Capítulo 16

ANÁLISIS PSICOMÉTRICO DE EXÁMENES TEORÍA DE MEDICIÓN CLÁSICA

Enrique Ricardo Buzo Casanova, Manuel García Minjares

"Hacer como el carpintero: medir dos veces, para cortar una vez."
(Refrán popular)

INTRODUCCIÓN

Sin duda, una de las consideraciones latentes al momento de diseñar un instrumento de evaluación, es que este arroje suficiente evidencia de validez que permita sustentar la interpretación de los resultados. Una pieza clave para alcanzar este objetivo es el análisis del instrumento, donde se puede constatar, por ejemplo: su dificultad, la factibilidad de los distractores, la calidad del diseño o la idoneidad de los niveles de respuesta, aspectos de diseño que son puestos a prueba cuando se someten a la inspección de este lente. Con el propósito de brindar criterios objetivos de evaluación, desde comienzos del siglo XIX, se buscó desarrollar métodos que permitieran de forma objetiva analizar la calidad psicométrica de diferentes instrumentos de evaluación, lo que se consolida a comienzos del siglo XX con el desarrollo de la *Teoría Clásica de los Tests* a partir de los trabajos de Spearman (Sartes, 2013) cuya aplicación, hoy en día, se extiende al análisis de exámenes objetivo. La Teoría Clásica de los Tests (TCT) es un análisis psicométrico que permite conocer el comportamiento estadístico de los reactivos que componen un examen como del instrumento en sí. Este tipo de análisis contribuye a que los docentes elaboren mejores instrumentos de evaluación que reflejen con mayor certeza el aprendizaje de sus estudiantes, además de que su aplicación es viable en todos los niveles educativos y de evaluación psicológica y hasta cierto punto su interpretación no resulta complicada.

La utilización emergente de la educación a distancia derivada del confinamiento por el COVID-19 mostró la necesidad de contar con métodos de enseñanza y evaluación que permitan realizar inferencias con evidencia de validez y el modelo de la TCT es una alternativa que permite alcanzar este objetivo.

En este capítulo se brindará las bases para analizar el funcionamiento de un examen objetivo con empleo de la TCT, y se habla sobre el Sistema de Análisis Psicométrico de Reactivos

(SISAPRE) diseñado en la Coordinación de Universidad Abierta, Innovación Educativa y Educación a Distancia (CUAIEED) de la Universidad Nacional Autónoma de México para calibrar un examen conforme a la TCT.

¿QUÉ ES LA TCT?

La TCT es un conjunto de conceptos y técnicas que dan sustento al desarrollo de numerosos instrumentos de evaluación (DeVellis, 2006), se fundamenta en el modelo lineal propuesto por el psicólogo británico Charles Spearman (1863–1945) en el que se plantea que la puntuación empírica de un test (X) es resultado de un error de medición (e) sobre la puntuación verdadera (V) es decir

$$X = V + e \qquad (1)$$

Donde *V* y *e* son desconocidos.

Para fortalecer al modelo (1) se establecen los siguientes supuestos:

- El valor esperado de la puntuación empírica es la verdadera (E(X) = V), es decir, si a un estudiante se le aplicara muchas veces el mismo instrumento el valor promedio sería su puntuación verdadera, lo que implica que a la larga se espera que el error de medición desaparezca.
- No existe asociación entre la puntuación verdadera y el error de medición ($\rho_{Ve} = 0$), esto significa que las puntuaciones elevadas o bajas de un estudiante no son consecuencia de un error de medición.
- Los errores de medición en exámenes distintos no guardan relación ($\rho_{e_j e_k} = 0$). Con esto no es de esperarse que estudiantes con errores de medición elevados en un examen lo vuelvan a presentar en otro.

El modelo también plantea como definición adicional el paralelismo de exámenes, es decir, la evaluación de un mismo constructo a través de instrumentos diferentes, lo que deriva en concebir a los exámenes paralelos como aquellos que miden lo mismo con diferentes reactivos. Por tanto, si el examen A es paralelo al B, sus puntuaciones verdaderas serán iguales ($V_A = V_B$) al igual que la variabilidad de sus errores de medición ($O^2_{e_A} = \sigma^2_{e_B}$).

A partir de este modelo clásico, en conjunto con sus supuestos, se deriva un resultado que es de mucha utilidad al momento de evaluar un instrumento: la confiabilidad.

Confiabilidad de un examen objetivo

Del modelo (1) puede observarse que conforme el error de medición es menor, la puntuación empírica (*X*) se acerca más a la verdadera (*V*), lo que implica que la confiabilidad de un examen radica en la cuantía del error de medición (*e*), sin embargo, existe la limitante de que

tanto la puntuación verdadera (*V*) como el error de medición (*e*) no se conocen. Para superar esta problemática se recurre a estimar de forma indirecta el error de medición (*e*) a partir de la correlación de dos exámenes paralelos. De acuerdo a la definición de formas paralelas, si ambos instrumentos se aplicaran a muestras de alumnos parecidas, se esperaría registrar las mismas puntuaciones verdaderas, lo que significaría tener una correlación perfecta, sin embargo, cualquier distanciamiento de esta situación, se debe al error de medición.

Supóngase que *X* y *X'* representan dos exámenes paralelos, la confiabilidad de *X* será la correlación de los exámenes paralelos ($\rho_{X X'}$) la cual es la proporción que representa la varianza de las puntuaciones verdaderas de las empíricas (Muñíz,1996), esto es:

$$\rho_{X X'} = \frac{\sigma_V^2}{\sigma_X^2} \tag{2}$$

de los supuestos del modelo (1), se desprende que:

$$\sigma_V^2 = \sigma_X^2 - \sigma_e^2 \tag{3}$$

al combinar (1) y (2) se obtiene:

$$\rho_{XX'} = \frac{\sigma_V^2}{\sigma_X^2} = \frac{\sigma_X^2 - \sigma_e^2}{\sigma_X^2} = 1 - \frac{\sigma_e^2}{o_X^2} \tag{4}$$

La expresión (4) muestra de manera más explícita que el nivel de confiabilidad de un examen está en función del error de medición expresado en su varianza, por otro lado, este resultado sirve de base para estimar el error típico de medición (SEM por sus siglas en inglés) y la puntuación verdadera (*V*).

Determinación del error típico de medición y la puntuación verdadera

Del modelo (1) se desprende que el error de medición es la diferencia entre la puntuación empírica y la verdadera, es decir, $e = X\text{-}V$, como ya se mencionó con anterioridad, *V* se desconoce, por lo que no es factible conocer *e*. En vez de intentar estimar el error de medición *e* se trabaja con la desviación estándar de esta variable (σ_e) a la cual se le conoce como el error típico de medición o SEM por sus siglas en inglés. Este error típico se estima al despejar (4) con la fórmula:

$$\sigma_e = \sigma_x \sqrt{1 - \rho_{X X'}} \tag{5}$$

donde la estimación depende de las puntuaciones empíricas.

Un supuesto que con frecuencia se incorpora al modelo (1) es que los errores de medición tienen una distribución normal con media 0 y varianza σ_e^2, de cumplirse lo anterior, entonces para algún valor *V*, la puntuación empírica sigue una distribución normal con media *V* y varianza σ_e^2, por tanto, la puntuación verdadera está contenida en el intervalo:

$$X \pm z_{1-\frac{\alpha}{2}}\sigma_e \tag{6}$$

Donde $z_{1-\frac{\alpha}{2}}$ es el cuantil de una distribución normal estandarizada que garantiza un intervalo que contiene la puntuación verdadera con una confiabilidad de *1-α (0 < α < 1)*.

Cálculo del coeficiente de confiabilidad

En el cálculo del error típico de medición que se determinó en (5) interviene el coeficiente de confiabilidad, el cual hasta el momento se considera como la correlación de dos exámenes paralelos, ahora bien, el cálculo de la confiabilidad de forma empírica puede llevarse a cabo con la aplicación del examen a la misma población en dos momentos diferentes (test–re test), o aplicar una mitad del examen a la mitad del grupo y la otra al resto (dos mitades) o utilizar formas equivalentes. Llevar a cabo alguna de estas dinámicas puede ocasionar inconvenientes de logística. Una alternativa para el cálculo de la confiabilidad de un examen objetivo con los resultados de la aplicación es la fórmula propuesta por Cronbach (1951), la cual se le denomina *alfa de Cronbach* cuyo valor normalmente es menor o igual al coeficiente de confiabilidad que se obtiene con exámenes equivalentes. (Muñiz, 1996). El *alfa de Cronbach* se calcula así:

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\Sigma_{i=1}^{n}\ \sigma_i^2}{\sigma\frac{2}{x}}\right) \tag{7}$$

donde n es el número de reactivos en el examen, σ_x^2 es la varianza de las puntuaciones empíricas σ_i^2 y es la varianza del i-ésimo reactivo. El valor de α dependerá de la manera en que se encuentren interrelacionados los reactivo s, obsérvese que la fórmula es semejante a (4). En la práctica se considera que un instrumento tiene una confiabilidad aceptable a partir de 0.7.

Análisis de los reactivos del examen

La TCT se enfoca tanto en el comportamiento del reactivo por sí mismo y su comportamiento como parte del instrumento en su totalidad. El análisis psicométrico de reactivos, llamado también calibración, engloba una serie de procedimientos con los que se evalúa, además del comportamiento del examen en sí, el funcionamiento de los reactivos y sus opciones de respuesta, lo que permite verificar de manera cuantitativa la dificultad del reactivo, su utilidad para determinar el grado de conocimiento de quienes contestan, además de inferir si su construcción favorece o complica su resolución. De esta manera, se cuenta con mayores elementos para el momento de tomar decisiones sobre el reactivo, si se tiene que cambiar,

mejorar o si es necesario eliminarlo del examen. En el análisis de los reactivos se revisan primordialmente tres aspectos: la dificultad, la discriminación y la correlación punto biserial.

Dificultad

La dificultad de un reactivo se denotará como p y se define como la proporción de alumnos que responden correctamente, se puede expresar así:

$$\rho_i = \frac{A}{N} \tag{8}$$

Donde:
p_i = Dificultad del i-ésimo reactivo (i=1,…,n).
A = Examinados que respondieron correctamente el reactivo. (A≤N)
N = Población expuesta al reactivo.

La dificultad de un reactivo puede tomar valores entre 0 y 1, mientras más cercano sea a uno significa que más alumnos lo respondieron de forma correcta y, por lo tanto, el reactivo es más fácil. Por el contrario, a medida que la dificultad se acerca a cero implica que una cantidad menor de estudiantes lo respondió correctamente, por lo que el reactivo es más difícil. Los reactivos pueden clasificarse de acuerdo a su dificultad, un criterio de agrupación es con una distribución uniforme, si la dificultad es menor a 0.2, el reactivo es muy difícil; entre 0.2 y 0.4 difícil; entre 0.4 y 0.6 regular; entre 0.6 y 0.8 difícil y mas de 0.8 muy fácil. En función del contexto de cada aplicación, es recomendable que un reactivo tenga un valor de dificultad que refleje la respuesta acertada de por lo menos la quinta parte de la población total, por ejemplo, un reactivo difícil debe tener al menos un 20% de respuestas correctas y, por el contrario, un reactivo fácil no debe exceder el 80% de dificultad.

Discriminación

La discriminación de un reactivo se traduce como su capacidad para distinguir entre los alumnos que cuentan con el conocimiento o habilidad respecto a los que no. Regularmente toma como base el 27% de los puntajes menores, a los que se les denomina el grupo bajo y el 27% de los mayores a quienes consideraremos como el grupo alto. Estos grupos se contrastan entre sí conforme a la siguiente fórmula:

$$D = \frac{A}{n_A} - \frac{B}{n_B} \tag{9}$$

Donde:
D=Discriminación.
A=Número de examinados del grupo alto que contestan bien al reactivo.
n_A=Número de examinados en el grupo alto.
B=Número de examinados del grupo bajo que contestan bien al reactivo.
n_B=Número de examinados en el grupo bajo.

La discriminación de un reactivo es la diferencia de la dificultad que observa en cada grupo, es decir, compara qué proporción de alumnos contestó correctamente el reactivo en cada grupo. Esta medida toma valores entre -1 y 1, es de esperarse que un reactivo bien elaborado sea contestado en mayor medida por los alumnos de mayor habilidad o conocimiento por lo que esta medida debiera ser, en un primer momento, positiva. Si la discriminación de un reactivo fuera negativa, es indicativo de que el funcionamiento del reactivo no es el adecuado porque lo contesta mejor el grupo de menor desempeño, y si tuviera un valor cercano a cero significa que el reactivo no está discriminando porque se contesta igual independientemente del nivel de habilidad de los alumnos. A partir de 0.2 se considera que el nivel de discriminación es aceptable.

Correlación punto biserial

Los coeficientes de correlación son medidas estadísticas que nos describen la relación lineal entre variables, toman valores entre -1 y 1. Un valor de correlación conforme más se acerque a 1 indica que los cambios en una variable afectan directamente a otra, conforme estos valores se acerquen a -1 indican lo contrario, cuando es cercano a cero indica que no existe relación lineal entre las variables.

Para el análisis de reactivos empleamos una correlación entre el puntaje del examen y la selección de la opción correcta del reactivo de interés. La fórmula resultante es la siguiente:

$$\rho_{pb} = \frac{\mu_A - \mu_x}{\sigma_x}\sqrt{\frac{p}{q}} \qquad (10)$$

donde:

μ_A = puntaje promedio en el examen de los alumnos que contestaron correctamente el reactivo.
μ_x = puntaje promedio en el examen de todos los alumnos.
σ_x = desviación estándar del puntaje del examen.
p = dificultad del reactivo.
q = 1-p.

Conforme a la expresión (9), la correlación punto biserial es una distancia estandarizada entre las medias de aciertos de los alumnos que contestan correctamente el reactivo y la media de la población, ajustada por la dificultad del reactivo. Si la correlación punto biserial arroja un valor inferior a cero significa que el desempeño de los alumnos que contestan de forma correcta el reactivo es inferior al de la población global, lo que implica que el reactivo no está funcionando porque de alguna manera *se requiere no saber para contestarlo*. Si la correlación del reactivo es cero, significa que no existe diferencia en el desempeño de los alumnos que lo responden bien respecto al resto, o bien, el ítem es muy difícil. Finalmente, si la correlación punto biserial es positiva, es señal que los alumnos que eligieron la respuesta correcta tuvieron un mejor desempeño en comparación al general. Como se espera

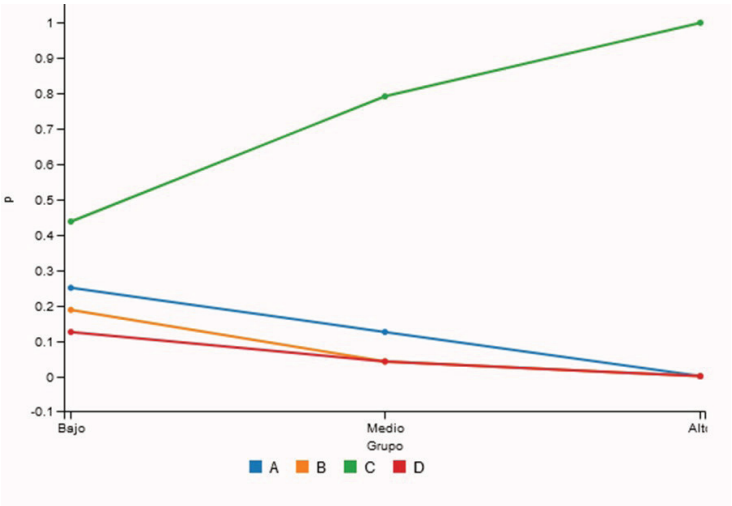
que un buen reactivo contribuya de manera importante en el puntaje del examen el valor de la correlación punto biserial debe ser positivo. Con base a la evidencia empírica de la CUAIEED, reactivos con correlaciones a partir de 0.15 muestran un funcionamiento aceptable, es deseable tener valores de 0.2 en adelante. La correlación punto biserial en cierta forma evalúa la calidad del diseño del reactivo, la experiencia de la Coordinación ha demostrado que reactivos con correlaciones minúsculas o negativas, incluso, están asociados a errores en el diseño de la base del reactivo o de alguno de los distractores.

Análisis de los distractores

Para analizar el comportamiento de los distractores es suficiente calcular para cada uno su discriminación y correlación punto biserial. A diferencia de la opción correcta, se espera que los distractores sean elegidos por los alumnos de menor habilidad o conocimiento, por lo que la proporción de alumnos del grupo bajo que los eligen supera a los del alto. Por otro lado, también se espera que los alumnos que seleccionan los distractores tengan un puntaje menor que el de la población. Por tanto, los distractores en un reactivo que funciona correctamente tienen una discriminación y correlación punto biserial negativa.

Si se graficara el comportamiento de las opciones de respuesta de acuerdo a la proporción de alumnos de los grupos de desempeño que las eligen, se esperaría que la respuesta correcta se eligiera más a medida que mejora el desempeño de los examinados, mientras que los distractores se esperaría un comportamiento inverso. La Figura 1 muestra un ejemplo de un reactivo que funciona de manera deseable.

Figura 1. Ejemplo de comportamiento deseable de las opciones de respuesta de un reactivo



Calibración realizada en el Sistema de Análisis Psicométrico de Reactivos (SISAPRE).
Fuente: Dirección de Evaluación Educativa de la CUAIEED.

En el reactivo ilustrado en la Figura 1, la opción correcta es la C, el cual se selecciona más conforme la habilidad del examinado es mayor, por el contrario, los distractores son menos factibles de ser seleccionados conforme mejora el desempeño de los estudiantes.

Para afianzar los temas que se han comentado a lo largo de este capítulo, se muestra a continuación tres ejemplos de resultados de una calibración de acuerdo a la TCT.

Ejemplo 1. Análisis global del instrumento

La Tabla 1 presenta resultados de la calibración realizada al examen diagnóstico de conocimientos al ingreso al bachillerato de la UNAM de la generación 2020 (Sánchez Mendiola, M. et. al., 2020) conforme a la Teoría Clásica.

Tabla 1. Estadísticos del examen diagnóstico de ingreso al bachillerato Generación 2020

Estadístico	Valor
Coeficiente de confiabilidad (alfa)	0.825
Error típico de medición (SEM)	4.966
Media de dificultad	0.551
Media del coeficiente de correlación punto biserial	0.174

Fuente: Sánchez Mendiola, M. et. al. (2020). *Evaluación diagnóstica de conocimientos. Resultados de los alumnos que ingresan al bachillerato de la UNAM. 2020.* CODEIC, UNAM. P 13.

La Tabla 1, muestra en primer lugar al coeficiente de confiabilidad del examen, a través del *alfa de Cronbach*, cuyo valor es 0.825, lo que habla de que el instrumento es confiable por ser mayor a 0.7; el error típico de medición de 4.966 dice en general que los errores se encuentran a esa distancia de su valor esperado; la media de dificultad y del coeficiente de correlación punto biserial respectivamente expresan que el examen tuvo un nivel de dificultad intermedio, ya que se espera que 55% de los alumnos respondan correctamente algún reactivo del que se conforma el instrumento, con un funcionamiento aceptable además.

Ejemplo 2. Análisis de reactivos

La Tabla 2 presenta los parámetros de dos reactivos que se calibraron conforme a la TCT.

Tabla 2. Parámetros de dos reactivos conforme a la TCT

Parámetro	Reactivo 1	Reactivo 2
Dificultad	0.263	0.558
Discriminación	0.032	0.286
CPB	-0.010	0.197

CPB=Correlación punto biserial.

Fuente: Dirección de Evaluación Educativa de la CUAIEED.

La Tabla 2 exhibe a dos reactivos con comportamientos diferentes, el primero es un reactivo difícil porque lo contestó acertadamente el 26% de los alumnos; el valor de discriminación indica que no hay diferencia importante en la forma en que lo responden los alumnos del grupo alto respecto a los del bajo; la correlación punto biserial muestra que no existe una diferenciación del desempeño de los alumnos que contestan bien el reactivo respecto al global. En contraste, el segundo reactivo tiene un nivel de dificultad intermedio, discrimina de manera clara y los alumnos que eligen la opción correcta tienen un mejor desempeño en el examen respecto a los que se inclinan por un distractor.

Ejemplo 3. Análisis de los distractores

Supóngase que el reactivo 1 del ejemplo anterior cuenta con cuatro opciones de respuesta y la proporción de alumnos que selecciona cada una, así como su correlación punto biserial son las que muestra la Tabla 3.

Tabla 3. Proporción de alumnos y correlación punto biserial de las opciones de respuesta

	Proporción			
Opción	**Total**	**Gpo. bajo**	**Gpo. alto**	**CPB**
A	0.276	0.317	0.254	-0.184
B	0.290	0.270	0.317	-0.028
C	0.157	0.143	0.143	-0.051
D*	0.263	0.238	0.270	-0.010

*Respuesta correcta; CPB=Correlación punto biserial

Fuente: Dirección de Evaluación Educativa de la CUAIEED.

La revisión de la proporción total de alumnos que seleccionó cada opción muestra que fueron seleccionadas en una cantidad semejante, con excepción de la C, confirmando la factibilidad de las alternativas de respuesta, obsérvese, sin embargo, que los distractores A y B se eligieron más que la respuesta correcta. Si se realiza una inspección por grupo de desempeño, se advierte que la proporción de alumnos del grupo alto que eligen los distractores B y C es mayor e igual, respectivamente, que los del grupo bajo, lo que confirma un mal funcionamiento de esos distractores, por otro lado, el distractor A funciona correctamente al ser mayor la proporción de alumnos del grupo bajo que lo elige. En cuanto a la respuesta correcta, aunque ciertamente la eligen más los alumnos del grupo alto, esta proporción es solo tres centésimas mayor que la del grupo bajo, por lo que en términos prácticos el reactivo no está discriminando, además el valor de la correlación punto biserial confirma el mal funcionamiento de la respuesta correcta, el resto de los distractores tiene un valor negativo en este parámetro. El análisis de los distractores invita a revisar cualitativamente este reactivo.

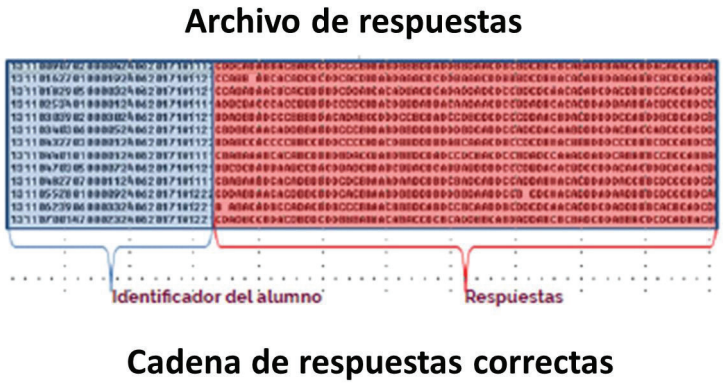
¿QUÉ NECESITO PARA REALIZAR UN ANÁLISIS CON TCT?

Realizar el análisis de un examen objetivo con TCT en general es un proceso ágil, ya que es factible disponer de una variedad de paquetería desarrollada para este fin, ejemplo de ello son *ITEMAN* (assess.com), *BILOG-MG* (ssicentral.com), *Jmetrik* (itemanalysis.com). En esencia, cualquiera de estos desarrollos requiere que se le alimente con las respuestas que los alumnos realizaron en el examen y la cadena de respuestas correctas, así como configurar las características del examen. A continuación, se hablará un poco del sistema desarrollado por la CUAIEED para apoyar a los interesados a realizar análisis psicométrico de exámenes objetivo.

SISAPRE

El Sistema de Análisis Psicométrico de Reactivos (SISAPRE) es un desarrollo de la CUAIEED destinado a apoyar a los interesados de cualquier parte del mundo a realizar un análisis psicométrico de un examen objetivo conforme a la TCT. Los insumos para realizar una calibración en el sistema son un archivo de texto plano, en formato dat o txt con las respuestas de los alumnos y la cadena de respuestas correctas. La Figura 2 muestra un ejemplo de los insumos.

Figura 2. Ejemplo de insumos para realizar una calibración en SISAPRE



Fuente: Manual de usuario del SISAPRE (sisapre.cuaieed.unam.mx)

Una vez que se ingresa los insumos que solicita el sistema, el usuario puede navegar de forma amigable para analizar el comportamiento de los reactivos que conforman, para mayor detalle sobre su uso, se recomienda abrir una cuenta. Se incorporan como recursos de este capítulo tres archivos: el primero con los resultados de un examen, el segundo con la cadena de respuestas correcta y el tercero es el manual del usuario.

CONCLUSIONES Y REFLEXIONES FINALES

A lo largo de este capítulo se brindaron los principales elementos para evaluar el funcionamiento de un examen objetivo y de sus reactivos conforme a la TCT. Desarrollada en principio para brindar objetividad a evaluaciones psicológicas, su aplicación en el análisis de exámenes objetivo contribuye al diseño de evaluaciones más justas, lo cual es uno de los fundamentos que se establecen en los estándares para pruebas educativas y psicológicas (AERA, APA, NCME, 2014) a través de instrumentos más confiables.

Una de las virtudes de mayor relevancia de la TCT es la sencillez del modelo en la que se sustenta, lo que contribuye a que no sea complicada su comprensión. A pesar de su antigüedad, no pierde vigencia, prueba de ello se reporta en la segunda encuesta a profesoras y profesores de la UNAM, en el marco de los retos educativos durante la pandemia de COVID-19 (De Agüero, M. et. al., 2021) donde se exhibe que 4 de cada 10 docentes encuestados de bachillerato o licenciatura emplearon los exámenes de opción múltiple como estrategia de evaluación durante el confinamiento, lo que abre una ventana de oportunidad no solo en su diseño e implementación en línea, sino también en su evaluación, a lo que la TCT pudiera responder de una forma eficaz.

También es cierto que la TCT presenta desventajas, discutidas desde hace tiempo en trabajos como los de Gulliksen (1950), Rasch (1961), Lord & Novick (1968) y Hambleton &

Slater (1997), donde se destaca el impacto de la muestra en los resultados de la evaluación, pues esta se convierte en un sustento de validez ya sea por su tamaño como por la preservación de sus características en aplicaciones subsecuentes del instrumento. Otras limitantes que se identifican en la TCT es que, si un mismo constructo se evalúa con instrumentos distintos, los resultados no se encuentran en la misma métrica, por lo que se vuelve necesario realizar una equiparación para hacerlos comparables; por otro lado, el modelo asume que el error típico de medición se mantiene constante en todos los examinados, cuando hay individuos que pueden observar resultados más consistentes que el resto (Hambleton & Swaminathan, 1985).

Es recomendable que el docente interesado en mejorar sus instrumentos de evaluación se familiarice, como un primer paso, en los resultados que arroja la TCT y posteriormente busque incorporar otros modelos en su análisis, esto le permitirá acercarse más elementos para tomar decisiones sobre el destino de los reactivos que conforman un examen, así como evaluar la confiabilidad de su instrumento.

Actividad sugerida

En el repositorio de recursos digitales de esta obra, se tiene acceso a un archivo de texto con las respuestas de los alumnos a un examen de 54 reactivos, el cual consta de dos componentes con 27 ítems en cada uno y un identificador de cinco caracteres; un archivo con la cadena de respuestas correctas y un manual del SISAPRE. Realizar la calibración del examen y analizar la salida.

Contacto
enrique_buzo@cuaieed.unam.mx
manuel_garcia@cuaieed.unam.mx

REFERENCIAS

AERA, APA, NCME (2014). *Standars for Educational and Psychological Testing*. AERA. USA.

CUAIEED UNAM. *Manual de usuario del SISAPRE*. Recuperado de: sisapre.cuaieed.unam.mx.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334.

De Agüero Servín, M. Benavides, M. A. Rendón, J. Pompa, M. Hernández, A. Martínez, A. M. Sánchez, M. (2021). Los retos educativos durante la pandemia de COVID-19: segunda encuesta a profesoras y profesores de la UNAM. *Revista Digital Universitaria* (RDU), *22*(5).

DeVellis, R. F. (2006). Classical Test Theory. *Medical Care*, *44*(11), 50-59.

Gulliksen, H. (1950). *Theory of mental test*. New York: Wiley.

Hambleton, R.K. & Swaminthan, H. (1985). *Item Response Theory: Principles and applications*. Boston MA: Kluwer Nijhoff.

Hambleton, R.K. & Slater, S. C. (1997). Item Response Theory Models and Testing Practices: Current international status and future directions. *European Journal of Psychological Assessment*, *13*(1), 21-28.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Muñíz, J. (1996). *Psicometría*. Editorial Universitas, S. A. Madrid.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. I, pp. 321-334). Berkeley, CA: University of Chicago Press.

Sánchez Mendiola, M. et al. (2020). *Evaluación diagnóstica de conocimientos. Resultados de los alumnos que ingresan al bachillerato de la UNAM. 2020*. CODEIC, UNAM.

Sartes, L. M. A. Souza-Formigoni, M. L. O. (2013). Avanços na Psicometria: Da Teoria Clássica dos Testes à Teoria de Resposta ao Item. *Psicologia: Reflexão e Crítica*, *26*(2), 241-250.