Anales Científicos. 81(2), 278-288 (2020)

Anales Científicos

ISSN 2519-7398 (Versión electrónica)

ARTÍCULO ORIGINAL – RESEARCH ARTICLE

http://dx.doi.org/10.21704/ac.v81i2.1638

EVALUACIÓN DE PRUEBAS INFORMATIZADAS APLICANDO LA TEORÍA CLÁSICA DE LOS TEST Y LA TEORÍA DE RESPUESTA AL ÍTEM

Evaluation of computerized tests applying the classical theory of the tests and the theory of response to the item

César Higinio Menacho Chiok[1]*; Jesús María Cano Alva Trinidad[1]

[1] Facultad de Economía y Planificación, Universidad Nacional Agraria La Molina, 15024, Lima, Perú.
* E-mail: cmenacho@lamolina.edu.pe

Recibido: 28/11/2019; Aceptado: 15/12/2020; Publicado: 31/12/2020

ABSTRACT

The objective of this study was to evaluate the reliability and validity of computerized tests via the web through the measurement of their psychometric and statistical properties by applying the Classical Test Theory (TCT) and the Item Response Theory (TRI). The TCT methodology was applied to assess the difficulty and discrimination of the test and the items. The data was adjusted to the TRI binary logistic models of one, two and three parameters. A computerized test of 30 questions was applied to 775 students enrolled in the Basic Statistics course in the 2016 II semester. The results indicated a good reliability of the test with a Cronbach's alpha of 0,833 and was corroborated with a correlation of 0,815. For the TCT the difficulty index identified three very easy questions (V7, V8 and V12) and the discrimination index did not find any questions to withdraw it. The assumption of unidimensionality with factor analysis was tested with an explained variance of the first factor of 24,7%. The binary logistic model of the three parameter TRI (3PL) was better adjusted to the data. For the calibration process with the 3PL model, questions V28 (discrimination index greater 0,65) were withdrawn; V8, V12, V16 and V18 (chance index greater than 0.4) and none with the difficulty index.

Keywords. Computerized tests; classic test theory; item response theory; binary logistic models; test calibration.

RESUMEN

El objetivo del presente estudio fue evaluar la confiabilidad y validez de las pruebas informatizadas vía web a través de la medición de sus propiedades psicométricas y estadísticas aplicando la Teoría Clásica del Test (TCT) y la Teoría de Respuesta al Ítem (TRI). Se aplicó la metodología de la TCT para evaluar la dificultad y de discriminación del test y los ítems. Se ajustaron los datos a los modelos logísticos binarios TRI de un, dos y tres parámetros. Un test informatizado de 30 preguntas se aplicó a 775 estudiantes matriculados en el curso de Estadística Básica en el semestre 2016 II. Los resultados indicaron una confiabilidad buena del test con un alfa de Cronbach de 0,833 y fue corroborada con una correlación de 0,815. Para la TCT el índice de dificultad identificó tres preguntas muy fáciles (V7, V8 y V12) y el índice de discriminación no encontró ninguna pregunta para retirarla. El supuesto de la unidimensionalidad con el análisis factorial fue probado con una variancia explicada del primer factor de 24,7%. El modelo logístico binario de la TRI de tres parámetros (3PL) se ajustó mejor a los datos. Para el proceso de

calibración con el modelo 3PL, se retiraron las preguntas V28 (índice de discriminación mayor 0,65); V8, V12, V16 y V18 (índice del azar mayor a 0,4) y ninguna con el índice de dificultad.

Palabras clave: Pruebas informatizadas; teoría clásica de los test; teoría de respuesta al ítem; modelos logísticos binarios; calibración de la prueba**.**

Forma de citar el artículo (Formato APA):
Menacho, C.H., & Alva, J. (2020). Evaluación de pruebas informatizadas aplicando la teoría clásica de los test y la teoría de respuesta al ítem. *Anales Científicos*. 81(2), 278-288. http://dx.doi.org/10.21704/ac.v81i2.1638

Autor de correspondencia (*): César Higinio Menacho Chiok. Email: cmenacho@lamolina.edu.pe

1. INTRODUCCIÓN

Las instituciones de educación superior están integrando cada vez más en sus ambientes educativos plataformas virtuales basada en la web (e-learning), con la finalidad de apoyar sus procesos de enseñanza y aprendizaje, así como implementar evaluaciones informatizadas que permitan medir con mayor precisión los conocimientos adquiridos por los estudiantes. Las pruebas informatizadas permiten muchos beneficios, tales como: generar un banco de preguntas, controlar y medir los procesos de aprendizaje, brindar información inmediata de los resultados a los profesores, mayor accesibilidad de evaluados, abaratar costos y el análisis de la validez de las puntuaciones y seguimiento de las propiedades psicométricas de la prueba (Olea et al., 1999). En las últimas décadas hay un creciente uso de la psicometría para construir pruebas (tests) educativas de calidad y aplicar modelos matemáticos y estadísticos que permitan medir y evaluar su confiabilidad y su validez; así como de las preguntas (ítems) que la conforman con el fin de mejorar el proceso de evaluación de los conocimientos adquiridos, las habilidades específicas y otras funciones cognitivas que los estudiantes realizan como parte de sus tareas de aprendizaje. En el entorno educativo universitario, los test informatizados, están permitiendo cambios sustanciales en la forma en que se está evaluando el aprendizaje de los estudiantes en sus diferentes cursos matriculados.

Dentro de la psicometría existen dos enfoques para la construcción y la medición de pruebas: la Teoría Clásica de los Test (TCT) y la Teoría de Respuesta al Ítem (TRI) (Martínez, 1990). La TCT usa un modelo simple para evaluar la habilidad cognitiva (rasgo latente) que se quiere medir en un test, mientras que la TRI emplea modelos matemáticos con supuestos más rigurosos y con análisis de resultados mayores que la TCT. Los modelos de la TRI usan una función monótona (gaussiana o logística) para establecer una relación no lineal entre la probabilidad de una respuesta correcta y la habilidad del sujeto evaluado. Asimismo, en la TRI se pueden definir varios modelos de uno (1PL), dos (2PL) y tres (3PL) parámetros los cuales están asociados a la dificultad (b), la discriminación (a) y la probabilidad de acertar por el azar (c) un ítem. Se aplican la TCT y la TRI, a un caso de estudio con los datos de un test informatizado aplicado en una plataforma virtual que consta de 30 preguntas (ítems) que se aplicó a 775 estudiantes de una institución educativa superior matriculados en el curso de

Estadística Básica durante el semestre 2016 Los test informatizados son implementados en sistemas computacionales que evalúan a millones de personas cada año (Davey, 2005). En Gonzáles et al. (2010) se propone un modelo logístico de tres parámetros (3LP) para medir pruebas computarizadas, incorporando un parámetro que relaciona el tiempo de respuesta a un ítem con el nivel de logro o competencia de los estudiantes en matemáticas; bajo la suposición: si un ítem requiere de mayor tiempo para su respuesta, esto implicará mayor dificultad y discriminación para el alumno. Los resultados con datos simulados indican la necesidad de incorporar al modelo logístico el tiempo de respuesta a un ítem como un factor diferenciador entre las puntuaciones de los estudiantes; de tal manera, que se mejora la precisión en los procesos de medición de los conocimientos matemáticos. En Bulut (2015), se aplican los modelos de la TRI al examen de admisión para estudios de posgrado en universidades de Turquía para un total de 142178 postulantes, de los cuales se extrajo una muestra aleatoria de 5000, a cuyos datos se

les aplicó las pruebas de razón máxima verosimilitud para determinar el mejor modelo TRI de 1PL, 2PL y 3PL. Los resultados indicaron que el modelo TRI de tres parámetros es el que proporciona el mejor ajuste de datos del modelo para el examen de ingreso a estudios de posgrado. Además, los resultados de este estudio destacan problemas potenciales que deben abordarse, como las altas tasas de omisión, la aceleración de la prueba y los comportamientos de adivinanzas aberrantes.

En Zanon et al. (2016), se aplica la TRI para desarrollar pruebas psicológicas para una muestra de 853 estudiantes universitarios entre las edades de 17 y 35 años. El análisis con la TRI muestra una mayor medición con ítems positivos para el índice discriminatorio de los encuestados por debajo del puntaje promedio y para la escala de afecto negativo también presentó ítems con discriminación moderada. Se concluye que la utilización de la TRI en las pruebas de evaluación, pueden tener como resultado una mejora de la medición de las escalas y permiten refinar las pruebas y aumentar la validez y confiabilidad en las medidas psicológicas. En Sudol & Studer (2010), se plantea la potencialidad de aplicar la TRI para construir prácticas calificadas para evaluar los conocimientos aprendidos de los estudiantes en un curso de informática. Para evaluar la confiabilidad de los ítems (preguntas) se analiza el índice de dificultad para varias CCI, con la finalidad de identificar las preguntas fáciles y difíciles. Las diferentes CCI permiten analizar las respuestas de los alumnos identificando patrones de las habilidades o rasgos que el instructor desea evaluar en las preguntas. En He & Tymms (2004), se presenta el desarrollo de un sistema de pruebas (CADATS) que involucra el diseño asistido por ordenador que puede ser utilizado por escuelas y otras organizaciones educativas para llevar a cabo pruebas y su evaluación por computadora. El sistema incorpora el modelo TRI Rasch (2LP) para facilitar la administración de las pruebas basadas en computadoras, el análisis de los ítems de prueba, el rendimiento de la prueba, además incluyen realizar el análisis del diagnóstico sobre el rendimiento de los estudiantes a nivel individual, identificando las áreas curriculares donde tienen bajo rendimiento académico.

A pesar de las diferencias teóricas entre la TCT y la TRI, se han hecho estudios cuantitativos para comparar las propiedades de estos modelos. Aunque como menciona Muñiz (1997), el modelo TRI proporciona las mejores soluciones a los problemas que presenta la TCT y representa un cambio del modelo de la teoría de los test; esta no llega a ser una teoría contrapuesta sino complementaria al modelo clásico. Una revisión completa de las ventajas e inconvenientes de la TRI con respecto a TCT se encuentra en Navas (1994). En Omobola & Adedoyin (2013), se realiza un estudio con muestra aleatoria de 10000 estudiantes seleccionados de una población de 36940 que rindieron una prueba de examen de matemáticas, con el propósito de evaluar la comparabilidad de los estimadores de los parámetros de los ítems entre los modelos de la TCT y TRI (3PL). Los resultados mostraron que los valores para la dificultad y la discriminación para la TCT y TRI resultaron positivamente correlacionados y no había ninguna diferencia estadística significativa entre los estimadores de estos parámetros.

En Progar et al. (2008), se estudia la comparabilidad de los parámetros sujeto/ítem, invariancia para diferentes grupos de participantes e invariancia para diferentes conjuntos de ítems en los modelos TCT y TRI. Para el análisis empírico se usó los datos del Tercer Estudio Internacional de Matemáticas de datos y Ciencias (TIMSS 1995). Los parámetros de los sujetos y los de dificultad para la TCT y TRI, resultaron con valores muy similares y en el caso de la discriminación de los sujetos hubo diferencias. La invariancia de los parámetros de los ítems, fueron similares para todo el conjunto de participantes. Hay un buen ajuste del modelo TRI, los parámetros de los ítems son generalmente más invariantes que los de la TCT, mientras que los parámetros del TCT son más invariantes en el caso de un pobre ajuste del modelo TRI.

El objetivo de la investigación fue evaluar la confiabilidad y validez de las pruebas informatizadas midiendo las propiedades psicométricas y estadísticas aplicando la TCT y la TRI.

2. MATERIALES Y MÉTODOS

La matriz de datos X, es de dimensión Nxp, correspondiente a los N estudiantes y p preguntas (ítems). Los datos de la matriz de datos son binarios:

Xij=1 ó 0, si el estudiante i respondió correcta o incorrectamente la pregunta. Sumando cada fila de la matriz de datos, se obtiene para cada estudiante su puntuación total de la prueba. Para el análisis estadístico de la TCT y la TRI, se aplicó el programa R: Package ltm para ajustar los datos a los modelos TRI logísticos binarios de un, dos y tres parámetros.

Teoría clásica de los test (TCT)

La TCT fue la primera teoría psicométrica para la construcción y evaluación de un test, propuesto por Spearman (1913). El modelo de la TCT asume que la puntuación que una persona obtiene en un test, denominada puntuación empírica (X), es una relación lineal formada por la suma de dos componentes hipotéticos desconocidos a priori: la puntuación verdadera (V) o habilidad real del sujeto evaluado y un error de medición (e) que puede ser debido a factores no controlados (factores aleatorios). Así, el modelo clásico lineal propuesto por Spearman se expresa: X = V + e. La principal limitación de la TCT se refiere a la dependencia de los examinados con el test (las puntuaciones obtenidas dependen del test utilizado y las propiedades de los tests dependen de los sujetos a quienes se les aplica). Además, en la TCT se calcula la dificultad de una pregunta en función de la cantidad de individuos que la responden correctamente (cuanto mayor sea el número de evaluados que responden bien a un ítem tanto más fácil se considerará la pregunta).

Evaluación del test y de los ítems en la TCT

Se supone que el test con su conjunto de ítems ha sido elaborado y formulado de manera lógica para que mida la variable o rasgo que interesa evaluar. En la TCT se puede comprobar y analizar estadísticamente la confiabilidad y validez del test y sus ítems a través de varios procedimientos.

La confiabilidad de un test

Para medir la confiabilidad o fiabilidad de un test se aplica el Coeficiente de Alfa de Cronbach, se expresa por:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum S_j^2}{S_x^2}\right)$$

Donde, k es el número de ítems,

$\sum S_j^2$ es la suma de las variancias de cada uno de los ítems

$S^2_{x_i}$ es la variancia del test (total de los ítems)

Según George & Mallery (1995), un Alfa de Cronbach dentro de [0; 0,5) muestra un nivel de fiabilidad no aceptable; entre [0,5; 0,6) un nivel pobre; entre [0,6; 0,7) un nivel débil; entre [0,7; 0,8) un nivel aceptable; entre [0,8; 0,9) un nivel bueno y
entre [0,9; 1] sería excelente.

Validez de un test

Un test es válido si mide realmente aquello que pretende medir. Para evaluar la validez de un test se aplica la técnica multivariada del Análisis Factorial (exploratorio o confirmatorio) sobre la matriz de correlaciones entre ítems, identificando las variables asociadas a los factores (dimensiones subyacentes).

Validez de los Ítems

Consiste en evaluar si cada ítem es válido para ser considerado en el test que pretende medir el rasgo o habilidad. En la TCT se usan: el alfa de Cronbach, el índice de dificultad y el índice de homogeneidad.

• *Coeficiente de alfa de Cronbach para un ítem*

El coeficiente alfa de Cronbach corregido se calcula retirando el ítem a ser evaluado y usando sólo los ítems restantes. El criterio es, retirar el ítem cuyo coeficiente alfa de Cronbach corregido supera al coeficiente global.

• *Índice de dificultad de un ítem (IDj*).

Cuantifica el grado de dificultad de cada ítem. El Dj para el ítem j, se calcula como el cociente entre el número de sujetos que lo han acertado (Aj) y el número total de sujetos que sólo lo contestaron (Nj). Esto es:

$$D_j = \frac{A_j}{N_j}$$

ID =El valor mínimo de IDj es 0 (ningún sujeto acierta el ítem) y el máximo 1 (todos los sujetos que lo intentaron lo acertaron).

Los niveles de aceptación son: [0,0 a 0,2) muy difícil (retirar el ítem); [0,2 a 0,3) difícil; [0,3 a 0,7) medio; [0,7 a 0,8) fácil y [0,8 a 1,0] muy fácil (retirar el ítem).

• *Índice de homogeneidad de un ítem (Hj).* El Hj llamado a también índice de discriminación, se calcula por la correlación entre las puntuaciones de los N sujetos en el ítem j y la puntuación total de todos los ítems del test (X). Se expresa por:

$$H_j = r_{jx}$$

El Hj mide la consistencia interna del test. Un ítem es bueno, si es acertado por los sujetos con mayor puntuación y no acertado con los de menores puntuaciones en el test (discrimina a los sujetos con mayores y menores puntuaciones en el test). Los ítems buenos tendrán un Hj alto y positivo (correlación positiva) y deberían eliminarse los ítems del test que tienen un Hj próximo a cero o cuando es negativo y alto.

Teoría de Respuesta al Ítem (TRI)

La TRI, surge como una reacción a los problemas y limitaciones que presenta la TCT (Lord, 1952). La TRI, supera la limitación de la TCT con modelos estadísticos orientados a los ítems (preguntas) y cuya ventaja es que la habilidad del examinado y la dificultad de los ítems se miden en la misma escala, facilitando la comparación. En la TRI cada ítem se define por una función matemática no lineal (logística o normal), expresada por la llamada Curva Característica del Ítem (CCI), que muestra la función de probabilidad de la respuesta correcta y la habilidad del sujeto evaluado. Los modelos de la TRI, establecen tres asunciones sobre los parámetros de los ítems: unidimensionalidad (los ítems sólo miden una única habilidad o rasgo), independencia local (cada ítem es independiente de todos los demás) e invarianza (la propiedades de los ítems no dependen de la habilidad o rasgo), independencia local (cada ítem es independiente de todos los demás) e invarianza (la propiedades de los ítems no dependen de la habilidad de los sujetos y el nivel estimado del evaluando).

Los modelos de la TRI se basan en establecer una relación entre los valores del rasgo o habilidad θ a evaluar y su probabilidad P(θ), por lo cual usa una función monótona creciente para valores asintóticos entre 0 y 1. La formulación general del modelo unidimensional (un sólo rasgo en estudio) y de respuesta dicotómica para cada ítem corresponde a la función logística binario para tres parámetros (3PL):

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}}, \quad 1,\ldots,n \qquad (1)$$

En (1), Pi(θ) representa la probabilidad de que un sujeto con nivel de habilidad o rasgo θ conteste correctamente el ítem i con un nivel de dificultad bi, un nivel de discriminación ai y un nivel del azar ci. Si ci = 0, se tiene el modelo logístico binario de dos parámetros (2PL) y cuando ci = 0 y ai = 0, el modelo logístico binario de un parámetro (1PL). El parámetro del test θ, es el nivel de habilidad (rasgo, conocimiento, aptitud, etc.) que un sujeto presenta al responder un test. En la práctica se usa una escala típica (estandarizada) con media cero y varianza uno y con un rango de valores entre -3 y 3. Los parámetros de los ítems definen la característica de cada ítem en los modelos de la TRI. El parámetro ai (índice de discriminación), mide el cambio (pendiente en la CCI) de la probabilidad de acertar el ítem conforme varíe el nivel de habilidad y toma valores entre 0 y 3, considera un ítem discriminatorio cuando es mayor a 1. El parámetro bi (índice de dificultad), es el valor de la abscisa en la escala de habilidad θ (máxima pendiente de la CCI), es el valor de θ para el cual P (θ = 0,5. Toma valores entre -4 y 4, cuanto mayor sea bi más difícil es el ítem. El parámetro ci (i índice de azar), representa la probabilidad que acierten los sujetos por el azar la respuesta correcta, es el valor P (θ) cuando θ tiende al su valor mínimo (– ∞). Toma valores entre 0 y 0,5.

Evaluación del test, los ítems y los modelos en la TRI

En la TRI existen varios procedimientos para validar el test y los ítems y probar la bondad de ajuste de sus modelos.

Curva característica del ítem (CCI)

La CCI es un gráfico que muestra la relación entre el nivel de habilidad y la probabilidad de responder correctamente cada uno de los ítems. Para cada nivel habilidad (θ) existe una probabilidad asociada de contestar correctamente al ítem i (Pi(θ)). Esta probabilidad es pequeña para sujetos con bajo nivel de rasgo y alta para sujetos con altos niveles. En la práctica la CCI, predice la probabilidad de acertar un ítem y suele representarse en una escala situada entre –4 y +4 puntos de habilidad. En un gráfico de la CCI, en el eje x se sitúan los valores de la habilidad θ y en el eje y las probabilidades asociadas Pi(θ) tomando valores de 0 a 1. En la Figura 1, se muestra un ejemplo de la CCI con los tres parámetros del ítem: el índice de dificultad (b), el índice de discriminación (a) y el índice de azar (c). Se observa que un sujeto con un nivel de habilidad θ=0 tiene una probabilidad de 0,65 de acertar el ítem, con un θ=-0,25 tiene 0,30 y con un θ=0,25 tiene 0,98. Esto indica que a mayor nivel de habilidad θ del sujeto, mayor es la probabilidad de acertar el ítem. El valor del índice de discriminación (a), es proporcional a la

inclinación de la recta tangente a la CCI en el punto máximo de la pendiente. El índice de dificultad es el valor de θ para la máxima pendiente de la CCI. El índice del azar expresa el valor de la P(θ) cuando tiende al -∞. Las curvas más a la izquierda en la parte superior, corresponde a ítems que son más discriminatorios. Las curvas paralelas corresponden a ítems con similar dificultad.

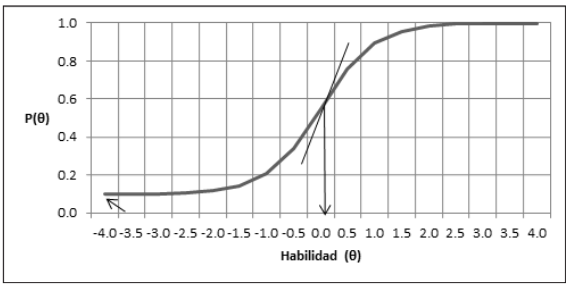

Figura 1. Curva característica del ítem.

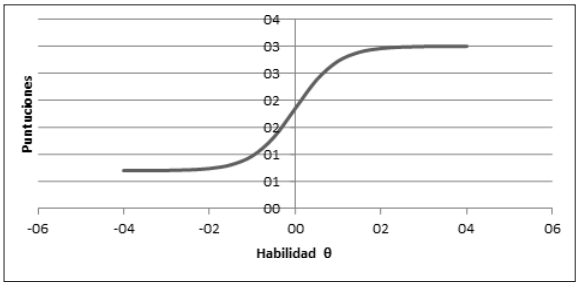

Figura 2. Curva característica del test

Curva característica del test (CCT)

El concepto de curva característica del test es similar al del CCI. Su interés es de dar la posibilidad de interpretar y comparar la TRI y la TCT las puntuaciones obtenidas por los sujetos. La CCT permite transformar las puntuaciones de θ a una nueva escala de puntuaciones verdaderas, que generalmente toma valores de 0,…,n (n= número de ítems). En la CCT, para obtener un determinado nivel de rasgo θ se suman los valores de las curvas características de los ítems Pi(θ) del test. La CCT permite transformar las puntuaciones θ en una nueva escala de las puntuaciones verdaderas (0 a n ítems). Para un test con n ítems la puntuación verdadera se expresarse:

$$PV_j = \sum_{i=1}^{n} P_i(\theta)$$

Sus valores indican la relación que existe entre el nivel en el rasgo latente θ de un determinado sujeto y el patrón de respuesta esperado en el test. En la Figura 2, se muestra un ejemplo de una CCT, dónde se observa como los valores de las para el rasgo o habilidad (θ) de los sujetos le corresponde la escala de puntuación de 0 a 3 (n=3 ítems).

Validar el supuesto de la unidimensionalidad del test

Para el supuesto que el test debe evaluar sólo un rasgo o una habilidad, se usa el análisis factorial con extracción de un factor. Existen varias criterios que se basan en el porcentaje de variación explicada por el primer factor: al menos el 40% (Carmines & Zeller, 1979), como un mínimo el 20% (Reckase, 1979), entre el 17 y 40% usando matrices de correlaciones phi y entre el 30% a 40% usando matrices de correlaciones tetracóricas (Zwick, 1987).

Medidas para comparar modelos de la TRI

Existen dos criterios para seleccionar el mejor modelo TRI que se ajusta a los datos:

1) Criterio de información de Akaike (AIC) propuesto por Akaike (1974), definido por: AIC = -2* log Lik + 2K;

2) Criterio de información bayesiano (BIC) derivado por (Schwarz, 1978), definido por: BIC = -2* log Lik + log (N)*K.

Los mejores modelos son los que presentan el menor AIC y BIC.

3. RESULTADOS Y DISCUSIÓN

Para la aplicación del presente estudio, se considera los datos de un test virtual aplicado a 775 estudiantes de una institución de educación superior matriculados en el curso Estadística General en el semestre 2016 II. La prueba de evaluación informatizada, está constituida por 30 preguntas (ítems) de respuesta múltiple (cuatro alternativas y una de ellas la respuesta), con la finalidad de evaluar a los estudiantes los conocimientos de la estadística descriptiva. El examen que deben responder los estudiantes en el aula de cómputo está constituido por 10 preguntas que son seleccionadas al azar y con una duración máxima de 40 minutos.

Prueba del supuesto de la unidimensionalidad del test

Para probar que el test está evaluando solo un rasgo o habilidad, se aplica el análisis factorial con la matriz de correlaciones biseriales. Se determinó que el primer factor explica el 24,7% de la variabilidad total del

conjunto de las 30 preguntas, entonces hay un factor dominante y por consiguiente la prueba informatizada cumple el supuesto de la unidimensionalidad.

Resultados del análisis estadístico para la TCT

Se aplica la TCT para analizar el test informatizado. Se evalúa la confiabilidad del test y para los ítems se usa el índice de dificultad y de discriminación.

Análisis de la confiabilidad del test

En la Tabla 1, se presentan los respectivos coeficientes de Alfa de Cronbach para todos los ítems del test y el ajustado para cada uno de los 30 ítems. Con un valor del alfa de Cronbach global de 0,833 se puede indicar una confiabilidad de la prueba con un nivel bueno. Se aprecia que ningún ítem presenta un Alfa mayor al valor global por consiguiente no se elimina ningún ítem y todos resultan adecuados.

Tabla 1. Coeficientes de Cronbach global y ajustados para cada ítem.

Global Ítems	0,8325				
Ítem	Alfa	Ítem	Alfa	Ítem	Alfa
V1	0,8272	V11	0,8296	V21	0,8275
V2	0,8299	V12	0,8305	V22	0,8289
V3	0,8270	V13	0,8290	V23	0,8257
V4	0,8255	V14	0,8222	V24	0,8243
V5	0,8250	V15	0,8302	V25	0,8278
V6	0,8253	V16	0,8280	V26	0,8253
V7	0,8288	V17	0,8286	V27	0,8301
V8	0,8283	V18	0,8281	V28	0,8308
V9	0,8291	V19	0,8267	V29	0,8286
V10	0,8252	V20	0,8302	V30	0,8276

Para el método de dos mitades para evaluar la confiabilidad del test, se dividió el conjunto de los 30 ítems en impares y pares para formar dos grupos cada uno de 15 ítems. Se calculó el coeficiente de correlación no paramétrica de Spearman-Brown:

$$r_{xx} = \frac{2r_{pI}}{1+r_{pI}} = \frac{2\times 0,74}{1+0,74} = 0,851$$

El valor alto de este coeficiente muestra que existe una consistencia interna entre las dos mitades de ítems y por ende una confiabilidad del test informatizado.

Análisis de la validación de los ítems

Se evalúan los índices de dificultad y de discriminación con la finalidad de identificar ítems inconsistentes, con posibilidad de ser eliminados. El coeficiente del índice de dificultad del ítem se halla con la tasa o porcentaje de aciertos. En la Tabla 2, se presenta el porcentaje de acierto y no acierto para cada uno de los ítems. Se observa que los ítems V7, V8 y V12 presentan porcentajes de acierto por encima del 80%, indicando que son los ítems fáciles; mientras que no hay ítems con un porcentaje de no acierto por encima del 80%, indicando la no existencia de ítems difíciles.

El índice de discriminación permite la elección de ítems que presentan consistencia interna. En la Tabla 3, se presentan los valores de las correlaciones biserales de cada una de las 30 preguntas (ítems) con la puntuación total de la prueba informatizada. Un valor negativo del coeficiente indicaría que los estudiantes que respondieron correctamente el ítem pertenecen al grupo de peor desempeño en la prueba y para los mayores valores son atribuidos a los ítems que tienen mayor número de respuestas correctas. Se observa que ninguno de los ítems presenta valores bajos o negativos para la correlación biserial, por lo que en este aspecto todos son ítems adecuados.

El resultado final de aplicar la TCT a la prueba informatizada, fue la eliminación de los ítems V7, V8 y V12 que se consideran muy fáciles al tener un porcentaje de acierto mayores al 80%. Por lo tanto, la TCT ha seleccionado solo 27 ítems.

Resultados del análisis estadístico para la TRI

Para poder llevar a cabo la comparación de ambos métodos de evaluación de la prueba informatizada, se consideró para el análisis estadístico con el TRI nuevamente la muestra de los 30 ítems (preguntas).

Análisis de los modelos logísticos binarios

Con la finalidad de seleccionar el mejor modelo de la TRI que se ajuste a los datos del test, se obtienen corridas para los modelos logísticos binarios de un parámetro (1PL), dos (2PL) y tres (3PL). En la Tabla 4, se presentan los resultados de las estadísticas usadas para comparar los modelos. Como se observa los p-valores resultaron significativos, mostrando con un nivel de significación de 0,05 que los datos los tres modelos son significativos.

Con la finalidad de identificar si el modelo 2PL es mejor que el 1PL, se realiza una prueba de bondad de ajuste del modelo 2PL con respecto a 1PL. En la Tabla 5, se presenta los resultados de la comparación de los modelos de 1PL y 2PL usando medidas de bondad de ajuste de razón de verosimilitud (AIC y BIC). Como el AIC del modelo de dos parámetros es más pequeño que el de un parámetro, entonces el 2PL se ajusta mejor que el modelo 1PL según el AIC, aunque el BIC dice lo contrario.

Tabla 2. Porcentaje de acierto y no acierto para cada ítem.

Ítem	Acierto	No acierto	Ítem	Acierto	No acierto	Ítem	Acierto	No acierto
V1	74,2	25,8	V11	75,7	24,3	V21	40,7	59,4
V2	65,7	34,3	V12	83,6	16,4	V22	33,0	66,9
V3	80,0	20,0	V13	74,5	25,6	V23	60,3	39,7
V4	53,8	46,2	V14	58,3	41,7	V24	55,9	44,1
V5	58,1	41,9	V15	66,1	33,9	V25	76,4	23,6
V6	58,2	41,8	V16	63,7	36,3	V26	75,7	24,3
V7	86,1	13,9	V17	68,7	31,4	V27	64,0	36,0
V8	84,9	15,1	V18	69,0	30,9	V28	53,7	46,3
V9	33,4	66,6	V19	61,4	38,6	V29	51,5	48,5
V10	68,1	31,9	V20	56,8	43,2	V30	54,8	45,2

Tabla 3. Correlaciones biseriales entre ítem y puntuación total

Item	Coeficiente	Item	Coeficiente	Item	Coeficiente
V1	0,425	V11	0,352	V21	0,423
V2	0,355	V12	0,310	V22	0,381
V3	0,431	V13	0,372	V23	0,469
V4	0,477	V14	0,558	V24	0,506
V5	0,489	V15	0,347	V25	0,407
V6	0,482	V16	0,408	V26	0,483
V7	0,369	V17	0,389	V27	0,351
V8	0,387	V18	0,402	V28	0,339
V9	0,376	V19	0,443	V29	0,396
V10	0,483	V20	0,352	V30	0,422

Así mismo, con la finalidad de identificar si el modelo 3PL es mejor que el 2PL, se realiza una prueba de bondad de ajuste del modelo 3PL con respecto al 2PL. En la Tabla 6, se presenta los resultados de la comparación de los modelos de 2PL y 3PL usando medidas de bondad de ajuste de razón de verosimilitud (AIC y BIC). El modelo 3PL se ajusta mejor a los datos de la prueba al tener menor valor de AIC. Aunque el BIC dice lo contrario.

Por lo tanto, los resultados anteriores indican que el modelo que mejor se ajusta a los datos de la prueba informatizada es el modelo logístico binario 3PL.

Tabla 4. Estadísticas para la comparación de los tres modelos

Estadística	1PL	2PL	3PL
log.LIK	-13241,97	-13178,41	-13144,77
AIC	26545,93	26476,82	26469,55
BIC	26690,17	26755,99	26888,31
Total inform. (%)	28,97	30,14	25,69
p-value	0,003	0.003	0.003

A continuación, se realizará una prueba para corroborar el ajuste de la prueba al modelo 3PL. En la Tabla 7, se presenta una prueba de bondad de ajuste con los valores Chi- Cuadrado y su respectivo P-valor para cada uno de los ítems del modelo 3PL. Entonces se puede concluir que todos los ítems son no significativos con α=0,05 y

por lo tanto la prueba informatizada se ajusta al modelo 3PL.

Análisis de los parámetros estimados del modelo 3PL

Consiste en la estimación de los parámetros, proceso que se conoce como la calibración del test. En la Tabla 8, se presenta las respectivas estimaciones de los parámetros del modelo logístico binario 3PL: los coeficientes de discriminación (a), los coeficientes de dificultad (b) y los coeficientes de azar (c) y sus respectivos errores estándar para cada uno de los ítems.

El parámetro estimado de discriminación del ítem (a). Se debe considerar que aquellos ítems con valores para "a" inferiores a 0,65 posiblemente estén asociados a otra dimensión del conocimiento, no contemplada en la elaboración del ítem o que los ítems fueron mal elaborados. Los ítems con estas características no deben estar en el proceso de calibrado. Se observa que el ítem V28 tienen un valor "a" menor a 0,65, por lo tanto, debe ser retirado del conjunto de ítems.

El parámetro estimado de dificultad del ítem (b). Según el criterio establecido, debe estar entre -3 y +3. En este caso se observa que todos los ítems presentan índices de dificultad dentro de ese rango, por lo tanto, todos los ítems son válidos.

Tabla 5. Estadísticas para la comparación de los modelos 1PL y 2PL.

Modelo	AIC	BIC	Log.Lik	LRT	g.l.	p-valor
1PL	26545,93	26690,17	-13241,97			
2PL	26476,82	26755,99	-13178,41	127,11	29	< 0,001

Tabla 6. Estadísticas para la comparación de los modelos 2PL y 3PL

Modelo	AIC	BIC	Log.Lik	LRT	g.l.	p-valor
2PL	26476,82	26755,99	-13178,41			
3PL	26469,55	26888,31	-13144,77	67,27	30	< 0,001

Tabla 7. Ajuste de los ítems al modelo 3PL.

Item	X^2	P-valor	Item	X^2	P-valor	Item	X^2	P-valor
V1	9,40	0,693	V11	10,20	0,465	V21	10,26	0,713
V2	19,89	0,109	V12	9,25	0,406	V22	14,63	0,465
V3	20,40	0,277	V13	10,33	0,446	V23	10,14	0,594
V4	13,37	0,436	V14	15,70	0,584	V24	12,82	0,564
V5	13,73	0,475	V15	3,63	1,000	V25	11,30	0,525
V6	11,30	0,485	V16	12,63	0,416	V26	11,17	0,772
V7	18,76	0,168	V17	17,08	0,089	V27	21,29	0,039
V8	10,50	0,396	V18	4,29	1,000	V28	13,23	0,228
V9	8,58	0,733	V19	19,57	0,099	V29	8,59	0,644
V10	19,44	0,168	V20	11,39	0,277	V30	4,52	1,000

Tabla 8. Estimaciones de los parámetros del modelo 3PL para cada ítem.

Item	(a)	(b)	(c)	Item	(a)	(b)	(c)	Item	(a)	(b)	(c)
V1	1,381	-0,540	0,297	V11	1,204	-0,429	0,396	V21	1,017	0,545	0,033
V2	0,726	-1,005	0,000	V12	1,940	-0,080	0,656	V22	1,764	1,170	0,165
V3	1,229	-1,441	0,001	V13	1,078	-0,683	0,283	V23	2,464	0,281	0,330
V4	1,566	0,198	0,168	V14	3,744	0,183	0,267	V24	1,370	-0,125	0,060
V5	1,281	-0,234	0,057	V15	1,047	0,063	0,341	V25	1,006	-1,403	0,000
V6	1,611	0,065	0,195	V16	3,086	0,383	0,428	V26	1,502	-0,879	0,136
V7	1,290	-1,641	0,144	V17	1,191	-0,218	0,305	V27	0,669	-0,952	0,000
V8	2,099	-0,534	0,556	V18	1,896	0,203	0,449	V28	0,602	-0,271	0,000
V9	0,799	0,973	0,000	V19	1,006	-0,562	0,001	V29	1,123	0,468	0,198
V10	1,313	-0,663	0,069	V20	0,793	0,116	0,171	V30	1,256	0,284	0,207

El parámetro estimado de acierto por el azar (c). Este índice se índice se refiere a la probabilidad de que un evaluando con baja habilidad responda correctamente a un ítem, respondiéndolo por el azar. Se espera ítems con valores de "c" pequeños, de lo contrario son considerados mal formulados que inducen al sujeto responder por el azar. Según el criterio, aquellos ítems con valores del parámetro "c " por encima de 0,4, deben ser retirados. El ítem V8 tiene un valor de c igual a 0,555, lo que significa que para este ítem, el evaluando que nada sabe tiene el 55.5% de posibilidad de acertarlo. Entonces los ítems V8, V12, V16 y V18 tienen valores de "c" mayores a 0.4, por lo que deben ser retirados.

Finalmente, según lo anterior se excluyen los ítems V8, V12, V16, V18 y V28 de la prueba informatizada analizada por no corresponder a los criterios propuestos de las estimaciones de los parámetros de los ítems.

Curva característica del ítem (CCI)

En la Figura 3, se han graficado las curvas características de los 25 ítems pero separados en dos grupos. Estas CCI muestran el comportamiento de la prueba informatizada rendida por los estudiantes para diferentes valores de sus habilidades (θ) y los tres parámetros de los ítems (la discriminación (a), la dificultad (b) y el azar (c)) que son considerados en el modelo 3PL. Además, se observa que los ítems que están más próximos a la parte superior izquierda son los más discriminantes: V14, V4, V22, V24, V1 y V26; puesto que permitirán distinguir a los estudiantes con mayor habilidad (mayor probabilidad de responder correctamente la pregunta) con los de menor habilidad. Los ítems más fáciles serán los que tienen mayor probabilidad, se identifican: V7, V3 y V17; mientras los más difíciles que tienen menor probabilidad son: V21 y V9.

La función de información del ítem (FII) y función de información del test (FIT)

La FII, es el gráfico de una curva que muestra que tanto brinda información un determinado ítem (precisión de su estimación). En la Figura 4, se presenta las curvas de la FII de los ítems. Se puede notar que los ítems que proporcionan la mayor información son: V1, V14, V23 y V24; los ítems con menor información son: V15, V9 y V29.

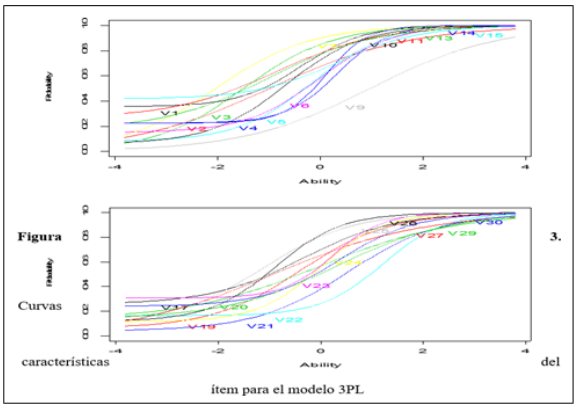

Figura 3. Curvas características del ítem para el modelo 3PL

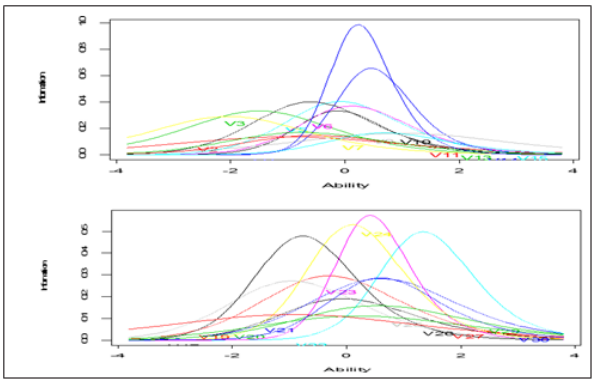

Figura 4. Curvas de información del ítem para el modelo 3PL

4. CONCLUSIONES

La aplicación de la TCT y la TRI al test informatizado, permiten obtener medidas psicométricas y estadísticas para evaluar su confiabilidad y validez del test y las preguntas. La evaluación de la consistencia interna, con el alfa de Cronbach para la prueba informatizada fue de 0,833, indicando una buena confiabilidad y para los ítems, ninguno resultó mayor al global por lo que no se eliminó ninguna pregunta. El método de dos mitades resultó con una correlación de Spearmen-Brown de 0,815, evidenciando una confiabilidad de la prueba. La TCT, con el índice de dificultad identificó a tres preguntas muy fáciles (más del 80% de los estudiantes acertaron), por lo que se retiraron: V7, V8 y V12; no se encontraron preguntas difíciles.

Conflictos de intereses

Los autores firmantes del presente trabajo de investigación declaran no tener ningún potencial conflicto de interés personal o económico con otras personas u organizaciones que puedan influir indebidamente con el presente manuscrito.

Contribuciones de los autores

Preparación y ejecución: CMC, JCAT; Desarrollo de la metodología: CMC, JCAT; Concepción y diseño: CMC, JCAT; Edición del artículo: CMC, JCAT; Supervisión del estudio: CMC, JCAT.

5. LITERATURA CITADA

- Akaike, H. (1974). A new look at the statistical identification model. IEEE Transactions on Automactic Control, 6. 716-723pp.
- Averaño, B.L. (2003). Teoría de Respuesta al Ítem. Otra alternativa para la Medición y Evaluación. Suma Psicológica, 10(2), 235-245.
- Bulut, O. (2015). Applying Item Response Theory Models to Entrance Examination for Graduate Studies: Practical Issues and Insights. Journal of Measurement and Evaluation in Education an Psychology, 6(2), 313-330.
- Carmines, E.G., & Zeller, R.A. (1979). Reliability and Validity Assessment. https://dx.doi.org/10.4135/9781412985642
- Davey, T. (2005). Computer-based testing. Encyclopedia of statistics in behavioral science. ISBN: 978-0-470-86080-9
- Fan, X. (1998). Item response theory and classical test theory: an empirical comparison of their item/person statistics. Educational and Psychological Measurem. June 1998 58(3), 357-382.
- George, D., & Mallery, P. (1995). SPSS/PC+ step by step: A simple guide and reference. 11.0 update (4th ed.). Boston: Allyn & Bacon.
- Gonzáles, J., Cabrera, E., Montenegro, E., Nettle, A., & Guevara, M. (2010). Condicionamiento del modelo logístico para la evaluación informatizada de competencias matemáticas. Ciencia, Docencia y Tecnología, 41, 173-191.
- Lord, F.M. (1952). A theory of test scores.Psychometric Monographs N° 7.
- Martínez, D. (1990). Psicometría: Teoría de los Tests Psicológicos y Educativos. Ed. Pirámide.
- Muñiz, J. (2010). Las Teoría de los Tests: Teoría Clásica y Teoría de Respuesta a los Ítems. Papeles del Psicólogo, 3 (1): 57-66.
- Muñiz, J., & Hambleton, R.K. (1992). Medio siglo de teoría de respuesta a los ítems. Anuario de Psicología, 52(1), 41-66.
- Navas, M.S. (1994). Teoría Clásica de los Test versus Teoría de Respuesta al ítem. Psicología 15. UNED, Madrid, pp. 175-208.
- Olea, J., Ponsoda, V., & Prieto, G. (1999). Tests Informatizados: Fundamentos y Aplicaciones. Colección Psicología. Madird, España. Ed. Prirámide.
- Omobola, O. A., & Adedoyin, J. A. (2013). Assessing the comparability between classical test theory (CTT) and item response theory (IRT) models in estimating test item parameters. Herald Journal of Education and General Studies, 2 (3), 107-114.
- Progar, S., Socan, G., & Pec, M. (2008). An empirical comparison of Item Response Theory and Classical Test Theory. Psihološka obzorja / Horizons of Psychology, 17(3), 5-24.
- Reckase, M.D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Education Statistic, 207-230.
- Schwarz, E. (1978). Estimating the dimension of a model. The Annals of Statistics, 461-464.
- Spearman, C. (1913). Correlations of sums and differences. Journal of Psychology, 5: 417-426.
- Sudol, L., & Studer, C. (2010). Analyzing Test Items:Using Item Response Theory to Validate Assessments. ACM, pp. 436-440.
- Zanon, C., Htz, C., Yoo, H., & Hambleton, R. (2016). An application of item response theory to psychological test development. 18-29.
- Zwick, R. (1987). Assessing the dimensionality of NAEP reading data. Journal of Educational Meusurement, 293-308.