Revista de Estudios y Experiencias en Educación

REXE

journal homepage: http://revistas.ucsc.cl/index.php/rexe

Estimación de la fiabilidad para instrumentos de medición adaptativos

José González Campos[a], Juan Aspeé Chacón[b] e Ida Sessarego-Espeleta[c]
Universidad de Playa Ancha[a]. Universidad Viña del Mar[b]. Universidad Técnica Federico Santa María[bc]. Chile

Recibido: 24 de mayo 2022 - Revisado: 08 de septiembre 2022 - Aceptado: 14 de septiembre 2022

RESUMEN

La estimación de la fiabilidad de pruebas evaluativas lineales y estandarizadas, de cualquier tipo, en ciencias sociales es relativamente simple con el uso del coeficiente de alpha de Cronbach. El problema ocurre con el uso de pruebas adaptativas que se ajustan a las respuestas que van entregando los sujetos. Ello se complejiza aún más cuando se utilizan pruebas adaptivas informatizadas, lo que tensiona el uso del alpha de Cronbach, puesto que requiere de estabilidad en su medición. Así, el presente trabajo expone el denominado método *Game-Adaptative* que permite determinar la fiabilidad de una prueba adaptativa considerando su naturaleza variable. Dicho método se somete a prueba con dos casos, uno simulado y otro real con base a una muestra de 50 estudiantes de postgrado de una universidad chilena, quienes se sometieron al *Oxford Online Placement Test*. Los resultados indican que la aplicación del método *Game-Adaptative* recoge adecuadamente la variabilidad de las respuestas de ítems utilizados en cada caso, pudiendo determinar sin mayor inconveniente la fiabilidad de los instrumentos en cuestión. En consecuencia, esta propuesta se transforma en una ruta plausible para analizar las pruebas de carácter adaptativo, contribuyendo a superar las limitaciones de los coeficientes tradicionales, al menos en términos de fiabilidad.

Palabras clave: Test adaptativo; fiabilidad; TICs; alpha de Cronbach; teoría clásica; teoría de respuesta al ítem.

*Correspondencia: José González Campos (J. González).
[a] https://orcid.org/0000-0003-4610-6874 (jgonzalez@upla.cl).
[b] https://orcid.org/0000-0003-3456-8414 (juan.aspee@usm.cl).
[c] https://orcid.org/0000-0001-7935-1775 (ida.sessarego@usm.cl).

Reliability estimation for adaptive measuring instruments

ABSTRACT

Estimating the reliability of linear and standardized evaluative tests of any type in the social sciences is relatively simple with the use of Cronbach's alpha coefficient. The problem occurs with the use of adaptive tests that adjust to the answers that the subjects are giving. This becomes even more complex when computerized adaptive tests are used, which stresses the use of Cronbach's alpha since it requires stability in its measurement. Thus, the present work exposes the so-called Game-Adaptive method that allows determining the reliability of an adaptive test considering its variable nature. This method is tested with two cases, one simulated and the other real, based on a sample of 50 graduate students from a Chilean university who submitted the Oxford Online Placement Test. The results indicate that the application of the Game-Adaptive method adequately collects the variability of the responses of the items used in each case, allowing us to determine without major inconvenience the reliability of the instruments in question. Consequently, this proposal becomes a plausible route to analyze adaptive tests, helping to overcome the limitations of traditional coefficients, at least in terms of reliability.

Keywords: Adaptive test; reliability; TICs; Cronbach's alpha; classical theory; item response theory.

1. Introducción

Las pruebas escritas de papel y lápiz han sido la forma más tradicional de evaluar masivamente a los estudiantes, usadas extensivamente en todos los sistemas y ciclos educativos. Sin embargo, hace aproximadamente 5 décadas atrás se crearon las pruebas utilizando computadores, conocidas en inglés como Computer Based Test (CBT) (Kingsbury y Houser, 1993). La forma más común de CBT es aquella medición lineal computarizada, con una duración fija. Esta presenta el mismo número de preguntas a cada persona evaluada, en un orden específico, cuyo puntaje usualmente depende del número de ítems que se responden correctamente. Una prueba lineal consiste en un completo rango de preguntas (de diferentes grados de dificultad) que están seleccionadas aleatoriamente desde una base más grande y que son similares para cada uno de los examinados. Este tipo de prueba imita una prueba de papel y lápiz tradicional, pero que se presenta en un formato digital, sin considerar las habilidades de cada individuo examinado (Triantafillou et al., 2008), es decir, sin adaptarse al nivel de cada persona sometida a la prueba.

En 1974 Weiss expone el primer test adaptativo en computador, llevando a la práctica las ideas descritas originalmente por Lord (1970). Las pruebas de evaluación adaptativas en computador, o Computerized Adaptive (CAT), consisten en la administración de pruebas que consideran el nivel de habilidad del examinado. La prueba adaptativa es un caso especial de prueba basada en computación, donde cada examinado rinde una prueba única que se hace a la medida de este (Triantafillou et al., 2008). Así, después de cada respuesta la estimación de la habilidad se actualiza y el próximo ítem es seleccionado en función de dicha estimación (Van der Linden y Glas, 2003). Esto significa que la prueba realiza evaluaciones del nivel de conocimiento que posee la persona y, en función de ello, asigna una pregunta. Esta ventaja es una consecuencia directa de usar algoritmos de selección de ítems, con estimaciones iguales o mejores del nivel de conocimiento del evaluado.

Los CAT son una contribución al proceso de evaluación del aprendizaje como una alternativa práctica a la evaluación tradicional en papel y lápiz (Kingsbury y Houser, 1993; Lane, 2014; Van der Linden y Glas, 2010). De hecho, Olea et al. (2004), indican que la principal ventaja de este tipo de prueba es su eficiencia, pues consiguen medidas precisas con menos ítems que las pruebas tradicionalmente aplicadas de forma lineal y universal, que no consideran las respuestas de los sujetos sometidos a ellas.

Olea et al. (2004) llaman a las evaluaciones adaptativas computarizadas como tests adaptativos informatizados (TAIs), indicando que su uso se hizo común en países del hemisferio norte en el ámbito psicológico y educativo. Desde estos autores los TAIs seleccionan de forma dinámica, mediante instrucciones programadas en algoritmos, los ítems más apropiados para cada sujeto sometido a una prueba, según lo que el mismo sujeto va manifestando en el transcurso de la prueba. Generalmente los TAIs se basan en modelos de la Teoría de Respuesta al Ítem. Por lo tanto, los TAIs son equivalentes a los CAT y basan su aplicación en algoritmos que adaptan los ítems a las respuestas de las personas sometidas a evaluación, ya sea psicológica o educativa.

Con el crecimiento del número de estudiantes en las salas de clases, los profesores se han visto en la obligación de evaluarlos usando formas de prueba estándares (casi exclusivamente), surgiendo el problema de cómo se puede discriminar entre niveles de conocimiento. Esto ha traído consigo largas pruebas que incluyen preguntas con diferentes grados de dificultad que generalmente se responden mediante selección de alternativas. No obstante, una prueba de alternativas múltiples mide el conocimiento de todos los estudiantes sin diferenciación, aunque consabido es que cada estudiante tiene distintas formas de aprender y procesar la información (Soflano et al., 2015). Por ello, para Conejo et al. (2004) la principal diferencia entre una prueba adaptativa y una tradicional de lápiz y papel es la capacidad del primero de adaptarse a las habilidades individuales de cada estudiante.

En tanto, la creación de pruebas adaptativas en computador se basa en la Teoría de Respuesta al Ítem que se usa para determinar la pregunta que sigue en la prueba y a la vez decidir cuándo termina la prueba (Cabrera et al., 2010; Guzmán y Conejo, 2005). Adicionalmente, las pruebas adaptativas deben controlar muy bien la exposición de los ítems utilizados, con el fin de evitar la sobreexposición de determinados ítems, o en su antípoda, la infrautilización de otros (Olea et al., 2004). De hecho, este es uno de los problemas más comunes de las pruebas adaptativas. Así las cosas, las pruebas adaptativas que se concentren en solo un conjunto de preguntas de su banco más amplio, tienen problemas de fiabilidad y de validez.

Consecuentemente, ya sea en pruebas lineales, en papel y lápiz, o en computadores, así como test adaptativos, se requiere imperiosamente que dichos instrumentos sean rigurosos desde el punto de vista de su validez y fiabilidad (Rizo, 2001). En este sentido, tradicionalmente para la calibración de las pruebas de selección múltiple, se ha usado la teoría de la generalización desarrollada por Cronbach et al. (1972), que es una extensión de la teoría clásica que atiende en forma más satisfactoria la problemática de la fiabilidad, substituyéndola por la noción de generalizabilidad (sic), también conocida como Teoría G. Esta teoría, en lugar del concepto de puntaje verdadero, usa el puntaje del universo, y en lugar de manejar el error de medición en forma global, identifica fuentes posibles de error y detecta su influencia gracias a técnicas estadísticas (Rizo, 2001).

La aplicación de pruebas estadísticas de fiabilidad en test lineales, estandarizados y universales no resulta muy problemática, especialmente si sus respuestas se encuentran dentro de alternativas preestablecidas. El problema se sucede en aquellas pruebas adaptativas que, por principio lógico, terminan siendo expresiones individuales de un conjunto mayor de alternativas interrogativas. Por ello, el principal ingrediente de una prueba adaptativa es una

base de datos con una gran cantidad de preguntas bien distribuidas. En este aspecto, los procedimientos para determinar la validez de una prueba adaptativa en computador son similares a los de las pruebas convencionales. Pero, cuando un CAT intenta ser equivalente a una prueba convencional, las dos pruebas son igualmente válidas solo si ellas han demostrado dar medidas equivalentes (Green et al., 1984). Además, para que una prueba adaptativa en computador sea fiable, el CAT debería ser capaz de recoger información del estudiante que sea pertinente para lograr hacer esta inferencia.

De acuerdo con Green et al. (1984) los índices de fiabilidad, validez y calidad del ítem son relevantes cuando todos quienes rinden una prueba pueden confrontar el mismo conjunto de preguntas. En este sentido, la fiabilidad, entendida como la consistencia interna de un instrumento de medición de variables operacionalizada en ítems de consultas, implica que dicha medición presente una estabilidad temporal (Prieto y Delgado, 2010), estabilidad de la que carecen las pruebas que se adaptan a cada examinado. Por ende, no sería posible, según el criterio de Green et al. (1984), lograr la fiabilidad, validez y calidad en pruebas adaptativas, ya sea aplicadas tradicionalmente o en computadores.

Lo precedente es extremadamente relevante, pues la determinación de la fiabilidad es crucial en las ciencias sociales en general y en la educación en particular, pues la medición de constructos metacognitivos es un proceso sumamente complejo, para los cuales no existen balanzas (González et al., 2016). Así, desde el punto de vista de un instrumento, la fiabilidad consiste en determinar el error de medición de este, que considera tanto la varianza sistemática como la varianza por el azar (Kerlinger y Lee, 2002). Pero una prueba adaptativa tiene preguntas diferentes para cada evaluado. Por ello, evaluar la fiabilidad de este tipo de prueba resulta crucial para poder utilizar sus resultados en la toma de decisiones de manera adecuada.

Considerando lo anterior, el coeficiente alpha de Cronbach (1951) es una de las herramientas estadísticas más usadas para estimar la fiabilidad de una escala contenida en un instrumento de recolección de datos (ejemplo: cuestionario). Yang y Green (2011) declaran que el alpha de Cronbach es fácil de interpretar, es objetivo y no requiere decisiones subjetivas, siendo útil para hacer revisión a escalas y decidir cuál ítem se puede dejar o eliminar. Esto es importante, ya que en la generación de conocimiento se busca determinar un buen nivel de fiabilidad, pues ello otorga la posibilidad de repetir la misma investigación con idénticos resultados o al menos consistentes (Bar, 2010). Con todo, este coeficiente requiere de la estabilidad de los elementos que son analizados por el mismo, estabilidad de la que carecen las pruebas adaptativas antes descritas. Por tal razón, se requiere de una nueva herramienta que permita recoger las bondades del alpha de Cronbach, pero en escenarios cambiantes.

En consecuencia, el presente trabajo pretende dar respuesta a la interrogante de: ¿cómo se puede medir la fiabilidad de una prueba adaptativa si cada persona contesta un número distinto de preguntas y las preguntas varían de un individuo a otro de acuerdo con su nivel de conocimiento? Así, y para responder esta pregunta, el presente trabajo describe y explica el denominado método *Game-adaptive* como modelo estadístico diseñado para calibrar una prueba adaptativa, específicamente respecto de la estimación de su fiabilidad. Así las cosas, esta propuesta coloca en pausa los hallazgos de Pedrosa et al. (2016); Abal et al. (2019); Lozzia et al. (2020), solo por nombrar algunos autores que, utilizando test adaptativos, basan sus resultados en un coeficiente de fiabilidad que no considera la adaptabilidad dentro de su construcción, ya que no fue concebido para tales propósitos. Ello no significa que sus conclusiones estén erradas, más bien se indica que lo tratado en este trabajo no estaba considerado cuando los autores señalados publicaron sus resultados.

2. Método

2.1 Tipo y alcance de la investigación

Investigación no experimental, cuantitativa, de tipo exploratoria. Es además propositiva, en la medida que se entrega a disposición de la comunidad científica un nuevo método para estimar la fiabilidad de una prueba considerando su adaptabilidad.

2.2 Universo y muestra

Para el análisis del método que se describirá se utilizaron dos casos de análisis, a saber:

a) Caso simulado: En una matriz de 25 casos simulados se especificó de manera aleatoria los tamaños de las pruebas para cada persona, que para el caso se entienden como estudiantes ficticios sometidos a pruebas adaptativas. Con dichos resultados se aplicó la metodología de análisis que se detallará en el siguiente apartado, de tal forma de clarificar el procedimiento de análisis.

b) Caso real: En un universo de 166 estudiantes de postgrado de una universidad regional del Consejo de Rectores de Universidades Chilenas, se determinó una muestra de 50 estudiantes elegidos de manera aleatoria. A dichos estudiantes, se les aplicó la prueba *Oxford Online Placement Test* que provee la información del nivel de inglés de las personas que se examinan. Esta prueba se rinde en computador y tiene la característica de ser adaptativa, ya que se acomoda al nivel de aptitud de cada persona que realiza la prueba, planteando las preguntas que se encuentren dentro del nivel de dificultad apropiado. Tiene aproximadamente 15 preguntas para medir la habilidad auditiva; y 35 preguntas para medir el uso de vocabulario, la gramática y la comprensión del significado en una conversación.

2.3 Método de análisis propuesto: Game-adaptive

Considere a la letra M como la denominación del vector fuente, mientras que n es el número total de ítems que constituyen el vector fuente. M es el vector desde el cual se extraen todos los ítems que constituyen cada una de las pruebas adaptativas. Además, entenderemos por tamaño de la prueba T el número de ítems que lo constituyen, que será simbolizado por #(T), por ejemplo, si la prueba T posee 7 ítems o preguntas, entonces #(T)=7 lo cual caracterizará su tamaño. Sea $n_{(j)}$ el estadístico de orden para los tamaños de las pruebas, luego $n_{(1)}$ representa el menor de los tamaños de las pruebas generadas desde M. Así, se observa que $n_{(j)} \leq n$ para todo j.

Seguidamente, para que una prueba adaptativa T pueda ser sometida a estimación de fiabilidad, que entenderemos como adaptativamente fiable, deben existir por lo menos tres pruebas adaptativas T_1, T_2 y T_3 tal que $\#(T_1 \cap T_2 \cap T_3) \geq 3$ y $T_1 \cup T_2 \cup T_3 = T$. De esta manera, para la cuantificación de la fiabilidad sobre test adaptativos, se utilizará como base una adecuación del coeficiente Alpha de Cronbach, denominada Alpha Game (González y Aspeé, 2021), cuya formulación indica lo siguiente:

Sea $\alpha_\ominus$ (Alpha-Game) la reformulación del coeficiente Alpha de Cronbach, definido por:

$$\alpha_\ominus = \frac{n}{n-1}\left(1 - \frac{1}{1 + \frac{\sum_{i \neq j} |Cov(I_i, I_j)|}{\sum_i V_i}}\right)$$

Donde V_i representa la varianza de las puntuaciones del i-ésimo ítem y $Cov(I_i, I_j)$ representa la covarianza entre las puntuaciones a dos ítems diferentes. Esta reformulación asegura un recorrido compacto dado por: $0 \leq \alpha_{\ominus} \leq 1$. Además, implica que el coeficiente $\alpha_{\ominus}$ caracteriza una cota superior para el coeficiente Alpha de Cronbach (α_c), esto es: $\alpha_c \leq \alpha_{\ominus}$. Esta adecuación se encuentra implementada en el software de referencia JAMOVI, por medio de la opción *Reverse scaled items*.

Por su parte, se denominará por test nuclear de tamaño k, o simplemente T^k, a la prueba que cumpla con ser adaptativamente fiable, siendo $k=\#(T^k)$. En este sentido, la expresión adaptativamente fiable, significa en principio que es posible medir la fiabilidad. Seguidamente, sea $T^{n(j)}$ una familia de pruebas nucleares, es decir, $T^{n(j)}$ representa a todos los test nucleares de tamaño $n_{(j)}$. Por tanto, se observa que cuando $n_{(j)}$ =n para toda j, se tiene el proceso de calibración convencional. Considerando ello, para definir la fiabilidad de una prueba adaptativa, es necesario razonar la siguiente notación:

Sean T^{kl} el l-ésimo test nuclear de la familia $T^{(k)}$, $\rho(T^{kl})$ la fiabilidad Game de T^{kl}, *h* el máximo tamaño de las pruebas nucleares, y c la cantidad de pruebas pertenecientes a la familia $T^{(k)}$. En función de esta notación se define a la fiabilidad de una familia de pruebas adaptativamente fiables (ρ_T) como:

$$\rho_T = \frac{1}{(h-2)c}\sum_{k=3}^{h}\sum_{l=1}^{c}\rho(T^{k_l})$$

Explicitando $\rho(T^{kl})$ en función del coeficiente de fiabilidad Game:

$$\rho_T = \frac{1}{(h-2)c}\sum_{k=3}^{h}\sum_{l=1}^{c}\rho(T^{k_l}) = \frac{1}{(h-2)c}\sum_{k=3}^{h}\sum_{l=1}^{c}\left(\alpha_{\ominus_{T^{k_l}}}\right)$$

$$= \frac{1}{(h-2)c}\sum_{k=3}^{h}\sum_{l=1}^{c}\left(\frac{k}{k-1}\left(1-\frac{1}{1+\frac{\sum_{i\neq j}|Cov_l(I_i,I_j)|}{\sum_i V_{il}}}\right)\right)$$

Donde $\sum_{i\neq j}|Cov_l(I_i,I_j)|$ representa la covarianza de las puntuaciones obtenidas a los ítems i y j de la prueba l, mientras que $\sum_i V_{il}$ representa la suma de las varianzas de cada uno de los ítems que constituyen la prueba l.

Seguidamente, se observa que $0 \leq \rho_T \leq 1$. Ello implica que la formulación del coeficiente Game-Adaptative puede ser entendida como una media o promedio de las estimaciones de las fiabilidades Alpha-Game de cada prueba nuclear constituyente de la familia de pruebas adaptativamente fiables. Por su parte, cuando el número de pruebas nucleares converge a uno y su orden a *n*, entonces $\rho_T \rightarrow \alpha_{\ominus}$, donde $\alpha_{\ominus}$ es el coeficiente de fiabilidad Alpha-Game (González y Aspeé, 2021). Por tanto, los coeficientes Apha de Cronbach y de Game, son situaciones particulares del coeficiente Game-adaptative, específicamente cuando las covarianzas son positivas o cuando las covarianzas no tienen restricción numérica respectivamente.

3. Resultados

3.1 Caso simulado

En la práctica, supongamos que se dispone de 5 ítems que constituyen el vector fuente, que se simboliza con la letra M. Desde este vector se extraen todos los ítems que constituyen cada una de las pruebas adaptativas. Consecuentemente, los ítems que identifican a los que fueron rendidos por una determinada persona están identificados con el número 1, y los que no, por 0, como se muestra en la tabla siguiente:

Tabla 1

Conjunto de pruebas adaptativas.

	Ítem 1	**Ítem 2**	**Ítem 3**	**Ítem 4**	**Ítem 5**
Persona 1	1	1	0	1	0
Persona 2	0	1	1	1	0
Persona 3	1	1	1	0	0
Persona 4	1	0	1	1	1
Persona 5	1	1	0	1	0
Persona 6	1	0	1	0	1
Persona 7	1	1	1	1	1
Persona 8	0	1	1	1	1
Persona 9	1	0	0	1	0
Persona 10	0	1	0	0	1
Persona 11	1	0	0	1	0
Persona 12	1	1	1	1	0
Persona 13	1	0	1	0	0
Persona 14	1	0	0	1	0
Persona 15	1	0	0	0	1
Persona 16	0	1	0	1	1
Persona 17	0	1	1	1	1
Persona 18	1	1	0	1	0
Persona 19	1	0	1	0	0
Persona 20	1	0	0	0	1
Persona 21	1	0	1	0	1
Persona 22	0	1	1	0	0
Persona 23	1	1	1	1	1
Persona 24	0	1	0	0	1
Persona 25	0	1	0	1	0

Fuente de elaboración propia.

Con base en la tabla 1, las familias de pruebas nucleares están caracterizadas por:

$$T^{n(3)} = \{(1,2,3),(1,2,4),(1,3,4),(1,3,5),(1,4,5),(2,3,4),(2,3,5),(2,4,5),(3,4,5)\}$$

$$T^{n(4)} = \{(1,2,3,4),(1,3,4,5),(2,3,4,5)\}$$

Cada uno de los números que caracterizan las agrupaciones antes indicadas, corresponde a los ítems que constituyen la prueba nuclear particular. Para cada prueba nuclear se estimó la fiabilidad utilizando el coeficiente de fiabilidad alpha-game, obteniéndose los resultados descritos en la tabla 2.

Tabla 2

Resultados simulados de alpha-game.

Test	Fiabilidad
Nuclear 1	0.63
Nuclear 2	0.54
Nuclear 3	0.75
Nuclear 4	0.46
Nuclear 5	0.71
Nuclear 6	0.70
Nuclear 7	0.71
Nuclear 8	0.85
Nuclear 9	0.62
Nuclear 10	0.79
Nuclear 11	0.87
Nuclear 12	0.78

Fuente: Elaboración propia.

Ahora utilizando el método descrito de *Game-Adaptative*, se tiene que la falibilidad de la prueba es de 0.70, pudiendo ser considerado como altamente fiable. Ello se muestra en la siguiente ecuación:

$$\rho_T = \frac{1}{(h-2)c}\sum_{k=3}^{h}\sum_{l=1}^{c}\left(\frac{k}{k-1}\left(1-\frac{1}{1+\frac{\sum_{i\neq j}\left|Cov_l(I_i,I_j)\right|}{\sum_i V_{il}}}\right)\right) = 0.70$$

La tabla 3 sintetiza de manera descriptiva la muestra aleatoria generada. La razón de presentar un sumario de los descriptivos clásicos, es dar la libertad al coeficiente *Game-Adaptative* de ser representado y complementado con otro estadístico de resumen, dada las limitaciones que puede tener la media (o promedio) de la muestra de fiabilidades de las pruebas nucleares. Por ejemplo, se sugiere la mediana en el caso de detectar situaciones atípicas o escapadas, o en el caso de tener una postura más conservadora, considerar el mínimo, de esa forma se exige que todos los test nucleares sean altamente fiables.

Tabla 3

Resúmenes descriptivos para Game-Adaptative simulado.

Descriptivos Clásicos	Fiabilidad
Media	0.701
Mediana	0.710
Moda	0.710
Desviación Estándar	0.122
Mínimo	0.460
Máximo	0.870

Fuente: elaboración propia.

Para ponderar estos resultados desde una perceptiva inferencial, se recurrió a la prueba t-*student*. Por supuesto que, antes de su uso, hubo un análisis de normalidad que permitió su despliegue (Shapiro-Wilk 0.961 p-valor 0.796; Kolmogorov-Smirnov 0.164, p-valor 0.904; y Anderson-Darling 0.224 p-valor 0.773). Así, considerando las fiabilidades de las pruebas nucleares como una muestra aleatoria y aplicando el teorema central del límite, se tomó como valor de prueba para la hipótesis nula el nivel de fiabilidad 0.7, considerando que la fiabilidad es alta a partir de este valor. El resultado de la aplicación de t-*student* se detalla en la tabla 4.

Tabla 4

Resultados de t-student.

T-Test para una muestra				Intervalo de Confianza 95%	
Student's t	statistic	df	p	Lower	Upper
	0.0237	11.0	0.982	0.623	0.778

Nota. H_a Media poblacional ≠ 0.7
Fuente: elaboración propia.

En la tabla 4 se observa que el intervalo de confianza contiene el valor de prueba 0.7, por tanto, los datos soportan evidencia a favor de la fiabilidad de la prueba, es decir, la prueba de la simulación puede ser considerada como altamente fiable.

3.2 Caso real

El instrumento aplicado para este estudio se conoce como *Oxford Online Placement Test*, prueba para determinar el nivel de inglés aplicada de manera adaptativa. Contiene aproximadamente 15 preguntas para medir la habilidad auditiva; y 35 preguntas para medir el uso de vocabulario, la gramática y la comprensión del significado en una conversación. Este instrumento se aplicó a 50 estudiantes de postgrado de una universidad tradicional chilena.

Para los estudiantes es una prueba fácil de rendir, y entrega información que se alinea con los estándares del Marco Europeo de Referencias de las Lenguas (Council of Europe, 2001), lo que permite asegurar que el curso de inglés que deben tomar sea el más ajustado a sus conocimientos. Por tanto, la calibración de un instrumento con estas características será el respaldo para la toma de decisiones y la derivación respectiva relacionada con el nivel de inglés efectivo de cada individuo.

Aplicado el instrumento en cuestión, se distinguieron 8 test nucleares $T^{n(40)}$,$T^{n(41)}$,$T^{n(42)}$,$T^{n(43)}$,$T^{n(44)}$, $T^{n(45)}$,$T^{n(46)}$ y $T^{n(47)}$, cuyas respectivas fiabilidades Game se describen en la tabla 5.

Tabla 5

Fiabilidad adaptativa de las pruebas nucleares.

Test Nuclear	$T^{n(40)}$	$T^{n(41)}$	$T^{n(42)}$	$T^{n(43)}$	$T^{n(44)}$	$T^{n(45)}$	$T^{n(46)}$	$T^{n(47)}$
Fiabilidad Game	0.670	0.692	0.699	0.701	0.726	0.850	0.939	0.961

Fuente: elaboración propia.

Ahora, utilizando Game-Adaptative, se tiene que ρ_T=0.779, lo que permite concluir que la prueba es altamente fiable. Los resúmenes descriptivos clásicos expuestos en la tabla 6 dan testimonio de la conclusión recién efectuada.

Tabla 6

Resúmenes descriptivos para Game-Adaptative real.

Descriptivos Clásicos	Fiabilidad
Media	0.780
Mediana	0.714
Desviación Estándar	0.119
Mínimo	0.670
Máximo	0.961
hapiro-Wilk p	0.035

Fuente: elaboración propia.

Por su parte, haciendo uso de pruebas inferenciales de comparación de grupos, previa aplicación de pruebas de normalidad (Shapiro-Wilk 0.808 p-valor 0.035; Kolmogorov-Smirnov 0.300 p-valor 0.391; y Anderson-Darling 0.732 p-valor 0.033), se consideró el instrumento como altamente fiable (>=0.7). El detalle de ello indicó que la estadística de Kolmogorov-Smirnov fue la única que dio evidencia a favor de la normalidad, por tanto, se desarrollaron pruebas paramétricas (t-student) y no paramétricas (Wilcoxon rank) como se presenta en la tabla 7. En dicha tabla, se observa que el intervalo de confianza contiene el valor de prueba 0.7, independiente de la estadística de prueba utilizada. Por ello, los datos soportan evidencia a favor de la fiabilidad de la prueba, es decir, *Oxford Online Placement Test* puede ser considerado de altamente fiable.

Tabla 7

Pruebas inferenciales de comparación de muestras.

Comparación para una Muestra				Intervalo de Confianza 95%	
Prueba	statistic	df	p	Lower	Upper
Student's t	1.90	7.00	0.099	0.681	0.879
Wilcoxon W	26.5		0.262	0.686	0.905

Note. H_a media poblacional ≠ 0.7

Fuente: elaboración propia.

4. Discusión y conclusiones

El propósito de las líneas precedentes se enmarcó en la necesidad determinar la fiabilidad de una prueba adaptativa, reconociendo que su misma naturaleza hace que el uso de las pruebas tradicionales de fiabilidad se vea limitada. Ello es especialmente sensible en pruebas informatizadas, ya que su programación de adaptabilidad las hace más sensible al sujeto que las está contestando, lo que otorga eficiencia al proceso de evaluación (Lane, 2014; Olea et al., 2004; Triantafillou et al., 2008; Van der Linden y Glas, 2010), pero al mismo tiempo dudas respecto de su fiabilidad. En efecto, se indicó que, para Prieto y Delgado (2010), la fiabilidad se entiende como la consistencia interna de un instrumento de medición de variables, operacionalizada en ítems de consultas, lo que involucra que dicha medición presente una estabilidad temporal. No obstante, las pruebas que se adaptan a cada examinado carecen de dicha estabilidad. A mayor abundamiento Olea et al. (2004) advierten que las pruebas adaptativas deben controlar la exposición de los ítems utilizados, para evitar la sobreexposición o la infrautilización de ítems, lo que resta validez y fiabilidad a la prueba.

Así, como se indicó, el coeficiente alpha de Cronbach (1951) es probablemente la herramienta más utilizada en las ciencias sociales para determinar la fiabilidad de los instrumentos de medición (Yang y Green, 2011). Empero, este coeficiente demanda la estabilidad de los elementos que son analizados por el mismo. En dicho sentido, el método de *Game-Adaptative* propuesto permite tomar la inestabilidad de las pruebas adaptativas para poder medir la fiabilidad, sin que dicha característica impida comprobar esa crucial característica.

Como se distinguió en ambas aplicaciones, tanto en el caso simulado como en el caso real, la variabilidad de los ítems no fue impedimento para poder medir la fiabilidad. De esta manera, considerar como una muestra aleatoria las estimaciones de la fiabilidad de las pruebas nucleares, permite usar el teorema central del límite, estableciendo intervalos de confianza para la estimación de la fiabilidad a partir de esta muestra, y, por tanto, realizar afirmaciones en términos inferenciales, tal como fueron descritas en las aplicaciones. Asimismo, la presentación de los resúmenes descriptivos otorga la opción de redefinir el coeficiente *Game-Adaptative*, dada las limitaciones que puede tener la media (o promedio) de la muestra de fiabilidades de las pruebas nucleares. Por ejemplo, se sugiere la mediana en el caso de detectar situaciones atípicas, o en su defecto, si se tiene una postura más conservadora, considerar el mínimo, de esa forma se exige que todos los test nucleares sean altamente fiables. Por ende, esta flexibilidad puede ser considerada como otra propiedad del coeficiente *Game-Adaptative*.

Desde que Cronbach (1951) generó un coeficiente de fiabilidad potente y útil para las ciencias sociales, este factor no ha presentados mayores inconvenientes en la investigación, y aunque existen cuestionamientos respecto de los criterios de decisión en su utilización y del contexto muestral, no los hay respecto de su estructura teórica y lógica. Sin embargo, el desarrollo de nuevos métodos evaluativos obliga a repensar esta herramienta. Así, en este trabajo se presentó una forma de comprobar la fiabilidad de una prueba adaptativa, es decir, una prueba no lineal que puede tener tantas manifestaciones como personas examinadas por la misma prueba. Esta propuesta complementa el coeficiente α de Cronbach (1951), respondiendo en parte a las inquietudes expuestas por Olea et al. (2004) respecto de los problemas de fiabilidad y validez de las pruebas adaptativas.

Para finalizar, se debe tener en cuenta que la rigurosidad científica de las ciencias sociales tiene mucho que ver con la pertinencia de los métodos que utiliza para establecer conclusiones, tomar decisiones, o incluso fundamentar acciones de corrección. En la educación esto es determinante, en la medida que, por ejemplo, establecer el grado de aprendizaje (grado de adquisición) de ciertos conocimientos o habilidades, está supeditado a la fiabilidad y validez

del instrumento evaluativo utilizado. Si dicho instrumento adolece de rigurosidad en dichos aspectos, todas las acciones y decisiones que deriven de él resultarán, a lo menos espurias, y a lo más, absolutamente falsas. La propuesta presentada es un avance en términos de medición de fiabilidad, la que se entrega a la comunidad científica y que está a la espera de sus observaciones.

Ahora bien, de la misma forma que es necesaria la fiabilidad en una prueba adaptativa, resulta necesario comprobar la validez de esta. No obstante, dicho aspecto no se trató en el presente trabajo. Ello pudiera ser el punto de inicio de una nueva propuesta, en el entendido que la validez de una prueba no puede disociarse o separarse de su fiabilidad, y menos de la dificultad del mismo instrumento (González et al., 2016). Igualmente, es necesario avanzar a una interpretación fundada en los datos de este coeficiente Game-Adaptative, de manera que su uso como criterio de decisión no sea consuetudinario, sino que científico. La propuesta queda abierta.

Agradecimientos

"El autor José González agradece a la Universidad de Playa Ancha y el apoyo del Ministerio de Educación a través del Plan de Fortalecimiento de Universidades Estatales, UPA 1799".

Referencias

Abal, F. J. P., Auné, S. E., y Attorresi, H. F. (2019). Construcción de un banco de ítems de facetas de neuroticismo para el desarrollo de un test adaptativo. *Psicodebate. Psicología, Cultura y Sociedad*, 1(1), 31-50. http://dx.doi.org/10.18682/pd.v1i1.854.

Bar, A. R. (2010). La metodología cuantitativa y su uso en américa latina. *Cinta de moebio*, (37), 1-14. https://dx.doi.org/10.4067/S0717-554X2010000100001.

Cabrera, E., González, J., Montenegro, E., Nettle, A., y Guevara, M. (2010). Test informatizados y el registro del tiempo de respuesta, una vía para la precisión en la determinación del nivel de logro de un saber matemático. *Estudios pedagógicos* (Valdivia), 36(1), 69-84. http://dx.doi.org/10.4067/S0718-07052010000100003.

Conejo, R., Guzmán, E., Millán, E., Trella, M., Pérez-De-La-Cruz, J. L., y Ríos, A. (2004). SIETTE: A web-based tool for adaptive testing. *International Journal of Artificial Intelligence in Education*, 14(1), 1-33. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.108.150&rep=rep1&type=pdf.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334. https://doi.org/10.1007/BF02310555.

Cronbach, L. J., Gleser, G. C. Nanda, H., y Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. New York, Wiley. https://doi.org/10.1126/science.178.4067.1275.

Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division. (2001). *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge University Press. https://rm.coe.int/16802fc1bf.

González, J., y Aspeé, J. (2021). Propuesta de estimador de la fiabilidad mediante Alfa-Game. *Revista Iberoamericana de Psicología*, 14(1), 1-10. https://doi.org/10.33881/2027-1786.rip.14101.

González, J., Viveros, F., y Carvajal, C. (2016). Coeficientes edumétricos para la validez y dificultad de un test: Propuesta. *Estudios pedagógicos* (Valdivia), 42(3), 467-481. https://dx.doi.org/10.4067/S0718-07052016000400025.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., y Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, 21(4), 347-360. https://doi.org/10.1111/j.1745-3984.1984.tb01039.x.

Guzmán, E., y Conejo, R. (2005). Self-assessment in a feasible, adaptive web-based testing system. *IEEE Transactions on Education*, 48(4), 688-695. https://doi.org/10.1109/TE.2005.854571.

Kingsbury, G. G., y Houser, R. L. (1993). Assessing the utility of item response modes: Computerized adaptive testing. *Educational Measurement: Issues and Practice*, 12(1), 21-27. https://psycnet.apa.org/doi/10.1111/j.1745-3992.1993.tb00520.x.

Kerlinger, F., y Lee, H. (2002). *Investigación del Comportamiento. Métodos de investigación en Ciencias Sociales*. McGraw Hill. México.

Lane, S. (2014). Validity evidence based on testing consequences. *Psicothema*, 26(1), 127-135. https://doi.org/10.7334/psicothema2013.258.

Lord, F. M. (1970). Some test theory for tailored testing. En W.H. Holtzman (Ed.), *Computer assisted instruction, testing and guidance*. (pp. 139-183). New York: Harper and Row.

Lozzia, G. S., Abal, F. J. P., Galibert, M. S., y Attorresi, H. F. (2020). Test Adaptativo Informatizado de Analogías Verbales: comparación de Criterios de Parada. *Revista de Psicología*, 38(1), 31-63. https://doi.org/10.18800/psico.202001.002.

Olea, J., Abad, F. J., Ponsoda, V., y Ximénez, M. C. (2004). Un test adaptativo informatizado para evaluar el conocimiento de inglés escrito: diseño y comprobaciones psicométricas. *Psicothema*, 16(3), 519-525. http://www.psicothema.com/pdf/3026.pdf.

Pedrosa, I., Suárez-Álvarez, J., García-Cueto, E., y Muñiz, J. (2016). A computerized adaptive test for enterprising personality assessment in youth. *Psicothema*, 28(4), 471-478. http://www.psicothema.com/pdf/4352.pdf.

Prieto, G., y Delgado, A. (2010). Fiabilidad y validez. *Papeles del psicólogo*, 31(1), 67-74. http://www.papelesdelpsicologo.es/pdf/1797.pdf.

Rizo, F. M. (2001). Evaluación educativa y pruebas estandarizadas. Elementos para enriquecer el debate. *Revista de la educación superior*, 30(120), 1-12. http://publicaciones.anuies.mx/pdfs/revista/Revista120_S3A3ES.pdf.

Soflano, M., Connolly, T. M., y Hainey, T. (2015). An application of adaptive games-based learning based on learning style to teach SQL. *Computers & Education*, 86, 192-211. https://doi.org/10.1016/j.compedu.2015.03.015.

Triantafillou, E., Georgiadou, E., y Economides, A. A. (2008). The design and evaluation of a computerized adaptive test on mobile devices. *Computers & Education*, 50(4), 1319-1330. http://dx.doi.org/10.1016/j.compedu.2006.12.005.

Van der Linden, W. J., y Glas, C. A. W. (2003). Preface. En W. J. Van der Linden y C. A. W. Glas (Eds.), *Computerised adaptive testing: theory and practice* (pp. vi–xii). Dordrecht, Boston, London: Kluwer Academic Publishers.

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

Weiss, D. J. (1974). *Strategies of adaptive ability measurement*. Research report 74-5. Dep. Of Psychology, University of Minnesota. https://files.eric.ed.gov/fulltext/ED104930.pdf.

Yang, Y., y Green, S. B. (2011). Coefficient alpha: A reliability coefficient for the 21st century? *Journal of Psychoeducational Assessment*, 29(4), 377-392. https://doi.org/10.1177%2F0734282911406668.