ISBN: 978-628-96613-3-0. ISSN: 2414-6390. Digital Object Identifier: https://dx.doi.org/10.18687/LACCEI2026.1.1.1407

Development and Simulation of a Computerized Adaptive Testing Algorithm Based on an Item Bank Measuring Verbal, Spatial, and Numerical Reasoning

Ghio, Fernanda Belén[1,2]; Batista, Isaias Wilson[2]; Garrido, Sebastian Jesús[1,2,3]; Azpilicueta, Ana Estefanía[1,2] y Cupani, Marcos[2,3]

[1]Universidad Siglo 21, [2]Facultad de Psicología, Universidad Nacional de Córdoba, [3]Instituto de Investigaciones Psicológicas (UNC-CONICET)
Contacto: fernanda.ghio@unc.edu.ar

Abstract– ***The main objective of this study is the development, specification, and subsequent simulation of the algorithm for a Computerized Adaptive Test (CAT) based on an Item Bank (IB) that measures numerical reasoning (NR), verbal reasoning (VR), and spatial reasoning (SR). The IB consists of 247 questions, of which 82 assess NR, 97 assess VR, and 68 assess SR. These items were calibrated in a previous study using the Rasch model. In this study, the difficulty parameters of the items were equated, and the final IB for each type of reasoning was established. Subsequently, specification rules were developed for the test algorithm, and finally, the CAT algorithm was simulated under different stopping rules based on the standard error of estimation. We present a preliminary analysis of the algorithm's performance by simulating 1,000 response patterns. Overall, this study allows us to conclude that this IB enables the precise estimation of an ability level of 0, using a stopping criterion of a standard error of estimation ≤ 0.5. Additionally, concerning this error and in comparison, to paper-and-pencil administration, the test length could be reduced by approximately 26 to 28%. While further studies in both real and simulated conditions are needed, the present study provides evidence of the benefits of incorporating an innovative technological system in Argentina for assessing different cognitive abilities in educational settings.***

Keywords-- Computerized Adaptive Testing, Cognitive Abilities, Simulation.

Desarrollo y simulación del algoritmo de un Test Adaptativo Informatizado (TAI) a partir de un Banco de Ítems (BI) que mide razonamiento verbal, espacial y numérico

Ghio, Fernanda Belén[1,2]; Batista, Isaias Wilson[2]; Garrido, Sebastian Jesús[1,2,3]; Azpilicueta, Ana Estefanía[1,2] y Cupani, Marcos[2,3]

[1]Universidad Siglo 21, [2]Facultad de Psicología, Universidad Nacional de Córdoba, [3]Instituto de Investigaciones Psicológicas (UNC-CONICET)
Contacto: fernanda.ghio@unc.edu.ar

Resumen– ***El objetivo principal del presente trabajo es el desarrollo, especificación y posterior simulación del algoritmo de un Test Adaptativo informatizado (TAI) a partir de un Banco de ítems (BI) que mide razonamiento numérico (RN), verbal (RV) y espacial (RE). El BI está compuesto por 247 preguntas, de las cuales 82 evalúan RN, 97 RV y 68 RE. Dichos ítems fueron calibrados en un estudio previo aplicando el modelo de Rasch. En este trabajo se equipararon los parámetros de dificultad de los ítems y se conformó el BI definitivo de cada tipo de razonamiento. Luego se generaron las reglas de especificación para la conformación del algoritmo del test y por último se simuló la aplicación del algoritmo del TAI mediante diferentes reglas de parada considerando el error estándar de estimación. Presentamos un análisis preliminar del funcionamiento del algoritmo del TAI simulando 1000 veces los patrones de respuesta. En líneas generales este estudio nos permite concluir que este BI posibilita estimar con precisión un nivel de habilidad 0, con un criterio de parada del error estándar de estimación ≤ 0.5. Además, en relación a dicho error y comparando la administración lápiz y papel nos permitiría reducir entre un 26 y 28% la longitud de la prueba. Si bien se requieren nuevos estudios en situaciones reales y simuladas el presente estudio aporta evidencia de los beneficios de la incorporación de un sistema innovador en aspectos tecnológicos en Argentina para evaluar diferentes capacidades cognitivas en el ámbito educativo.***

Palabras clave-- Test Adaptativo Informatizado, Capacidades Cognitivas, Simulación.

I. Introducción

El avance de la tecnología computacional ha motivado la integración de herramientas tecnológicas en diversos ámbitos como por ejemplo en el educativo. La utilización de estos recursos modernos permitió incrementar la motivación de los estudiantes en el proceso de aprendizaje y obtener medidas de rendimiento académico más efectivas [1][2]. En este panorama, aparecen los Test Adaptativos Informatizados (TAIs) permitiendo evaluar de una manera más eficiente y flexible el rendimiento académico de los estudiantes [3].

En la educación superior, especialmente en contextos que integran tecnologías de evaluación, los Test Adaptativos Informatizados (TAI) han demostrado ser herramientas eficientes para medir habilidades con mayor precisión y menor longitud de prueba en comparación con instrumentos tradicionales. Esto contribuye a optimizar procesos de enseñanza y aprendizaje mediados por tecnologías digitales [4].

Los primeros desarrollos de los TAIs se dieron en el marco de los test de inteligencia y de aptitudes cognitivas, debido a que los factores cognitivos resultan buenos predictores del rendimiento académico (Vedel & Poropat, 2017), lo que se demuestra en los elevados niveles de correlación entre las medidas de inteligencia y el rendimiento académico (r = .50) [5].

La inteligencia como constructo teórico ha sido definido desde diversas teorías y modelos que se focalizaron en aspectos biológicos, cognitivos o psicométricos [6]. Particularmente las perspectivas que coinciden con la concepción de inteligencia que en este estudio se adopta son la cognitivas y psicométricas quienes dividen el constructo inteligencia en dos grandes modelos teóricos. Por un lado, los que postulan un factor general (g) de inteligencia y por el otro los que establecen una definición a partir de la presencia de habilidades cognitivas relativamente independientes [7] [8]. Existe evidencia que postulan que diversas aptitudes específicas (razonamiento verbal, abstracto, numérico, espacial, entre otros) resultarían igualmente idóneas para predecir el rendimiento de los estudiantes en determinadas áreas de conocimiento [9].

En Argentina contamos con instrumentos adaptados para medir inteligencia y diversas habilidades cognitivas específicas, reconocidos a nivel internacional y con amplio respaldo psicométrico. Entre ellos podemos mencionar el test de Matrices Progresivas de Raven [10], la escala de Wechsler (en sus diferentes versiones para adultos y para niños) [11] [12], el Test de Aptitudes Diferenciales (DAT) [13]. Sin embargo, dichos instrumentos presentan algunas limitaciones, uno de

ellas remite a que en general no pueden ser administradas nuevamente a los examinados por el efecto de aprendizaje y práctica [14]. Además, gran parte de ellas se construyeron a partir de la Teoría Clásica de los Test (TCT), lo que produce que las puntuaciones de un participante dependan del instrumento utilizado [15].

Hoy en día, para construir test de aptitudes se intentan vincular modelos teóricos y avances en tecnología psicométrica [16]. Los modelos de TRI cada vez se utilizan más en el desarrollo de pruebas de ejecución máxima, ya que permiten evaluar con un mayor nivel de rigurosidad y objetividad a los participantes en una determinada habilidad, aptitud o rendimiento [17]. Los TAIs ofrecen una forma atractiva de construir instrumentos a medida ya que cada administración de ítems se adaptará a las respuestas que el participante vaya otorgando al ítem que se le presenta.

Uno de los requisitos para la construcción de un TAI es contar con un (BI) adecuadamente calibrado desde algún modelo de TRI. El uso de los TAIs permite obtener estimaciones de las habilidades de los participantes más eficientes que las pruebas administradas en lápiz y papel dado que, a medida que avanza la prueba se irán seleccionando ítems con un nivel de información acorde al nivel de habilidad del participante [18]. En efecto un TAI permite estimar la habilidad de un participante a partir de un número variable de ítems y ya no con un instrumento de longitud de ítems fija [16].

El desarrollo de los TAIs posibilita obtener medidas con mayor nivel de precisión combinando teoría y tecnología. Como mencionamos, para su funcionamiento requiere de un BI con reactivos que presenten adecuadas propiedades psicométricas. Otro de los requisitos remite a que deben especificarse una serie de criterios fundamentales para su correcta aplicación, dichas especificaciones refieren a delimitar reglas de inicio de la prueba, selección de los próximos ítems a administrar, también se debe especificar el método de estimación de la habilidad, y un criterio para finalizar la prueba [19]. La lógica de funcionamiento de un TAI es que la prueba comience con un ítem de dificultad moderada, si el participante acierta dicho ítem, se presentará un ítem con una dificultad superior, por el contrario, se aplicará un ítem más fácil si lo responde incorrectamente; la prueba se detendrá cuando se cumpla el criterio establecido para terminarla [20].

Una de las principales ventajas de los TAIs es que logran el mismo grado de precisión que otras pruebas administrando un número menor de ítems. Esto se debe a que, al establecer un algoritmo con reglas definidas, cuando comience la prueba y en base a las respuestas de un participante, los ítems que se presentan se irán seleccionando acorde a su nivel de habilidad. Cabe aclarar que los trabajos de investigación que se realizan para el desarrollo de un TAI pueden realizarse en situaciones reales o simuladas. Esta última estrategia tiene la ventaja de ofrecer una primera aproximación a los datos con mayor rapidez y economía que una aplicación real del TAI. Además, el investigador puede fijar el verdadero nivel de habilidad otorgando datos con un menor nivel de sesgo y error típico [16].

Si bien las pruebas adaptativas computarizadas se utilizan desde hace tiempo en la medición de la inteligencia y aptitudes cognitivas, actualmente también se utilizan para obtener certificaciones en determinadas profesiones, uno de ellos es el que se aplica en el servicio de las fuerzas armadas de EE. UU (ASVAB, U.S. Department of Defense, 1984). De igual forma se utilizan en pruebas de admisión como el Graduate Record Examination (GRE) y el Graduate Management Admission Test (GMAT) [21]. En España, hoy día se reconoce al TRASI [22] como uno de los pocos TAIs comercializados que evalúan la capacidad de razonamiento secuencial e inductivo. Otro TAI reconocido en la lengua española como medida del nivel de comprensión del inglés escrito es el eCAT [17].

Ahora bien, en Argentina existen estudios de construcción de BI desarrollados en el marco de la TRI, pero no hay evidencia de aplicación de TAIs en el ámbito educativo. Particularmente un estudio que sienta las bases para el desarrollo de un TAI en el ámbito educativo es el propuesto por [23], quienes construyeron un BI para medir analogías verbales y definieron el algoritmo que se utilizará para la posterior del TAI; sin embargo, hasta el momento no se reportan estudios de aplicación de pruebas que evalúan diversas capacidades cognitivas en situaciones reales o simuladas.

Para el desarrollo de un TAI resulta fundamenta realizar estudios de simulación en cada una de las etapas de construcción, desde la etapa de planificación, desarrollo del algoritmo, construcción del conjunto de ítems y el control de calidad del mismo. Dichos estudios simulados deben replicarse a medida que se avanza en el proceso de construcción final de un TAI [24]. Por eso partiendo de la calibración realizada desde el modelo de Rasch de un banco de ítems que miden razonamiento numérico, verbal y espacial [25], en este trabajo se equipararán las dificultades de los ítems de cada BI de razonamiento y se construirá el algoritmo para un TAI por tipo de aptitud. Además, se utilizará un diseño de simulaciones de patrones de respuesta para estimar la cantidad de ítems necesarios del BI para estimar un nivel de habilidad 0 variando el criterio de parada del TAI.

Objetivos

El objetivo principal del presente trabajo de investigación es el desarrollo y especificación del algoritmo de un TAI a partir de un BI que mide razonamiento verbal, numérico y espacial. Luego de crear el algoritmo se utilizará un diseño de simulaciones para determinar la cantidad de ítems de este BI que se necesitan para estimar un nivel de habilidad 0 a través de diferentes reglas de parada. Considerando que el BI fue calibrado mediante el modelo de Rasch, se llevará a cabo la equiparación de los parámetros de dificultad mediante el método de equiparación media-sigma. Luego se utilizó la interfaz R studio para generar el algoritmo del TAI a través del paquete catR. Se definieron las reglas fundamentales de un TAI (ítems de inicio, selección del próximo ítem, estimación de la habilidad del participante y criterio de parada) y se generaron a través del comando randomCAT la simulación de un patrón de respuesta. Por último, se programó un ciclo repetitivo en el que

se estimó 1000 veces este patrón mediante la función "*for*" (comando que nos permite programar ciclos repetitivos) en un determinado nivel de habilidad con diferentes reglas de parada, establecidas por el error estándar (SE) de estimación (theta).

II. MÉTODO

Banco de ítems de razonamiento numérico, verbal y espacial [25].

El BI quedó conformado por 247 ítems que presentaron un buen ajuste al modelo. 82 preguntas evalúan Razonamiento Numérico (RN), 97 preguntas Razonamiento Verbal (RV) y 68 Razonamiento Espacial (RE). Estos ítems se almacenaron en un BI para ser utilizados en la construcción del TAI [25]. Para construir este BI se seleccionaron y tradujeron un conjunto de preguntas propuestas por [26] que permiten medir aptitudes específicas de inteligencia (razonamiento verbal, espacial y numérico). Para evaluar y calibrar estos ítems se construyeron diversas versiones de las pruebas. Cada forma se conformó por 15 ítems libres y 10 ítems comunes a cada tipo de razonamiento. Luego se administraron los ítems en formato lápiz y papel, a una muestra de N = 1140 estudiantes universitarios y desde allí se obtuvieron las propiedades psicométricas de los ítems, el ajuste al modelo y el cumplimiento de dos de los supuestos fundamentales de TRI (unidimensionalidad e independencia local). Específicamente los análisis se realizaron desde el modelo de TRI de un parámetro (Modelo de Rasch). Luego del proceso de análisis se seleccionaron los ítems que no presentaron desajuste al modelo Rasch.

Análisis de datos

E*quiparación*

Una de las ventajas de los modelos de TRI es que a partir de diversos procedimientos podemos establecer la invariancia de las medidas. En la práctica resulta fundamental encontrar métodos que nos permitan llevar a una métrica común los parámetros de dificultad de un conjunto de ítems que miden un mismo rasgo pero que fueron aplicados en diferentes muestras. Es por eso que, a partir de un diseño de anclaje, como el que se utilizó para la construcción de la prueba de razonamiento numérico, verbal y espacial podemos equiparar mediante reescalamiento lineal los parámetros de dificultad de los ítems de las diferentes versiones de a prueba. Este proceso de equiparación ubica en una misma métrica las dificultades de los ítems de diferentes test que midan un constructo [27]. En este trabajo se utilizó el método media-sigma para definir el escalamiento de cada forma de la prueba. Este método, a partir de los dos primeros momentos de la estimación de parámetros, define los valores de las constantes de la pendiente de la recta (k) y la ordenada en el origen (d). Cabe remarcar que la equiparación siempre se realizará a partir de los ítems anclas y se optará por equiparar las puntuaciones a la versión de la prueba que presente las mejores propiedades psicométricas. Cabe agregar que este método es uno de los procedimientos más recurrentes por su simplicidad y porque provee estimaciones de los parámetros más estables. En este procedimiento se utilizan las medias y desviaciones estándar de los parámetros de dificultad de los ítems anclas [28].

Construcción y simulación del algoritmo del TAI

Para construir un TAI se requiere de dos aspectos. Por un lado, necesitamos de un BI adecuadamente calibrado desde algún modelo de TRI, por el otro se necesita especificar el algoritmo del TAI que queremos desarrollar. En el trabajo realizado por [25] se evaluaron las propiedades psicométricas de los ítems que evalúan razonamiento numérico, verbal y espacial. A partir de dicho trabajo seleccionamos los ítems con mejores propiedades psicométricas y equiparamos los parámetros de dificultad de todos los ítems que conforman el BI de cada tipo de razonamiento. Para construir el algoritmo se definieron cada uno de los siguientes argumentos: la regla de inicio del test, el criterio para la selección del próximo ítem, el método estadístico que se utilizará para estimar el nivel de habilidad y la regla que determinará la finalización del test [29].

En este trabajo utilizamos el paquete catR de la interfaz R Studio y el comando randomCAT para generar, a partir de las especificaciones del algoritmo, patrones de respuesta para cada tipo de razonamiento [30]. Los argumentos del algoritmo especificados fueron: *start*, *test*, final y *stop*. Se construyeron los cuatro argumentos mencionados de la siguiente manera: (a) start: se delimitó que la prueba inicie con niveles de habilidad de entre -1.5 a 1.5 (theta), con un mínimo de administración de 4 ítems antes que finalice la prueba, con el propósito de garantizar la representatividad del contenido (b) test: para la estimación de la habilidad provisional se utilizó el método estimación ponderada de verosimilitud ("WL") [31] y para la selección del próximo ítem el criterio "MFI" (*Maximum Fisher Information*), (c) final: como criterio para la estimación de la habilidad final se utilizó la estimación esperada a posteriori (EAP) [32],(d) *stop*: la regla de parada se estableció por medio de criterio de longitud variable, establecido por la regla de "*precision*", a partir de un error estándar de la habilidad provisional (theta) menor o igual al valor especificado.

En la generación de cada patrón de respuesta se replicaron los análisis en las diferentes reglas de parada por cada tipo de razonamiento: SE (θ) $\leq$ 0.2, SE (θ) $\leq$ 0.3, SE (θ) $\leq$ 0.4, SE (θ) $\leq$ 0.5 y SE (θ) $\leq$ 0.6. Por último, a partir de la programación de un ciclo repetitivo del comando *randomCAT* se estimaron 1000 veces, por medio de la función "*for*", cada patrón de respuesta en cada una de las reglas de parada, considerando individualmente los ítems de cada tipo de razonamiento. Cabe mencionar que el paso final en el proceso de aplicación de un TAI nos otorga la estimación del nivel de habilidad de un participante, utilizando el patrón de respuesta completo de la prueba adaptativa. Se evaluó el desempeño del TAI mediante, longitud del test (media y desviación estándar), error estándar promedio (SE), sesgo (bias), raíz del error cuadrático medio (RMSE) y función de información del BI de cada tipo de razonamiento.

III. RESULTADOS

Equiparación

Razonamiento Numérico (RN)

El BI de RN quedó conformado con 82 ítems. En la tabla 1 se presentan los valores equiparados de los parámetros de dificultad de los ítems. Las dificultades de los ítems de la forma B, C, D y E se equipararon a la forma A ya que fue la versión de la prueba que presentó mejores propiedades psicométricas en la mayoría de los ítems.

Tabla 1

Equiparación de las dificultades del BI de RN

ítem	b	ítem	b	ítem	b	ítem	b
A2	0.52	A24	0.35	C10	0.11	D22	0.38
A3	-1.49	A25	-0.26	C11	-0.60	D23	0.28
A4	0.57	B6	-0.96	C12	1.45	D24	1.49
A5	-0.87	B7	0.75	C13	1.05	D25	1.66
A6	-0.47	B8	-1.11	C19	0.06	E6	0.32
A7	-1.07	B9	-0.93	C20	1.11	E7	0.31
A8	0.40	B10	0.23	C21	-0.23	E8	0.32
A9	-0.47	B11	-0.02	C22	0.06	E9	0.31
A10	-0.10	B12	0.28	C23	-0.23	E10	0.29
A11	1.11	B13	0.15	C24	0.51	E11	0.32
A12	0.23	B19	-0.29	C25	1.00	E12	0.32
A13	-0.71	B20	0.05	D6	-1.48	E13	0.34
A14	1.54	B21	0.09	D7	-0.91	E19	0.37
A16	-0.03	B22	0.80	D8	0.83	E20	0.34
A17	0.23	B23	1.73	D9	-0.82	E21	0.37
A18	0.60	B24	1.28	D10	0.32	E22	0.35
A19	-0.28	B25	1.08	D11	0.32	E23	0.35
A20	-0.22	C6	0.18	D13	-0.42	E24	0.31
A21	1.42	C7	-0.68	D19	0.11	E25	0.37
A22	0.30	C8	0.18	D20	0.75		
A23	0.95	C9	0.23	D21	0.83		

Nota. b= parámetro de dificultad del ítem, los ítems anclas aparecen en negrita.

Razonamiento Verbal (RV)

Noventa y siete ítems se almacenaron en el BI de RV. Las formas A, B, C y D se equipararon a la forma E del cuestionario. En la tabla 2 se presentan las dificultades de los ítems luego del proceso de equiparación de los parámetros de dificultad.

Tabla 2.

Equiparación de las dificultades del BI de RV

Ítem	*b*	Ítem	*b*	Ítem	*b*	Ítem	*b*	Ítem	*b*
1E	-0.54	23E	-0.36	8B	-1.38	19C	0.11	24D	-1.44
2E	-2.10	24E	-0.75	9B	-0.88	20C	-1.02	25D	-0.89
4E	-1.44	25E	1.85	10B	-1.59	21C	-0.11	6F	-0.81
5E	0.98	6A	-1.73	11B	-0.09	22C	0.30	7F	-0.81
6E	0.61	7A	-0.64	12B	-1.19	23C	0.20	8F	1.20
7E	-0.39	8A	0.90	13B	-0.09	24C	-0.80	9F	-0.61
8E	-0.51	9A	-0.24	19B	-1.38	25C	1.22	10F	-1.70
9E	1.91	10A	0.99	20B	-0.80	6D	-1.57	11F	-1.37
10E	1.09	11A	-1.52	21B	1.29	7D	-1.44	12F	-1.70
11E	0.64	12A	0.34	22B	0.27	8D	0.72	13F	1.45
12E	0.06	13A	0.20	23B	-1.25	9D	-0.78	19F	-1.37
13E	-0.54	19A	0.70	24B	-1.35	10D	1.35	20F	-0.46
15E	-1.29	20A	0.23	25B	0.05	11D	-1.23	21F	-1.11
16E	0.06	21A	-1.05	6C	-1.39	12D	-0.03	22F	2.06
17E	-1.08	22A	-1.44	7C	-1.26	13D	-1.52	23F	-1.29
18E	-0.09	23A	0.15	8C	-1.49	19D	1.31	24F	-1.40
19E	-0.75	24A	-0.76	9C	0.79	20D	-1.23	25F	-0.91
20E	-0.21	25A	-0.09	10C	-1.18	21D	1.43		
21E	0.81	6B	-1.82	11C	0.15	22D	0.26		
22E	0.98	7B	-1.65	13C	-0.87	23D	-1.11		

Nota. b= dificultad de los ítems equiparados, los ítems anclas aparecen en negrita

Razonamiento Espacial (RE)

El BI de RE quedó conformado con 68 ítems. En la tabla 3 se presentan las dificultades de los ítems equiparados a la forma B que presentó un mayor número de ítems con adecuadas propiedades psicométricas.

Tabla 3.

Equiparación de las dificultades del BI de RE

ítem	b	íte m	b	íte m	b	íte m	b	íte m	b
B2	-0.01	**B17**	-0.73	A11	-1.04	C10	0.57	D9	-0.16
B3	1.05	**B18**	-0.19	A12	0.23	C11	-027	D10	0.09
B4	-0.28	B19	0.28	A13	-0.99	C12	-0.85	D11	0.08
B5	0.41	B20	0.39	A19	1.12	C13	-0.89	D12	-0.07
B6	0.61	B21	0.24	A20	0.76	C19	-0.26	D13	-0.03
B7	0.19	B22	0.28	A21	-0.34	C20	0.29	D19	-0.07
B8	0.09	B23	0.32	A22	0.58	C21	-0.05	D20	0.07
B9	-0.75	B24	0.53	A23	0.48	C22	0.18	D21	-0.01
B10	-0.57	B25	-0.41	A24	1.35	C23	0.51	D22	0.05
B11	-1.08	A6	1.65	A25	1.39	C24	0.66	D23	0.13
B12	-0.01	A7	0.76	C6	-0.47	C25	0.21	D24	0.17
B13	-0.22	A8	0.12	C7	0.15	D6	-0.11	D25	0.05
B14	-0.17	A9	-0.99	C8	-0.37	D7	-0.30		
B16	0.79	A10	0.58	C9	-0.22	D8	0.12		

Nota. b= dificultad de los ítems equiparados, los ítems anclas aparecen en negrita.

Construcción y simulación del algoritmo del TAI

Razonamiento numérico (RN)

Luego de especificar los argumentos del algoritmo del TAI, se generó un patrón de respuesta mediante el comando *random*CAT y se estimó, mediante un diseño de simulaciones, 1000 veces ese patrón de respuesta considerando las dificultades de los ítems del BI y los niveles de habilidad en cada simulación. En la tabla 4 se observa que a medida que

disminuye el Error Estándar (SE) de estimación (theta) se requiere de un mayor número de ítems para estimar la habilidad final del examinado. En general se observa que la media de los ítems aumentó de manera gradual a medida que la regla de parada establecía un criterio más estricto.

En la tabla 3 se observa que la media del error estándar de la habilidad varió en un rango entre -0.02 a 0.03, dicho rango se mantuvo cercano al nivel de habilidad 0 a través de todas las reglas de parada. Los ítems que componen este BI presentan limitaciones en la información del banco para un criterio de parada SE (theta) ≤ 0.2.

En términos de precisión, los valores de sesgo (bias) se mantuvieron cercanos a cero en todas las condiciones, indicando ausencia de sesgo sistemático en la estimación de la habilidad. Por su parte, el RMSE disminuyó progresivamente al reducir el umbral de error estándar, pasando de 0.44 (SE ≤ 0.6) a 0.22 (SE ≤ 0.2), lo que refleja una mejora en la exactitud de las estimaciones. No obstante, se observó que la reducción del RMSE no es proporcional al aumento en la longitud del test. Por ejemplo, al pasar de SE ≤ 0.3 a SE ≤ 0.2, la longitud del test aumentó considerablemente (de 48 a 82 ítems), mientras que la mejora en RMSE fue relativamente moderada.

Tabla 4

Características del TAI del BI de RN según la regla de parada

Regla de parada	Número de ítems utilizados		Media de SE(θ)	Bias	RMSE
	M	*DE*			
SE(θ) ≤ 0.6	13	0.51	0.02	0.006	0.44
SE(θ) ≤ 0.5	18	0.45	0.01	0.013	0.39
SE(θ) ≤ 0.4	27	0.37	0.02	0.002	0.35
SE(θ) ≤ 0.3	48	0.29	0.01	0.001	0.26
SE(θ) ≤ 0.2	82	0.23	0.01	0.009	0.22

Nota. EE (θ): error estándar de estimación

En relación con la función de información del test, en la figura 1 se observa que se concentra la información en torno al nivel medio superior de habilidad

Figura 1.

Función de información del BI de Razonamiento numérico

Razonamiento verbal (RV)

En la tabla 5 se muestran los resultados de la simulación de 1000 patrones de respuesta del algoritmo del TAI estimado para un nivel de habilidad 0 con una regla de parada basado en las diferentes reglas de parada de la prueba. Se observa que la media y DE de los ítems aumenta a medida que SE (theta) de la regla de parada disminuye. Además, observamos que con un SE (theta) ≤ 0.2 el banco de ítems no resultó adecuado para la estimación de la habilidad final en el presente estudio de simulación. En general se observa que a medida que el criterio de precisión fue más estricto la media de los ítems aumentó y la media del SE (theta) se mantuvo cercana a los niveles de habilidad cercanos a 0. Cabe mencionar que a medida que aumenta el SE (theta) necesitamos un menor número de ítems para estimar la habilidad final. Si consideramos los resultados obtenidos, observamos que en promedio se necesitaron, en la aplicación del TAI, un promedio 18 ítems para estimar una habilidad media con un SE (theta) ≤ 0.5.

En relación con el sesgo, los valores de bias se mantuvieron próximos a cero en todas las condiciones (entre 0.03 y 0.00), lo que indica que el algoritmo no presenta sesgos sistemáticos en la estimación de la habilidad. En cuanto a la exactitud, el RMSE mostró una disminución progresiva al exigir mayores niveles de precisión. Sin embargo, esta mejora no se corresponde linealmente con el aumento en la longitud del test. En particular, al comparar los criterios SE (θ) ≤ 0.3 y SE (θ) ≤ 0.2, se observa un incremento considerable en la cantidad de ítems administrados, mientras que la mejora en la precisión resulta relativamente moderada, evidenciando rendimientos decrecientes.

Tabla 5

Características del TAI del BI de RV según la regla de parada

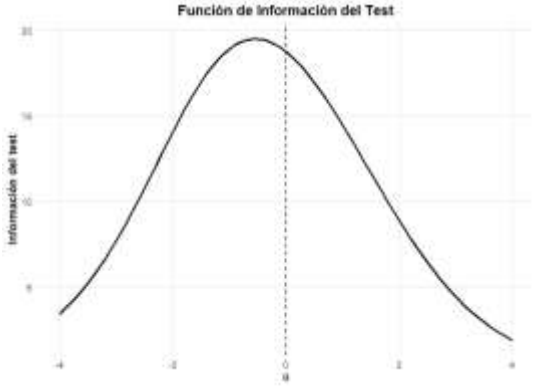
Regla de parada	Número de ítems utilizados		Media de *SE* (θ)	Bias	RMSE
	M	*DE*			
SE(θ) ≤ 0.6	13	0.51	0.02	0.012	0.52
SE(θ) ≤ 0.5	18	0.45	-0.02	0.035	0.40
SE(θ) ≤ 0.4	28	0.37	-0.00	0.007	0.34
SE(θ) ≤ 0.3	49	0.29	0.00	0.000	0.27
SE(θ) ≤ 0.2	96	0.23	0.01	0.006	0.22

Nota. EE (θ): error estándar de estimación

La función de información del BI de razonamiento verbal nos indica su mayor precision en los niveles medios-bajos de habilidad (Figura 2).

Figura 2.

Función de información del BI de Razonamiento verbal



Razonamiento Espacial (RE)

A partir de la simulación del BI de RE, en la tabla 6 observamos que la media del error estándar de estimación varía de 0.02 a 0.03 manteniendo una estimación del nivel de habilidad cercano a 0. En términos de precisión, el error estándar de estimación disminuye a medida que se reduce el valor permitido de SE (θ), lo que refleja una mejora en la precisión de las estimaciones. Asimismo, los valores de sesgo (bias) se mantuvieron cercanos a cero en todas las condiciones (entre -0.02 y 0.03), lo que indica ausencia de sesgo sistemático en la estimación de la habilidad. En relación con la exactitud, el RMSE disminuyó progresivamente al exigir criterios de parada más estrictos, evidenciando una mejora en la calidad de las estimaciones. No obstante, se observa que esta mejora no es proporcional al incremento en la longitud del test.

Con un criterio de parada de SE (θ) ≤ 0.5 se requiere un promedio de 18 (*DE* = 0.45) ítems para finalizar la estimación de dicho nivel de aptitud. Cuando se estableció un criterio de parada con un SE (theta) ≤ 0.2 el BI de ítems no resultó suficiente para la estimación de la habilidad final en dicha regla de parada, por lo cual en la simulación utilizó el pool completo de ítems del BI de RE. Esto sugiere que el BI presenta limitaciones en la información disponible para alcanzar altos niveles de precisión de manera eficiente.

Tabla 6.

Características del TAI del BI de RE según la regla de parada

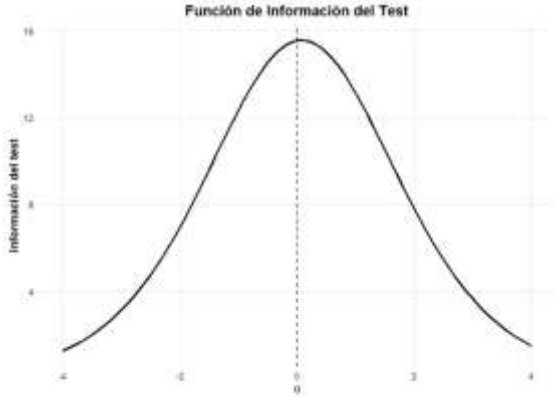
Regla de parada	Número de ítems utilizados		Media de *SE*(θ)	Bias	RMSE
	M	*DE*			
SE(θ) ≤ 0.6	13	0.51	0.01	0.003	0.43
SE (θ) ≤ 0.5	18	0.45	0.03	0.013	0.40
SE(θ) ≤ 0.4	27	0.37	0.02	0.004	0.34
SE(θ) ≤ 0.3	47	0.29	-0.00	0.008	0.27
SE(θ) ≤ 0.2	67	0.37	-0.02	0.010	0.24

Nota. EE (θ): error estándar de estimación

En la Figura 3 se presenta la función de información del BI de razonamiento espacial, mostrando su mayor precisión en los niveles medios de habilidad.

Figura 3.

Función de informacion del BI de razonamiento espacial



IV. DISCUSIÓN

El objetivo de este trabajo fue desarrollar el algoritmo de un TAI a partir de los BI que evalúan razonamiento numérico, verbal y espacial. Para esto, se utilizaron los ítems que fueron sometidos a un proceso de evaluación psicométrica y presentaron un adecuado ajuste al modelo de Rasch en el proceso de calibración de los ítems [25]. En nuestro país se observa que han aumentado las investigaciones relacionadas al desarrollo de BI desde los modelos de TRI, pero aún no existe evidencia de trabajos que desarrollen algoritmos de TAI y sometan dichas especificaciones a un proceso de simulación de patrones de respuestas en pruebas que evalúan diferentes capacidades cognitivas.

Considerando que uno de los requisitos del desarrollo de un TAI es contar con un BI adecuadamente calibrado, en este trabajo presentamos la equiparación de las puntuaciones de las dificultades de los ítems para la construcción definitiva del BI. Luego se especificó el modo de funcionamiento del TAI para construir el algoritmo del test. Este algoritmo guió cada ciclo de repetición para determinar la selección de los ítems que debieron aplicarse en cada examinado, considerando su nivel de habilidad y estableciendo a partir de cada regla de parada, el criterio de finalización del TAI. En este trabajo presentamos un análisis preliminar del funcionamiento del algoritmo del TAI simulando los patrones de respuesta. Evaluamos la precisión del BI de cada tipo de razonamiento a través de las diferentes reglas de parada.

En líneas generales observamos que, al establecer como criterio de parada el error estándar de estimación de la habilidad, el presente BI nos permite estimar con precisión un nivel de habilidad medio con un SE (theta) menor o igual a 0.5. En este punto, comparando la administración en el formato lápiz y papel [25], el TAI permite reducir el número de ítems administrados entre un 26% y un 28% aproximadamente en cada BI. Los resultados nos permiten observar que aplicando el algoritmo desarrollado necesitaremos por ejemplo entre 47 y 49

ítems para estimar la habilidad de un participante con un nivel de habilidad 0 en un criterio de parada de SE (θ) ≤ 0.3, mientras que para estimar un nivel de habilidad 0 con un SE (theta) ≤ 0.5 un participante necesitará en promedio responder a 18 ítems antes que se ejecute el criterio de parada del TAI. Cabe aclarar que, si bien es deseable un criterio de parada con el menor error posible, con un SE (theta) ≤ 0.5 estaríamos garantizando un criterio de precisión aceptable, acortando considerablemente la cantidad de ítems para estimar un nivel de habilidad media [33]. Este resultado es consistente con los valores observados de RMSE, que muestran mejoras moderadas en la exactitud al aumentar la exigencia del criterio de parada.

En el estudio de simulación realizado en los tres tipos de razonamiento observamos que las tres reglas de paradas que podrían resultar más precisas para estimar la habilidad final de un participante son aquellas que presentaron un error estándar de estimación menor o igual a 0.4, 0.5 y 0.6 dichos resultados fueron similares a los reportados por el estudio realizado por [34], en el cual obtuvieron mayor precisión en reglas de parada con simulaciones del error estándar de estimación similares a los que se presentan en este estudio. Respecto a la regla de parada con un error menor o igual que 0.2 en los tres tipos de razonamiento la prueba se detuvo aplicando la totalidad de los ítems del BI, en este caso dicho comportamiento podría deberse a que el BI no fue suficiente para estimar la habilidad final con este criterio de parada [35].

En el marco de la medición de las aptitudes cognitivas la utilización de los TAIs permite evaluar de manera más eficiente a los estudiantes y generar pruebas de longitud variable adecuándolas al nivel de habilidad de cada participante. En consecuencia, se reduce la cantidad de ítems y el tiempo de administración de la prueba. Cabe mencionar que los estudios de simulación resultan fundamentales en cada una de las etapas de construcción de un TAI [36]. Es por esto que, si bien este estudio muestra un procedimiento para generar un algoritmo de un TAI y luego simula los patrones de respuesta del BI de razonamiento numérico, verbal y espacial variando las reglas de parada, cada estudio deberá adecuar los procedimientos considerando las acciones específicas a implementar en un TAI.

Cabe agregar que, si bien se realizó un estudio simulado, futuros estudios deberían explorar el funcionamiento del algoritmo en situaciones prácticas con participantes reales. Esto resulta necesario porque existen factores que pueden afectar las respuestas de los participantes tales como el tiempo de respuesta y el contexto de administración [34]. Además, estudios próximos deberían observar la precisión del TAI en otros niveles de habilidad y probar especificaciones del algoritmo del test con otros métodos de selección de ítems y estimación de habilidad. De igual forma resta realizar estudios más profundos respecto a la fiabilidad y validez del TAI de razonamiento numérico, verbal y espacial.

Conclusión

La construcción y simulación del algoritmo del TAI que evalúa razonamiento numérico, verbal y espacial sienta las bases para la aplicación de un sistema innovador en aspectos tecnológicos en Argentina que pueden ser aplicados en el ámbito educativo. La utilización de estas herramientas posibilitará ofrecer mayor precisión, objetividad y economía en la administración de pruebas de capacidades cognitivas. Tal es así que su aplicación motivará a los estudiantes a responder pruebas que no dependan de la administración de un número determinado de ítems sino más bien que las evaluaciones estén adaptadas al rendimiento de cada estudiante, logrando reducir la longitud de la prueba y tal como se mencionó adecuar la estimación de la habilidad del participante al patrón de respuesta en dicha prueba adaptativa.

Referencias

[1] M. Gusev and G. Armenski, "E-assessment systems and online learning with adaptive testing," in *E-Learning Paradigms and Applications*, M. Ivanovic and L. C. Jain, Eds. Heidelberg, Germany: Springer, 2014, pp. 145–168.

[2] H. K. Wu, C. Y. Kuo, T. H. Jen, and Y. S. Hsu, "What makes an item more difficult? Effects of modality and type of visual information in a computer-based assessment of scientific inquiry abilities," *Computers & Education*, vol. 85, pp. 35–48, 2015.

[3] S. W. Choi, T. Podrabsky, and N. McKinney, "Firestar-'D': Computerized adaptive testing simulation program for dichotomous item response theory models," *Applied Psychological Measurement*, vol. 36, no. 1, pp. 67–68, 2012.

[4] J. Suárez-Álvarez, R. Fernández-Alonso, F. García-Crespo, and J. Muñiz, "El uso de las nuevas tecnologías en las evaluaciones educativas: La lectura en un mundo digital," *Papeles del Psicólogo*, vol. 43, no. 1, pp. 36–47, 2022. doi: 10.23923/pap.psicol.2986

[5] U. Neisser, G. Boodoo, T. J. Bouchard Jr., A. W. Boykin, N. Brody, S. J. Ceci, and S. Urbina, "Intelligence: Knowns and unknowns," *American Psychologist*, vol. 51, no. 2, pp. 77–101, 1996. doi: 10.1037/0003-066X.51.2.77.

[6] A. R. A. Conway and K. Kovacs, "New and emerging models of human intelligence," *WIREs Cognitive Science*, vol. 6, pp. 419–426, 2015. doi: 10.1002/wcs.1356.

[7] J. P. Guilford, *The Nature of Human Intelligence*, New York: McGraw-Hill, 1967.

[8] B. Spinath, H. H. Freudenthaler, and A. C. Neubauer, "Domain-specific school achievement in boys and girls as predicted by intelligence, personality and motivation," *Personality and Individual Differences*, vol. 48, no. 4, pp. 481–486, 2010.

[9] R. J. Sternberg, "Implicit theories of intelligence, creativity, and wisdom," *Journal of Personality and Social Psychology*, vol. 49, no. 3, pp. 607–627, 1985.

[10] J. C. Raven, J. H. Court, and J. Raven, *Test de Matrices Progresivas: Escalas Coloreada, General y Avanzada. Manual*, Madrid: TEA Ediciones, 1993.

[11] D. Wechsler, *Wechsler Abbreviated Scale of Intelligence*, San Antonio, TX: Psychological Corporation, 1999.

[12] D. Wechsler, *Manual for the Wechsler Intelligence Scale for Children*, 4th ed. San Antonio, TX: Psychological Corporation, 2005.

[13] G. K. Bennett, A. G. Wesman, and H. G. Seashore, *DAT-5: Tests de aptitudes diferenciales. Versión 5. Manual*, Madrid: TEA Ediciones, 2000.

[14] T. P. Hogan, *Pruebas psicológicas: Una introducción práctica*. México: Editorial El Manual Moderno, 2015. ISBN: 978-607-448-501-1

[15] K. Jiménez and E. Montero, "Aplicación del modelo de Rasch en el análisis psicométrico de una prueba de diagnóstico en matemática," *Revista Digital Matemática*, vol. 13, no. 1, pp. 1–24, 2013.

[16] J. Olea, V. Ponsoda, and G. Prieto, *Tests informatizados: Fundamentos y aplicaciones*, Madrid: Pirámide, 1999.

[17] F. J. P. Abal, G. S. Lozzia, D. Blum, M. Galibert, M. E. Aguerri, and H. F. Attorresi, "Análisis de ítems de un test de altruismo a partir del modelo logístico de un parámetro," *Perspectivas en Psicología: Revista de Psicología y Ciencias Afines*, vol. 7, no. 1, pp. 16–23, 2010.

[18]H. Wainer, *Computerized Adaptive Testing: A Primer*, 2nd ed. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.
[19]M. Phankokkruad, "Association rules for data mining in item classification algorithm: Web service approach," in *Proc. 2nd Int. Conf. Digital Information and Communication Technology and Its Applications (DICTAP)*, 2012, pp. 463–468. doi: 10.1109/DICTAP.2012.6215408.
[20]S. Kanjanawasee, *Modern Test Theory*, 4th ed. Bangkok: Chulalongkorn University Press, 2012.
[21]H. H. Chang, "Understanding computerized adaptive testing: From Robbins–Monro to Lord and beyond," in *The SAGE Handbook of Quantitative Methodology for the Social Sciences*, D. Kaplan, Ed. Thousand Oaks, CA: Sage, 2004, pp. 321–340.
[22] V. Rubio and J. Santacreu, *TRAS-I: Test adaptativo informatizado para la evaluación del razonamiento secuencial y la inducción*, Madrid: TEA Ediciones, 2004.
[23]G. S. Lozzia, F. J. P. Abal, G. D. Blum, M. E. Aguerri, M. S. Galibert, and H. F. Attorresi, "Construcción de un banco de ítems de analogías verbales como base para un test adaptativo informatizado," *Revista Mexicana de Psicología*, vol. 32, no. 2, pp. 134–148, 2015.
[24]K. C. T. Han, "Conducting simulation studies for computerized adaptive testing using SimulCAT: An instructional piece," *Journal of Educational Evaluation for Health Professions*, vol. 15, art. no. 20, 2018. doi: 10.3352/jeehp.2018.15.20.
[25]F. Ghio, V. Morán, S. Garrido, A. Azpilicueta, F. Cortéz, and M. Cupani, "Calibración de un banco de ítems mediante el modelo de Rasch para medir razonamiento numérico, verbal y espacial," *Avances en Psicología Latinoamericana*, vol. 38, no. 1, pp. 157–171, 2020. doi: 10.12804/revistas.urosario.edu.co/apl/a.7760.
[26]K. Russell and P. Carter, *Ultimate IQ Tests: 1000 Practice Test Questions to Boost Your Brainpower*, London: Kogan Page, 2015.
[27]M. J. Navas, "Equiparación de puntuaciones," in *Psicometría*, J. Muñiz, Ed. Madrid: Universitas, 1996, pp. 345–372.
[28]M. J. Kolen and R. L. Brennan, *Test Equating, Scaling, and Linking: Methods and Practices*, 3rd ed. New York: Springer, 2014. doi: 10.1007/978-1-4939-0317-7.
[29]G. S. Lozzia and H. F. Attorresi, "Especificación del algoritmo para un test adaptativo informatizado de analogías verbales," *SUMMA Psicológica UST*, vol. 9, no. 2, pp. 15–23, 2012.
[30]D. Magis and G. Raîche, "Random generation of response patterns under computerized adaptive testing with the R package catR," Journal of Statistical Software, vol. 48, no. 8, pp. 1–31, 2012. doi: 10.18637/jss.v048.i08.
[31]T. A. Warm, "Weighted likelihood estimation of ability in item response models," *Psychometrika*, vol. 54, pp. 427–450, 1989. doi: 10.1007/BF02294627.
[32]R. D. Bock and R. J. Mislevy, "Adaptive EAP estimation of ability in a microcomputer environment," *Applied Psychological Measurement*, vol. 6, pp. 431–444, 1982. doi: 10.1177/014662168200600405.
[33]F. J. P. Abal, S. E. Auné, and H. F. Attorresi, "Construcción de un banco de ítems de facetas de neuroticismo para el desarrollo de un test adaptativo," *Psicodebate*, vol. 19, no. 1, pp. 31–50, 2019. doi: 10.18682/pd.v1i1.854.
[34]Q. Tan, Y. Cai, Q. Li, Y. Zhang, and D. Tu, "Development and validation of an item bank for depression screening in the Chinese population using computerized adaptive testing: A simulation study," *Frontiers in Psychology*, vol. 9, art. no. 1225, 2018. doi: 10.3389/fpsyg.2018.01225.
[35]J. M. Linacre, "Computer-adaptive testing: A methodology whose time has come," in *Development of Computerized Middle School Achievement Test*, S. Chae, U. Kang, E. Jeon, and J. M. Linacre, Eds. Seoul, South Korea: Komesa Press, 2000.
[36]K. C. T. Han, "Conducting simulation studies for computerized adaptive testing using SimulCAT: An instructional piece," *Journal of Educational Evaluation for Health Professions*, vol. 15, art. no. 20, 2018. doi: 10.3352/jeehp.2018.15.20.