PARAMETROS EMERGENTES EN LA EVALUACIÓN COMPUTARIZADA EL TIEMPO DE RESPUESTA

Felipe Tirado
Universidad Nacional Autónoma de México
Martín Rosas
Universidad Autónoma de Baja California
Norma Larrazolo
Métrica Educativa

Resumen

En este reporte se presenta la primera fase de un estudio progresivo, en el que se analizaron las respuestas dadas a un examen de selección (Examen de Competencias Básicas – Excoba) por 1711 aspirantes a ingresar al bachillerato de una universidad pública. El Excoba fue aplicado en tres versiones suministradas por un generador computarizado de exámenes (GenerEx). Se analizaron los niveles de acierto y el tiempo de respuesta obtenidos por cada sustentante en cada uno de los 120 reactivos que componen el examen. Los resultados indican que entre otros factores, la dificultad está relacionada con la competencia de los sustentantes, dado que a mayor dificultad mayor el tiempo de respuesta. Se concluye que el tiempo de respuesta es un parámetro emergente que puede contribuir de manera importante a mejorar los indicadores psicométricos de los instrumentos de evaluación.

Palabras clave: evaluación computarizada, tiempo de respuesta, niveles de dificultad.

INTRODUCCIÓN

El tiempo de respuesta o tiempo de reacción, se ha considerado un parámetro sustantivo para analizar los fenómenos psicológicos, en tanto se considera que corresponde a un correlato intrínseco a fenómenos tales como la percepción, la memoria, el pensamiento o la inteligencia. Uno de los primeros en estudiar sistemáticamente el tiempo de reacción como una unidad de análisis fue Wilhem Wundt, frecuentemente reconocido como el padre de la psicología científica, quien en el siglo XIX realizó los primeros trabajos experimentales con el tiempo de reacción como unidad de análisis, con lo cual se obtenían evaluaciones cronométricas susceptibles de ser definidas y medidas con toda precisión (Rieber y Robinson, 2001). Desde entonces se han realizado y reportado cientos de miles de estudios relacionados con el tiempo de reacción o tiempo de respuesta, aunque no son exactamente lo mismo.

Las potencialidades de los sistemas computarizados están impactando de manera decisiva en todos los campos del conocimiento, la evaluación educativa no es una excepción. Además de muchas otras aplicaciones del cómputo a la evaluación, la posibilidad que ahora se tiene para medir con toda precisión el tiempo de respuesta en los exámenes que operan en una plataforma computarizada, ofrece información muy valiosa que permite analizar la validez, los niveles de dificultad, la precisión y confiabilidad de los exámenes, incluso de aquellos de aplicación a gran escala, lo que antes sería impensable el considerar medir cuánto tiempo tarda en contestar a cada pregunta, cada uno de los sustentantes, lo que significa una tarea imposible de cronometrar manualmente y que requiere hacer cientos de miles de registros.

Esto hace que el tiempo de respuesta sea un parámetro emergente en los análisis psicométricos de exámenes de aplicación a gran escala en México. Ciertamente en el mundo, particularmente en los Estados Unidos, hay un sin número de estudios que han utilizado este parámetro. Para observar la vigencia del tema, se puede mencionar que en el programa del Annual Meeting 2015 del National Council on Measurement in Education, se presentaron 17 trabajos sobre estudios relacionados con el tiempo de respuesta, constituye un tema que forma parte de la corriente principal (mainstream) de la psicometría contemporánea.

En el principio de una prueba psicométrica se asume que hay atributos latentes que no son directamente observables, como qué tantos conocimientos tiene una persona sobre determinada área temática, pero este conocimiento se puede hacer manifiesto y observable cuando se dan respuestas a determinadas preguntas. La calidad de las respuestas y el tiempo

que se tarda en responder (tiempo de respuesta) una persona, forma parte del dominio del campo de conocimientos a evaluar, tiempo de respuesta puede ser medido, registrado y analizado, ofreciendo valiosa información para comprender y mejorar los proceso psicométricos. Esto es lo que constituye el propósito de este trabajo.

MARCO TÉORICO

El postulado central que hace al tiempo de respuesta un parámetro de interés en el campo de la evaluación educativa, es que se puede asumir que cuando a una persona se le formula una pregunta, esto le demanda que razone o piense en torno a una respuesta, lo que le toma algún tiempo en integrar su contestación. En el postulado se asume que si una pregunta demanda un razonamiento profundo, ésta tomará más tiempo en poder integrar una respuesta a otra en cuyo caso la solución es simple. De aquí que se entienda que debe haber una relación directamente proporcional del tiempo de respuesta con el nivel de dificultad de la pregunta. Pero se debe considerar que en la función tiempo-dificultad también entra otro factor en juego que la transforma, este corresponde a la habilidad o aptitud de una persona para responder determinada pregunta. De aquí que no es una relación lineal simple entre el tiempo de respuesta y la dificultad de la pregunta, porque ésta se altera en función de las competencias de la persona para responderla.

La Teoría de Respuesta al Ítem (TRI) (en inglés Item Response Theory – IRT) revolucionó en los años sesenta muchos de los planteamientos de la teoría clásica de la evaluación, y se ha vuelto un referente central para la psicometría contemporánea. El la teoría clásica la concepción básica es que hay una probabilidad de respuesta que tiene una distribución normal o de Gauss, las respuestas se distribuyen dentro de un continuo que asume constituyen una escala, donde la dificultad se prorratea en toda la prueba; en cambio la TRI, asume que el nivel de dificultad es propio de cada ítem, ya que no están en la misma escala, sino son independientes, se evalúa la habilidad de cada examinado en cada ítem, lo cual constituye un cambio central en el esquema de análisis, por lo cual el tiempo de respuesta que tarda cada sustentante en responder cada ítem, agrega un nuevo parámetro de extraordinaria valía.

Las pruebas psicométricas tradicionales a gran escala se han basado en el modelo de respuestas de opción múltiple, donde el examinado tiene que seleccionar la respuesta correcta dentro de varios distractores que constituyen errores, el planteamiento es dicotómico: respuestas correctas e incorrectas (errores). Hay otros modelos para hacer el planteamiento de una pregunta,

donde la respuesta puede ser parcialmente correcta, por ejemplo: México tiene fronteras con Estados Unidos es correcto, pero es mejor indicar que tiene fronteras con Estados Unidos y Guatemala, y aún mejor si se refiere a Belice, de manera tal que la calificación puede ser ponderada y determinar así con mayor precisión el valor o peso de la respuesta dada.

La Teoría de Respuesta al Ítem puede hacer el análisis de ítems polivalente (varios valores) y no restringirse a valores dicotómicos (correcto – incorrecto; sí – no; 1 - 0), lo que constituye otro parámetro psicométrico de gran relevancia para obtener indicadores mejor mesurados, en tanto se establece una función matemática entre las competencias personales (inteligencia, conocimientos, actitudes) y los parámetros del ítem (el rango dificultad, discriminación – correlación habilidad y respuesta). Se trata de estimar la probabilidad de que una persona con determinado nivel de competencias pueda contestar correctamente una pregunta, con procedimientos calibrados con niveles de precisión dentro de un rango en una escala.

Uno de los usos del análisis de tiempo de respuesta está en poder identificar patrones propios de los estudiantes que responden en un examen por azar o adivinación. En el Annual Meeting 2015 del National Council on Measurement in Education hubo una sesión dedicada a investigaciones de la adivinación relacionada al tiempo de respuesta (Investigations in Examinee Guessing and Response Time). Keun Im (2015) plantea un estudio sobre un nuevo modelo de tiempo de respuesta para hacer frente a los efectos de una prueba, usando datos reales y simulados, para cuantificar los errores en una estimación en función de diferentes condiciones de ensayo, al comparar con un modelo de tiempo de respuesta. Pokropek (2015) platea un modelo de respuesta para identificar el comportamiento propio de respuestas rápidas de adivinación con datos del tiempo y el modelo IRT, indicando que los sujetos se comportan diferencialmente formando patrones propios. Widiatmo y Wright (2015) plantean dos modelos de medición, el jerárquico de van der Linden y el Q-difusion de van der Maas, utilizan las respuestas y el tiempo de respuesta para la calibración de habilidades, y contrastar los valores conocidos con los estimados.

PROCEDIMIENTO

Este estudio se trata del reporte de proyecto progresivo en desarrollo en torno al tiempo de respuesta como parámetro emergente, cuyo propósito es lograr procesos más equilibrados y justos en los exámenes de selección, en tanto éstos puedan estar más cercanos o apegados al

valor real del parámetro a evaluar, en este caso las competencias de los aspirantes a ingresar a bachillerato.

En este estudio se reporta la etapa descriptiva del comportamiento del parámetro de tiempo de respuesta en un examen (Examen de Competencias Básicas – Excoba), aplicado a 1711 estudiantes que aspiran a ingresar al bachillerato de una universidad pública. Se aplicaron tres versiones del Examen de Competencias Básicas (Excoba), la versión 1° fue aplicada a 431 sustentantes, la versión 2° a 470 y la 3° a 810. Cada versión contiene 120 reactivos. 20 de Habilidades Verbales, 20 de Habilidades Cuantitativas, 20 de competencias en Español, 20 de Matemáticas, 20 de Ciencias Naturales (6 de Biología, 6 de Física, 8 de Química), y 20 de Ciencias Sociales (6 de Geografía, 8 de Historia y 6 de Formación cívica y ética).

El examen está concebido para que cada pregunta sea respondida en un minuto en promedio (dos horas), pero el tiempo de aplicación es de120 minutos, por lo que se le hace saber a los sustentantes que cuentan con dos horas para responder el examen.

El examen se aplica en una plataforma computarizada especializada (GenerEx), que es generadora del Examen de Competencias Básicas (Excoba) en sus diferentes versiones, en la cual califica y registrar en bases de datos cuando ingresa uno de los estudiante en cada reactivo y computa al salir el número de segundo que permaneció, registrando de este modo el parámetro de tiempo de respuesta.

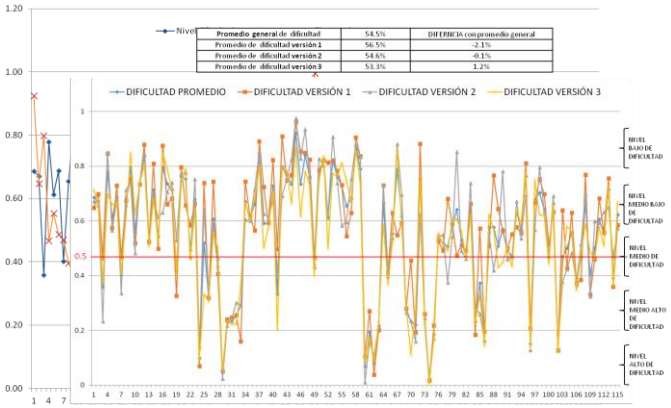
Se procesaron en el GererEx los 1711 registros, se obtuvo para cada examinado en cada reactivo los dos parámetros objeto de este estudio: puntaje y tiempo de respuesta por reactivo, formando así una base integrada por 410,640 datos. Se obtuvo la media de puntajes y tiempos de respuesta para cada registro y cada reactivo, se indexaron en orden descendente y se procesaron gráficos para observar el comportamiento de los datos, analizarlos e interpretarlos.

RESULTADOS

Lo primero que se evaluó fue el nivel de dificultad general de la prueba, de cada reactivo y de cada una de las versiones. Se consideró como niveles de dificultad baja (muy fácil) cuando el 80% o más de los participantes obtuvieron la respuesta correcta, de dificultad media baja (fácil) cuando menos del 80% pero más del 60% contestaron de manera correcta, de dificultad media cuando menos del 60% o más de 40% fueron respuestas correctas, de dificultad media alta (difícil) en los casos que menos del 40% pero más del 20% obtuvieron acierto, y de niveles de dificultad alta (muy difícil) cuando no más del 20% lograron dar la respuesta correcta.

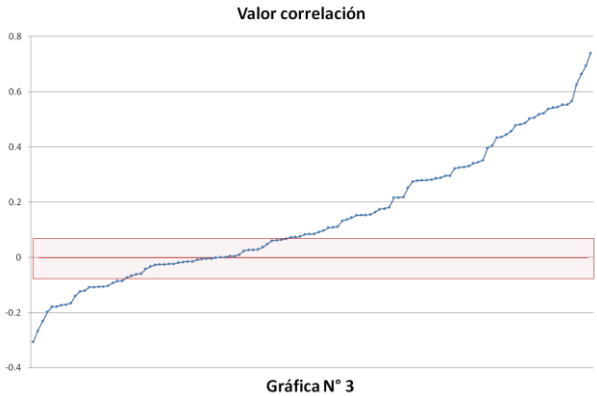
La media general de aciertos fue de 54.5% de aciertos, en la 1° versión obtuvo 56.5%, en la 2° 54.6% y la 3° 53.3% de respuestas correctas en promedio, lo que implica que las medias se encuentran en un rango de variación con respecto a la media general de +/- 2 puntos entre versiones, lo que es muy aceptable (Ver gráfica N° 1). En el mismo gráfico se puede observar la variación de la media de dificultad de cada reactivo en las tres versiones y la media general, observan un paralelismo que dibuja la regularidad del comportamiento de los datos en las tres versiones aplicadas.

Posteriormente se comparó el grado de dificultad y el tiempo de respuesta promedio en cada uno de los reactivos, para hacerlos comparables se relativizaron los valores en una misma escala (0 a1), observando una variación que es heterogénea. En el 75% de los casos (90 de 120) el nivel de dificultad es más alto que el tiempo de respuesta, sólo en 30 casos (25%) la relación es inversa (Ver gráfica N° 2). Por ejemplo, en el recuadro A de la gráfica se observa que en el reactivo 12 la dificultad promedio fue baja (muy fácil – 0.84) y el tiempo de respuesta promedio fue corto (0.32), lo que es de esperarse porque una pregunta fácil no requiere mucho pensarse; en cambio el recuadro B se aprecia que en el ítem 60 la dificultad promedio fue alta (muy difícil – 0.07) ya que sólo un 7% obtuvo acierto, y el tiempo de respuesta fue alto (0.86), lo que implica que los sustentantes pensaron mucho la respuesta sin conseguir la mayoría responder correctamente, relación que también es consistente con lo esperado.



Gráfica N° 1

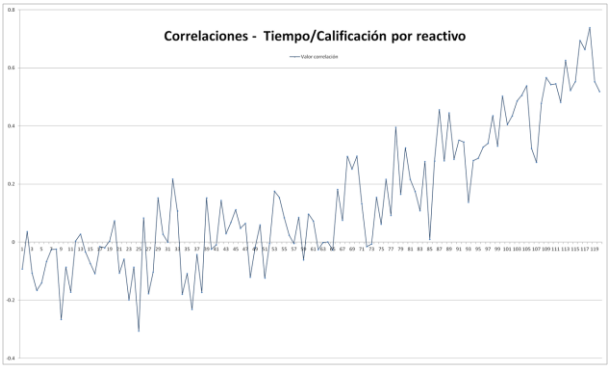
Otro análisis que se hizo fue correlacionar el tiempo de respuesta y el nivel de dificultad de cada reactivo. Se observa que esta relación no es lineal, sino muy heterogénea (Ver gráfica N° 3). Hay correlaciones que son negativas, al haber muchos aciertos y poco tiempo o a la inversa, mucho tiempo pocos aciertos. Se



Gráfica N° 3

observa una banda de correlaciones débiles (+/- 0.10) que no son significativas. Esta diversidad se entiende, porque hay respuestas difíciles que tiene un tiempo de respuesta grande, como sería de esperarse, pero en otras ocurre lo contrario, porque también ocurre que al comprender el sustentante que no sabe o no entiende la pregunta (le es difícil) en poco tiempo la abandona, evitando así malgastar su tiempo. También puede haber preguntas fáciles que les tome mucho tiempo a algunos, y esto puede ser por diversas razones, como distraerse.

Al graficar las correlaciones obtenidas en cada reactivo es interesante apreciar una tendencia a ser más altas hacia el final de la prueba, en los últimos reactivos (Ver gráfica No 4). Una razón podría ser que dado que el último componente de la prueba son preguntas de Ciencias Sociales, éstas tomen menos tiempo en ser contestadas, o bien que el orden de presentación cuenta, que al final se les dedica menos tiempo (por la fatiga).



Gráfica N° 4

Podría haber comportamientos paradójicos, tiempo de respuesta cortos ante preguntas de dificultad alta, porque podría esperarse que cuando un estudiante está ante una pregunta difícil, al apreciar su complejidad la abandone pronto, sin pensar mucho. Es muy interesante que esto ocurriera poco, ya que sólo en el 4.5% de los casos (9211) (Ver gráfica N° 4), los estudiantes abandonaron la pregunta antes de transcurrir 10 segundos, lo que hace apreciar que la mayoría fueron persistentes, contestaron con responsabilidad al esmerarse para tratar de obtener la mejor calificación posible, lo que constituye un indicador que fortalece la validez del examen.

Otro hallazgo interesante al observar el tiempo de respuesta, es que los sustentantes responden en el orden numérico de presentación de los reactivos, al inicio del examen le dedican menos tiempo en promedio a las preguntas, comparado al final. Pareciera que hacia el final de la prueba se sintieran aliviados del apremio del tiempo, lo que ofrece indicadores de que el orden de presentación constituye otro factor que entra en juego en un examen. También cómo va avanzando la prueba se observa que va incrementando el abandono, en el 4.8% de los casos (9913) (Ver gráfica N° 5), los sustentantes no ingresaron a la pregunta (tiempo de respuesta = 0), lo que fue en aumento a medida que avanzaron y llegaban a la parte final del examen, lo que también es congruente y explicable, porque un alumno que hacia el final de la prueba se siente que no está teniendo éxito, que las preguntas las aprecia difíciles y considera que no va a alcanzar el límite de aciertos requeridos para ingresar al bachillerato, empieza a presentar este comportamiento propio del abandono (ya no ingresar a las preguntas). Este comportamiento se empieza a incrementar en las últimas 20 preguntas.

Gráfica N° 5

El tiempo de respuesta hace una diferencia. El mejor promedio obtuvo 75.5% de respuestas correctas y resolvió el examen en 1 hora 58 minutos, el mejor desempeño concluyó en sólo 1 hora con 18 minutos con 69.3% de aciertos. Otros dos estudiantes empataron con 70 % de aciertos, pero uno resolvió la prueba en 2 horas 9 minutos, en tanto el otro en 1 hora 38 minutos, lo que indica que uno es más ágil que el otro.

CONCLUSIÓN

Hay que empezar por precisar que el tiempo de respuesta no es un correlato unívoco de un proceso cognitivo. Cuando una pregunta es difícil de contestar, requiere tiempo para pensar en su solución, por el contrario, si es fácil toma un tiempo de respuesta corto. Esto dibuja un parámetro que va de los que más a los que menos tardan. Se perfilará una distribución más o menos normal, donde la mayoría (68.2%) estará dentro del rango de una desviación estándar por arriba o por abajo de la media. Entre más grande sea el promedio del tiempo de respuesta,

podemos considerar que la pregunta fue difícil y a la inversa, si demanda poco tiempo estimamos que es fácil. Quien tardan poco y obtiene una repuesta correcta en una pregunta difícil nos invita a considerar que es competente y hábil, si es incorrecta la respuesta probablemente se trata de un adivinador porque no la pensó mucho, pero si tarda mucho y la respuesta es incorrecta consideramos que no tiene la competencia para responder. Claro que esto no es así necesariamente, el tiempo de respuesta no puede ser atribuido de manera unívoca a un correlato de que estuvo pensando en cómo responder, se pude distraer el sustentante, tener desinterés, fatiga o cansancio, es decir, hay otras fuentes de varianza que componen el tiempo de respuesta.

Lo anterior no anula la relevancia del parámetro del tiempo de respuesta para poder analizar y mejorar los indicadores psicométricos en los exámenes de aplicación a gran escala. Constituye un parámetro emergente en México porque hasta ahora no ha sido aprovechado. Si el tiempo de respuesta es un parámetro que puede contribuir a optimizar los niveles de validez, confiabilidad, precisión y certeza en la evaluación, se debe empezar a utilizar sistemáticamente.

Bibliografía y Referencias

Annual Meeting Program 2015. National Council on Measurement in Education, Chicago, Illinois.

Balivada, A., Chen, J., & Abraham, J. A. (1996). Analog testing with time response parameters. IEEE Design & Test of Computers, 13(2), 18-25.

Keun Im, S. (2015). Hierarchical Modeling of Item Responses and Response Times for Testlets. University of Kansas. Proceeding Annual Meeting 2015 - National Council on Measurement in Education.

Rieber, Robert W., and David Robinson, eds (2001). Wilhelm Wundt in history: The making of a scientific psychology. New York: Springer Science & Business Media.

Pokropek, A. (2015). Response Model for Rapid-Guessing Behaviors in Data With Timing Information. Polish Academy of Sciences (IFiS). Proceeding Annual Meeting 2015 - National Council on Measurement in Education.

Widiatmo, H. y Wright, D. (2015). Comparing Two Item Response Models That Incorporate Response Times. ACT, Inc. Proceeding Annual Meeting 2015 - National Council on Measurement in Education.