Revista Estudios Psicológicos
Vol. 1 Núm. 4 Págs. 80-101 (2021)
ISSN: 2788-6506 / ISSN-L 2788-6492
Editada por: Instituto Universitario de Innovación Ciencia y Tecnología Inudi Perú

REVISTA ESTUDIOS PSICOLÓGICOS

Artículo original

Dificultad asignada y observada de las preguntas de una prueba de rendimiento académico

Assigned and Observed Difficulty of the Questions on an Academic Achievement Test

Dificuldade atribuída e observada das perguntas em um teste de desempenho acadêmico

Luis Hurtado
Universidad de Ingeniería y Tecnología (Perú)
https://orcid.org/0000-0001-5373-556X

DOI: https://doi.org/10.35622/j.rep.2021.04.004

Recibido 27/05/2021/ Aceptado 16/11/2021

RESUMEN. Los encargados de construir pruebas de rendimiento académico proponen un conjunto de preguntas en tres niveles de dificultad: baja, media y alta. Cuando no se cuenta con un banco de preguntas calibradas, los constructores apelan a su experiencia para estimar la dificultad de las preguntas que proponen. De este modo, previo a la aplicación, tenemos un nivel de dificultad asignado a cada pregunta de la prueba. Posteriormente, con la data recogida del patrón de respuestas de los examinados, podemos obtener el índice de dificultad de cada pregunta. Así, luego de la aplicación, tenemos un nivel observado para cada pregunta. El objetivo del presente estudio es determinar el grado de coincidencia entre la dificultad asignada y la observada en cada una de las preguntas en una prueba de rendimiento académico.

PALABRAS CLAVE: Índice de dificultad, índice de discriminación, indicador de evaluación, outfit, infit.

ABSTRACT. Those in charge of constructing academic achievement tests propose a set of questions at three levels of difficulty: low, medium, and high. When a calibrated question bank is not available, the builders draw on their experience to estimate the problem of the questions they propose. In this way, before the application, we have a level of difficulty assigned to each test question. Subsequently, with the data collected from the response pattern of the examinees, we can obtain the difficulty index of each question. Thus, after application, we have an observed level for each question. The objective of this study is to determine the degree of coincidence between the assigned difficulty and that observed in each of the questions in an academic achievement test.

KEYWORDS: Difficulty index, discrimination index, evaluation indicator, outfit, infit.

RESUMO. Os responsáveis pela construção dos testes de desempenho acadêmico propõem um conjunto de questões em três níveis de dificuldade: baixo, médio e alto. Quando um banco de perguntas calibrado não está disponível, os construtores baseiam-se em sua experiência para estimar a dificuldade das perguntas que propõem. Desta forma, antes da aplicação, temos um nível de dificuldade atribuído a cada questão do teste. Posteriormente, com os dados coletados do padrão de resposta dos examinados, podemos obter o índice de dificuldade de cada questão. Assim, após a aplicação, temos um nível observado para cada questão. O objetivo deste estudo é determinar o grau de coincidência entre a dificuldade atribuída e aquela observada em cada uma das questões de um teste de desempenho acadêmico.

PALABRAS CLAVE: Índice de dificuldade, índice de discriminação, indicador de avaliação, outfit, infit.

1. INTRODUCCIÓN

La educación es un derecho fundamental de toda persona. En su diccionario Egg (2012) la concibe como un proceso continuo a lo largo de toda la vida, pero señala que educación alude a un conjunto de actividades y procedimientos que un educador realiza intencional, sistemática y metódicamente sobre sus educandos. Bajo este enfoque la educación no es algo involuntario, sino que persigue una finalidad. En el Diccionario Enciclopédico de Educación (2003) al ampliar la definición del término, se señala que la educación es un proceso intencional que exige la presencia de una finalidad.

En el Perú la Ley General de Educación señala que la educación básica busca, entre otros, el desarrollo de capacidades, conocimientos, actitudes y valores fundamentales que la persona debe poseer para actuar adecuada y eficazmente en los diversos ámbitos de la sociedad (Ley 28044, art.29). Uno de los objetivos es desarrollar aprendizajes en el campo de las ciencias (art.31). "Aprender es la capacidad de adquirir el conocimiento de alguna cosa por el estudio, la observación o la experiencia" (Egg, 2012, p.19).

Un criterio usado para definir el aprendizaje es el cambio perdurable en la conducta. Según Schunk (1997) el término aprendizaje es empleado cuando alguien se vuelve capaz de hacer algo distinto de lo que hacía antes. Desarrollar aprendizajes se convierte en uno de los principales fines de las instituciones educativas y la evaluación de los estudiantes permite recoger los resultados de estos aprendizajes. La Evaluación Censal de Estudiantes (ECE), que es una evaluación estandarizada realizada por el Ministerio de Educación del Perú, busca conocer los logros de aprendizaje alcanzados por los estudiantes peruanos. Evaluar los aprendizajes supone emitir un juicio de valor de la información recogida al ser comparada con los criterios de evaluación establecidos previamente. Los criterios de evaluación son establecidos por el agente evaluador tomando como base documentación curricular vigente. La información se recoge a partir de los instrumentos de evaluación como los exámenes o pruebas de rendimiento académico, también llamados test de aprovechamiento.

Al evaluar el aprendizaje lo hacemos, sobre todo, basándonos en las expresiones verbales,

los escritos y las conductas (Schunck, 1997). Por su practicidad y economía, los docentes suelen aplicar exámenes para medir los logros de aprendizaje de sus estudiantes. Sin embargo, debido a que los aprendizajes no son observables directamente, estos deben ser inferidos a partir de las respuestas dadas por los examinados a las preguntas del examen. Para Cano (1991) uno de los usos de los exámenes sería como un medio para provocar o estimular la presencia de las conductas a ser evaluadas. Bajo este uso, en tanto instrumento de evaluación educativa, el examen puede definirse como "un conjunto estructurado y probado de reactivos y estímulos a través de los cuales podemos obtener información válida y confiable sobre determinada variable o variables, objetos o fenómenos de interés" (Cano, 1991, p.126). Para esta autora la construcción del examen, sea por selección, adaptación o creación de las preguntas que lo comprenden, debe llevarse a cabo en la forma más científica posible. Las interpretaciones que se hagan a partir de los resultados del examen obligan al constructor, o equipo de constructores, a tomar seriamente esta tarea.

Gronlund (1999) señala que los resultados del aprendizaje medidos por un test deberán reflejar fielmente los objetivos de enseñanza del curso. Lo que requiere, en primer lugar, enunciar claramente los resultados generales del aprendizaje y, en segundo lugar, hacer una lista de las conductas específicas que se aceptarán como prueba de que se han logrado los resultados. En línea con lo señalado por Gronlund diremos que lo segundo correspondería al listado de indicadores de evaluación. Un indicador de evaluación es un enunciado claro y preciso cuya redacción posibilite observar y medir el nivel de conocimiento del contenido o destreza en la habilidad, de aquello que se pretende evaluar.

Un examen es un tipo de test de rendimiento académico. Según Paz (1996) estos van dirigidos a evaluar el grado en que los sujetos dominan un campo de conocimiento. En un examen las preguntas no son vacías, ellas tienen un contenido y son formuladas intencionalmente para que, bajo un esquema conductista, cada una sea un estímulo que busque generar una respuesta del examinado. Usamos el término pregunta como sinónimo de ítem, esto es cada una de las partes de que se compone un examen y que tiene asignada una puntuación propia. Las preguntas del examen permiten hacer operables los indicadores de evaluación de modo que, en su conjunto, el examen evalúe aquello que pretende evaluar. Para hacer operativo un indicador, su redacción debe permitir que este sea observable y

medible. Su enunciado debe presentar verbo, contenido y condición. Consideremos, por ejemplo, el siguiente indicador: "Ind1: Resuelve inecuaciones racionales de una variable, en contexto intra-matemático". El verbo resuelve indica la acción, el contenido es inecuaciones racionales de una variable y la condición es que dicha inecuación debe presentarse en contexto intra-matemático.

Las siguientes tres preguntas son una muestra del universo de preguntas que operan el Ind1:

Preg1: Resolver: $\frac{7}{x-3} < 0$

Preg2: Resolver: $\frac{7}{x-3} \leq 1$

Preg3: Resolver: $\frac{x+4}{x-3} < 1$

Resolver una inecuación supone encontrar su conjunto solución. Las preguntas Preg1, Preg2 y Preg3 responden al contenido "inecuaciones racionales de una variable" y al "contexto intra-matemático" que enuncia el Ind1. Sin embargo, Preg1 y Preg2, a pesar de responder al mismo indicador y ser muy parecidas en su formulación, no miden exactamente lo mismo. Si bien se pueden pensar en distintos procedimientos, un análisis de la inecuación en Preg1 llevaría a concluir que el denominador debe ser negativo y por tanto que $x < 3$, esto es el conjunto solución CS1: $x \in]-\infty; 3[$. En Preg1 el signo $<$ nos lleva a asociarlo con un intervalo abierto y, en este caso, con un solo punto crítico 3. Mientras que en Preg2 debemos resolver la inecuación equivalente $\frac{7}{x-3} - 1 \leq 0$ o, lo que es lo mismo, $\frac{10-x}{x-3} \leq 0$ que lleva al análisis de los signos del numerador y denominador, considerar dos puntos críticos 3 y 10, así como determinar que 10 si pertenecería al conjunto solución pero 3 no, lo que nos lleva a CS2: $x \in]-\infty; 3[\cup [10; +\infty[$. La resolución de la inecuación en Preg2 requiere de la puesta en acción de más habilidades matemáticas que en la inecuación de Preg1. Lo mismo podríamos decir de Preg3 respecto a Preg1, a pesar que ambas inecuaciones son equivalentes. En la medida que Preg2 demanda un mayor número de procedimientos y análisis que Preg1, diremos que Preg2 es más difícil que Preg1.

La construcción de una prueba de rendimiento académico demanda, previamente, la elaboración de su tabla de especificaciones. Si bien existen distintos tipos de tablas de

especificaciones, una de las más comunes contiene, en las filas, el listado de contenidos temáticos del área a evaluar y, en las columnas, los niveles de complejidad. De este modo cada pregunta de la prueba está asociada con un área, un contenido temático y un nivel de complejidad. Los niveles de complejidad se refieren a categorías que indican las fases, etapas o niveles por los que atraviesan las operaciones cognitivas que debe mostrar el examinado al realizar la tarea establecida por la pregunta. Pueden referirse a capacidades, habilidades, niveles de logro de competencia o verbos tomados de alguna taxonomía que permitan hacer operables los objetivos planteados en el área. De aquí que suela asociarse el nivel de dificultad con el nivel de complejidad de la pregunta.

La dificultad de una pregunta está asociada con el nivel de conocimientos o habilidades observadas en un examinado al responder conscientemente dicha pregunta. Bajo el marco de la teoría clásica, la dificultad no es una característica per sé de la pregunta, sino que es dependiente del grupo de examinados. Y, para un mismo grupo de examinados, dependerá del momento en que la pregunta le es formulada. Así una misma pregunta puede ser considerada difícil para un grupo de examinados y fácil para otro; difícil en un momento dado y fácil en otro momento. Canales (2009) señala que, en sentido estricto, la dificultad de una pregunta no está en su naturaleza misma, sino en el aprendizaje que la pregunta evalúa. Una pregunta como “Calcular 237 + 541”, que evalúa la suma de dos números de tres cifras, puede ser difícil para estudiantes de primaria que solo han operado con números de dos cifras, pero puede resultar muy fácil para estudiantes de la secundaria acostumbrados a operar con números más grandes.

La dificultad de una prueba de rendimiento académico depende de la dificultad de las preguntas que ella comprende. Según el Diccionario de la Real Academia Española (DRAE), dificultad es embarazo, inconveniente, oposición o contrariedad que impide conseguir, ejecutar o entender algo bien y pronto. Piscoya (2005) señala que la dificultad está relacionada con el “grado de complejidad de los contenidos, procedimientos y mecanismos que se examinan o por la mayor o menor familiaridad de los usuarios con los temas que constituyen su contenido” (p.68). Respecto a una pregunta de una prueba de rendimiento académico y atendiendo a la definición del DRAE, diremos que una pregunta será fácil si presenta un alto porcentaje de acierto y será difícil si presenta un bajo porcentaje de acierto.

La teoría clásica del test (TCT) provee el marco teórico y los índices de más sencilla aplicación y análisis para el docente. Sin embargo, una de sus limitaciones es que los índices de dificultad calculados bajo la TCT son dependientes del grupo de personas a las que se aplicó la prueba y, por tanto, no gozan de la propiedad de invarianza. Así, una misma pregunta resultaría fácil si es aplicada a un grupo con alto dominio en el tema, pero resultaría difícil si se tratase de un grupo con poco dominio. Si se pretende una medición rigurosa, Muñiz (2010) indica que la ausencia de invarianza de las propiedades del test, respecto a las personas utilizadas para estimarlas, resulta algo inadmisible. El problema de la invarianza se va a resolver con la teoría de respuesta al ítem (TRI).

Yen y Fitzpatrick (2006) señalan que, usando la TRI, se ha podido calibrar un conjunto de ítems que miden un dominio particular, lo que ha posibilitado contar con bancos de ítems. Debido a que los ítems están debidamente calibrados, para Barbero (1996) un banco de ítems bien construido permitirá diseñar el test más adecuado para cada objetivo. En nuestro medio, si bien muchos docentes e instituciones educativas cuentan con bancos de preguntas, por lo general estos no incluyen preguntas cuyos niveles de dificultad estén calibrados en una escala común. Esta carencia, unida al desconocimiento de las teorías del test o el poco interés en aplicarlas, lleva a que cuando busca construir una prueba calibrada, el docente lo haga basado en su experiencia imaginando lo dificultoso que resultaría para sus examinados contestar correctamente cada pregunta de la prueba.

La dificultad de una pregunta suele medirse con el índice de dificultad (IDif) el cual es "la expresión numérica del grado en el que una pregunta resulta difícil de responder correctamente para el grupo al cual se le aplica" (Hurtado, 2018). Bajo el enfoque de la TCT y para una pregunta dicotómica, el IDif viene dado por la diferencia $1 - C/N$, donde N representa el total de examinados y C los que, de ellos, responden correctamente la pregunta. Una pregunta que es respondida correctamente por todos los examinados ($C = N$) tiene IDif igual a cero, mientras que, si no es respondida correctamente por ninguno ($C = 0$), tiene IDif igual a uno. De acuerdo a lo anterior, para medir la dificultad de una pregunta, a través del IDif, debemos registrar y contabilizar las puntuaciones otorgadas a las respuestas dadas por los examinados. Llamaremos a la dificultad calculada de este modo como Dificultad Observada y quienes proporcionan la información para su cálculo es el

grupo de examinados. Con frecuencia se usa la razón C/N como indicador de la dificultad de una pregunta (Reynolds et al., 2009; García y Fidalgo, 2005; Canales, 2005). De acuerdo a la razón C/N, cuanto mayor sea el número de personas que responden correctamente la pregunta, mayor sería su dificultad lo que resulta contrario al sentido de dificultad. Atendiendo a su significado, sería más correcto denominar a la razón C/N como índice de facilidad.

Una pregunta puede clasificarse como fácil, de dificultad media o difícil, atendiendo al IDif observado. Podemos encontrar distintas clasificaciones para el nivel de dificultad de las preguntas. El IDif varía de 0 a 1 y, dividiendo este intervalo en otros más pequeños, podemos valorar una pregunta según el sub-intervalo al que pertenezca su IDif. Esta valoración dependerá de la fórmula que se haya usado para calcular el IDif. Considerando como indicador de dificultad la razón C/N, Canales (2005) propone la clasificación de la tabla 1.

Tabla 1

Valoración de la dificultad de las preguntas según Canales

Índice de dificultad	Valoración
0,00 – 0,20	Muy difícil
0,21 – 0,40	difícil
0,41 – 0,60	Regular o media
0,61 – 0,80	Fácil
0,81 – 1,00	Muy fácil

Fuente: elaboración propia

En la evaluación psicométrica de las preguntas y pruebas crecer 96, Bazán (2000) mide la dificultad en base al porcentaje de aciertos en la pregunta, es decir bajo la fórmula C/N, y considera las categorías mostradas en la tabla 2.

Tabla 2

Clasificación de la dificultad de las preguntas según Bazán

Clasificación	Índice de dificultad
Muy fácil	0,75 – 1
Fácil	0,55 – 0,74
Intermedio	0,45 – 0,54
Difícil	0,25 – 0,44
Muy difícil	0,00 – 0,24

Fuente: elaboración propia

Una de las modalidades para asignar las calificaciones de los examinados es la referida a

normas. De acuerdo a Cueto (2007) las evaluaciones basadas en normas permiten definir cómo se ubica el rendimiento de un estudiante o grupo de estudiantes frente a otros (de la misma edad, grado de estudios u otro rasgo en común). Este tipo de modalidad permite hacer un ranking de los examinados. Ordenando, de mayor a menor el puntaje total obtenido, se podía identificar a los examinados de más alto rendimiento. El índice de discriminación es la expresión numérica de la medida en que una pregunta separa a los examinados de más alto rendimiento de los de más bajo rendimiento (Hurtado, 2018). Debido a que los examinados con puntajes altos en la prueba tienden a responder correctamente la pregunta, mientras que aquellos con puntajes bajos tienden a fallarla, uno de los métodos para medir el índice de discriminación es por medio de la correlación biserial pregunta-prueba. Cuanto mayor sea la correlación del puntaje en la pregunta y el puntaje total, al que previamente se le restó el puntaje de la pregunta, mayor será la discriminación.

El índice de discriminación (IDisc) puede tomar valores entre –1 y 1. Una pregunta con IDisc igual a 1 discrimina totalmente, mientras que, en el otro extremo, una pregunta con IDisc igual a –1 discrimina erróneamente y con IDisc igual a 0 no discrimina en absoluto. En la teoría tradicional del test, la alta discriminación es interpretada como una característica deseable de un ítem y un indicador clave de la calidad del ítem (Masters, 1988). El IDif y el IDisc de una pregunta están relacionados. Considerando sus valores extremos, diremos que una pregunta con IDif igual a 0 o a 1 tiene IDisc igual a 0, esto debido a que, si todos los examinados o ninguno de ellos responden correctamente la pregunta, esta no permite distinguir a los examinados de mayor y menor rendimiento. Mientras que una pregunta con IDif igual a 0,5 presenta una discriminación perfecta al tener un IDisc igual a 1.

Hay una medida de la dificultad de la pregunta obtenida a partir de las respuestas del grupo de examinados. Pero también, desde el lado del que selecciona la pregunta, hay una percepción de su dificultad. Basado en su experiencia, quien construye una prueba de rendimiento académico puede asignar un nivel de dificultad esperado para cada una de las preguntas que propone. Distintos indicadores para evaluar un mismo contenido pueden llevarnos a distintos niveles de dificultad. Así, por ejemplo, consideremos estas tres preguntas del contenido ecuaciones cuadráticas:

- Preg4: Resolver la ecuación $x^2 - 2x - 3 = 0$
- Preg5: Resolver la ecuación $x^2 - 2x = 1$
- Preg6: Encontrar el número real positivo cuyo cuadrado excede a su doble en 1.

Preg4 donde se presenta una ecuación cuadrática completa factorizable, suele ser más fácil de resolver que otra similar que involucre una expresión cuadrática no factorizable como la incluida en Preg5 y, a su vez, Preg5 suele ser más fácil que su equivalente Preg6 donde el examinado debería primero plantear la ecuación. Llamaremos Dificultad Asignada al nivel de dificultad otorgada por quien propone la pregunta. El nivel de dificultad asignado es una valoración subjetiva de la dificultad basada en cierto criterio o, generalmente, en la experiencia del docente que la propone. Backhoff et al., (2000), haciendo referencia al manual del EXHCOBA –examen de habilidades y conocimientos básicos aplicado a gran escala en México–, señalan que el nivel medio de dificultad del examen debe oscilar entre 0,5 y 0,6 distribuyéndose el total de reactivos en 5% fáciles, 20% medianamente fáciles, 50% con dificultad media, 20% medianamente difíciles y 5% difíciles. En la praxis resulta difícil que, solo basado en su experiencia, el docente constructor logre tal dosificación.

La taxonomía de Bloom describe procesos de pensamiento desde niveles inferiores hasta los superiores, contando con categorías y subcategorías que suelen servir de guía para asignar niveles de dificultad en las preguntas propuestas. Para Tristán (2001) "en la evaluación de conocimientos, cada división de la escala correspondería a un reactivo, dosificado en términos de la dificultad" (p.6). Si todos contestaran correctamente una pregunta, esta no permitiría establecer diferencias entre los examinados. Lo mismo podríamos decir de una pregunta que no es contestada correctamente por ninguno de los examinados. En este tipo de preguntas la varianza sería igual a cero.

En 2005, en su análisis de los ítems, García y Fidalgo refieren a Thorndike (1989) al señalar que, para que un ítem resulte de utilidad, es imprescindible que genere variabilidad entre las personas que lo responden. En una pregunta dicotómica la varianza es máxima cuando su índice de dificultad es 0,5. Por ello Reynolds et al., (2009) consideran a 0,5 como el nivel de dificultad óptimo. Esto indica que el 50% de los examinados respondieron correctamente la pregunta y el otro 50% la respondieron incorrectamente, o la dejaron sin responder. Buscando que el examen se comporte como una escala y mida con la mayor precisión posible, lo recomendable es que las preguntas cubran un amplio rango de niveles de

dificultad, procurando una distribución simétrica por encima y por debajo del nivel óptimo.

Según Muñiz (1997), el modelo de Rasch, es sin duda el modelo más popular de TRI. Con este modelo logístico de un parámetro es posible la construcción de una escala de intervalo conformada por preguntas según su nivel de dificultad y obtener medidas invariantes del nivel de habilidad de los examinados (Hurtado, 2012). Debido a que los niveles de habilidad (Bs) y dificultad (Di) están medidos en las mismas unidades, el modelo de Rasch señala que la probabilidad de responder correctamente una pregunta depende de la diferencia Bs–Di. Si se cuenta con un banco de preguntas, cuyos Di han sido estimados, es posible calibrar la prueba de rendimiento y estimar de una manera más fina los Bs de los examinados. En caso contrario, a partir del patrón de respuestas de los examinados y usando programas como el Winsteps, podemos estimar conjuntamente los Bs y Di. Un examinado con nivel de habilidad Bs debería responder correctamente aquellas preguntas tales que Di<Bs y responder incorrectamente, o dejar de responder, aquellas donde Di>Bs. Este es un modelo de comportamiento deseable, las respuestas de los examinados deberían ajustarse a este comportamiento ideal. Estadísticos de ajuste interno como el infit, sensibles al comportamiento inesperado que afecta a las preguntas cercanas al nivel de habilidad del examinado, o de ajuste externo como el outfit, sensibles al comportamiento inesperado que afecta a las preguntas alejadas del nivel de habilidad, permiten valorar que tan bien los datos cumplen los requerimientos del modelo de Rasch. Los valores de la media cuadrática MNSQ del infit y outfit tienen un valor esperado de 1. Según Schumacker (2004), los valores superiores a 1 indican ruido y los valores inferiores a 1 indican una falta de ajuste estocástico al modelo de Rasch. Schumacker hace referencia a la recomendación de Smith en usar el residuo estandarizado ZSTD para interpretar el ajuste de una pregunta, señalando que una buena regla para evaluar el ajuste del ítem ha sido descartar cualquier pregunta con un valor ZSTD mayor que 2 o menor que -2.

El objetivo general de este estudio fue determinar la coincidencia entre la dificultad asignada y la dificultad observada de cada una de las preguntas comprendidas en un examen. Se tomó en cuenta los tres niveles de dificultad (baja, media y alta) descritos. El estudio buscó principalmente responder la pregunta: ¿existe coincidencia total entre la dificultad asignada y la dificultad observada en las preguntas del examen? Y, en caso de no ser total, ¿cuál es

el grado de coincidencia?, ¿cuáles serían las posibles razones de la no coincidencia?

2. MÉTODO

Participantes

Para el presente estudio se tomaron las respuestas de 679 escolares a un examen que constituía la primera etapa de un proceso de selección en un concurso de becas para seguir estudios universitarios. La población estuvo conformada por estudiantes del quinto de secundaria, 49,5% mujeres y el 51,5% restante hombres, de colegios de alto rendimiento (COAR) de las distintas regiones del Perú inscritos voluntariamente en el concurso.

Instrumento

El examen comprendió 50 preguntas divididas en pruebas de cuatro áreas académicas según: 15 preguntas de Matemática, 15 de Física, 10 de Química y 10 de Comunicación. Todas las preguntas fueron de opción múltiple, con cuatro alternativas de respuesta, donde solo una era la opción correcta. La duración del examen fue de 120 minutos y la modalidad de aplicación online. Los estudiantes rindieron el examen en un laboratorio de cómputo del COAR al que pertenecían bajo la supervisión de un docente de la misma institución.

Procedimiento

Las preguntas del examen fueron construidas por un equipo de profesores de cada área. A ellos se les pidió asignar un nivel de dificultad a cada pregunta propuesta. Considerando el total de examinados, el criterio para asignar una pregunta como de dificultad alta fue considerar que, a lo más, el tercio superior podría responderla correctamente y, para la dificultad baja, que al menos el tercio inferior podría responderla correctamente. En la tabla 3 se describe el criterio para clasificar las preguntas según el nivel de dificultad asignada.

Tabla 3

Criterio para asignar el nivel de dificultad de las preguntas

Dificultad	Descripción
Baja	Si considera que más de las dos terceras partes de los examinados podrían responder correctamente.
Media	Si considera que al menos la tercera parte, pero a lo más las dos terceras partes de los examinados, podrían responder correctamente

Alta	Si considera que menos de la tercera parte de los examinados podrían responder correctamente.

Fuente: elaboración propia

En el anexo se muestra la tabla de especificaciones de la prueba de Matemática. Las 15 preguntas del área fueron etiquetadas según: P1 para la primera pregunta, P2 para la segunda pregunta y así hasta P15 para la última del área. De la tabla de especificaciones se puede observar que, en el diseño, hubo cierto sesgo del equipo constructor en direccionar la prueba de Matemática hacia un alto nivel de dificultad.

Para establecer rangos para el nivel de dificultad, según la descripción presentada en la tabla 3, usaremos la fórmula IDif= $1 - C/N$ (Hurtado, 2018). Donde C representa el número de examinados que respondieron correctamente la pregunta y N el número total de examinados. Además, consideraremos el hecho que, por su propia naturaleza C es una cantidad entera no negativa no mayor que N, esto es $0 \leq C \leq N$.

- Dificultad baja: Si $\frac{2}{3}N < C \leq N$ $\rightarrow$ $\frac{2}{3} < \frac{C}{N} \leq 1$
  $\rightarrow$ $-1 \leq -\frac{C}{N} < -\frac{2}{3}$
  $\rightarrow$ $0 \leq 1 - \frac{C}{N} < \frac{1}{3}$
  $\rightarrow$ $0 \leq$IDif$< \frac{1}{3}$
- Dificultad media: Si $\frac{1}{3}N \leq C \leq \frac{2}{3}N$ $\rightarrow$ $\frac{1}{3} \leq \frac{C}{N} \leq \frac{2}{3}$
  $\rightarrow$ $-\frac{2}{3} \leq -\frac{C}{N} \leq -\frac{1}{3}$
  $\rightarrow$ $\frac{1}{3} \leq 1 - \frac{C}{N} \leq \frac{2}{3}$
  $\rightarrow$ $\frac{1}{3} \leq$IDif$\leq \frac{2}{3}$
- Dificultad alta: Si $0 \leq C < \frac{1}{3}N$ $\rightarrow$ $0 \leq \frac{C}{N} < \frac{1}{3}$
  $\rightarrow$ $-\frac{1}{3} < -\frac{C}{N} \leq 0$
  $\rightarrow$ $\frac{2}{3} < 1 - \frac{C}{N} \leq 1$
  $\rightarrow$ $\frac{2}{3} <$IDif≤ 1

Las respuestas de los examinados a las preguntas del examen fueron registradas en una hoja de respuesta virtual que fue enviada y guardada en un documento Excel. La data se procesó y analizó, en conjunto y por cada área, obteniéndose los índices de dificultad y discriminación de las preguntas, así como otros índices psicométricos. Para medir la

confiabilidad de los resultados se calculó el coeficiente alpha de Cronbach con ayuda del SPSS y para medir el infit y outfit de las preguntas se usó Winsteps.

Cada respuesta correcta se calificó con 1 punto, las respuestas incorrectas o preguntas no respondidas fueron calificadas con 0 puntos. Para cada una de las preguntas se contabilizó el número de aciertos (C) y luego se aplicó la fórmula IDif= $1 - C/N$ con $N = 679$. Según el IDif obtenido, las preguntas fueron clasificadas siguiendo los rangos de la tabla 4, la misma que es la expresión numérica de la presentada en la tabla 3.

Tabla 4

Clasificación del nivel de dificultad observada de una pregunta

Dificultad	IDif
Baja	0 ≤ IDif < 1/3
Media	1/3 ≤ IDif ≤ 2/3
Alta	2/3 < IDif ≤ 1

Fuente: elaboración propia

3. RESULTADOS

La tabla 5 muestra los estadísticos descriptivos de las cuatro pruebas del examen, así como sus respectivos coeficientes alpha de Cronbach.

Tabla 5

Estadísticos descriptivos de las pruebas del examen

Área	Número de preguntas	Puntaje máximo	Puntaje mínimo	Media de los puntajes	Desviación estándar	Alpha de Cronbach
Matemática	15	14	0	7,47	2,30	0,457
Física	15	14	0	4,75	2,24	0,424
Química	10	8	0	3,76	1,63	0,227
Comunicación	10	9	0	5,02	1,67	0,317

Fuente: elaboración propia

En el presente estudio la exposición se centra en el análisis de las preguntas de la prueba de matemática. Ningún examinado obtuvo el puntaje máximo de 15 puntos, pero solo uno registró el puntaje mínimo de 0 puntos. La media de los puntajes fue de 7,47 puntos con una desviación estándar de 2,30 puntos. En Matemática podemos observar que los examinados no dejaron de responder ninguna de las preguntas del área. Hay preguntas con un alto número de aciertos, como P9 y P10, que podemos caracterizar como muy fáciles y otras muy difíciles como P1 y P8. Sin embargo, la media de los porcentajes de acierto ($100C/N$)

fue de 49,8%, lo que nos indica que podemos caracterizar a la prueba del área de matemática como de dificultad media. Si ordenamos las preguntas en forma decreciente a su porcentaje de acierto vemos que este va desde un 85,4%, para el caso de P10, hasta un 9,9% para el caso de P1.

La pregunta P1 fue respondida correctamente por 67 de los 679 examinados. Aplicando la fórmula IDif=$1 - C/N$, para $C = 67$ y $N = 679$, obtenemos un índice de dificultad igual a 0,90 para P1 que, de acuerdo con la clasificación de la tabla 4, corresponde a una pregunta con un nivel de dificultad alto. La tabla 6 muestra, para cada pregunta, el número de examinados que la respondieron correctamente, así como su respectivo IDif.

Tabla 6

Índices de dificultad de las preguntas de la prueba de Matemática

Pregunta	C	IDif
P1	67	0,90
P2	447	0,34
P3	195	0,71
P4	290	0,57
P5	297	0,56
P6	498	0,27
P7	448	0,34
P8	104	0,85
P9	530	0,22
P10	580	0,15
P11	275	0,59
P12	252	0,63
P13	264	0,61
P14	459	0,32
P15	363	0,47

Fuente: elaboración propia

Atendiendo a la tabla de especificaciones de las preguntas del área de Matemática y la clasificación presentada en la tabla 3 para la dificultad asignada, se obtuvo la distribución mostrada en la tabla 7.

Tabla 7

Distribución de las preguntas de la prueba de Matemática según la dificultad asignada

Dificultad	Descripción
Baja	P7, P9 y P11
Media	P1, P2, P3, P10, P13, P14 y P15
Alta	P4, P5, P6, P8 y P12

Fuente: elaboración propia

Atendiendo a los IDif mostrados en la tabla 6 y la clasificación señalada en la tabla 4 para la dificultad observada, se obtuvo la distribución mostrada en la tabla 8.

Tabla 8

Distribución de las preguntas de la prueba de Matemática según la dificultad observada

Dificultad	Descripción
Baja	P6, P9, P10 y P14
Media	P2, P4, P5, P7, P11, P12, P13 y P15
Alta	P1, P3 y P8

Fuente: elaboración propia

La tabla 9 muestra los IDisc de las preguntas de la prueba de matemática. Observamos que P1 presenta un IDisc negativo, mientras que P8 y P6, ambas con IDisc menor que 0,10, carecen de utilidad para discriminar. Sin considerar estas tres preguntas, el promedio del IDisc de la prueba de matemática es de 0,17 que consideramos con un bajo poder discriminatorio. La confiabilidad de las 15 preguntas de la prueba de matemática, obtenida por medio del coeficiente alpha de Cronbach, fue de 0,457 y sin considerar las preguntas P1, P6 y P8 esta aumenta a 0,487.

Tabla 9

Índices de discriminación de las preguntas de la prueba de Matemática

Pregunta	IDisc
P1	-0,07
P2	0,11
P3	0,22
P4	0,18
P5	0,28
P6	0,08
P7	0,23
P8	0,05
P9	0,22
P10	0,18
P11	0,13
P12	0,14
P13	0,17
P14	0,10
P15	0,21

Fuente: elaboración propia

La tabla 10 muestra los valores del infit y outfit de las preguntas de la prueba de matemática. Observamos que P1 presenta outfit con un valor ZSTD de 4,53 lo que indica un inesperado patrón de respuesta en P1 al ser respondida correctamente por examinados cuyos niveles de habilidad estaban alejados del nivel de dificultad de esta pregunta. La pregunta P5 presenta infit y outfit con valores ZSTD de -3,28 y -2,85, respectivamente, lo que indica una falta de ajuste estocástico esperado.

Tabla 10

Valores MNSQ del INFIT y OUTFIT de las preguntas de la prueba de Matemática

Pregunta	Infit		Outfit	
	MNSQ	ZSTD	MNSQ	ZSTD
P1	1,11	1,13	1,86	4,53
P2	1,03	0,98	1,04	0,76
P3	0,95	-1,17	0,91	-1,51
P4	0,99	-0,38	0,98	-0,49
P5	0,91	-3,28	0,89	-2,85
P6	1,05	1,14	1,13	1,74
P7	0,94	-1,68	0,92	-1,52
P8	1,06	0,85	1,21	1,80
P9	0,95	-1,02	0,86	-1,65
P10	0,96	-0,47	0,90	-0,81
P11	1,03	0,93	1,04	0,99
P12	1,01	0,35	1,02	0,53
P13	1,00	-0,14	1,00	-0,05
P14	1,03	0,87	1,10	1,67
P15	0,96	-1,52	0,93	-1,78

Fuente: elaboración propia

4. DISCUSIÓN

De las 15 preguntas analizadas de la prueba de Matemática, observamos en las tablas 7 y 8 que en cinco de ellas existe coincidencia entre la dificultad observada y la dificultad asignada. Con el fin de obtener una medida del grado de coincidencia entre la dificultad asignada y observada en las preguntas de cada prueba del examen, calcularemos la razón del número de preguntas coincidentes al número total de preguntas del área. Multiplicando por 100 el valor de dicha razón tendríamos el porcentaje de coincidencia en la prueba del área. Así, por ejemplo, para el caso de la prueba de Matemática, el cálculo sería (5/15) ×100 que corresponde a un 33,3% de coincidencia. Si en una prueba, al comparar la dificultad asignada y la dificultad observada, todas las preguntas coincidieran, tendríamos un 100%

de coincidencia. Por el contrario, si ninguna pregunta coincidiera, tendríamos un 0% de coincidencia.

En el área de Matemática la coincidencia se dio en P9 para el nivel de dificultad baja; en P2, P13 y P15 para el nivel medio y en P8 para el nivel de dificultad alto. En particular P8 presentó un alto IDif pero un IDisc positivo cercano a cero, por lo que carecía de utilidad para discriminar. Respecto a las otras cuatro preguntas coincidentes P2, P9, P13 y P15, observamos que presentan bajos índices de discriminación, pero estadísticos de ajuste interno y externo dentro de los rangos aceptables.

Las diez preguntas donde no se encontró coincidencia fueron P1, P3, P4, P5, P6, P7, P10, P11, P12 y P14. De este grupo hay tres preguntas en las que sus índices indican ruido estadístico. P1, con un IDisc negativo, presenta un comportamiento erróneo, ya que fueron los examinados de más bajo rendimiento, y no los de más alto rendimiento, los que respondieron correctamente la pregunta. Esto es consistente con el comportamiento inesperado que indica su alto valor ZSTD outfit, P1 fue respondida correctamente por examinados con niveles de habilidad alejados de su nivel de dificultad, indicando que hubo aciertos por efecto de la adivinación. P5 fue la pregunta con el más alto valor del IDisc en la prueba de matemática, pero al mismo tiempo presentó valores negativos del infit y outfit, menores que -2, que indicaron falta de ajuste al modelo. Se esperaba que P5 fuera una pregunta difícil para el grupo, pero resultó de dificultad media. P6, con un IDisc bajo que lo hace carecer de utilidad para discriminar, presentó un comportamiento contrario a lo esperado. Fue asignada como de dificultad alta, pero resultó de dificultad baja. Sus valores de ajuste interno y externo se encuentran dentro de los rangos aceptables.

En las otras siete preguntas, si bien presentan bajos índices de discriminación, no encontramos desajustes con el modelo de Rasch. En cuatro de ellas, P4, P10, P12 y P14, observamos desplazamientos positivos en tanto fueron asignadas en un nivel de dificultad mayor que el observado. Consideramos positivo el desplazamiento de una pregunta que se asignó con dificultad media y se observó dificultad baja (P10 y P14) o aquella en la que se asignó dificultad alta y se observó dificultad media (P4 y P12). Las otras tres preguntas presentaron desplazamientos negativos al asignarse un nivel de dificultad menor que el

observado. P3 fue asignada con dificultad media, pero se observó dificultad alta; mientras que P7 y P11 fueron asignadas con dificultad baja y se observó dificultad media.

Siguiendo el procedimiento descrito líneas arriba, se compararon las coincidencias de las preguntas de las otras tres pruebas del examen y calcularon sus respectivos porcentajes de coincidencia. En la tabla 9 se presentan los resultados obtenidos.

Tabla 9

Porcentaje de coincidencia de las pruebas del examen

Área	%
Matemática	33,3
Física	46,7
Química	70,0
Comunicación	30,0

Fuente: elaboración propia

5. CONCLUSIONES

Se concluye que, para el examen descrito, el nivel de dificultad asignada al momento de seleccionar, editar o construir una pregunta no se corresponde con el nivel de dificultad observada de la misma pregunta, medida a través del IDif, por aplicación del examen. Nuestro análisis no profundizó en las razones del bajo porcentaje de coincidencia. Podemos señalar cuatro posibles razones de esta distorsión.

Asignar como de nivel de dificultad alta a una pregunta que resultó de dificultad baja, o viceversa, indicaría poco conocimiento de los docentes constructores acerca de la población escolar que rindió el examen. De este modo, en el diseño de la prueba, se habría subestimado o sobreestimado el nivel de dominio de los examinados en algunos contenidos evaluados. Los desplazamientos positivos o negativos en los niveles de dificultad de las siete preguntas descritas respaldarían esta posibilidad.

Las condiciones de aplicación del examen habrían afectado la confiabilidad de los resultados. Los estudiantes debieron descargar y abrir el documento con las preguntas de cada prueba, el mismo que mantuvieron a la vista en la pantalla de la PC durante el examen. Las respuestas se registraron en una tarjeta virtual y luego fueron enviadas al término del examen. Si bien los examinados estaban informados del modo de aplicación, este no era

algo a lo que estén acostumbrados y podría haber afectado su rendimiento. Los bajos valores del alpha de Cronbach de cada prueba respaldarían esta posibilidad.

A pesar del cuidado puesto en la construcción de las preguntas de las pruebas, siguiendo recomendaciones técnicas, la redacción de los enunciados de las preguntas podrían no haber sido claros para los examinados. Los bajos índices de discriminación en las preguntas soportarían esta posibilidad ya que la discriminación de las preguntas depende en gran medida de su calidad técnica. Por otro lado, la confiabilidad depende, entre otras cosas del número de preguntas en la prueba. Es mayor cuanto más altos son los índices de discriminación de sus preguntas y cuanto más homogénea sea la prueba. Las cuatro pruebas cubrieron diversos contenidos, lo que habría afectado su confiabilidad. Haber incluido solo diez preguntas que cubrían una variedad de contenidos podría explicar la baja confiabilidad en la prueba de Química, aunque no deja de llamar la atención que esta área presentó el mayor porcentaje de coincidencia.

La no coincidencia entre el nivel de dificultad asignado y el observado podría deberse al efecto de la adivinación. En las instrucciones del examen se indicaba que las preguntas respondidas incorrectamente no tendrían puntaje en contra, lo que habría animado a marcar una opción sin estar seguro de ella. Esto habría posibilitado que algunos examinados, o bien por azar o bien por intuición, hayan marcado la opción correcta. Desajustes observados a partir del outfit, como en P1, respaldarían esta posibilidad.

REFERENCIAS

Ander-Egg E. (2012). Diccionario de educación. 1ra edición, Córdoba – Editorial Brujas. (p.19, p.84)

Alvaro Page, Mariano et al. (1990) Hacia un modelo causal del rendimiento académico. Centro de Publicaciones del Ministerio de Educación y Ciencia CIDE, Madrid.

Backhoff, E., Larrazolo, N. y Rosas, M. (2000). Nivel de dificultad y poder de discriminación del Examen de Habilidades y Conocimientos Básicos (EXHCOBA). Revista Electrónica de Investigación Educativa, 2 (1). Consultado el día 27 de setiembre de 2019 en: https://redie.uabc.mx/redie/article/view/15

Barbero, I. (1996). Bancos de ítems, en Psicometría, Muñiz, J. (Coord.). Madrid, Editorial Universitas, S.A. (pp.139-170).

Canales, I. (2009). Evaluación Educacional. Programa de Licenciatura para profesores sin Título Pedagógico en Lengua Extranjera. Facultad de Educación de la Universidad Nacional Mayor de San Marcos, Lima. (p.58)

Cano, Y. (1991). Hacia una evaluación científica. Lima, Grafimag.

Cueto, S. (2007). Las evaluaciones nacionales e internacionales de rendimiento escolar en el Perú: balance y perspectivas. En Investigación, políticas y desarrollo en el Perú. Lima: GRADE. p. 405-455. Disponible en: http://www.grade.org.pe/upload/publicaciones/archivo/download/pubs/InvPolitDesarr-10.pdf

Ebel, R. (1997). Fundamentos de la Medición educacional. Buenos Aires, Editorial Guadalupe.

Educación. (2003). En Diccionario Enciclopédico de Educación, Barcelona, España. Grupo Editorial Ceac S.A. (p.147)

García-Cueto, E. y Fidalgo, A. (2005). Análisis de los ítems. En Muñiz, Fidalgo, García-Cueto, Martínez y Moreno (2005). Análisis de los ítems. Cuadernos de Estadística, Madrid, Editorial La Muralla. (pp.53-131)

Gómez Castro, J.L. "Rendimiento escolar y valores interpersonales: Análisis de resultados en EGB con el cuestionario SIV de Leonardo V. Gordon". Bordón, n° 262, pp. 257-275, 1986.

Gronlund, N. (1999). Elaboración de tests de aprovechamiento. México, Editorial Trillas (p.28, p.32)

Hurtado, L. (2012). Análisis de pruebas de rendimiento usando el modelo de Rasch. San Bernandino, Editorial académica española.

Hurtado, L. (2018). Relación entre los índices de dificultad y discriminación. Revista Digital de Investigación en Docencia Universitaria, 12(1),273-300. doi: http://dx.doi.org/10.19083/ridu.12.614

Masters, G. (1988). Item discrimination: When more is worse. Journal of Educational Measurement, 25(1), 15-29. doi: https://doi. org/10.1111/j.1745-3984. 1988.tb00288.x

Miljánovich, M. (2005). Nuevo modelo de prueba de admisión de la Universidad Nacional Mayor de San Marcos, en Calidad de las Pruebas de Admisión de la Universidad Peruana, Piscoya L. (Ed.) Lima, Asamblea Nacional de Rectores (pp.103-112).

Muñiz, J. (1997). Introducción a la teoría de respuesta a los ítems. Madrid, Ediciones Pirámide (p.34).

Muñiz, J. (2010). Las teorías de los tests: teoría clásica y teoría de respuesta a los ítems. Papeles del psicólogo, vol 31, núm. 1, pp.57-66. ISSN:0214-7823. Disponible en https://www.redalyc.org/pdf/778/77812441006.pdf

Paz, M. (1996). Validez, en Psicometría, Muñiz, J. (Coord.). Madrid, Editorial Universitas, S.A. (pp.50-103).

Piscoya, L. (2005). Cuánto saben nuestros maestros. Fondo Editorial de la Universidad Nacional Mayor de San Marcos, Lima.

Rama, C. (2005). El acceso a la educación superior en América Latina y el Caribe, en Calidad de las Pruebas de Admisión de la Universidad Peruana, Piscoya L. (Ed.) Lima, Asamblea Nacional de Rectores (pp.17-46).

Reynolds, Livingston y Wilson (2019). Measurement and Assessment in Education, 2da. Edición. Pearson. (p.149)

Schunck, D. (1997). Teorías del aprendizaje. Pearson educación, México. (p.2, p.3)

Schumacker, R. (2004). Rasch Measurement: The dichotomous model, en Introduction to Rasch Measurement. Smith, E. y Smith, R. (Editores). Minessota, JAAM Press, (pp. 226-257).

Tristán, A. (2001). Análisis de Rasch para todos. México, Ceneval.

Yen, W. y Fitzpatrick, A. (2006). Item Response Theory, en Educational Measurement, Fourth Edition. Brennan, R. (Editor). American Council on Education and Praeger Publishers, (pp.111-154)

Anexo

Tabla de Especificaciones del área de Matemática

Indicador	Nivel de dificultad			Total
	Baja	Media	Alta	
Evalúa numéricamente fórmulas relacionadas con la ingeniería redondeando el resultado a dos cifras decimales.		P1		1
Modela fórmulas a partir de enunciados sobre relaciones de proporcionalidad entre magnitudes.		P2		1
Resuelve inecuaciones racionales, lineales y cuadráticas, de una variable, en contexto intramatemático.	P11			1
Resuelve problemas de contexto real que involucren la aplicación del teorema de Pitágoras, semejanza de triángulos o triángulos notables.		P13		1
Calcula perímetros y áreas de figuras geométricas compuestas descomponiéndolas en figuras geométricas conocidas.		P3	P4	2
Calcula volúmenes de sólidos compuestos descomponiéndolos en sólidos geométricos conocidos.	P6		P5	2
Resuelve problemas de contexto real que impliquen la resolución de triángulos usando razones o leyes trigonométricas.		P14		1
Identifica la variable independiente y dependiente en situaciones reales que pueden ser modeladas por medio de una función.			P7	1
Describe las principales características de una función de una variable a partir de su representación gráfica.			P8	1
Representa gráficamente una función de una variable a partir de su representación algebraica.	P9			1
Plantea funciones lineales, cuadráticas o definidas por tramos a partir de enunciados y condiciones.		P15	P12	2
Modela una situación real mediante funciones lineales o cuadráticas, a partir de enunciados y condiciones.		P10		1
Total	3	7	5	15