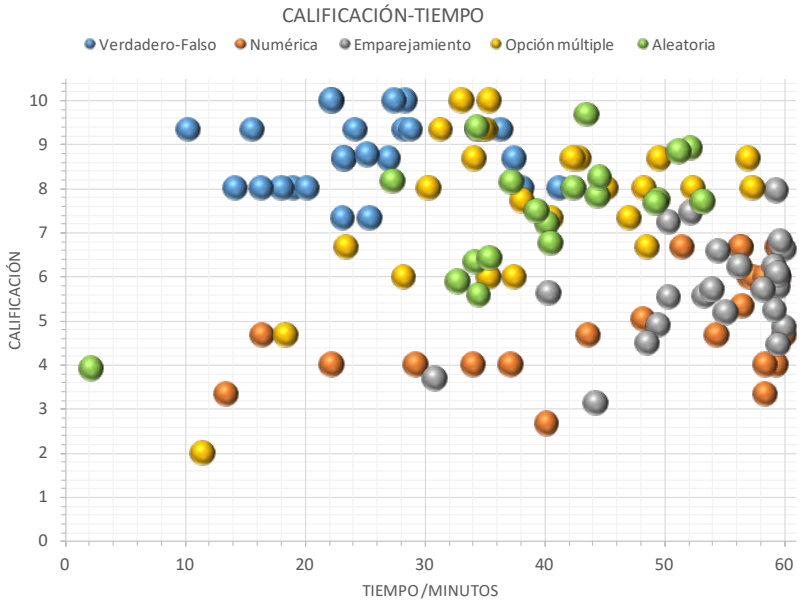

Fig. 4. *Calificación-tiempo de todos los participantes en los cuestionarios de la asignatura de Química Analítica*

Las calificaciones obtenidas en cualquiera de los cuestionarios presentan una moderada dispersión con respecto al promedio, con una DS en el rango de 8-20%, siendo la más baja para la prueba verdadero/falso y ligeramente más elevada para las pruebas de opción múltiple y aleatoria. Respecto a la distribución de las calificaciones, definida por los parámetros de sesgo (desviación horizontal) y curtosis (desviación vertical), se obtienen valores típicos para una distribución simétrica y aplanada, como es el caso de la prueba de emparejamiento con sesgo, -0.23 y curtosis, 0.13 o bien, ligeramente desplazada y con valores concentrados en calificaciones elevadas como es la prueba aleatoria con un sesgo de -0.80 y curtosis de 0.83. Sin embargo, se obtiene un valor alejado de los valores de referencia únicamente para la prueba de opción múltiple, con un valor de sesgo de -1.45 y curtosis de 2.89, lo que indica que hay gran número de calificaciones por encima del promedio y concentradas a un valor elevado de calificación. Sin embargo, esta prueba es la que presenta el CCI más elevado, con una confiabilidad del 73.53%, arrojando, por tanto, una baja tasa de error del 51.45% debida a procesos aleatorios en la prueba. Se puede decir, por

tanto, que la prueba de opción múltiple es la que proporciona diferencias significativas entre los estudiantes y por tanto, presenta carácter evaluativo, al igual que la prueba de tipo aleatorio con un CCI del 69.28% y una tasa de error del 55.42%. Un CCI ligeramente inferior presenta la prueba de emparejamiento con un 53.12%, la cual entra dentro del valor de referencia que se propone como prueba evaluativa (Valero, 2022). Sin embargo, las pruebas de verdadero/falso y numérica son las que presentan menor confiabilidad, con un CCI del 17.70 y 29.52% respectivamente y, por tanto, no sería factible su utilización como prueba evaluativa de la asignatura. Este comportamiento se encuentra respaldado con el análisis del IF y ID de cada uno de los ítems (Fig. 5).

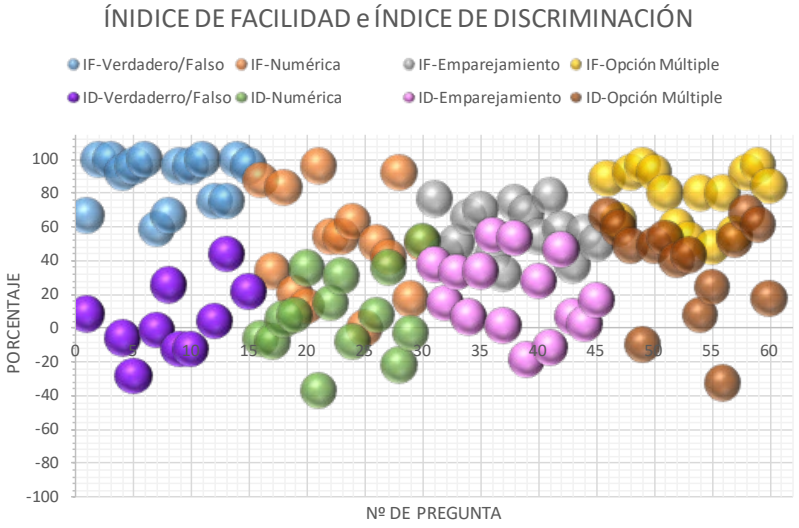

Fig. 5. *Índice de Facilidad (IF) e Índice de Discriminación (ID) de los diferentes ítems que forman parte de los cuatro cuestionarios en la asignatura de Química Analítica*

En la Fig. 5 se puede observar que la mayoría de los ítems verdadero/falso presentan un IF elevado, próximo al 100%, indicando una dificultad baja, y dando un ID bajo, inferior al 20%, y por tanto, no serían buenas para discriminar las calificaciones de los estudiantes. Los ítems de tipo numérico y emparejamiento presentan un comportamiento parecido, con la diferencia que la mayoría de los ítems presentan una dificultad media, con un IF en torno al 50%, lo que hace que el ID sea bajo en la mayoría de los ítems, aunque algo mejor para los de tipo emparejamiento comparado con los de verdadero/falso. En definitiva, entre estos tipos de ítems solo la prueba de emparejamiento sería adecuada para realizar una prueba evaluativa compuesta únicamente por estos ítems. No obstante, los ítems de opción mútiple funcionan aún mejor, dando una distribución en el IF y el ID adecuado a estas pruebas online (Valero, 2022). El IF se distribuye ente un 50-90% y el ID en un rango de 20-60% indicando que la dificultad en las preguntas es moderada y que éstas son adecuadas para distinguir las distintas capacidades entre los estudiantes. Este comportamiento cambia ligeramente si se considera el análisis del IF e ID de todos los ítems en el cuestionario realizado en modalidad aleatoria (Fig. 6).

En el caso de modalidad aleatoria, hay que tener en cuenta que es el último cuestionario que realizan los estudiantes, por tanto, ya conocen los ítems que pueden aparecer en dicha prueba, porque previamente han formado parte de las anteriores experiencias. Este hecho se refleja en los resultados obtenidos para todos los ítems, en la Fig. 6. La dificultad de la mayoría de los ítems ha disminuido con un IF elevado cercano al 100%, no solo para los ítems de verdadero/falso como era de esperar, sino también para los de tipo opción múltiple, dando lugar a un ID bajo o casi nulo. Solamente los ítems de tipo emparejamiento

son los que presentan un IF moderado en un rango de 40-80% al igual que un ID del mismo orden, aunque habría que revisar algunas preguntas por presentar un valor negativo.

En resumen, la experiencia docente realizada en la asignatura de Química Analítica arroja resultados similares a los obtenidos en la asignatura de Química Física General. Los ítems verdadero/falso y numérico no serían adecuados para realizar pruebas evaluativas en ninguna de las dos modalidades, único o aleatorio, siendo los ítems de emparejamiento y opción múltiple los más convenientes para realizar una evaluación online, en cualquiera de las modalidades. Además, se ha comprobado que la prueba aleatoria, al tratarse de un prueba repetitiva pues utiliza los mismos ítems conocidos por los participantes en las anteriores pruebas, provoca que el IF sea más elevado incluso para los ítems de opción múltiple, siendo los ítems de emparejamiento los que mejor funcionarían incluso en pruebas reiterativas.

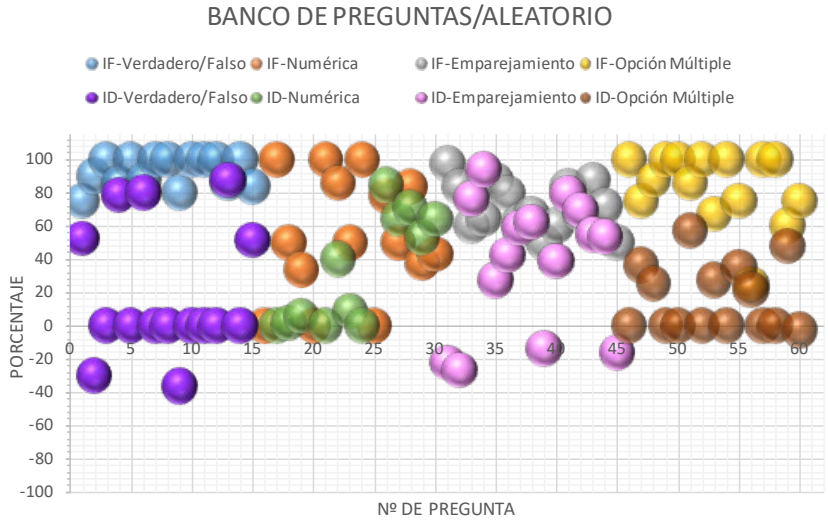

Fig. 6. *Índice de Facilidad (IF) e Índice de Discriminación (ID) de los ítems que forman el banco de preguntas en el cuestionario aleatorio de la asignatura de Química Analítica*

5. Conclusiones

Se han analizado los parámetros estadístico-psicométricos de los cuestionarios realizados durante el curso académico 2022-2023 en las asignaturas de Química Física General y Química Analítica del Grado en Química e Ingeniería Química, respectivamente. Han sido cinco cuestionarios, cuatro de ellos contienen ítems de un mismo tipo, verdadero/falso, numérico, emparejamiento o multi-respuesta (opción múltiple) y el quinto cuestionario se realiza de forma aleatoria con todos los ítems creados en el banco de preguntas dentro de la plataforma Moodle.

Los resultados obtenidos para ambas asignaturas han sido muy similares. La prueba de verdadero/falso permite obtener calificaciones elevadas en poco tiempo de realización, presentan generalmente poca dificultad, con un IF elevado y, por tanto, un ID bajo, dando como resultado una prueba con una confiabilidad baja y con una tasa de aleatoriedad elevada en las calificaciones. Se trataría de un tipo de ítem no adecuado para realizar pruebas de conocimiento de tipo evaluativo, ya que no es capaz de discriminar entre las distintas capacidades o habilidades de los alumnos. Lo mismo se puede concluir para la prueba de tipo numérico. En este caso, se obtiene una calificación baja, a pesar de emplearse un tiempo de realización mayor, debido a que muchas preguntas no son contestadas o se responden erróneamente por los estudiantes. Esto hace que los ítems se repartan entre dificultad elevada y fácil, presentando un IF

moderado, pero con un ID relativamente bajo, resultando una prueba con confiabilidad (CCI) baja. Por tanto, una prueba de conocimiento con ítems de tipo numérico no sería aconsejable como evaluación de la asignatura.

Sin embargo, las pruebas realizadas con ítems de tipo emparejamiento y opción múltiple son las que han arrojado resultados satisfactorios, como para ser consideradas en la evaluación de una materia educativa. En ambos casos, la mayoría de los ítems presentan una dificultad y discriminación media, con un IF y un ID moderado, dando como resultado pruebas con elevada confiabilidad y tasa de aleatoriedad baja. Al igual ocurre con una prueba realizada de modo aleatorio donde se integran todos los tipos de ítems, aunque en este caso, hay que resaltar que las preguntas verdadero/falso y numérica son las que presentan menor discriminación entre los estudiantes, mientras que las de opción múltiple y de emparejamiento son las que funcionan mejor para diferenciar las habilidades y conocimientos de los estudiantes más brillantes respecto a la media.

No obstante, en futuros trabajos docentes se pretende verificar si este comportamiento que presentan los diferentes ítems analizados en este estudio se mantiene, o por el contrario se ve afectado por el propio diseño de los cuestionarios y de otras variables como son un mayor número de ítems empleados en los cuestionarios, o la utilización de un banco de preguntas más extenso para el caso de pruebas aleatorias.

Agradecimientos. Este trabajo se enmarca dentro del Proyecto Docente PIE22-070, correspondiente a la Convocatoria de Proyectos de Innovación Educativa 2022-2024 de la Universidad de Málaga.

Referencias

AIKEN (1996). Test psicológicos y evaluación. Mexico: Prentice-Hall.

BLANCO, M. y GINOVART, M. (2012). "Los cuestionarios del entorno Moodle: su contribución a la evaluación virtual formativa de los alumnos de matemáticas de primer año de las titulaciones de Ingenieria" en RUSC. Universities and Knowledge Society Journal, vol. 9, issue 1, p.166-183. http://rusc.uoc.edu/ojs/index.php/rusc/article/view/v9n1-blanco-ginovart/v9n1-blanco-ginovart

BROWN, F.G. (2000). Principios de la Medición en Psicología y educación. México: Manual Moderno.

CABRERA, I.M. (2013). El análisis de ítems del módulo de cuestionario Moodle en la asignatura Medición y Evaluación Psicológica. Recuperado en https://www.uned.ac.cr/academica/edutec/memoria/ponencias/cabrera_25.pdf

CANO, F. (2004). "Construcción de pruebas de conocimiento" en el Seminario Internacional: Compromiso de la evaluación objetiva con el mejoramiento de la calidad de la educación superior. Bogotá. ACOFI-Asociación Latinoamericana de Psicología.

CREWS, T.B. y CURTIS, D.F. (2011). "Online course evaluations: Faculty perspective and strategies for improved response rates" en Assessment & Evaluation in Higher Education, vol 36, issue 7, p.865-878. https://www.learntechlib.org/p/70044/.

DOWNING, S.M. (2003). Validity: on the meaningful interpretation of assessment data. Med. Educ. 37, 830-837. https://doi.org/10.1046/j.1365-2923.2003.01594.x.

FERRAO, M. (2010). "E-assessment within the Bologna paradigm: evidence from Portugal" en Assessment & Evaluation in Higher Education, vol. 35, issue 7, p. 819-830. https://doi.org/10.1080/02602930903060990

GUILFORD, J.P. (1975). Psychometric methods. Bombay, Nueva Delhi. Editorial Tata McGraw-Hill.

HALADYNA, T.M. (2004). Developing and Validating Multiple-Choice Test Items. 3rd Ed. Mahwah, N.J.: Lawrance Erlbaum associates, Inc. Publishers 2004, Chapter 1:3-18.

JURADO-NÚÑEZ, A., FLORES-HERNÁNDEZ, F., DELGADO-MALDONADO, L., SOMMER-CERVANTES, H., MARTÍNEZ-GONZÁLEZ, A., SANCHÉZ-MENDIOLA, M. (2013). "Distractores en preguntas de opción múltiple para estudiantes de medicina: ¿cuál es su comportamiento en un examen sumativo de altas consecuencias?" Investigación en Educación Médica 2, 202-210. https://doi.org/10.1016/S2007-5057(13)72713-3

KANE, M. (2001). Current concerns in validity theory. Journal of Educational Mesurement, 38, 319-342. https://doi.org/10.1111/j.1745-3984.2001.tb01130.x

LÓPEZ TOCÓN, I. (2021). "Moodle Quizzes as a Continuous Assessment in Higher Education: An Exploratory Approach in Physical Chemistry" Education Sciences 11, no. 9: 500. https://doi.org/10.3390/educsci11090500

LORD, F. y NOVICK, M. (1968). Statistical theories of mental test scores. USA, Editorial Addison-Wesley.

McCOUBRIE, P. (2004). Improving the fairness of multiple-choice questions: a literature review. Med. Teach. 26, 709-712. https://doi.org/10.1080/01421590400013495

McGARTLAND, D.; BERG-WEGER, M.; TEBB, S.S.; LEE, E.S.; RAUCH, S. (2003). Objectifying Content Validity: Conducting a content validity study in social work. Research Social Work Research, 27, 94-104. https://doi.org/10.1093/swr/27.2.94

MESSICK, S. (1995). Validity of Psycological Assessment: Validation of Inferences from Persons-Response and Performances as Scientific Inquiry into Score Meaning. American Psychologist, 50, 741-749. https://doi.org/10.1037/0003-066X.50.9.741

MUÑIZ, J. (2017). Teoría Clásica de los Tests. Madrid: Editorial Pirámide.

MUÑIZ, J. y FONSECA-PEDRERO, E. (2019). "Diez pasos para la construcción de un test". Psicothema, 31, 7-16. https://doi.org/10.7334/psicothema2018.291

PAZ, M. (1996). Validez en Muñiz, J. (Ed) Psicometría. Madrid. Universitas.

PARSHALL, C.G., HARMES, J.C., DAVEY, T. y PASHLEY P. (2010). Innovative items for computerized testing, en W.J. van der Linden y C.A. Glas, elements of adapting testing (pp. 215-230). Londres: Springer.

VALERO, G. (2022). Reporte de estadísticas de examen en https://docs.moodle.org/all/es/Reporte_de_estad%C3%ADsticas_de_examen