APLICACIÓN DEL MODELO DE RESPUESTA GRADUADA A UNA ESCALA DE VOLUNTAD DE TRABAJO*

APPLICATION OF THE GRADED RESPONSE MODEL TO A WILL-TO-WORK SCALE

HORACIO FÉLIX **ATTORRESI****, FACUNDO JUAN PABLO **ABAL*****, MARÍA SILVIA **GALIBERT******, GABRIELA SUSANA **LOZZIA******* Y MARÍA ESTER **AGUERRI******

*Trabajo realizado con subsidios de la Universidad de Buenos Aires (UBACyT P043) y de la Agencia Nacional de Promoción Científica y Tecnológica (ANPCyT PICT Nº 20909).

**Licenciado en Ciencias Matemáticas. Profesor Titular de la Cátedra II de Estadística de la Facultad de Psicología de la Universidad de Buenos Aires (UBA). Director de los proyectos ANPCyT PICT 20909 y UBACyT P043.
Rivera Indarte 132, 1er Piso Dpto. A - (C1406DXD) Ciudad Autónoma de Buenos Aires - República Argentina. E–Mail: hattorre@psi.uba.ar

***Licenciado en Psicología. Becario del Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) y Ayudante de Primera de las Cátedras de Estadística de la Facultad de Psicología de la Universidad de Buenos Aires (UBA).

****Magister Scientiae en Biometría. Profesora Adjunta Regular de la Cátedra II de Estadística de la Facultad de Psicología de la Universidad de Buenos Aires (UBA). Codirectora del Proyecto UBACyT P043 e Investigadora en el Proyecto ANPCyT PICT 20909 de la Agencia Nacional de Promoción Científica y Tecnológica (ANPCyT).

*****Licenciada y Profesora en Psicología. Becaria del Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) y Jefa de Trabajos Prácticos de las Cátedras de Estadística de la Facultad de Psicología de la Universidad de Buenos Aires (UBA).

******Magister Scientiae en Biometría. Profesora Adjunta Regular de la Cátedra I de Estadística de la Facultad de Psicología de la Universidad de Buenos Aires (UBA). Codirectora del Proyecto UBACyT P043 e Investigadora en el Proyecto ANPCyT PICT 20909 de la Agencia Nacional de Promoción Científica y Tecnológica (ANPCyT).

RESUMEN

Se presentan los resultados obtenidos con la aplicación de un modelo de la *Teoría de Respuesta al Ítem* (TRI) a los reactivos de una prueba que mide la *Voluntad de Trabajo* (VT). La VT es un rasgo de la personalidad que describe la tendencia de un individuo a asumir sus obligaciones con responsabilidad, automotivación y sin dilaciones, aun cuando estas pudieran no ser de su agrado. La escala que mide la VT se compone de 9 ítemes con formato de respuesta ordenada de 4 valores. Se aplicó el *Modelo de Respuesta Graduada* (MRG) de Samejima a datos provenientes de la administración de la escala de VT a una muestra de 1.141 universitarios. Previamente se verificó la condición de unidimensionalidad de los ítemes, requerida por

el MRG. El análisis de los datos se realizó operando el programa MULTILOG. La estimación de los parámetros de este modelo se efectuó por máxima verosimilitud marginal. Sólo uno de los ítemes mostró un comportamiento inadecuado. La mayoría de los parámetros de localización tendieron a ubicarse en niveles medios bajos de la escala. Los parámetros de discriminación adoptaron valores entre moderados y altos. La Función de Información del Test evidenció que la escala es más precisa para discriminar individuos con niveles medios-bajos del rasgo evaluado. Los resultados revelaron los problemas que presenta la escala de VT y permitieron obtener información útil para orientar la construcción de nuevos reactivos.

Palabras clave: Modelo de Respuesta Graduada; Modelo de Samejima; Voluntad de Trabajo; Teoría de Respuesta al Ítem.

Abstract

This study aims to present the findings obtained from the application of an *Item Response Theory* model (IRT) to the reactive of a *Will-to-Work* Measuring Test (WW). *WW* is defined as the individuals' tendency to generate efficient volatile processes that allow them to commit themselves to starting a task and to persisting in its execution by overcoming obstacles until they achieve its materialization with precision and without procrastination. WW is a personality trait that describes the predisposition of an individual to assume duties in a responsible, self-motivated and prompt manner even when such duties may be unappealing. The scale measuring the WW comprises 9 items in polychromous response format (four-point rating scale), with the response options graded. Accordingly, one of the IRT most widely used models was applied for the monetization of ordered polychromous responses: Samejima's *Graded Response Model* (GRM). The GRM is a generalization of the Two Parameter Logistic Model of Birnbaum. In GRM, a person's probability of responding in category *j* to a specific item *i*, $P_{ij}(\theta)$, is obtained by subtracting the probability of responding in or below category *j*-1 from the probability of responding in or below category *j*. Through the Item Response Category Curves, the GRM allows for the representation of an individual's likelihood to choose each of the item categories based on the level of the latent trait measured. The data from this psychometric test was obtained from a sample of 1,141 university students. The one-dimensional assumption required by the GRM was corroborated through an exploratory analysis of the data factor structure. The local independence assumption was considered to be satisfied after proving the scale one-dimensionality. All analyses based on the IRT were performed by operating the MULTILOG software program. The GRM parameters estimation was carried out through marginal maximum likelihood procedures. A discrimination parameter (a) and three location parameters (b_1, b_2 and b_3) corresponding to thresholds separating the 4 response categories were estimated for each item. The model's goodness-of-fit was studied on an item basis by examining the residue of observed and expected proportions for each of the ordered response categories. The residue obtained was the same as or lower than .01, which led to the conclusion that the model adjustment to the data was satisfactory for all reactive.

Despite this, one of the items showed inappropriate behavior. The value of its location parameters turned out to be very different from the expected one and showed high estimation errors when compared against the values obtained for the rest of the items.

Most of the location parameters showed mid-low WW values and discrimination parameters showed mid-high values (0.73-1.73). The instrument's reliability was acceptable if we consider the .75 marginal reliability coefficient obtained from IRT. However, local accuracy measures showed that the test is less reliable when measuring the WW highest levels. In other words, the measure error increases as we attempt to discriminate strongly willful individuals. This means that the WW scale is useful to measure mid-low levels of WW, but less accurate when it comes to individuals whose trait level is higher. It is therefore concluded that it is necessary to raise the number of WW scale items to optimize the instrument quality. It is particularly important to identify construct indicators allowing for a more accurate detection of the

highest trait levels. The shortage of the application of IRT models to personality tests as well as the difficulty that the achievement of their exigent assumptions were discussed. The findings showed the problems posed by the WW scale and allowed us to obtain useful information to guide the building of new items.

Key words: Graded Response Model; Samejima´s model; Will-to-Work; Item Response Theory.

La *voluntad de trabajo* (VT) es un rasgo de la personalidad que describe la predisposición de una persona a abocarse a sus obligaciones con autodisciplina y responsabilidad, aun cuando estas pudieran no ser de su agrado. Se define como la capacidad de un individuo para generar procesos volitivos que le permiten comprometerse activamente en el inicio de un trabajo, persistir en su concreción y concluirla con precisión y sin dilaciones (Lozzia, Abal, Aguerri, Galibert & Attorresi, 2007).

Dadas las características de este constructo, resultó de interés construir un instrumento que permitiera su medición en el ámbito académico. Abal, Lozzia, Galibert y Aguerri (2006) delimitaron el alcance de esta variable y elaboraron una escala para su medición según los supuestos del Modelo Lineal de la Teoría Clásica de Tests (TCT). Las evidencias obtenidas acerca de la validez factorial de la escala indicaron que la misma combina en una única dimensión estadística, aspectos como la perseverancia, la responsabilidad y la motivación intrínseca y extrínseca. En consonancia con lo esperado teóricamente, la prueba demostró capacidad para discriminar los universitarios que trabajan de aquellos que no trabajan. Así también, se asoció moderadamente con la tendencia a asumir la responsabilidad de ciertos actos aunque esto conlleve consecuencias negativas (Abal, Lozzia, Aguerri, Galibert & Attorresi, 2007).

A pesar de los avances alcanzados en la construcción de la escala VT en el marco de la perspectiva clásica, en los estudios instrumentales actuales se recomienda incorporar evidencias de validez proveniente de fuentes que ponen el acento en el análisis detallado de los ítems (Elosua, 2003). La TCT presenta las limitaciones que surgen de utilizar indicadores globales en la evaluación de la calidad de los ítemes. Este análisis resulta de gran utilidad, pero está muy lejos de ser exhaustivo.

En el marco de la Teoría de Respuesta al Ítem (TRI) se han propuesto numerosos modelos probabilísticos para predecir y explicar la respuesta de un individuo a un ítem a partir de su nivel en un rasgo inobservable (Hambleton & Swaminathan, 1985; Lord, 1980; Muñiz, 1997). Los modelos de la TRI postulan la existencia de una relación directa entre el comportamiento de un individuo frente a un ítem y el rasgo que genera esta conducta. La formalización de esta relación adopta la forma de una función matemática que vincula la probabilidad de dar una determinada respuesta a un ítem (i.e., elegir la opción - clave) para cada nivel del rasgo medido por éste.

Para la elección del modelo se debe tener en cuenta la cantidad de opciones que presenta el reactivo (dos o más de dos opciones) y el procedimiento de puntuación utilizado para codificar las respuestas (nominal u ordinal). En la TRI se considera a cada uno de los reactivos que componen el test como unidad de análisis. Por ende, el valor asignado a la opción elegida por la persona es el único dato con el que cuenta el investigador para formalizar la relación entre la respuesta al ítem y el rasgo latente.

Acorde con el formato de respuesta tipo Likert de los ítemes de la Escala VT se consideró efectuar el estudio psicométrico con el Modelo de Respuesta Graduada (MRG) de Samejima (1969, 1997). Éste es el primer modelo generado para el análisis de ítemes con respuesta politómica ordenada y, aunque surgieron nuevos modelos similares, el MRG presenta una notable vigencia en el área. Su elección para el presente estudio se

justifica en las múltiples investigaciones que mostraron la utilidad del MRG tanto para el análisis de ítemes de tests de rendimiento típico (Asún & Zúñiga, 2008; Flannery, Reise & Widaman, 1995; Gray-Little, Williams & Hancock, 1997) como de rendimiento máximo (e.g., Lane, Stone, Ankemann & Liu, 1995; Samejima, 1972).

Modelo de Respuesta Graduada

El MRG constituye la generalización del Modelo Logístico de dos parámetros (ML2p) de Birnbaum (1968) a ítemes con formato de respuesta politómica ordenada. El MRG describe el comportamiento de un ítem *i* mediante un parámetro de discriminación (a_i) y una serie de parámetros de umbral (b_{ij} = 1,..., *m*) que se ubican entre las categorías contiguas del ítem politómico (*j* = 0, ..., *m*). En consecuencia, el umbral b_{i1} está localizado entre la primera categoría (*j* = 0) y la segunda (*j* =1) mientras que el valor del umbral más alto (b_{im}) se encuentra entre la anteúltima (*j* = *m* - 1) y la última categoría (*j* = *m*). Uno de los objetivos del MRG es determinar la localización de cada uno de esos valores de umbral b_{ij} en el continuo del rasgo latente.

La extensión del ML2p propuesta por Samejima (1997) para ítemes politómicos se basa en una segmentación acumulativa de las categorías de respuesta. Es decir, en el MRG se descompone la respuesta politómica en una serie de dicotomías que separa para cada umbral b_{ij}, las categorías menores a *j* de las iguales o mayores a *j*, lo que permite aplicar el ML2p en cada una de las particiones. De esta manera, el MRG permite representar la probabilidad de que la respuesta de un examinado al ítem *i* esté en o por encima de un umbral b_{ij} en función del nivel del rasgo latente (θ) a partir de la ecuación:

$$P_{ij}^{*}(\theta) = \frac{1}{1 + e^{-a_i(\theta - b_{ij})}}$$

Donde

j = 1, ..., *m*

Como resultado de aplicar esta fórmula para cada una de las particiones que separan las categorías segmentadas acumulativamente se obtienen las denominadas *Curvas Características de Operación* (*Operating Characteristic Curves* - Embretson & Reise, 2000). Los parámetros de estas curvas características son similares a los del ML2p, aunque su interpretación resulta distinta en virtud de que presenta más de un parámetro de localización *b*. Cada parámetro b_{ij} es un valor de umbral que refleja, en la misma métrica de θ, la cantidad de rasgo necesaria para tener una probabilidad igual a .50 de responder en o por encima de la opción *j*. Esto significa que, por ejemplo, el b_{i1} es el mínimo nivel de rasgo que necesita un individuo para tener más chance de optar por la categoría *j* = 1 o una mayor (*j* = 2, 3, ..., *m*) en lugar de elegir la categoría *j* = 0. El parámetro de discriminación a_i representa el grado en que las categorías de respuestas diferencian entre los niveles del rasgo y se mantiene constante para todos los umbrales de las categorías de un mismo ítem.

La obtención de las Curvas Características de Operación $P^{*}_{ij}(\theta)$ es un paso intermedio para alcanzar las Curvas Características de las Categorías de Respuesta del Ítem (CCRI - Samejima, 1969). Estas últimas representan la probabilidad que tiene un individuo de optar por la categoría *j* en función de su nivel en el rasgo latente y se denotan como $P_{ij}(\theta)$.

Siendo *j* una categoría cualquiera del ítem y *j*+1 la categoría siguiente, el procedimiento aplicado para el cálculo de la probabilidad de elegir la categoría *j* dado un nivel de θ consiste en una sustracción de las probabilidades acumuladas a derecha:

$$P_{ij}(\theta) = P_{ij}^{*}(\theta) - P_{i(j+1)}^{*}(\theta)$$

Samejima (1969) definió que la probabilidad acumulada de responder en o por en-

cima de la categoría más baja $j = 0$ es $P^*_{i0} = 1$ mientras que la probabilidad de responder por encima de la última categoría $j = m$ es $P^*_{i(m+1)} = 0$. Por ende,

$$P_{i0}(\theta) = P^*_{i0}(\theta) - P^*_{i1}(\theta) = 1 - P^*_{i1}(\theta)$$

$$P_{ik}(\theta) = P^*_{ik}(\theta) - P^*_{i(k+1)}(\theta)$$

para k tal que $1 \leq k \leq m - 1$

$$P_{im}(\theta) = P^*_{im}(\theta) - P^*_{i(m+1)} = P^*_{im}(\theta) - 0$$

A partir del MRG también es posible obtener medidas locales de precisión, las cuales se hacen operativas mediante las Funciones de Información (FI) de los ítemes y del test. La FI de un ítem muestra la precisión con que un reactivo mide el rasgo latente a lo largo de todo su rango de valores. En virtud de la aditividad de las FIs de los ítemes para todos los niveles del rasgo, es posible obtener una FI del test completo así como una función de error típico de la estimación. Esto permite identificar para qué niveles de rasgo la escala resulta más o menos confiable.

Por otra parte, la confiabilidad marginal es otra medida de precisión del test que provee la TRI (Thissen, 1991). A diferencia de las medidas locales, este coeficiente es un indicador global de la confiabilidad del instrumento. Se calcula como un promedio de la fiabilidad a través de todos los niveles del rasgo ponderado por la proporción de individuos que corresponden a cada nivel. Esta característica lo convierte en un estimador próximo a la confiabilidad clásica, por lo que su representatividad sería mayor cuanto más uniforme sea la FI del test.

Objetivos

El objetivo general de este informe es presentar los resultados obtenidos con la aplicación del Modelo de Respuesta Graduada de Samejima (1969, 1997) al análisis de ítemes que componen la Escala de Voluntad de Trabajo.

Los objetivos específicos fueron los siguientes:

a.- Examinar la calidad de los ítemes de la Escala de VT a partir de los parámetros de las Curvas Características de las Categorías de Respuesta del Ítem obtenidas con el MRG.
b.- Obtener evidencias acerca de la precisión con que los ítemes y la escala total VT miden este constructo sobre la base de los indicadores propuestos desde la TRI.

Método

Participantes

Se contó con la colaboración de 1.141 estudiantes que cursaban el segundo año de la carrera de Psicología de la Universidad de Buenos Aires. El 17% fueron varones y el 83% fueron mujeres. La edad de los participantes osciló entre 18 y 69 años siendo la mediana de 20 años y la amplitud intercuartil de 4 años.

Instrumento

Escala de Voluntad de Trabajo

La escala administrada contenía 21 enunciados con formato de respuesta politómica de cuatro opciones tipo Likert (*casi nunca, a veces, con frecuencia, casi siempre*). Sin embargo, para el presente estudio sólo se consideraron nueve ítemes que reunieron criterios de calidad psicométrica basados en la TCT (Abal et al., 2006). El análisis de consistencia interna de la escala arrojó un Alpha de Cronbach de .73.

Procedimientos de recolección de datos

Se llevaron a cabo administraciones grupales durante el horario de clases que estuvie-

ron coordinadas por integrantes del equipo de investigación. Para motivar a los participantes en la realización de la tarea se efectuó previamente una charla en la que se les explicó la finalidad de la actividad y la futura utilización de los datos recogidos en una investigación. Los alumnos respondieron de forma voluntaria y no recibieron recompensa por su participación. A fin de favorecer la sinceridad en las respuestas y garantizar la confidencialidad de los datos, el inventario fue respondido de forma anónima.

ANÁLISIS DE DATOS

Para corroborar el supuesto de unidimensionalidad requerido por el MRG se efectuó un estudio exploratorio de la estructura factorial de los datos. Se empleó el método de componentes principales partiendo de un criterio que a priori supuso la unidimensionalidad del constructo analizado (Hair, Anderson, Tatham & Black, 1999). El supuesto de independencia local se dará por satisfecho si se comprueba la unidimensionalidad de la escala (Lord & Novick, 1968).

Todos los análisis basados en la TRI se realizaron operando el programa MULTILOG™ (Thissen, 1991). La bondad de ajuste del modelo se estudió ítem a ítem a partir de examinar los residuos de las proporciones observadas y las esperadas por el modelo para cada categoría de respuesta. La estimación de los parámetros del modelo se efectuó por máxima verosimilitud marginal. Asimismo, este programa también permitió obtener medidas de precisión basadas en la TRI tales como las Funciones de Información (FIs) de los ítemes y del test y el coeficiente de confiabilidad marginal.

RESULTADOS

COMPROBACIÓN DEL SUPUESTO DE UNIDIMENSIONALIDAD

El índice KMO de .80 y la prueba de esfericidad de Bartlett ($\chi^2 = 1060.21$, $p < .0001$) mostraron que la matriz de intercorrelaciones era adecuada para efectuar el análisis factorial de los datos. Mediante el método de componentes principales se obtuvo una solución con un único factor dominante que describe el 32% de la variancia. Los pesajes de los ítemes resultaron adecuados dado que oscilaron entre .45 y .67. Estos resultados corroboran la unidimensionalidad de los datos según los criterios propuestos por Reckase (1979) y Kaiser (1960).

AJUSTE DEL MODELO

Los residuos obtenidos para las cuatro categorías de cada ítem fueron iguales o inferiores a .01. Esto implica que, según la información suministrada por MULTILOG, el ajuste de los datos al modelo fue satisfactorio para todos los reactivos.

ESTIMACIÓN DE PARÁMETROS

La Tabla 1 exhibe los resultados obtenidos en la estimación de los parámetros del modelo y el valor promedio de los mismos considerando los nueve reactivos. Para cada ítem se estimaron un parámetro de discriminación (a) y tres parámetros de localización (b_1, b_2 y b_3). Los reactivos de la escala VT fueron respondidos en una escala Likert de cuatro categorías, por lo que el parámetro b_1 debe ser entendido como el mínimo valor de rasgo necesario para tener una probabilidad mayor de .50 de contestar la opción *a veces*. Por lo tanto ese valor es el umbral que separa las categorías *casi nunca* y *a veces*. De igual modo, b_2 separa las categorías *a veces* y *con frecuencia* y b_3 diferencia las opciones *con frecuencia* y *casi siempre*. Cuando el ítem es de sentido opuesto se invierten las opciones para definir los umbrales.

A modo ilustrativo, la Figura 1 exhibe las CCRI del Ítem 9 (*Doy muchas vueltas antes de ponerme a trabajar*) cuyo comportamiento resultó óptimo. La discriminación de las categorías de respuesta fue elevada y sus parámetros de localización se ubicaron en los va-

lores medios - bajos del rasgo. A su vez, la distancia entre los umbrales fue adecuada como para discriminar distintos niveles del rasgo con cada una de las categorías. Es decir, los parámetros de localización no se encontraron próximos entre sí como para suponer que alguna de las opciones es innecesaria.

En contraposición, el Ítem 4 (*Prefiero un menor esfuerzo a un mejor resultado*) mostró un funcionamiento deficiente. En la Tabla 1 se observa que su parámetro de discriminación adoptó un valor moderado-bajo y sus parámetros b_1 y b_2 están por fuera del rango razonable (entre -3 y 3) para el rasgo latente. Considerando que este ítem se puntúa en sentido inverso, el valor del parámetro b_2 indica que un individuo con un nivel extremadamente bajo de VT ($\theta = -4.57$) tendría una probabilidad superior a .5 de elegir la opción *a veces* o *casi nunca*. Incluso a un valor relativamente bajo de $\theta = -1.55$ comienza a ser más probable la opción *casi nunca*. La Figura 2 muestra la representación de las CCRI de este ítem. Allí puede apreciarse que la categoría *casi nunca* es la más probable en gran parte del espectro medio y alto de la variable.

Medidas de precisión

La Tabla 2 exhibe los valores que alcanzan las FIs de los ítemes y del test para algunos niveles de VT. La mayoría de los ítemes presentó una FI con valores relativamente bajos y uniformes a lo largo de todos los niveles del rasgo. Son excepciones los ítemes 5, 7 y 9 por tener FIs más elevadas y su mayor información la aportan en los niveles medio - bajos del rasgo. No obstante, ninguno de los ítemes analizados maximiza la información en los niveles altos de VT. Los ítemes 4 y 6 son los que presentaron las curvas más planas y los niveles de información más bajos; esto último es explicado por tener los parámetros de discriminación más pequeños.

La Figura 3 muestra que la FI del test resultó ligeramente decreciente y alcanzó su valor máximo en $\theta = -2$. Como es de esperar dada su relación inversa, el error de medida alcanzó en este punto su valor mínimo de .46. Esta función muestra que el test proporciona estimaciones de VT más precisas en los niveles de rasgo *theta* de -2 a 0, mientras que el error en la medición crece hacia los niveles más altos de VT.

El índice de confiabilidad marginal obtenido a partir de la aplicación del MRG fue de .75. Este indicador arroja información útil considerando la relativa uniformidad que presentó la FI del test.

Discusión

Una parte sustancial del proceso de construcción de un instrumento de evaluación psicológica se basa en la recolección de evidencias de validez y confiabilidad que permitan garantizar la calidad de la medida. Los desarrollos generados desde la perspectiva de la TRI permiten comprobar el adecuado funcionamiento de los ítemes contribuyendo a la fase de validez interna - estructural con información específica (Simms, 2008).

La aplicación de la TRI a la modelización de variables de personalidad es un área muy poco explorada incluso a nivel internacional. Así lo confirmaron revisiones actuales de la literatura psicométrica (Morizot, Ainsworth & Reise, 2007; Reise & Henson, 2003). Según Embretson y Reise (2000), si se observa el transcurso de la evolución de la TRI podría afirmarse que fue utilizada casi exclusivamente en tests de rendimiento, habilidades y aptitudes.

Reproduciendo la situación mundial, entre los escasos desarrollos efectuados en nuestro medio también primó la aplicación de la TRI en pruebas de ejecución máxima con modelos dicotómicos (e.g. Cortada de Kohan, 1998) y politómicos (e.g. Galibert, Aguerri, Pano, Lozzia & Attorresi, 2005). Sólo recientemente, Abal y colaboradores (2008) usaron el Modelo Logístico de Dos Parámetros para el análisis de ítemes de un test que mide altruismo. Por consiguiente, la presente aplicación del MRG a la Escala de VT es uno de los primeros aportes locales en el que se utiliza un modelo politómico para

el estudio psicométrico de un test de comportamiento típico.

Si bien los ítemes analizados habían superado un análisis de calidad desde la TCT, el Ítem 4 en particular no tuvo un comportamiento acorde a lo esperable en el estudio con TRI. Más allá del estudio del ajuste realizado sobre la base de los residuos provistos por MULTILOG, existen evidencias que hacen sospechar de un inadecuado funcionamiento. Como aseguraron Gray-Little, Williams y Hancock (1997), la estimación de parámetros con valores poco razonables y con errores comparativamente elevados constituye indicadores indirectos de que el ajuste del modelo a los datos no es satisfactorio. Por ende, será necesario eliminar este ítem o modificar su formulación a fin de que las estimaciones de los umbrales sean plausibles.

Este resultado también deja en evidencia el problema metodológico que acarrea la TRI en temas de la evaluación del ajuste. Ostini y Nering (2005) reseñaron una serie de aproximaciones matemáticas basadas en distintos métodos de medición de ajuste que incluyen gráficos y estadísticos. Sin embargo, la relativa novedad de los modelos politómicos genera que la mayoría de los *software*s del mercado aún no hayan sido lo suficientemente probados como para garantizar una aplicación segura. En cambio, programas como MULTILOG (Thissen, 1991) o PARSCALE (Muraki & Bock, 1997) para los que sí son conocidas las condiciones óptimas de aplicación, suelen ofrecer indicadores de ajuste pobres. Como aseguraron Revuelta, Abad y Ponsoda (2006), este tema es uno de los aspectos de la TRI que necesita una mayor investigación.

En relación con los parámetros de discriminación, Reise y Waller (1990) especificaron que suelen oscilar entre .5 y 1.5 para tests de comportamiento típico. Estos valores son más amplios de los que podrían exigirse para ítemes de habilidad dado que se reconoce el mayor grado de ambigüedad que pueden presentar los reactivos de personalidad. Por lo tanto, si se toman en consideración estos criterios es posible afirmar que los parámetros *a* obtenidos para los reactivos de VT resultaron con una capacidad discriminatoria entre moderada y alta.

Como se pudo observar, los errores de estimación de los parámetros de localización b_1 de todos los ítemes fueron superiores a los de los otros valores de umbrales. La explicación de este resultado se encuentra en la marcada asimetría negativa de la distribución de las observaciones en la escala Likert. Las categorías que denotan menor VT fueron poco elegidas por los participantes, lo que deriva en una deficiente estimación de los parámetros b_1.

Una de las falencias de la Escala VT descubierta por los resultados del presente estudio es la ausencia de ítemes que discriminen en los niveles de rasgo más elevado. Los umbrales de las categorías más altas (b_3) presentaron valores que, en promedio, están a .26 desvíos por encima de la media de VT. Esto significa que la escala carece de reactivos que permitan medir una importante parte del espectro superior del rasgo.

El mismo problema ha sido detectado al analizar la FI del test. Estas medidas locales de precisión mostraron que la escala es menos fiable para medir los niveles más altos de VT. Es decir, el error de medida aumenta cuando se pretende discriminar a los individuos muy voluntariosos. La disminución de la precisión de la medida se explica justamente por la ausencia de ítemes que discriminen en los niveles más altos de la variable. Aun así, la fiabilidad del instrumento es aceptable si se toman los indicadores globales. En este sentido, la consistencia interna de .73 obtenida desde la TCT en una investigación anterior (Abal et al., 2006) concuerda con el coeficiente de confiabilidad marginal del presente estudio.

La ausencia de ítemes que discriminen los niveles más altos de VT encuentra una explicación posible en la depuración efectuada previamente desde la TCT. La TRI necesita que los reactivos discriminen en todo el espectro del constructo incluidos los extremos. En cambio, desde la teoría clásica se requiere que los ítemes discriminen en torno a un valor central del rasgo medido para potenciar la representatividad de los indicadores globales

que utiliza. Los criterios de selección de los reactivos responden a este requerimiento, descartando aquellos cuya contribución no sea elevada con respecto a ese valor medio. En el caso de la Escala de VT, este valor medio aparece corrido hacia valores bajos y medios del constructo tal como lo demuestra la FI. De aquí la importancia de utilizar medidas locales de precisión.

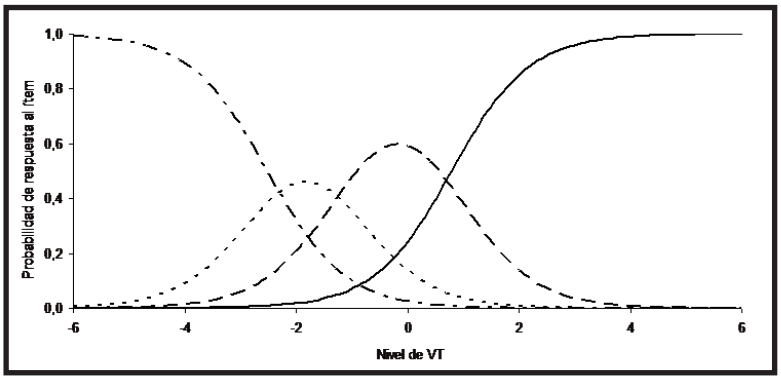
En un intento por compensar la deficiencia de ítemes que discriminen en los niveles altos del rasgo se consideró la posibilidad de analizar con la TRI los 21 elementos que componían la escala antes de la depuración clásica. Pero esta versión administrada incluyó ocho ítemes redundantes con el objetivo de elegir la formulación que mejor funcionara en términos de las propiedades psicométricas de la escala. Es muy frecuente encontrar que los cuestionarios que evalúan rasgos de la personalidad contengan varias formulaciones ligeramente diferentes (o completamente opuestas) del mismo ítem. Si bien la TCT es más permisiva con respecto a la inclusión de estos ítemes redundantes para aumentar la consistencia interna de la escala, desde la perspectiva de la TRI se encuentra un fuerte límite con el supuesto de independencia local. Este requerimiento de los modelos de la TRI exige que, para cada nivel del rasgo latente, las respuestas a distintos ítemes deben ser estadísticamente independientes. En consecuencia, se tornó inviable reanalizar con TRI los 21 ítemes administrados originalmente.

En virtud de lo expuesto, se impone la necesidad de aumentar la cantidad de ítemes de la Escala VT para optimizar la calidad del instrumento. En particular, los esfuerzos deben concentrarse en identificar indicadores del constructo que permitan detectar con mayor precisión a los niveles más altos del rasgo. En la medición de una habilidad puede resultar relativamente más fácil predecir el nivel de dificultad que tendrá un ítem y así anticipar su capacidad para discriminar en un nivel particular del rasgo latente. Sin embargo, la experiencia de Reise y sus colegas mostró que esta tarea es más compleja al construir ítemes para inventarios de personalidad (Flannery et al., 1995; Morizot et al., 2007; Reise & Waller, 1990). Asimismo, se debe procurar que los nuevos enunciados generados no resulten redundantes, lo cual también es complejo si se tiene en cuenta que sólo existe un conjunto finito de indicadores representativos del constructo (Reise & Henson, 2003). Esto es, no existen tantas maneras distintas de preguntarle a una persona si se percibe como voluntariosa sin caer en formulaciones equivalentes. Al mismo tiempo, la incorporación de otros indicadores que permitan ampliar la cantidad de ítemes puede traer aparejada una pérdida de la unidimensionalidad obtenida con la presente versión de la escala, o en su defecto, una redefinición de la variable.

Además de incorporar nuevos ítemes, futuras investigaciones ensayarán el ajuste de los datos con otros modelos politómicos. El hecho de alcanzar predicciones similares a partir de formulaciones distintas constituye una estrategia factible para obtener una poderosa evidencia de validez de constructo.

Los resultados de esta aplicación del MRG al análisis de ítemes de la Escala de VT no deben considerarse como concluyentes. En líneas generales, han servido para orientar las modificaciones que requiere la escala. La TRI proporciona un estatus de rigurosidad y objetividad a la medición de constructos psicológicos imposible de alcanzar desde la perspectiva clásica. Las poderosas herramientas de análisis de ítemes que provee inauguran un campo de investigación psicométrica con una gran cantidad de desafíos.

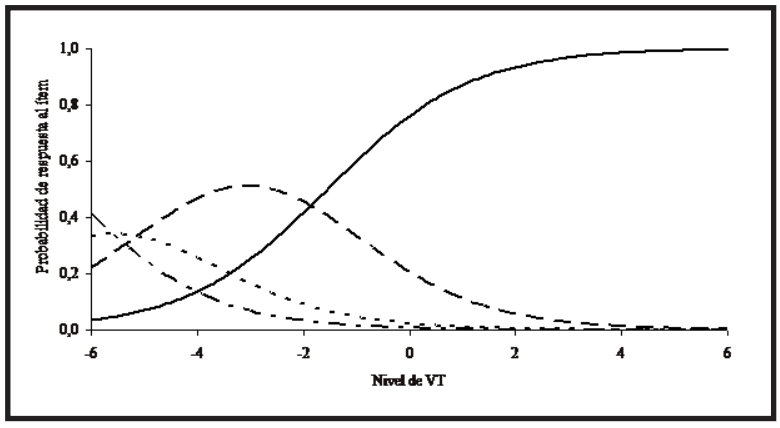
FIGURA 1
CURVAS CARACTERÍSTICAS DE LAS CATEGORÍAS DE RESPUESTA DEL ÍTEM 9



Notación

- - - Casi siempre
- - - - Con frecuencia
— — A veces
——— Casi nunca

FIGURA 2
CURVAS CARACTERÍSTICAS DE LAS CATEGORÍAS DE RESPUESTA DEL ÍTEM 4



Notación

- - - Casi siempre
- - - - Con frecuencia
— — A veces
——— Casi nunca

FIGURA 3
FUNCIÓN DE INFORMACIÓN DEL TEST Y ERROR ESTÁNDAR

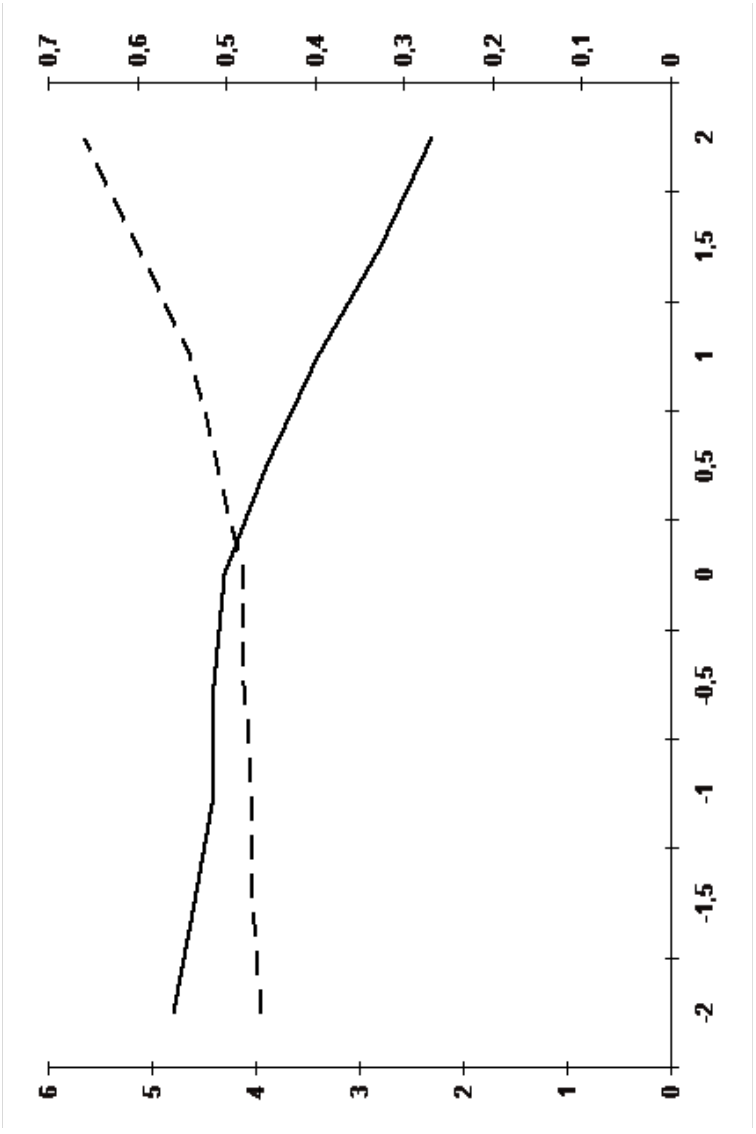


TABLA 1
PARÁMETROS DE DISCRIMINACIÓN Y DE LOCALIZACIÓN CON SUS ERRORES DE ESTIMACIÓN

Item	*a(Se)*	b_1*(Se)*	b_2*(Se)*	b_3*(Se)*
It1	1.12 (.10)	-4.27 (.40)	-2.27 (.19)	1.48 (.13)
It2	1.10 (.09)	-3.18 (.27)	-1.64 (.14)	0.72 (.09)
It3	1.04 (.10)	-3.73 (.34)	-2.34 (.20)	-0.80 (.10)
It4	0.75 (.10)	-6.47 (.92)	-4.57 (.58)	-1.55 (.21)
It5	1.60 (.11)	-2.95 (.21)	-1.81 (.11)	-0.06 (.06)
It6	0.73 (.08)	-2.50 (.28)	-1.12 (.16)	0.97 (.15)
It7	1.73 (.12)	-3.18 (.26)	-2.02 (.12)	-0.23 (.06)
It8	1.01 (.08)	-3.32 (.28)	-0.65 (.10)	1.03 (.11)
It9	1.44 (.10)	-2.52 (.17)	-1.13 (.09)	0.79 (.08)
Promedio (*DE*)	1.17 (.35)	-3.57 (1.22)	-1.95 (1.13)	.26 (.99)

TABLA 2
FUNCIONES DE INFORMACIÓN DE LOS ÍTEMES Y DEL TEST

Ítem / Nivel de rasgo θ	-2	-1.5	-1	-.5	0	.5	1	1.5	2
It1	**.33**	.29	.24	.22	.23	.26	.31	.32	.29
It2	**.36**	.35	.33	.32	.33	.33	.31	.26	.19
It3	**.33**	**.33**	.31	.28	.24	.18	.13	.08	.05
It4	**.16**	.15	.14	.12	.10	.08	.06	.05	.03
It5	**.75**	.70	.66	.67	.66	.53	.34	.18	.09
It6	.16	**.17**	**.17**	**.17**	.16	.16	.15	.14	.12
It7	**.84**	.76	.72	.78	.73	.52	.29	.14	.06
It8	.27	.27	.29	**.30**	**.30**	**.30**	.29	.25	.21
It9	**.60**	**.60**	.59	.55	.54	.55	.52	.41	.27
Test	**4.8**	4.6	4.4	4.4	4.3	3.9	3.4	2.8	2.3
SE	**.46**	.47	.47	.48	.48	.51	.54	.60	.66

Nota
Se destacan en negrita los valores máximos de la función de información.

REFERENCIAS BIBLIOGRÁFICAS

Abal, F., Lozzia, G. & Galibert, M.S. (2008). Aplicación del modelo logístico de dos parámetros en una escala de altruismo [Application of the two-parameter logistic model to the item analysis of an altruism scale]. *Memorias de las XV Jornadas de Investigación y Cuarto Encuentro de Investigadores en Psicología del Mercosur, 2,* 453-454.

Abal, F., Lozzia, G., Aguerri, M.E., Galibert, M. S. & Attorresi, H. (2007). Construcción de una escala de voluntad de trabajo [Construction of a will to work scale]. *Investigaciones en Psicología, 12,* 7-16.

Abal, F., Lozzia, G., Galibert, M.S. & Aguerri, M.E. (2006). Delimitación del constructo para elaborar una escala unidimensional de voluntad de trabajo según los supuestos del modelo lineal clásico [Demarcation of the construct for building an unidimensional scale of Will-to-Work according to the assumptions of Classical Lineal Model]. *Memorias de las XIII Jornadas de Investigación y Segundo Encuentro de Investigadores en Psicología del Mercosur, 3*, 19-21.

Asún, R. & Zúñiga, C. (2008). Ventajas de los modelos politómicos de la teoría de respuesta al ítem en la medición de actitudes sociales. El análisis de un caso [Advantages of polytomous models of item response theory in measuring social attitudes. A case study]. *Psykhe, 17*, 103-115.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. En F.M. Lord & M.R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-424). Reading, MA: Addison Wesley.

Cortada de Kohan, N. (1998). La teoría de respuesta al ítem y su aplicación al "Test Verbal Buenos Aires" [The item response theory and application in the "Verbal Test Buenos Aires"]. *Interdisciplinaria, 15*(1-2), 101-129.

Elosua, P. (2003). Sobre la validez de los tests [About test validity]. *Psicothema, 15*(2), 315-321.

Embretson, S.E. & Reise, S. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum Publishers.

Flannery, W.P., Reise, S.P. & Widaman, K.F. (1995). An item response theory of the general and academic scales of the Self-Description Questionnaire II. *Journal of Research in Personality, 29*, 168-188.

Galibert, M.S., Aguerri, M.E., Pano, C., Lozzia, G. & Attorresi, H. (2005). Análisis de ítem de analogías verbales mediante la aplicación de un modelo politómico de la teoría de respuesta al ítem [Analysis of items on verbal analogies through the use of a polytomic model of the theory on the response to items]. *Revista Mexicana de Psicología, 22*, 419-431.

Gray-Little, B., Williams, V.S.L. & Hancock, T.D. (1997). An item response theory analysis of the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin, 23*, 443-451.

Hair, J.F., Anderson, R.E., Tatham, R.L. & Black, W.C. (1999). *Análisis multivariante* [Multivariate analysis]. Madrid: Prentice Hall.

Hambleton, R. & Swaminathan, H. (1985). *Item response theory. Principles and applications.* Boston: Kluwer Nijhoff Publishing.

Kaiser, H.F. (1960) The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141-151.

Lane, S., Stone, C.A., Ankenmann, R.D. & Liu, M. (1995). Examination of assumptions and properties of graded item response model: An example using a mathematics performance assessment. *Applied Measurement in Education, 8*, 313-340.

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

Lord, F.M. & Novick, M.R. (1968). *Statistical theories of mental tests scores*. Reading, MA: Addison-Wesley.

Lozzia, G., Abal, F., Aguerri, M.E., Galibert, M.S. & Attorresi, H. (2007). Delimitación del constructo voluntad de trabajo [Demarcation of the will-to-work construct]. *Summa Psicológica UST, 4*, 137-148.

Morizot, J., Ainsworth, A.T. & Reise, S.P. (2007). Toward modern psychometrics. Application of item response theory models in personality research. En R.W. Robins, R.C. Fraley & R.F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 407-423). New York: Guilford Press.

Muñiz, J. (1997). *Introducción a la teoría de respuesta a los ítems* [Introduction to item response theory]. Madrid: Pirámide.

Muraki, E. & Bock, R.D. (1997). *PARSCALE. IRT item analysis and test scoring for rating-scale data.* Chicago: Scientific Software.

Ostini, R. & Nering, M.L. (2005). *Polytomous item response theory models.* Newbury Park, CA: Sage.

Reckase, M.D. (1979). Unifactor latent trait models applied to multifactor test: Results and implications. *Journal of Educational Statistics, 4*, 207-230.

Reise, S.P. & Henson, J.M. (2003). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment, 81*, 93-103.

Reise, S.P. & Waller, N.G. (1990). Fitting the two-parameter model to personality data. *Applied Psychological Measurement, 14*, 45-58.

Revuelta, J., Abad, F.J. & Ponsoda, V. (2006). *Modelos politómicos de respuesta al ítem* [Polytomous item response theory models]. Madrid: La Muralla.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph, 17*.

Samejima, F. (1972). A general model por free-response data. *Psychometrika, Monograph Supplement, 18.*

Samejima, F. (1997). Graded response model. En W.J. Van der Linden. & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85-100). New York: Springer.

Simms, L.J. (2008). Classical and modern methods of psychological scale construction. *Social and Personality Psychology Compass, 2,* 414-433.

Thissen, D. (1991). *MULTILOG™. User's guide. Multiple, categorical item analysis and test scoring using item response theory*. Chicago: Scientific Software.

Instituto de Investigaciones
Facultad de Psicología,
Universidad de Buenos Aires (UBA)
Ciudad Autónoma de Buenos Aires
República Argentina

Fecha de recepción: 26 de mayo de 2010
Fecha de aceptación: 10 de agosto de 2010