APLICACIÓN DE MODELOS POLITÓMICOS DE LA TRI PARA LA ESPECIFICACIÓN DE LOS FORMATOS DE RESPUESTA TIPO LIKERT EN UN TEST DE COMPORTAMIENTO TÍPICO

Director: *Attorresi, Horacio Félix*
Codirectora: *Aguerri, María Ester*
Becario: *Abal, Facundo Juan Pablo*
E-mail: *fabal@psi.uba.ar*

El objetivo general de este proyecto es contribuir al desarrollo de los modelos politómicos de la TRI en nuestro país a través de una de sus aplicaciones más importantes: el estudio de las características óptimas de las escalas tipo Likert. Se confeccionó un protocolo que contiene dos instrumentos validados en estudios previos que miden los constructos Confianza y Utilidad pertenecientes a la dimensión afectiva del aprendizaje de nociones matemáticas. La variable Confianza evalúa las creencias que un estudiante tiene sobre sus posibilidades y dificultades para responder a las habilidades requeridas en la actividad matemática. La variable Utilidad examina las creencias con respecto de la importancia atribuida a la Matemática tanto para la carrera (utilidad presente) como para el ejercicio profesional (utilidad futura). Los ítems de ambas pruebas aparecen en el protocolo con 3 formatos de respuesta experimentales: a) el formato original con 6 categorías (*totalmente en desacuerdo*, *en desacuerdo*, *más bien en desacuerdo*, *más bien de acuerdo*, *de acuerdo* y *totalmente de acuerdo*), b) un formato de 5 opciones que resume las dos centrales y cuyo anclaje es *indiferente*, c) un formato con 3 categorías para estudiar si el efecto de una reducción considerable de opciones repercute en los indicadores de calidad psicométrica (*en desacuerdo*, *indiferente* y *de acuerdo*). Se administró el protocolo a 940 estudiantes de Psicología. Se aleatorizó el orden de exposición en que los participantes contestaron los distintos formatos experimentales y se incluyeron otras pruebas intercaladas para reducir el efecto de la memoria en las respuestas. Estos instrumentos adicionales también serán utilizados para obtener evidencias externas de validez basadas en la relación con otras variables. Se prevé el análisis de los datos obtenidos con los modelos de Respuesta Nominal de Bock, Respuesta Graduada de Samejima y Crédito Parcial de Masters. Las evidencias de validez y confiabilidad recolectadas con los 3 modelos y la teoría clásica permitirán determinar el formato óptimo y más parsimonioso para la escala Likert de ambas escalas. La concreción del proyecto permitirá llevar a la práctica un planteo metodológico innovador en nuestro medio así como también realizará un aporte tecnológico a partir del perfeccionamiento de herramientas válidas, fiables y útiles.

ESTUDIOS DE SIMULACIÓN PARA EL ANÁLISIS DEL FUNCIONAMIENTO DIFERENCIAL DEL ÍTEM Y APLICACIONES DE LA TEORÍA DE RESPUESTA AL ÍTEM A UN TEST DE RAZONAMIENTO ANALÓGICO

Director: *Attorresi, Horacio F.*
Becario: *Blum, G. Diego.*
E-mail: *blumworx@gmail.com*

Objetivo.
Estudiar el Error de Tipo I y la potencia de las pruebas de Breslow-Day y de reglas que las combinan con el procedimiento estándar de Mantel-Haenszel aplicadas a la detección del funcionamiento diferencial del ítem (Differential Item Functioning, DIF) y evaluar su desempeño comparando con otros métodos. Delimitar conceptual y operacionalmente el constructo Razonamiento Analógico, administrar una escala de Analogías Figurales (AF), modelizarla con los componentes de la Teoría Clásica de Tests (TCT) y de la Teoría de Respuesta al Item (TRI) y aplicar los resultados de la simulación a dicha escala.

Informe Final. Objetivos Alcanzados.
Se estudió el concepto de Funcionamiento Diferencial del Ítem (DIF) y se desarrollaron simulaciones para estudiar las tasas de Falsos Positivos (FP) y de Identificaciones Correctas (IC) de los métodos Breslow Day global (BD) y de la tendencia (BDT), reglas que los combinan con el procedimiento Mantel-Haenszel estándar (MH) mencionadas como MHoBD y MHoBDT, Mantel-Haenszel modificado (MHmo) por Mazor, Clauser y Hambleton (1994) y Regla de Decisión Combinada (RDC) de Penfield (2003). Se simularon las respuestas a un test de 75 ítems, 20 de los cuales poseyeron parámetros de discriminación (*a*) y de dificultad (*b*) que son resultado de combinar cuatro niveles de *a* (0.25, 0.6, 0.9 o 1.25) con cinco de *b* (-1.5, -1, 0, 1 o 1.5) mientras que el parámetro de acierto por azar (*c*) se consideró igual a 0.20 en todos los casos. Para mayor realismo, los 55 ítems restantes se tomaron de una calibración sobre datos reales entre los publicados por Clauser, Mazor y Hambleton (1994). Se investigaron cuatro condiciones, que son el resultado de combinar dos tamaños de los grupos Focal y de Referencia (GR=GF=1,000 y GR=GF=200) con dos situaciones de impacto (presencia o ausencia). Para simular impacto se consideró que μ_R-μ_F=1 y σ_R=σ_F=1, siendo GR perteneciente a una población normal estándar. Se utilizó PARDSIM® (Yoes, 1997) para simular 100 repeticiones de respuestas por condición y Bday (Prieto-Marañón, 2005) para estudiar las tasas de FP e IC sobre estas repeticiones, al 1% y al 5%. Se estudió si la tasa de FP cumplía la condición liberal o la estricta de Bradley (1978).

En primer lugar, se simuló el DIF no Uniforme Paralelo (Hanson, 1998) en los 20 ítems artificiales. El parámetro *b* del ítem en el GF se obtuvo sumando 1 al parámetro *b* del ítem en el GR (Δb = 1.00). Según el estudio, sólo la prueba global de BD satisfizo la condición liberal de Bradley cuando el tamaño de los grupos es 1,000, excepto al 1% y en presencia de impacto. Las tasas de FP de los métodos restantes resultaron infladas tanto al 5% como al 1%. BD presentó las tasas de IC más bajas mientras que MHoBDT y MHmo presentaron las tasas de IC más altas en todas las condiciones. Las altas tasas de falsos positivos podrían ser atribuibles, entre otras razones, a la utilización de un programa computacional que no es bietápico, como sucede con Bday. Por lo tanto se decidió continuar con la evaluación de los métodos pero en situación de mínima contaminación. Por tal razón, en las simulaciones siguientes se generó DIF en un único ítem del test. La inspección general de las tasas de FP y de IC para la situación con mínima contaminación indica que cuando el tamaño de los grupos es 200, MH es el único que posee tasas de FP que cumplen al menos la condición liberal de Bradley. Frente a tamaños de 1,000, las tasas que cumplen esta condición en numerosas situaciones corresponden a los métodos BD, BDT, MH y RDC. Las tasas de IC varían de acuerdo al diseño y al tipo de DIF estudiado.

Se estudió el constructo Razonamiento Analógico (RA. Blum, Abal, Lozzia, Picón Janeiro, & Attorresi, 2011) y se lo operacionalizó mediante un test con 36 ítems, cuyo contenido es principalmente figural. Cada reactivo consiste en una matriz de 2x2 con un elemento ausente, del estilo de problemas A:B::C:?, más cinco distractores y una opción correcta. Las reglas utilizadas para relacionar pares de figuras dentro de la matriz fueron la rotación, la traslación, la distorsión del tamaño, la distorsión de la forma, la adición y la sustracción. Se elaboraron ítems con una regla, con dos reglas y con tres reglas y se los administró a una muestra de estudiantes de Estadística de la Facultad de Psicología de la Universidad de Buenos Aires. Se realizó un análisis de datos sobre 475 protocolos depurados de dicha muestra. La consistencia interna del test es elevada (α = .91) y se confirma la unidimensionalidad por medio de una serie de indicadores, deduciendo de ella la independencia local. Aplicando el Modelo Logístico de Tres Parámetros (ML3P) de la Teoría de Respuesta al Ítem (TRI), no se rechaza el ajuste global al 5% mediante la prueba χ^2. Existe una buena potencia discriminatoria general (*a*: Media = 1.02), un nivel de dificultad medio (*b*: Media = -0.03) y un nivel de acierto por azar ligeramente inferior a lo esperable con 6 opciones de respuesta (*c*: Media = .14). Estos resultados han sido desarrollados por Blum, Galibert, Abal, Lozzia y Attorresi (2011).

Luego se modificaron 15 de los reactivos debido a limitaciones tales como la frecuencia asimétrica de contestación de sus distractores y la mala regulación de la dificultad a medida que crece el número de reglas en el ítem. El nuevo test fue administrado a una muestra de estudiantes de la Facultad de Arquitectura, Diseño y Urbanismo y del Instituto Universitario Nacional del Arte, de la cual se consideraron 1,129 protocolos depurados. Las razones principales que llevaron a realizar el estudio con la segunda muestra fueron: 1) afinidad de los alumnos a los conceptos visuales y espaciales del test dado el contenido de las carreras, 2) nivelación de los porcentajes según el género y 3) superar algunas limitaciones del primer estudio, a saber, a) el problema de la velocidad interviniendo en las respuestas, ya que en el primer estudio la correlación entre el tiempo total y el puntaje total resulta significativa al 1%, y b) el tamaño de la muestra acorde para un Modelo Logístico de Tres Parámetros (ML3P), ya que se recomienda que $n > 1000$ (Hanson & Beguin, 2002; Yen, 1987). En este segundo estudio, el índice de consistencia interna resulta más elevado (α = .93) y vuelve a confirmarse la unidimensionalidad. La potencia discriminatoria general aumenta (*a*: Media = 1.26), el nivel de dificultad continúa siendo medio (*b*: Media = -0.09) y el nivel de pseudoazar general sigue siendo el esperable (*c*: Media = .15). Estos resultados pueden encontrarse en Blum, Lozzia, Abal y Attorresi (2014, en prensa), así como recomendaciones para futuras construcciones de ítems de analogía figural.

Se estableció un análisis del DIF sobre los datos reales del test, en orden de indentificar y eliminar aquellos reactivos que hayan mostrado una tendencia comprobable a funcionar a favor de un determinado grupo, independientemente del nivel de habilidad de estos individuos. El objetivo de esta parte del estudio fue aplicar los métodos más apropiados para detectar DIF en los ítems del test. Se consideraron tres matrices de datos: aquélla que reúne los 475 protocolos de la primera muestra (M1), la que reúne los 1,129 protocolos de la segunda muestra (M2) y la que reúne los 1,604 protocolos de ambas muestras considerando sólo los 21 ítems no modificados de muestra en muestra (M3). Para M1 y M2, los grupos fueron divididos por género (GR = mujeres; GF = varones), y para M3 se dividieron por orientación académica (GR = M2; GF = M1). Según los resultados descritos por Penfield (2003), RDC es un método efectivo para la detección del DIF tanto Uniforme como no Uniforme; por este motivo, se lo escogió para el presente estudio. Además, se consideraron libres de DIF los ítems donde MH D-DIF no superó a 1 como valor absoluto. También se estudió el DIF con la Prueba Normal para la Diferencia de los *b* considerando un nivel de confianza del 95%, para contar con un tercer dato que respalde la presencia de DIF. Luego se calculó la Diferencia de Proporciones Estandarizadas (DPE) para determinar la dirección del DIF en los ítems así detectados con los métodos anteriores. Finalmente, se calculó el Índice de Impacto (I) de los ítems bajo estudio restando la proporción de respuestas correctas del GR a las del GF ($p_F - p_R$). Un resultado con signo positivo indica impacto a favor del GF. Sobre la base de los resultados, puede concluirse que el único ítem que posee DIF con respaldo en todos los métodos considerados, es el ítem 9 de M2, y es un DIF que favorece al GF (varones). Además, su DPE puntuó 0.07, favoreciendo al GF, y su E = 1.57. El cálculo del impacto en este reactivo brindó un I = 0.17, favoreciendo nuevamente al GF.