Test Adaptativo Informatizado de Analogías Verbales: comparación de Criterios de Parada[1]

Gabriela Susana Lozzia[2], Facundo Juan Pablo Abal[3], María Silvia Galibert[4], Horacio Félix Attorresi[5]

Instituto de Investigaciones de la Facultad de Psicología de la Universidad de Buenos Aires-Argentina[2,3,4,5] *y Consejo Nacional de Investigaciones Científicas y Técnicas-Argentina*[3]

Este trabajo busca actualizar los estudios psicométricos en la Argentina. Se desarrolló un Test Adaptativo Informatizado (TAI) a partir de un Banco de Ítems (BI) de Analogías Verbales y se programó con FastTEST Pro versión 2.0. Se contó con una muestra de 108 estudiantes universitarios. Se compararon los resultados obtenidos a partir de tres criterios de finalización distintos del TAI (longitud fija de 32 ítems, longitud variable al alcanzar un error de .40 y un error de .30). Las tres variantes obtuvieron correlaciones significativamente altas ($r > .90$; $p < .001$) con el nivel de rasgo estimado a partir del BI. El TAI de longitud fija de 32 ítems presentó el balance óptimo entre precisión y longitud dadas las características del BI.
Palabras clave: test adaptativo informatizado; banco de ítems; teoría de respuesta al ítem; razonamiento verbal; estudiantes universitarios.

Verbal Analogies Computerized Adaptive Test: Comparison of Stopping Rules

This study seeks to update Argentine psychometric studies. A Computerized Adaptive Test (CAT) was developed from a Verbal Analogies Item Bank (IB); it was programmed with FastTEST Pro version 2.0. A sample of 108 undergraduate students was assessed.

[1] Esta investigación fue financiada con los subsidios de la Universidad de Buenos Aires (UBACyT2018 20020170100200BA y 20020170200001BA) y de la Agencia Nacional de Promoción Científica y Tecnológica (PICT-2017-3226).

[2] Doctora y profesora en Psicología. Profesora adjunta de la Facultad de Psicología, Universidad de Buenos Aires. Dirección postal: Nemesio Trejo 5044, (1407), Ciudad Autónoma de Buenos Aires, Argentina. Contacto: glozzia@psi.uba.ar. https://orcid.org/0000-0001-7753-6303

[3] Doctor en Psicología. Investigador asistente, Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET). Profesor adjunto de la Facultad de Psicología, Universidad de Buenos Aires. Dirección postal: Zuviría 5691, (1439), Ciudad Autónoma de Buenos Aires, Argentina. Contacto: fabal@psi.uba.ar. https://orcid.org/0000-0001-7023-5380

[4] Doctora en Psicología. MSc Biometría. Prof. Matemática. Profesora adjunta de la Facultad de Psicología, Universidad de Buenos Aires. Dirección postal: Tejedor 555, (1424), Ciudad Autónoma de Buenos Aires, Argentina. Contacto: galibert@psi.uba.ar. https://orcid.org/0000-0002-7476-4105

[5] Licenciado en Ciencias Matemáticas. Profesor consulto titular de la Universidad de Buenos Aires. Dirección postal: Rivera Indarte 132, 1°A, (1406), Ciudad Autónoma de Buenos Aires, Argentina. Contacto: horacioattorresi@gmail.com. https://orcid.org/0000-0002-3027-1069

https://doi.org/10.18800/psico.202001.002

Results were compared according to three different CAT termination criteria (fixed length of 32 items, variable length when reaching an estimation error of less than or equal to .40 and to .30). Significantly high correlations were obtained ($r > .90$; $p < .001$) between the level of trait estimated from the BI and each of the three CAT variants. Given the BI characteristics, the fixed-length of 32 items for the CAT presented the optimal balance between precision and length.
Keywords: computer adaptive testing; item bank; item response theory; verbal reasoning; college students.

Teste Adaptativo Informatizado de Analogias Verbais: Comparação de Critérios de Parada

Este trabalho busca atualizar os estudos psicométricos na Argentina. Um Teste Adaptativo Informatizado (TAI) foi desenvolvido a partir de um Banco de Itens (BI) de Analogias Verbais e programado com FastTEST Pro versão 2.0. Participaram da pesquisa 108 estudantes universitários. Os resultados obtidos foram comparados a partir de três critérios de parada diferentes do TAI (comprimento fixo de 32 itens, comprimento variável quando atingiu um erro de 0,40 e um erro de 0,30). As três variantes obtiveram correlações significativamente altas ($r > .90$; $p < .001$) com o nível de habilidade estimado a partir de BI. O TAI de comprimento fixo de 32 itens apresentou o equilíbrio ideal entre precisão e comprimento, dadas as características do BI.
Palavras chave: teste adaptativo informatizado; banco de itens; teoria de resposta ao item; analogias verbais; raciocínio verbal; estudantes universitários.

Test Adaptatif Informatisé d'Analogies Verbales: Comparaison de Critères d'Arrêt

Ce travail cherche à actualiser les études psychométriques en Argentine. Un Test Adaptatif Informatisé (TAI) a été développé à partir d'une Banque d'Items (BI) d'Analogies Verbales et a été programé en FastTEST Pro version 2.0. On a travaillé avec une échantillon de 108 étudiants universitaires. On a comparé les résultats obtenus à partir de trois critères d'arrêt du TAI (longitude fixe de 32 items, longitude variable au moment d'atteindre un erreur de .40 et de .30). Les trois variantes ont obtenu des corrélations significativement élevées ($r > .90$; $p < .001$) avec le niveau de trait à partir de la BI. Vu les caractéristiques de la BI, le TAI de longitude fixe de 32 items a présenté le bilan optimal entre précision et longitude.
Mots-clés: test adaptatif informatisé; banque d'items; théorie de la réponse à l'item; raisonnement verbal; étudiants universitaires.

La psicometría siempre ha estado interesada en aportar un soporte metodológico óptimo para alcanzar procedimientos de evaluación más ajustados a las características de cada uno de los examinados. Sin embargo, llevar adelante este tipo de evaluación tenía limitaciones dentro del marco de la Teoría Clásica de Tests (TCT) por la imposibilidad de comparar las puntuaciones de las personas obtenidas con de diferentes conjuntos de ítems. Los avances de la tecnología informática posibilitaron aplicar los nuevos modelos psicométricos de la Teoría de Respuesta al Ítem (TRI) a la construcción de Bancos de Ítems y obtener, a partir de ellos, instrumentos que presentaran únicamente los reactivos que fueran altamente informativos para estimar el nivel de habilidad de cada individuo (Chang, 2015; Drasgow, 2015; Olea, Ponsoda & Prieto, 1999; van der Linden & Glas, 2010; Wainer et al., 2000). Así surgieron los Tests Adaptativos Informatizados (TAI, traducción de la expresión inglesa Computerized Adaptive Test, CAT).

Como señalan Olea y Ponsoda (2013), los TAI son pruebas para la evaluación psicológica o educativa cuyos ítems se presentan y responden mediante una computadora. Su singularidad radica en que los ítems se seleccionan mediante un algoritmo computacional teniendo en cuenta el nivel de rasgo que progresivamente va manifestando la persona al responderlo. Si la respuesta dada es correcta, el programa presentará un ítem más difícil. Si es incorrecta, presentará un ítem más fácil. La administración de los ítems continúa hasta que se alcanza un número de ítems especificado o un valor determinado de precisión en la estimación del nivel de rasgo del evaluado (Wainer et al., 2000). De esta manera, se consigue una evaluación precisa presentando el menor número posible de ítems (Olea & Ponsoda, 2013). Esta es justamente su mayor ventaja. La misma ha sido demostrada empíricamente en variadas investigaciones que indican que, a pesar de ser en promedio un 50% más corto que un Test Convencional (TC), un TAI posee igual o mayor nivel de

precisión (Embretson & Reise, 2013). Asimismo aporta una mayor precisión de la medida en todos los niveles del rasgo, a diferencia de un TC que tiene su máxima precisión en los niveles de habilidad cercanos a su dificultad promedio. Otra de las ventajas que presenta un TAI está relacionada con la seguridad de la prueba, ya que como los individuos reciben distintos ítems, no sabrán a priori sobre qué contenidos deberán responder (Olea & Ponsoda, 2013). A su vez, redunda en los beneficios de ahorro del tiempo invertido (reduce los problemas de fatiga, desatención, aburrimiento, apatía y descuido) y en la satisfacción de los evaluados, ya que al enfrentarse a pruebas acordes con su nivel se minimizan los aspectos frustrantes que lleva aparejada toda evaluación.

El funcionamiento de los TAIs se basa sobre dos componentes: un *Banco de Ítems* (BI) calibrados a partir de uno de los modelos de la TRI y un *algoritmo adaptativo informatizado* que ejecuta los procedimientos de inicio, estimación provisional del nivel del rasgo, selección dinámica de los ítems en función del nivel de rasgo que va manifestando el evaluado al completar el test y finalización del TAI. Según cómo se programen estos puntos del algoritmo se obtendrán distintos tipos de TAIs. Existen especificaciones adicionales que dependerán del diseño y de la finalidad de cada TAI en particular (e.g. omitir y corregir respuestas, tasa de exposición de los ítems, restricciones de contenido).

El BI es un conjunto de reactivos que miden una misma variable, que puede ser un dominio de conocimiento o rasgo. Las propiedades psicométricas de los ítems deben ser conocidas; es decir, sus parámetros deben estar estimados en una misma escala (calibrados) mediante un modelo de la TRI determinado (Barbero, 1996). Algunas de las características más importantes del TAI estarán condicionadas por el BI (e.g., el rango de valores del nivel de rasgo que permite evaluar adecuadamente y la precisión alcanzada en la estimación de los distintos niveles del rasgo, la necesidad de balance de contenido, el criterio de finalización). Por ello se dice que de la calidad del BI dependerá la calidad del TAI. En este sentido, la Función de Información (FI) del BI impondrá una cota a la máxima precisión que puede obtenerse mediante un TAI (Olea, Abad, Ponsoda & Ximénez, 2004).

Actualmente son muchos los TC para los cuales existen versiones adaptativas y es frecuente, tanto en Estados Unidos como en Europa, su uso en diferentes ámbitos de aplicación (Beckmann et al., 2015; Chang, 2015; Devine et al., 2016; Drasgow, 2015; Educational Testing Service, 2016; Gibbons, Weiss, Frank & Kupfer, 2016; Hol, Vorst & Mellenbergh, 2008; Su, 2016; van der Linden, 2016; van der Linden & Glas, 2010; Wang, Zheng & Chang, 2014). Algunos ejemplos de TAIs que evalúan conocimientos son el *Test of English as a Foreign Language* (TOEFL), el *Graduate Management Admissions Tests* (GMAT), y el *Graduate Record Exam* (CAT-GRE). También se encuentran TAIs de selección de personal (e.g., CAT-ASVAB, el TAI para la selección de programadores de la empresa *State Farm*); de admisión a centros educativos (e.g., *Law School Admission Test*); de evaluación y certificación educativa (e.g., COMPASS *placement tests*, NCLEX/CAT, sistema CARAT); y, aunque en menor medida, de actitudes, rasgo de personalidad y diagnóstico cognitivo (e.g., CAT-MMPI-2, Anxiety-CAT, Cognitive Diagnosis Computerized Adaptive Testing, CAD-MDD: Computerized Adaptive Diagnostic Test for Major Depressive Disorder, PROMIS: Patient-Reported Outcomes Measurement Information System).

En el contexto iberoamericano, España va a la vanguardia, al menos en las investigaciones teóricas ya que todavía el uso de los TAIS no es tan habitual (Barrada, 2012; Barrada, Abad & Olea, 2014; Hernández, Tomás, Ferreres & Lloret, 2015). Entre los TAI producidos en España se encuentran los que evalúan conocimientos y habilidades (e.g., García, Abad, Olea & Aguado, 2013; López-Cuadrado, Pérez, Vadillo & Gutiérrez, 2010) y los aplicados a las evaluaciones en el ámbito de la salud (e.g., Fonseca-Pedrero, Menéndez, Paino, Lemos-Giráldez & Muñiz, 2013; Kaplan, de la Torre & Barrada, 2015; Suárez-Álvarez & Pedrosa, 2016). Por otra parte, en Brasil se llevan adelante distintos proyectos de aplicación de la TRI y se han encontrado algunas publicaciones sobre TAIs (da Cunha & Nogueira, 2015; Junior & Pinto, 2015; Moreira Junior, Tezza, Andrade & Bornia, 2013; Piton-Gonçalves & Aluísio, 2015; Veldkamp & Matteucci, 2013). También se ha iniciado

el desarrollo de TAIs en México (Toledo, Mezura Godoy, Cruz Ramírez & Benítez Guerrero, 2013), en Colombia (Abuchar & Simanca, 2013; Jiménez & Herrera, 2016; Simanca & Abuchar, 2014), en Chile (Salcedo, Ferreira & Barrientos, 2013) y en Uruguay (Sistema Nacional de Educación Pública, 2011). Mientras que en Perú hay un inicio de uso de la TRI (Escurra Mayaute & Salas Blas, 2014).

En Argentina, las aplicaciones de la TRI son escasas e infrecuentes (Attorresi, Lozzia, Abal, Galibert & Aguerri, 2009; Tornimbeni, Pérez & Olaz, 2008) y las investigaciones sobre TAIs son prácticamente inexistentes. En investigaciones anteriores se calibró un Banco de Ítems de Analogías Verbales adaptado al contexto local que sirve de base para generar diversos tipos de tests (e.g., TAIs, Tests Paralelos, Tests referidos al Criterio, Tests con Características Prefijadas) y permite evaluar a estudiantes universitarios en su habilidad para reconocer y discriminar relaciones (para más información ver Lozzia, Abal, Blum, Aguerri, Galibert & Attorresi, 2015). Esta aptitud se correlaciona con el factor ideativo de la comprensión verbal (Thurstone, 1938, 1940), que es común al razonamiento deductivo, serial y probabilístico, de clasificación y de resolución de problemas (Yela, 1987). Las investigaciones sobre inteligencia humana y razonamiento han encontrado que el rendimiento en analogías representa una de las mejores medidas de la comprensión verbal y el pensamiento analítico (Gentner, Holyoak & Kokinov, 2001; Sternberg, 1985, 2001, 2015). Numerosos estudios indican que se trata de una capacidad crítica para el éxito tanto académico como profesional (Hey, Linsey, Agogino & Wood, 2008; Jones & Estes, 2015; Kuncel & Hezlett, 2007; Kuncel, Hezlett & Ones, 2004; Meagher, 2012; Wendler & Bridgeman, 2014; Young, Klieger, Bochenek, Li & Cline, 2014). Por ello, los tests de analogías verbales son frecuentemente utilizados en Estados Unidos para la admisión universitaria, otorgamiento de becas de estudios, orientación vocacional y selección de personal.

A pesar del avance que supone la construcción del Banco de Ítems de Analogías Verbales para la psicometría argentina, su implementación práctica usando una estrategia adaptativa aún no ha sido examinada.

El desarrollo de un TAI de Analogías Verbales posibilitaría agilizar las tareas evaluativas de profesionales de la psicología y de la educación al permitir mediciones más precisas en menor tiempo. Pero un paso previo indispensable consiste en llevar adelante estudios para definir las especificaciones del algoritmo a fin de optimizar la medición del constructo. Esto significa que la eficiencia de un TAI debe ser demostrada empíricamente.

En esta línea, el objetivo del presente trabajo es analizar la eficiencia del TAI de Analogías Verbales comparando su aplicación con tres criterios de parada distintos: a) longitud fija al administrar 32 ítems (supone aplicar la mitad del BI), b) longitud variable al alcanzar un error de estimación menor o igual a 0.4 (correspondiente a una confiabilidad clásica de .84), y c) longitud variable al alcanzar un error de estimación menor o igual a 0.3 (esta última variante se consideró para indagar cómo funciona el TAI con un criterio más exigente, equivalente a una confiabilidad de .91).

Al tratarse de una investigación de desarrollo instrumental (Montero & León, 2005) y por lo tanto no experimental, no pueden plantearse las tradicionales hipótesis. Sin embargo, es posible plantear los siguientes resultados que razonablemente podrían esperarse:

1. Los niveles de habilidad obtenidos por los evaluados al responder al BI completo correlacionarán positiva y fuertemente con los niveles de habilidad estimados a partir de la administración del TAI (con independencia del criterio de parada utilizado). Se espera que la mayor correlación se presente en el caso del TAI de longitud variable al alcanzar un error de estimación menor o igual a 0.3.

2. Los errores de estimación en los niveles de habilidad de los evaluados al completar BI completo correlacionarán positiva y fuertemente con los errores de estimación obtenidos mediante el TAI de longitud fija de 32 ítems.

3. La cantidad de ítems presentados en las administraciones de los TAI de longitud variable será inferior a la cantidad de ítems que posee el BI.

Método

Participantes

Los participantes fueron 108 cursantes del segundo año de la Facultad de Psicología de Universidad de Buenos Aires. El 18% del total de individuos fueron varones mientras que el 82% fueron mujeres. La edad varió entre 18 y 52 años, con media de 23.17 años (*DE*=5.39), mediana de 21 y amplitud semi-intercuartil de 2 años.

Medición

Cuestionario de variables sociodemográficas. Recaba información acerca de características tales como género y edad.

Banco de Ítems de Analogías Verbales. El banco está compuesto por ítems llamados de Analogías Verbales o de Relaciones ya que miden la capacidad para reconocer y discriminar relaciones entre palabras (Attorresi, Pano, Fernández Liporace & Cayssials, 1993). Cada ítem está formado por un par de palabras base que poseen una relación entre ellas y cuatro opciones de pares de palabras. Su resolución consiste en elegir entre las opciones el par que presenta la relación más parecida a la que existe entre las palabras del par base (Galibert, Aguerri, Pano, Lozzia & Attorresi, 2005; Lozzia, Picón Janeiro & Galibert, 2008). Un ejemplo de los ítems elaborados es el siguiente:

JINETE – CABALLO
a) arqueólogo – museo
b) director – escuela
c) administrador – consorcio
d) conductor – camión
La respuesta correcta para este ítem es la opción d)

El BI consta de 64 ítems unidimensionales calibrados con el Modelo Logístico de Tres Parámetros, sin funcionamiento diferencial por género, con adecuada capacidad discriminativa y un nivel de acierto por azar cercano al esperable para ítems con cuatro opciones

de respuesta (Tabla 1). El BI contiene una cantidad suficiente y variada de ítems que permite evaluar con precisión los niveles de habilidad comprendidos entre -1.75 y 3.00. Este BI cumple con las características que debe tener para ser utilizado como base de un TAI: incluir ítems informativos a lo largo de todo el rango del rasgo. Para más detalles sobre la construcción del BI ver Lozzia et al., 2015.

Tabla 1

Propiedades psicométricas del Banco de Ítems de Analogías Verbales

	Índices del Modelo de Tres Parámetros		
	a	b	c
Media	0.83	0.20	0.24
Desvío	0.14	0.98	0.02
Mínimo	0.65	-2.42	0.20
Máximo	1.16	2.35	0.26

Nota. a = Parámetro de Discriminación; *b* = Parámetro de Dificultad; *c* = Parámetro de Aciertos por Azar.

La administración adaptativa del BI se programó con la versión 2.0 del FastTEST Professional Testing System (Weiss, 2008). En la determinación del algoritmo adaptativo se tuvieron en cuenta las características de BI, los objetivos de evaluación, la población por evaluar y las características del *software* (para más detalles ver Lozzia & Attorresi, 2012). Se utilizaron las siguientes especificaciones en su diseño: a) un procedimiento de inicio aleatorio (para evitar que se repita la secuencia inicial en diferentes estudiantes) entre niveles levemente inferiores a la media del rasgo en el rango -1.0 a -0.5 (para asegurar una primera experiencia satisfactoria que disminuya la ansiedad ante la evaluación), b) el método de Máxima Verosimilitud Condicional (Lord, 1980) para estimar después de cada respuesta el nivel de rasgo (simbolizado con la letra griega θ) y el error asociado a dicha estimación, y c) selección sucesiva de los ítems con el Método de Máxima Información de Fisher (Lord, 1980) que permite elegir dentro del conjunto de los reactivos aún no presentados el más apropiado para el nivel θ estimado.

Para definir los criterios de finalización posibles se tuvieron en cuenta: a) la Función de Información del Banco para garantizar que el error fijado como punto de corte pueda ser alcanzado por la mayoría de los evaluados, b) la precisión alcanzada en las versiones convencionales del test para fijar el error de estimación máximo tolerable (en las pruebas de calibración se obtuvieron índices de confiabilidad (α de Cronbach) entre .77 y .85 equivalentes a un error de entre .39 y .48 aproximadamente), y c) el número de ítems administrados en las versiones convencionales del test para fijar la cantidad máxima de reactivos (los TC utilizados en los estudios de calibración del BI presentaron entre 30 y 38 ítems). Se determinaron tres criterios de parada: 1) longitud fija al administrar 32 ítems, 2) longitud variable al alcanzar un error de estimación menor o igual a .4, y 3) longitud variable al alcanzar un error de estimación menor o igual a .3. Se consideró probar esta última variante para verificar cómo funcionaba el TAI con un criterio más exigente.

Procedimiento

Los estudiantes evaluados respondieron a todos los ítems del Banco Completo de Analogías Verbales en una administración informatizada y adaptativa a través de una computadora personal portátil que disponía del software completo para una sesión de evaluación (FastTEST Pro 2.0 de Weiss, 2008). Por lo tanto, se trató de una administración individual bajo la supervisión del evaluador. Se brindó el tiempo suficiente para completar adecuadamente la evaluación. Para motivar a los participantes en la realización de la tarea se efectuó previamente una charla en donde se les explicó la finalidad de la actividad y la futura utilización de los datos recogidos en una investigación. Los alumnos firmaron su consentimiento y respondieron de forma voluntaria y anónima. No recibieron recompensa por su participación. Se han seguido las normas éticas pertinentes al tipo de procedimiento y población (AERA, APA & NCME, 1999; Colegio Oficial de Psicólogos e ITC, 2000).

Dado que un TAI debería proporcionar, con un número reducido de ítems, un nivel de habilidad aproximado al que obtendría la persona si respondiera a todos los ítems del BI (Bartram & Hambleton, 2006; Olea y Ponsoda, 2013; van der Linden & Glas, 2010), se decidió aplicar a un mismo grupo de sujetos el Banco de Ítems de Analogías Verbales en formato adaptativo (Eggen, 2004; van der Linden & Glas, 2010). Esto se consiguió programando el TAI con un criterio de parada fijo por el cual finalizaría al presentar todos los ítems que componían el BI (64 reactivos). De esta manera, para cada evaluado se obtendría la estimación de su nivel de rasgo (y error) no sólo al completar el BI completo sino también al alcanzar las diferentes variantes de criterios de parada: longitud fija de 32 ítems, longitud variable con error de .40 y longitud variable con error de .30.

Análisis de datos

Antes de realizar los estudios sobre las propiedades psicométricas del TAI, se examinaron los reportes de cada evaluado y se eliminaron los detectados como anómalos (fallo en las estimaciones y patrón de respuesta incoherente). Fueron considerados patrones de respuesta incoherente los gráficos de rendimiento que mostraban: a) aciertos en ítems difíciles mientras que se fallaban ítems fáciles, b) error de estimación constante o creciente y c) menos de 30% de aciertos a partir del quinto ítem (la pauta normal es acertar aproximadamente el 50% de los reactivos) (Gershon & Bergstrom, 1995). De cada reporte se tomaron los siguientes datos:

- Edad y sexo.
- Nivel de rasgo asignado al evaluado tras completar el BI completo (64 reactivos), simbolizado como θ_{64} y su correspondiente error de estimación EEE(θ_{64}).
- Nivel de rasgo asignado al evaluado tras completar los primeros 32 ítems del TAI, simbolizado como θ_{32} y su correspondiente error de estimación EEE(θ_{32}). Estos valores fueron utilizados para evaluar la condición: TAI de longitud fija de 32 reactivos.

- Nivel de rasgo asignado al evaluado cuando el TAI consiguió un error de estimación igual o inferior a 0.40, simbolizado como $\theta_{0.4}$ y la cantidad de ítems administrados hasta ese momento de la evaluación, $m_{0.4}$. Estos valores fueron utilizados para evaluar la condición: TAI de longitud variable fijando un nivel de error ≤ 0.40.
- Nivel de rasgo asignado al evaluado cuando el TAI consiguió un error de estimación igual o inferior a 0.30, simbolizado como $\theta_{0.3}$ y la cantidad de ítems administrados hasta ese momento de la evaluación, $m_{0.3}$. Estos valores fueron utilizados para evaluar la condición: TAI de longitud variable fijando un nivel de error ≤ 0.30.
- Proporción de respuestas correctas obtenido por el evaluado tras completar el BI completo (64 reactivos), simbolizado como P_{64}.

Estos datos se analizaron mediante: estadísticos descriptivos, diferencias de medias para muestras independientes y para muestras relacionadas, y correlaciones *r* de Pearson.

La TRI proporcionó el nivel de precisión obtenido por cada nivel de rasgo. De esta manera, se pudo estudiar la eficacia de la evaluación realizada tanto a través del BI completo como a través del TAI mediante el error de estimación obtenido para cada evaluado. Siguiendo la recomendación de Olea y Ponsoda (2013), se obtuvo el error de estimación medio como dato ilustrativo de la eficacia del TAI cuando se emplea un criterio de parada de longitud fija. Para los TAIs de longitud variable se obtuvo como indicador de su eficiencia la cantidad media de reactivos que se requirió para alcanzar el nivel de error prefijado. Para determinar si cada variante del TAI brindaba un nivel de habilidad aproximado al que obtenía la persona al responder a todos los ítems del BI, se correlacionaron por un lado los niveles de habilidad estimados a partir de las respuestas al BI completo con los estimados teniendo en cuenta los distintos criterios de parada considerados para el TAI. También se correlacionaron los errores de estimación de los niveles de habilidad obtenidos mediante el BI con los errores de estimación del TAI de longitud fija de 32 ítems.

Resultados

En primer lugar, se examinaron los 108 reportes (gráficos de rendimiento) obtenidos de la administración del TAI a cada estudiante para evaluar la progresión temporal de las respuestas, la evolución de la estimación de θ y su EEE(θ) en cada momento del proceso (Bergstrom & Gershon, 1992; Stocking, 1997). Se eliminaron un total de 12 casos que presentaron un patrón de respuestas incoherente.

Para la muestra depurada de 96 evaluados se obtuvieron los resultados presentados en la Tabla 2. Se puede observar que en todos los casos se pudo obtener, además de los resultados de administrar el BI completo, el nivel de rasgo estimado y su error bajo la modalidad de TAI de longitud fija de 32 ítems. Como los errores para la estimación de θ fueron diferentes para cada evaluado, no todos los examinados registraron EEE(θ) menores a .40. Al 91% (n = 87) de los participantes se les asignó un θ con un nivel de error inferior a .40 (TAI con error ≤ .40). Mientras que sólo el 56% (n = 54) presentó un error de estimación inferior al .30 y pudo obtener una estimación de θ mediante el TAI con error ≤ .30.

En cuanto a los resultados de la administración del BI completo (64 ítems), se obtuvo una estimación de θ media de .04 (DE=.86), oscilando entre los valores -1.43 y 2.24. Su correlación con la proporción de respuestas correctas fue de .99 (p<.001). Esta última tuvo una media de .59 (DE=.15) con un valor mínimo de .30 y un valor máximo de .92. El valor mínimo fue superior a la posibilidad de acierto por azar para los ítems de elección múltiple con cuatro alternativas. Es decir, el evaluado menos habilidoso obtuvo un 30% de respuestas correctas al test. Tanto con el puntaje obtenido en el marco de la TRI como con la proporción de respuestas correctas se observó que la muestra de evaluados tuvo un desempeño en torno al término medio de la escala. No se encontró una diferencia significativa en el rendimiento en analogías verbales entre varones y mujeres (prueba t para muestras independientes: Puntaje TRI $t(82) = -1.31$, $p = .19$; Puntaje Clásico $t(82) = -1.56$, $p = .12$).

Tabla 2

Resultados de la administración del BI completo y de las distintas versiones del TAI

	BI-AV Completo 64 ítems			TAI-AV 32 ítems		TAI-AV Con error ≤ 0.4		TAI-AV Con error ≤ 0.3	
	θ_{64}	EEE(θ_{64})	P_{64}	θ_{32}	EEE(θ_{32})	$\theta_{0.4}$	$m_{0.4}$	$\theta_{0.3}$	$m_{0.3}$
Media	.04	.31	.59	.12	.36	.31	21.66	.42	45.65
Desvío	.86	.04	.15	.92	.04	.89	5.22	.56	7.10
Mínimo	-1.43	.25	.30	-1.44	.31	-1.33	16	-.40	33
Máximo	2.24	.44	.92	2.09	.45	2.05	41	1.58	63
N° casos*	96			96		87		54	

Nota. Nivel de rasgo estimado para el evaluado al completar el BI = θ_{64} y su error de estimación EEE(θ_{64}).

Nivel de rasgo estimado para el evaluado al completar los primeros 32 ítems del TAI = θ_{32} y su error de estimación EEE(θ_{32}). Nivel de rasgo estimado para el evaluado al conseguir un error de estimación ≤ .40 = $\theta_{0.4}$ y la cantidad de ítems administrados = $m_{0.4}$. Nivel de rasgo estimado para el evaluado al conseguir un error de estimación ≤ .30 = $\theta_{0.3}$ y la cantidad de ítems administrados = $m_{0.3}$. P_{64} = proporción de respuestas correctas al BI. * Se refiere a la cantidad de evaluados que cumplieron con cada condición.

El error en la estimación de θ obtenido al responder el BI completo indicó el error mínimo que se podía obtener en la evaluación para cada participante, ya que la mejor estimación de su nivel de habilidad (en términos de precisión) es la que se logra al administrar todos los ítems que componen un BI. Para esta muestra el EEE(θ_{64}) mostró una media de .31 (*DE* = .04) con un mínimo y un máximo de .25 y .44 respectivamente. Se podría decir que la evaluación mediante el BI presentó aproximadamente una confiabilidad clásica promedio de .90, cuyos mínimo y máximo fueron .81 y .94 respectivamente.

La Figura 1 exhibe el diagrama de dispersión de los EEE(θ_{64}) con respecto a cada uno de los niveles estimados de θ_{64}. Para los niveles de rasgo centrales el error fue menor y hacia los extremos este fue

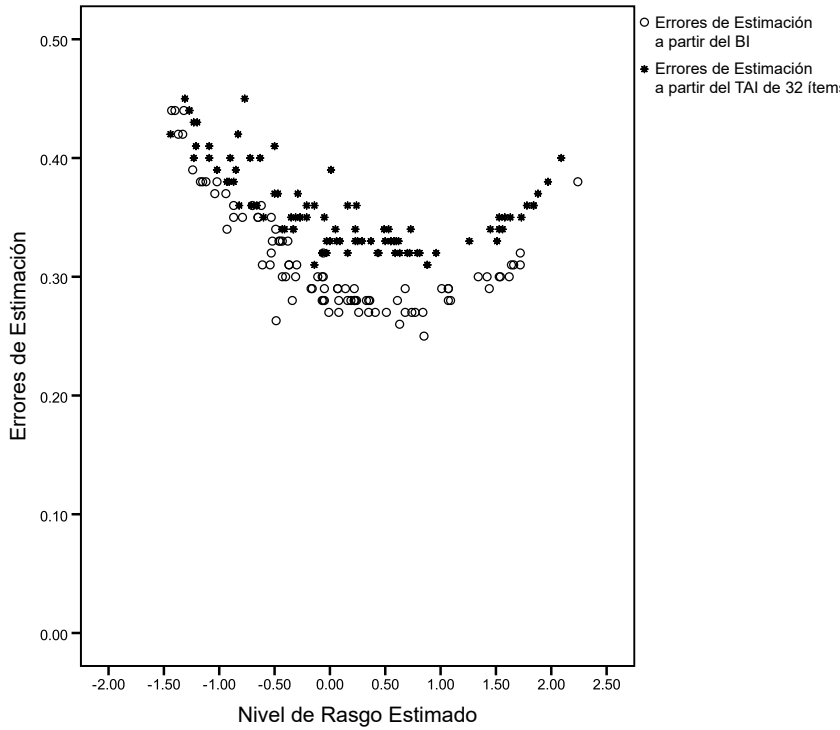

Figura 1. Diagrama de dispersión de los errores de estimación con respecto a los niveles de rasgo estimados a partir de la administración del BI completo y de la administración del TAI de longitud fija de 32 ítems.

aumentando, aunque no en forma simétrica ya que el BI tenía más ítems informativos de dificultad alta que baja. Un error ≤ .30 se verificó para los niveles de θ entre -.40 y 1.60.

Para el TAI de 32 ítems, se encontró un nivel estimado de rasgo medio de .12 (*DE* = .92), oscilando entre -1.44 y 2.09. La media del error de estimación fue .36 (*DE* = .04) y varió entre .31 y .45 (Figura 1). Al administrar una menor cantidad de reactivos (50% menos) se verificó un error levemente superior al obtenido con los 64 ítems del BI (.36 contra .31). No obstante, se alcanzó un nivel de precisión muy bueno ya que como mínimo le correspondería una confiabilidad clásica

de .80 y la confiabilidad promedio sería de .87. Además, las correlaciones fueron altas y positivas entre los valores θ estimados con 32 y 64 ítems (r = .97, p < .001) y, también, entre sus errores de estimación (r = .95, p < .001). Esto se debió a que los reactivos presentados en primer lugar eran los más adecuados para evaluar a cada persona.

Con respecto al nivel de rasgo estimado cuando el TAI alcanzó un error ≤ .40, se halló un θ medio de .31 (*DE* = .89, mínimo = -1.33 y máximo = 2.05). Se necesitaron en promedio 21.66 (*DE*=5.22) reactivos para alcanzar este criterio de parada, con un mínimo de 16 y un máximo de 41 ítems. Fueron 87 participantes los que cumplieron con este nivel de precisión (95%) y la mayoría de ellos (n = 78) requirieron menos de 30 reactivos (Figura 2). Los evaluados con menores niveles

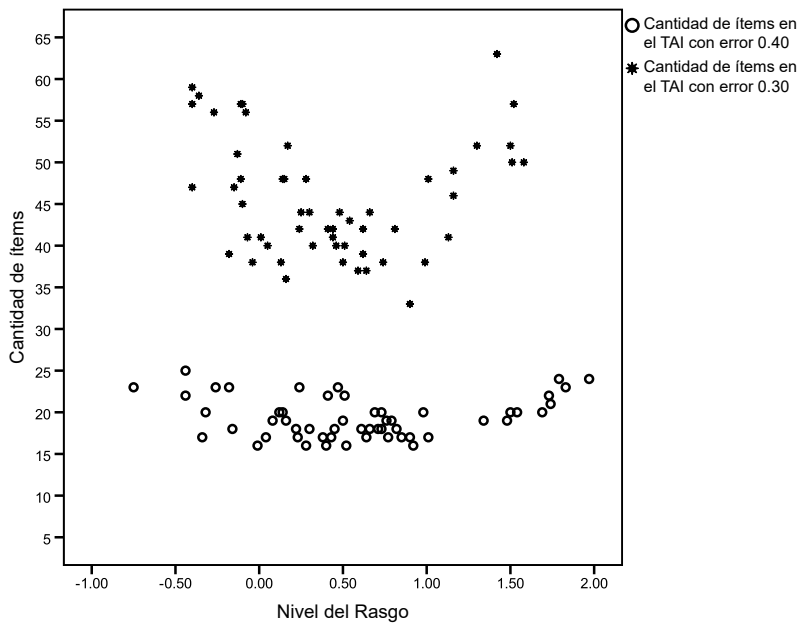

Figura 2. Diagrama de dispersión de la cantidad de ítems necesaria para alcanzar el criterio de parada con respecto a los niveles de rasgo estimados en el TAI de longitud variable con error ≤ 0.40 y con error ≤ 0.30.

de habilidad fueron los que necesitaron más ítems. Mientras que para los valores centrales de θ se requirieron menos reactivos. Los niveles de rasgo estimados al alcanzar un error ≤ .40 correlacionaron de manera positiva y alta con los estimados con el BI (r = .90, p < .001).

Al revisar los gráficos de rendimiento de las cinco personas que no alcanzaron un EEE(θ_{64}) ≤ .40 al finalizar el BI, se encontró que el error se mantenía estable entre la administración del ítem número 32 y del ítem número 64. Esto indicaría que para estos evaluados continuar presentando ítems no tenía sentido, ya que no aportaba más precisión a la evaluación. Se trataba de personas con bajo nivel de habilidad (θ < -1.3) para los cuales no había suficientes ítems informativos. Como el TAI presentaba en primer lugar los reactivos más informativos, se llegaba rápidamente a una meseta en el EEE(θ).

El criterio de parada más exigente (TAI con error ≤ .30) obtuvo las estimaciones del nivel de habilidad que más se acercaron a las estimadas a partir de las respuestas al BI completo (r = .98, p < .001). No obstante, esta alta precisión meta no pudo ser alcanzada en todos los casos. Sólo el 56% de la muestra evaluada cumplió este criterio (n = 54) y necesitaron responder en promedio 45 ítems (*DE*=7.10, mínimo=33 y máximo=63). Ninguno de los evaluados consiguió este nivel de precisión fijando el TAI en 32 ítems. Sólo el 25% lo logró con la presentación de 40 reactivos o menos (Figura 2). Para los valores centrales de θ se alcanzó la precisión meta administrando menos reactivos que para los extremos. La media en el nivel de rasgo fue 0.42 (*DE*=.56, mínimo=-.40 y máximo=1.58).

Cuando se comparó el criterio de parada al alcanzar un EEE(θ) ≤ .40 con el de longitud fija en 32 ítems, se constató que 83 personas hubieran terminado el TAI antes de responder a 32 reactivos. Sin embargo, para estos evaluados se obtenían mejores estimaciones de su nivel de habilidad (i.e., con menor error) si se utilizaba el criterio de parada fijo en 32 ítems. Se verificó que en este punto las estimaciones para estas personas registrarían un error en promedio de .35 (*DE*=.02) con un valor mínimo de 0.31 y un valor máximo de 0.40. Evidentemente, el criterio de finalización fijo de 32 ítems resulta más eficiente

que el de longitud variable con un error de 0.40. Además, para las cuatro personas que alcanzaron este nivel de error con la presentación de más reactivos (entre 33 y 41 ítems), se constató que tras la presentación de 32 ítems obtuvieron estimaciones de θ con errores de .41 y .42. Esto apoyaría la decisión de finalizar el TAI con un criterio de longitud fija en 32 ya que el beneficio de continuar la presentación de ítems no fue importante en cuanto a mejorar la precisión de la evaluación.

Discusión

La administración del TAI modificando su especificación en cuanto al criterio de parada permitió evaluar la precisión de las estimaciones para el nivel de rasgo de los participantes bajo distintas condiciones de parada y determinar cuál sería la más adecuada dadas las características del BI de Analogías Verbales. En concreto, se compararon las estimaciones obtenidas a partir de responder al BI completo, del TAI de longitud fija de 32 ítems, del TAI de longitud variable fijando un error de estimación ≤ .40 y del TAI de longitud variable fijando un error de estimación ≤ .30.

Se constató un adecuado funcionamiento de todos los módulos que conformaban el TAI (instrucciones, test propiamente dicho, finalización y reporte). En todos los casos, el TAI presentó como primer ítem uno de dificultad baja según lo especificado. El proceso de selección de los ítems se llevó adelante sin inconvenientes.

Con respecto a los casos en los que el programa FastTEST Pro no pudo estimar θ y/o su EEE(θ), se encontraron distintas explicaciones. Para algunos evaluados se constató un patrón de respuestas incoherente (i.e., se aciertan ítems difíciles y se fallan ítems fáciles) y para otros se observó que su proporción de respuestas correctas era inferior a la que se obtendría por responder aleatoriamente a los reactivos. Ambas circunstancias no eran lo esperado por el modelo de la TRI utilizado para la estimación de θ. El ML3P indica, por un lado, que cuanto más fácil es un reactivo más probable es dar la respuesta correcta y, por el otro,

que aún para niveles muy bajos de habilidad es probable responder correctamente al ítem por azar. Como la imposibilidad de estimar θ sólo se presenta cuando se utiliza el ML3P (Hambleton, Swaminathan & Rogers, 1991), esto condujo a que fallara el proceso de estimación. Otra explicación podría hallarse en el *stepsize* utilizado por el programa FastTEST Pro para forzar un patrón de respuesta mixta. Si se da una respuesta correcta, θ se establece en 4; mientras que para las respuestas incorrectas, θ se establece en -4. Esta estrategia, que el programa no permite modificar, se repite hasta que el patrón de respuestas del evaluado deje de ser constante y pueda implementarse la estimación por máxima verosimilitud. Aunque van der Linden y Pashley (2010) señalaron que las formas de estimar θ en los inicios del TAI sólo repercuten en tests con menos de 10 ítems, Dodd (1990) encontró que los casos en que no se alcanzaba la convergencia en la estimación de θ eran mayores con este procedimiento ya que el θ estimado puede exceder el rango de dificultad del BI con la administración de algunos pocos reactivos.

También hubo otros casos en los cuales se alcanzó una estimación de θ pero su EEE(θ) era demasiado grande o no decrecía con la presentación de los siguientes ítems. Los errores de estimación grandes se evidenciaron en los evaluados que presentaron θ estimados inferiores a -2. Esto se debió a que el BI no disponía de la cantidad de ítems suficientemente informativos para el extremo inferior del rasgo. Por otro lado, el EEE(θ) se mantenía constante o subía y bajaba en forma alternada en los evaluados con patrones de respuesta incoherentes. Como señalaron Bock y Mislevy (1982) y Embretson y Reise (2013), las buenas propiedades estadísticas del procedimiento de estimación por máxima verosimilitud dependen del supuesto de que las respuestas del evaluado se ajusten al modelo. En todos los casos, la inconsistencia del patrón de respuestas determina que disminuya la verosimilitud para todos los niveles de θ con respecto a ese patrón. Era importante tener en cuenta esto al momento de evaluar la interpretabilidad del puntaje θ estimado para una persona y debido a ello estos casos fueron eliminados de los ulteriores análisis. Los patrones de respuesta incoherentes suelen presentarse cuando se responde al test por azar o cuando las respuestas

se ven afectadas por otras variables (e.g., distracción, descuido, cansancio, falta de motivación o compromiso). Como esta evaluación no implicaba ninguna consecuencia para los participantes podría haber sucedido que algunos de ellos la completaran sin prestar la suficiente atención, aún cuando se ofrecía la posibilidad de cesar en cualquier momento su colaboración voluntaria. Tampoco se pudo volver a evaluar a los sujetos con patrones incoherentes porque la prueba se respondió de manera anónima.

En todos los casos con puntajes θ válidos se pudo obtener, además de los resultados de administrar el BI completo, el nivel de rasgo estimado y su error bajo la modalidad de TAI de longitud fija de 32 ítems. Diferente fue la situación cuando se buscó alcanzar un determinado nivel de precisión (criterios con errores ≤ .30 y .40). Para cada evaluado, tanto las estimaciones de θ provisionales como la final se realizaban con distinto error. Cuanto más se alejaba el patrón de respuesta del evaluado del esperado por el ML3P, más grande resultaba el error de estimación. Asimismo, si el evaluado poseía un nivel de rasgo para el cual el BI disponía de pocos ítems informativos, la estimación de su nivel de rasgo también presentaba mayor error. Por lo tanto, no todos los casos alcanzaron el nivel de precisión prefijado. El 91% de los individuos pudo ser evaluado con un nivel de error ≤ .40; mientras que sólo el 56% cumplió con el criterio más exigente de un error ≤ .30. Este hecho no debe interpretarse como una limitación de la evaluación mediante TAIs sino como una limitación del BI (no disponer de la cantidad suficiente de ítems informativos para los niveles inferiores del rasgo). Todo TAI depende del BI en el que se sustenta (i.e., de sus propiedades psicométricas) y nunca podrá obtener mayor precisión que la obtenida a través de la administración del BI completo. En conclusión, el EEE(θ) obtenido al responder al BI significó el error mínimo (i.e., la precisión máxima) que se podía alcanzar en la evaluación de cada participante.

Al administrar el BI completo, tanto los valores θ estimados para la habilidad como los puntajes clásicos indicaron un rendimiento promedio en torno al término medio de la escala y similar al obtenido por

las muestras utilizadas para la calibración del BI (Lozzia et al., 2015). También, en consonancia con los resultados de las fases de calibración del BI, no se halló una diferencia significativa en el desempeño en analogías verbales entre varones y mujeres.

Se analizó cada una de las variantes del TAI teniendo en cuenta que un TAI eficiente debería cumplir los siguientes requisitos (Muñiz & Hambleton, 1999): a) nivel de habilidad estimado aproximado al que obtendría la persona al responder a todos los ítems del BI, b) error de estimación adecuado (i.e., los valores estimados para θ eran precisos), c) presentación de igual o menor cantidad de ítems que los TC utilizados en los estudios de calibración del BI, d) criterio de parada alcanzable por todos los evaluados.

Como se esperaba según los estudios clásicos (Bartram & Hambleton, 2006; Davey & Pitoniak, 2006; van der Linden & Glas, 2010; Walter & Holling, 2008), se confirmó que todas las variantes del TAI de Analogías Verbales aquí analizadas reproducían los niveles de habilidad estimados a partir de las respuestas al BI (requisito a). Al comparar cada uno de los niveles de θ estimados bajo las distintas condiciones, se encontró que cada una de estas correlacionaba positiva e intensamente ($r > .90$) con las estimaciones del BI. Sin embargo, algunos criterios de parada resultaron más convenientes que otros. Con respecto al requisito a), las variantes TAI de longitud variable con error ≤ .30 y TAI de 32 ítems obtuvieron las mejores correlaciones ($r = .98$ y $r = .97$ respectivamente). Estos resultados están en la línea de lo considerado como correcto por Thompson (2009), quien espera correlaciones superiores a .95. Por otro lado, los niveles estimados bajo la condición de longitud variable al alcanzar un error ≤ .40 fueron los más alejados de los resultados obtenidos con el BI ($r = .90$). Esto era esperable ya que esta condición era la menos rigurosa. Es decir, requería la presentación de menos ítems (en promedio 21 y para el 90% de los casos menos de 30 reactivos) y se obtenía una estimación de θ con un EEE(θ) igual para todos los evaluados pero en el límite de lo aceptable.

Todas las variantes del TAI obtuvieron estimaciones precisas de los niveles de θ (requisito b). Este requerimiento fue, por definición,

cumplido (aunque en distinta medida) en los TAIs cuyo final implicaba alcanzar un determinado nivel de error: TAI de longitud variable con error ≤ .30 y TAI de longitud variable con error ≤ .40. Por lo tanto, era importante confirmar el requisito b) en el caso del TAI de 32 ítems, ya que los TAIs de longitud fija brindan estimaciones de θ con diferente nivel de error para cada evaluado. Bajo esta condición el EEE(θ) medio fue .36, variando entre .31 y .45. Entonces, ningún individuo fue evaluado con confiabilidad clásica menor a .80.

Mientras que el TAI de 32 ítems cumplió por definición con el requisito c) de presentar una cantidad de ítems similar a las versiones en formato convencional, fue necesario confirmar la adecuación de las variantes de longitud variable. El TAI de longitud variable con error ≤ .40 requirió la presentación de la menor cantidad de ítems (menos de 30 en el 90% de los casos y en promedio 21 reactivos). Pero, el TAI de longitud variable con error ≤ .30 falló en alcanzar este objetivo, ya que necesitó más ítems (en promedio 45) para alcanzar su criterio de parada más exigente. Relacionado con esto se encuentra el hecho de que sólo el 56% de los evaluados alcanzó este criterio de finalización. Por lo tanto, esta condición tampoco cumplió el requisito d). Al ser la precisión meta menos exigente en el TAI de longitud variable con error ≤ .40, se constató que sólo un 5% de los participantes no lo alcanzaban.

En síntesis, la metodología TAI aquí implementada proporciona, con la administración de una parte de los ítems que componen el BI, estimaciones precisas de los niveles de habilidad de los evaluados que reproducen los resultados que se obtendría al responder al BI completo. Todas las variantes obtuvieron estimaciones de θ cercanas a las obtenidas al responder al BI completo y con EEE(θ) considerados adecuados.

El TAI de longitud variable con error ≤ .30 fue el criterio de parada más exigente. Por ello, si bien era el más preciso y el que mejor reprodujo los niveles de θ estimados por el BI, requirió presentar demasiados reactivos y muchos evaluados no llegaron a cumplir este criterio de finalización. Esto no se debió a una deficiencia del TAI sino, como ya se explicó, a una limitación del BI en cuanto a la distribución de su FI.

Aunque el TAI de longitud variable con error ≤ .40 fue el que menos ítems requirió, esto sucedió porque era el menos preciso. Por este motivo fue el que más se alejó de las estimaciones de θ obtenidas al responder al BI completo. Esta condición permitió verificar que, en términos clásicos, la confiabilidad del TAI (.84) era similar a la que se obtuvo en los TC administrados para la calibración del BI pero requirió administrar muchos menos ítems (45% menos). Esto fue consistente con los hallazgos teóricos y empíricos que demostraban que un TAI, administrando aproximadamente la mitad de reactivos que su equivalente convencional, era igual de eficiente (McBride & Martin, 1983; Segall & Moreno, 1999).

En conclusión, el TAI de 32 ítems cumplió con los cuatro requisitos y mostró un balance óptimo entre precisión y cantidad de reactivos presentados. Con la administración de la mitad de los ítems del BI, sus estimaciones de θ y sus EEE(θ) fueron muy cercanos a los obtenidos al responder al BI completo (64 ítems). La precisión alcanzada fue mayor que la conseguida en las versiones de lápiz y papel utilizadas en los estudios de calibración del BI. Mientras que para estas pruebas los análisis indicaron errores de estimación entre .37 y .60, el TAI de 32 ítems presentó valores entre .31 y .45. Asimismo, al analizar los gráficos de rendimiento de todos los participantes se comprobó que continuar presentando más ítems no conducía a una mejora sustancial en la precisión de la evaluación (la diferencia media fue de .04 y la máxima de .09). Cada reactivo adicional aportaba muy poca información ya que los ítems más informativos para el nivel de rasgo de cada evaluado fueron los que el TAI presentó en primer lugar. En especial, se constató que los casos que no alcanzaban un nivel de precisión aceptable bajo esta condición tampoco lo lograban al completar el BI. Un hallazgo similar reportan los constructores del CAT-ASVAB (McBride, Wetzel & Hetter, 1997). Al realizar un estudio de simulación no encontraron ventajas del criterio de parada de longitud variable por sobre el de longitud fija. Los ítems más informativos se encontraban en un rango de θ restringido, por lo que los evaluados con θ por fuera de este rango tendían a recibir tests más largo y, sin embargo, cada ítem adicional

aportaba muy poca información. Ellos indicaron que, en estos casos, un TAI de longitud variable significaba un uso ineficiente del tiempo y del esfuerzo del evaluado.

Una de las limitaciones del BI que parece afectar al TAI es que la FI no es uniforme a lo largo de todo el espectro de la habilidad. Por lo tanto, la evaluación de las personas con muy bajo nivel en el rasgo se efectuará con mayor nivel de error. Frente a este hecho sería conveniente agregar más ítems al BI para los niveles que disponen de pocos ítems informativos. Algo similar sucedió en el desarrollo de otros TAIs (e.g., Hetter & Sympson, 1997; Ponsoda, Olea & Revuelta, 1994; Olea, Abad, Ponsoda & Ximénez, 2004)

Otra limitación importante remite a las características particulares de la muestra utilizada en el presente estudio. La homogeneidad de la muestra afecta la posibilidad de generalizar los resultados obtenidos a otras poblaciones. Por ello, sería interesante trabajar con muestras de otras poblaciones. Esto, también, permitiría obtener indicadores de la invarianza de los parámetros estimados en otras poblaciones metas.

Futuros estudios tratarán otras variantes en el algoritmo adaptativo. Por ejemplo, probar diferentes criterios de inicio para el TAI. También sería conveniente realizar comparaciones con otros procedimientos de finalización como el criterio de parada mixto.

Finalmente, los resultados de esta primera implementación del TAI de Analogías Verbales fueron alentadores ya que demostraron que se puede evaluar el constructo de Analogías Verbales de una manera más rápida y precisa. Es importante señalar que los hallazgos aquí relatados no son válidos para todos los TAIs sino sólo para los TAIs que sustenten en el mismo BI y tengan los mismos objetivos de evaluación. Ambos puntos definen las características que tendrán los TAIs que se diseñen y también los beneficios y limitaciones que tendrá la implementación de este tipo de evaluaciones.

Referencias

Abuchar, A. A. & Simanca, F. (2013). e-learning en procesos de evaluación académica; Pruebas Saber Pro. *Vínculos, 10*(1), 360-372.

American Educational Research Association (AERA), American Psychological Association (APA) & National Council on Measurement in Education (NCME). (1999). *Standars for educational and psychological testing.* Washington: APA.

Attorresi, H., Lozzia, G., Abal, F., Galibert, M. & Aguerri, M. (2009). Teoría de Respuesta al Ítem. Conceptos básicos y aplicaciones para la medición de constructos psicológicos. *Revista Argentina de Clínica Psicológica, 18*, 179-188.

Attorresi, H., Pano, C., Fernández Liporace, M. & Cayssials, A. (1993). Evaluación de la habilidad para identificar y discriminar relaciones. *Anuario de Investigaciones, 3*, 27-34.

Barbero, M. (1996). Banco de ítems. En J. Muñiz (Ed.), *Psicometría* (pp. 139-170). Madrid: Universitas.

Barrada, J. (2012). Tests adaptativos informatizados: Una perspectiva general. *Anales de Psicología, 28*, 289-302.

Barrada, J., Abad, F. J. & Olea, J. (2014). Optimal number of strata for the stratified methods in computerized adaptive testing. *Spanish Journal of Psychology, 17*, e48. https://doi.org/10.1017/sjp.2014.50

Bartram, D. & Hambleton, R. (2006). *Computer-based testing and the internet: Issues and advances*. Chichester, West Susex: Wiley. https://doi.org/10.1002/9780470712993

Beckmann, J., Hung, M., Bounsanga, J., Wylie, J., Granger, E. & Tashjian, R. (2015). Psychometric evaluation of the PROMIS Physical Function Computerized Adaptive Test in comparison to the American Shoulder and Elbow Surgeons score and Simple Shoulder Test in patients with rotator cuff disease. *Journal of Shoulder and Elbow Surgery, 24*(12), 1961-1967. https://doi.org/10.1016/j.jse.2015.06.025

Bergstrom, B. & Gershon, R. (1992). *Computer adaptive testing: Using individual student maps to understand test performance.* Trabajo presentado en Annual Meeting of American Educational Research Association, San Francisco, California.

Bock, R. D. & Mislevy, R. (1982). Adaptative EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431-444. https://doi.org/10.1177/014662168200600405

Chang, H. H. (2015). Psychometrics Behind Computerized Adaptive Testing. *Psychometrika, 8,* 1-20. https://doi.org/10.1007/s11336-014-9401-5

Colegio Oficial de Psicólogos & International Test Commission (ITC). (2000). Directrices internacionales para el uso de los tests. *Infocop, 77*, 21-32.

da Cunha, S. M. A. & Nogueira, C. R. T. (2015). Aplicação de Testes Adaptativos Computadorizados em Modelos de Desdobramento Graduado Generalizados. *Blucher Mathematical Proceedings, 1*(1), 766-778.

Davey, T. & Pitoniak, M. J. (2006). Designing computerized adaptive tests. En S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development*. Mahwah,NJ: LEA.

Devine, J., Fliege, H., Kocalevent, R., Mierke, A., Klapp, B. & Rose, M. (2016). Evaluation of Computerized Adaptive Tests (CATs) for longitudinal monitoring of depression, anxiety, and stress reactions. *Journal of Affective Disorders, 190*, 846-853. https://doi.org/10.1007/s11336-014-9401-5

Dodd, B.G. (1990). The effect of ítem selection procedure and stepsize on computerized adaptive attitude measurement using the rating scale model. *Applied Psychological Measurement, 14*, 355-366. https://doi.org/10.1177/014662169001400403

Drasgow, F. (2015). *Technology and testing: Improving educational and psychological measurement*. Nueva York, NY: Taylor and Francis Inc. https://doi.org/10.4324/9781315871493

Educational Testing Service (2016). *GRE information and registration bulletin*. Princeton, NJ: Author.

Eggen, T. J. H. M. (2004). *Contributions to the theory and practice of computerized adaptive testing.* Amsterdam: Citogroep.

Embretson, S. E. & Reise, S. P. (2013). *Item response theory*. Psychology Press. https://doi.org/10.4324/9781410605269

Escurra Mayaute, Miguel & Salas Blas, Edwin (2014). Construcción y validación del cuestionario de adicción a redes sociales (ARS). *Liberabit*, *20*(1), 73-91. Recuperado en 18 de febrero de 2017, de http://www.scielo.org.pe/scielo.php?script=sci_arttext&pid=S1729-482720140001 00007 &lng=es&tlng=es.

Fonseca-Pedrero, E., Menéndez, L.F., Paino, M., Lemos-Giráldez, S. & Muñiz, J. (2013) Development of a Computerized Adaptive Test for Schizotypy Assessment. *PLoS ONE, 8* (9). https://doi.org/10.1371/journal.pone.0073201

Galibert, M., Aguerri, M., Pano, C., Lozzia, G. & Attorresi, H. (2005). Análisis de Ítem de Analogías Verbales mediante la Aplicación de un Modelo Politómico de la Teoría de Respuesta al Ítem. *Revista Mexicana de Psicología, 22*, 419-431.

García, P., Abad, F., Olea, J. & Aguado, D. (2013). A new IRT-based standard setting method: Application to eCAT-Listening. *Psicothema, 25*, 238-244.

Gentner, D., Holyoak, K. J. & Kokinov, B. N. (2001). *The analogical mind: Perspectives from cognitive science*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/1251.001.0001

Gershon, R. & Bergstrom, B. (1995). *Does cheating on CAT pay: NOT!* Trabajo presentado en Annual Meeting of American Educational Research Association, San Francisco, CA.

Gibbons, R., Weiss, D., Frank, E. & Kupfer, D. (2016). Computerized Adaptive Diagnosis and Testing of Mental Health Disorders. *Annual Review of Clinical Psychology, 12*, 83-104. https://doi.org/10.1146/annurev-clinpsy-021815-093634

Hambleton, R., Swaminathan, H. & Rogers, H. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage.

Hernández, A., Tomás, I., Ferreres, A. & Lloret, S. (2015). Tercera evaluación de tests editados en España. *Papeles del Psicólogo, 31*(1), 1-8.

Hetter, R. & Sympson, J. (1997). Item exposure control in CAT-ASVAB. En W. Sands, B. Waters y J. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 141-144). Washington: APA. https://doi.org/10.1037/10244-014

Hey, J., Linsey, J., Agogino, A. M. & Wood, K. L. (2008). Analogies and metaphors in creative design. *International Journal of Engineering Education, 24*(2), 283.

Hol, A., Vorst, H. & Mellenbergh, G. (2008). Computerized adaptive testing of personality traits. *Journal of Psychology, 216,* 12-21. https://doi.org/10.1027/0044-3409.216.1.12

Jiménez, J. & Herrera, A. (2016, mayo). Test adaptativo informatizado para invidentes. Trabajo presentado en el *Primer Congreso Colombiano de Teoría de Respuesta al Ítem*. Bogotá, Colombia.

Jones, L. & Estes, Z. (2015). Convergent and divergent thinking in verbal analogy. *Thinking & Reasoning, 21*, 1-28. https://doi.org/10.1080/13546783.2015.1036120

Junior, M. & Pinto, A. (2015). *Uso do tempo de resposta para melhorar a convergência do algoritmo de testes adaptativos informatizados*. Tesis de Maestría. Universidade de Brasília, Brasil.

Kaplan, M., de la Torre., J. & Barrada, J. (2015). New item selection methods for cognitive diagnosis computerized adaptive testing. *Applied Psychological Measurement, 39*, 167-188. https://doi.org/10.1177/0146621614554650

Kuncel, N. & Hezlett, S. (2007). Standardized tests predict graduate students' success. *Science, 315*(5815), 1080-1081. https://doi.org/10.1126/science.1136618

Kuncel, N., Hezlett, S. & Ones, D. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all? *Journal of Personality and Social Psychology, 86*(1), 148-161. https://doi.org/10.1037/0022-3514.86.1.148

López-Cuadrado, J., Pérez, T. A., Vadillo, J. Á. & Gutiérrez, J. (2010). Calibration of an item bank for the assessment of Basque language knowledge. *Computers & Education, 55*(3), 1044-1055. https://doi.org/10.1016/j.compedu.2010.04.015

Lord, F. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: LEA.

Lozzia, G., Abal, F., Blum, D., Aguerri, M., Galibert, M. & Attorresi, H. (2015). Construcción de un Banco de Ítems de Analogías Verbales como base para un Test Adaptativo Informatizado. *Revista Mexicana de Psicología, 32*(2), 134-148.

Lozzia, G. & Attorresi, H. (2012). Especificación del algoritmo para un Test Adaptativo Informatizado de Analogías Verbales. *SUMMA Psicológica UST, 9*(2), 15-23. https://doi.org/10.18774/448x.2012.9.90

Lozzia, G., Picón Janeiro, J. & Galibert, M. S. (2008). La Evaluación del Razonamiento Verbal mediante el Formato de Analogías Verbales. *Memorias de las XV Jornadas de Investigación y 4° Encuentro de Investigadores en Psicología del Mercosur.* Facultad de Psicología, UBA. Tomo II, 474-476.

McBride, J. & Martin, J. (1983). Reliability and validity of adaptive ability tests in a military setting. En D. Weiss (Ed.), *New horizons in testting: Latent trait test theory and computerized adaptive testing* (pp. 223-236). NY: Academic Press. https://doi.org/10.1016/B978-0-12-742780-5.50022-6

McBride, J., Wetzel, C. & Hetter, R. (1997). Preliminary psychometric research for CAT-ASVAB: Selecting an adaptive testing strategy. En W. Sands, B. Waters & J. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 83-95). Washington: APA. https://doi.org/10.1037/10244-008

Meagher, D. (2012). *Miller analogies test: Reliability and Validity*. San Antonio, TX: NCS Pearson.

Montero, I. & León, O. (2005). Sistema de clasificación del método en los informes de investigación en Psicología. *International Journal of Clinical and Health Psychology, 5*, 115-127.

Moreira Junior, F., Tezza, R., Andrade, D. & Bornia, A. (2013). Algoritmo de um teste adaptativo informatizado com base na teoria da resposta ao item para a estimação da usabilidade de sites de e-commerce. *Production, 23* (3), 525-536. https://doi.org/10.1590/S0103-65132012005000095

Muñiz, J. & Hambleton, R. (1999). Evaluación psicométrica de los tests informatizados. En J. Olea, V. Ponsoda & G. Prieto (Eds.), *Tests informatizados: Fundamentos y aplicaciones*. (pp. 23-52). Madrid: Pirámide.

Olea, J., Abad, F., Ponsoda, V. & Ximénez, M. (2004). Un test adaptativo informatizado para evaluar el conocimiento del inglés escrito: diseño y comprobaciones psicométricas. *Psicothema, 16*, 519-525.

Olea, J. & Ponsoda, V. (2013). *Tests adaptativos informatizados*. Madrid: Ediciones UNED.

Olea, J., Ponsoda, V. & Prieto, G. (1999). *Tests informatizados: fundamentos y aplicaciones.* Madrid: Pirámide.

Piton-Gonçalves, J. & Aluísio, S. (2015). Teste Adaptativo Computadorizado Multidimensional com propósitos educacionais: princípios e métodos. *Ensaio: Avaliação e Políticas Públicas em Educação, 23*(87), 389-414. https://doi.org/10.1590/S0104-40362015000100016

Ponsoda, V., Olea, J. & Revuelta, J. (1994). ADTEST: A Computer Adaptive Test Based on The Maximum Information Principle. *Educational and Psychological Measurement, 54*, 680-686. https://doi.org/10.1177/0013164494054003011

Salcedo, P., Ferreira, A. & Barrientos, F. (2013). A Bayesian Model for Lexical Availability of Chilean High School Students in Mathematics. En *Natural and Artificial Models in Computation and Biology* (pp. 245-253). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-38637-4_25

Segall, D. & Moreno, K. (1999). Development of the Computerized Adaptive Testing Version of the Armed Service Vocational Aptitude Battery. En F. Drasgow & J. Olson-Buchanan (Eds.),

Innovations in computerized assessment (pp. 35-65). Mahwah, NJ: LEA.

Simanca, F. & Abuchar, A. (2014). *AEI - Algoritmo de Evaluación Inteligente. En Virtual Educa Innovación, competitividad y desarrollo*. Lima, Perú: Feijóo. Recuperado de http://www.virtualeduca.org/ponencias2014/129/AlgortimoevaluacioninteligenteAEI-Peru.pdf

Sistema Nacional de Educación Pública (2011). *Hacia la construcción de una agenda para la mejora educativa*. Disponible en http://educacion.mec.gub.uy/innovaportal/file/12416/1/ informe_agenda_mejora_educativa_ 5_diciembre_2011.pdf

Sternberg, R. (1985). *Beyond IQ: A triarchic theory of human intelligence.* Cambridge: Cambridge University Press.

Sternberg, R. (2001). *How to Prepare for the MAT-Miller Analogies Test.* Nueva York, NY: Barron's Educational Series.

Sternberg, R. (2015). Multiple intelligences in the new age of thinking. In *Handbook of Intelligence* (pp. 229-241). Nueva York, NY: Springer. https://doi.org/10.1007/978-1-4939-1562-0_16

Stocking, M. (1997). Revising item responses in computerized adaptive tests: A comparison of three models. *Applied Psychological Measurernent, 21*, 129-142. https://doi.org/10.1177/01466216970212003

Su, Y-H. (2016). A Comparison of Constrained Item Selection Methods in Multidimensional Computerized Adaptive Testing. *Applied Psychological Measurement,40*(5), 346-360. https://doi.org/10.1177/0146621616639305

Suárez-Álvarez, J. & Pedrosa, I. (2016). Evaluación de la personalidad emprendedora: situación actual y líneas de futuro. *Papeles del Psicólogo, 37*(1), 62-68.

Thompson, N. (2009). *Ability estimation with item response theory.* St. Paul, MN: Assessment Systems Corporation.

Thurstone, L. (1938). *The primary mental abilities*. Chicago, IL: University of Chicago Press.

Thurstone, L. (1940). Experimental Study of Simple Structure. *Psychometrika, 5*, 153-168. https://doi.org/10.1007/BF02287873

Toledo, G., Mezura Godoy, C., Cruz Ramírez, N. & Benítez Guerrero, E. (2013). Modelo de evaluación adaptativa y personalizada mediante razonamiento probabilista. *Conferencias LACLO, 4*(1), 283-294.

Tornimbeni, S., Pérez, E. & Olaz, F. (2008). *Introducción a la psicometría*. Buenos Aires: Paidós.

van der Linden, W. J. (Ed.) (2016). *Handbook of item response theory: Models, statistical tools, and applications* (*Vols.1-3*). Boca Raton, FL: Chapman & Hall/CRC. https://doi.org/10.1201/b19166

van der Linden, W. J. & Glas, C. E. W. (2010). *Elements of adaptive testing*. Nueva York, NY: Springer. https://doi.org/10.1007/978-0-387-85461-8

van der Linden, W. J. & Pashley, P. J. (2010). Item selection and ability estimation in adaptive testing. En W. J. van der Linden & C. E. W. Glas (Eds.), *Elements of adaptive testing* (pp. 3-30). Nueva York, NY: Springer. https://doi.org/10.1007/978-0-387-85461-8_1

Veldkamp, B. P. (2013). Ensuring the future of CAT. En T. J. H. M. Eggen & B. P. Veldkamp (Eds.), *Psychometrics in practice at RCEC* (pp. 137-150). Enschede: RCEC.

Veldkamp, B. & Matteucci, M. (2013). Bayesian computerized adaptive testing. *Ensaio: Avaliação e Políticas Publicas em Educação, 21*, 57-82. https://doi.org/10.1590/S0104-40362013005000001

Wainer, H., Dorans, N., Eignor, D., Flaugher, R., Green, B., Mislevy, R., Steinberg, L. & Thissen, D. (2000). *Computerized Adaptive Testing: A Primer*. (2a. Ed.). Mahwah, NJ: Erlbaum. https://doi.org/10.4324/9781410605931

Walter, O. & Holling, H. (2008). Transitioning from fixed-length questionnaires to computer-adaptive versions. *Zeitschrift für Psychologie / Journal of Psychology, 216*, 22-28. https://doi.org/10.1027/0044-3409.216.1.22

Wang, C., Zheng, C. & Chang, H. H. (2014). An Enhanced Approach to Combine Item Response Theory With Cognitive Diagnosis

in Adaptive Testing. *Journal of Educational Measurement, 51*(4), 358-380. https://doi.org/10.1111/jedm.12057

Weiss, D. (2008). *Manual for the FastTEST Professional Testing System, Version 2*. St. Paul, MN: Assessment Systems Corporation.

Wendler, C. & Bridgeman, B. (2014). *The Research Foundation for the GRE revised General Test: A compendium of studies*. Princeton, NJ: Educational Testing Service.

Yela, M. (Ed.). (1987). *Estudios sobre inteligencia y lenguaje*. Madrid: Pirámide.

Young, J., Klieger, D., Bochenek, J., Li, C. & Cline, F. (2014). The Validity of Scores from the GRE revised General Test for Forecasting Performance in Business Schools: Phase One. *ETS Research Report Series, 2014*(2), 1-10. https://doi.org/10.1002/ets2.12036

Recibido: 26 de enero, 2018
Revisado: 10 de junio, 2019
Aceptado: 27 de junio, 2019