

Terapia psicológica  
Sociedad Chilena de Psicología Clínica  
sochpscl@entelchile.net  
ISSN (Versión impresa): 0716-6184  
CHILE

2007

René Gempp Fuentealba / Sergio Chesta Saffirio  
ERRORES ESTÁNDAR DE MEDIDA CONDICIONALES PARA LAS NORMAS  
METROPOLITANAS DE LA ADAPTACIÓN CHILENA DEL EPQ-R: APLICACIÓN DE  
UN MODELO BINOMIAL A UN TEST DE PERSONALIDAD

*Terapia psicológica*, junio, año/vol. 25, número 001

Sociedad Chilena de Psicología Clínica

Santiago, Chile

pp. 51-62

# Errores Estándar De Medida Condicionales para las Normas Metropolitanas de la Adaptación Chilena Del EPQ-R: Aplicación de un Modelo Binomial a un Test de Personalidad

## Conditional Standards of the Error of Measurement for the Chilean Norms of the EPQ-R: Applying a Binomial Model to a Personality Test

René Gempp Fuentealba

Escuela de Psicología, Universidad Alberto Hurtado, Chile

Sergio Chesta Saffirio

Escuela de Psicología, Universidad Mayor, Chile

(Rec: 17 de Abril 2007 Acep: 17 de Mayo 2007)

### Resumen

El Error Estándar de Medida (*EEM*) es un índice de la precisión de la puntuación obtenida por una persona en un test. El *EEM*, sin embargo, no es constante a través de todo el rango de puntuaciones. Estudios teóricos y empíricos indican que el *EEM* es más pequeño para las puntuaciones cercanas al extremo de la escala y mayor alrededor del centro de la escala. Un valor de *EEM* que aplica a un nivel de puntuación específica es denominado *Error Estándar de Medida Condicional* ( $EEM_{COND}$ ). Este trabajo utiliza un Modelo Beta Binomial (Lord, 1964, 1965) para estimar los  $EEM_{COND}$  de las normas chilenas del EPQ-R (N=1666), desarrolladas por Kaplan y Liberman (1992). Se concluye que los  $EEM_{COND}$  proveen información psicométrica más útil sobre las escalas del EPQ-R que el *EEM* tradicional. Adicionalmente, los resultados muestran la validez del Modelo Beta Binomial como enfoque psicométrico para analizar un test de personalidad. Se entregan recomendaciones prácticas para el uso de los  $EEM_{COND}$ .

*Palabras Clave:* Error Estándar de Medida Condicional, Error Estándar de Medida, Modelo Beta Binomial, EPQ-R, Personalidad.

### Abstract

The Standard Error of Measurement (*SEM*) is an index of precision of an examinee's test score. The *SEM*, however, is not constant throughout the full range of scale scores. Both theoretical and empirical studies indicate that the *SEM* is smaller for scores near the extremes of the scoring scale and larger near the middle of the scale. A value of the *SEM* that applies to a specific score level is referred to as *Conditional Standard Error of Measurement* ( $SEM_{COND}$ ). This paper uses a Beta Binomial Model (Lord, 1964, 1965) to estimate the  $SEM_{COND}$  for the Chilean norms of the EPQ-R (N=1666), developed by Kaplan and Liberman (1992). It is concluded that the  $SEM_{COND}$  provides more psychometric insight about the EPQ-R scales than the traditional *SEM*. In addition, the results show the validity of the Beta Binomial Model as a psychometric framework for analyzing a personality test. Recommendations are given for the practical use of  $SEM_{COND}$ .

*Keyword:* Conditional Standard Error of Measurement, Standard Error of Measurement, Beta Binomial Model, EPQ-R, Personality.

El propósito de un test psicológico es proveer evidencia replicable (i.e. "fiable") que permita formular inferencias relevantes y fundadas (i.e. "válidas") sobre las personas que lo responden. Estas inferencias se obtienen a partir de una muestra objetiva de la conducta de esas personas. De este modo, un test permite utilizar las respuestas a un conjunto reducido de ítems (una muestra de conducta) para hacer inferencias sobre el universo posible de conductas de un

individuo, en un campo determinado de su funcionamiento psicológico (e.g. personalidad, inteligencia, salud mental, entre otras). Como el resultado de un test es sólo una muestra de los posibles resultados que una persona podría obtener, tiene asociado un margen de imprecisión que en Psicometría es denominado "error de medida". En breve, el error de medida puede concebirse como la discrepancia entre el resultado de una evaluación particular y el promedio

de todos los resultados que una persona podría hipotéticamente obtener (Feldt y Brennan, 1989). Aunque este error es imposible de calcular para una persona específica, sí es factible estimar la desviación estándar de los errores de medida para un grupo de personas. Este último indicador es llamado *error típico de medida* o *Error Estándar de Medida (EEM)*. En la práctica, el *EEM* puede obtenerse a partir del coeficiente de fiabilidad (e.g. Alfa de Cronbach) y de la desviación típica [*DT*] de las puntuaciones observadas en un estudio normativo, a partir de la conocida ecuación:

$$EEM = SD\sqrt{1 - \text{fiabilidad}}$$

Esta ecuación puede encontrarse en la mayoría de los manuales de evaluación psicológica, mientras un tratamiento más formal de la fiabilidad y el *EEM* puede revisarse en textos de psicometría como el de Muñiz (2001). Un artículo publicado recientemente en esta misma Revista (Gempp, 2006) ofrece una aproximación didáctica al tema y explica el uso del *EEM* para estimar la Puntuación Verdadera. En el mismo trabajo se presenta un pequeño programa para simplificar estos cálculos, disponible gratuitamente en la dirección web <http://www.sigmas.cl/soft/winerror/>.

El uso del *EEM* para cuantificar la incertidumbre asociada a los resultados de un test es ampliamente recomendado en textos básicos de evaluación psicológica (e.g. Anastasi & Urbina, 1998) y promovido resueltamente por normativas técnicas como los *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA] & National Council on Measurement in Education [NCME], 1999), los *ETS Standards for Quality and Fairness* (Educational Testing Service [ETS], 2002) o las *International Guidelines for Test Use* (International Test Commission [ITC], 2000). Sin embargo, su aplicación ha estado restringida casi exclusivamente a las pruebas de medición educativa y, en menor proporción, a los tests de inteligencia y aptitudes, mientras que en evaluación clínica o de la personalidad es escasamente utilizado. De hecho, mientras muchos manuales de pruebas de rendimiento máximo reportan el *EEM* de sus resultados, esta práctica rara vez es imitada en la documentación de las pruebas de personalidad o psicopatología más populares. Esta evidencia advierte que en el área de la evaluación clínica y de la personalidad los usuarios están menos prevenidos de las limitaciones técnicas de los instrumentos que emplean.

Un problema conceptual con *EEM* convencional es que induce a suponer que el error de medida es constante para todos los niveles de puntuación en la prueba. No obstante, desde hace más de 50 años se descubrió que el *EEM* no es igual para distintas puntuaciones en el test, sino que varía para diferentes puntajes (Mollenkopf, 1949; Thorndike, 1951). Desde entonces se ha acumulado una gran cantidad de teoría y evidencia empírica demostrando que la distribución de los “*EEM Condicionales*” ( $EEM_{COND}$ ) a

cada nivel de puntaje tiene habitualmente la forma de una U invertida: las puntuaciones cercanas a la media tienen más error de medida que las puntuaciones extremas de la escala (e.g. Blixt & Shama, 1986; Feldt, Steffen & Gupta, 1985; Lord, 1955, 1957, 1959, 1984; Qualls-Payne, 1992). La consecuencia práctica de estos hallazgos es que las decisiones diagnósticas tomadas a partir de una puntuación cercana al promedio grupal contienen más error de medida que aquellas realizadas a partir de una puntuación muy alta o muy baja. Esta proposición es perfectamente congruente con la intuición de muchos usuarios de instrumentos psicométricos: entre más extremo (alto o bajo) sea un resultado, más fiable es.

Así como muchas fuentes especializadas sugieren el uso del *EEM* en la estandarización, corrección, interpretación e informe de los tests psicológicos, la necesidad de tomar en cuenta los  $EEM_{COND}$  también ha sido reconocida en varios textos avanzados y normativas técnicas sobre el uso de tests. Por ejemplo, los *Standards for Educational and Psychological Testing* recomiendan el cálculo y uso de los  $EEM_{COND}$  desde su primera edición (1954) y en todas las revisiones y actualizaciones posteriores (1955, 1966, 1974, 1985 y 1999). La última versión de los *Standards* dedica un apartado explícito al  $EEM_{COND}$ , declarando que:

#### Standard 2.14

*Los Errores Estándar de Medida Condicionales se deben reportar para varios niveles de puntuación (...). Cuando se especifiquen puntos de corte para selección o clasificación, los errores estándar de medida se deben reportar en la vecindad de cada punto de corte.*

*Comentario: La estimación de los Errores Estándar de Medida Condicionales es generalmente factible incluso con los tamaños muestrales que se utilizan habitualmente para los análisis de fiabilidad. Si se asume que el error estándar es constante para un rango amplio de niveles de puntuación, el fundamento de este supuesto debe ser presentado. (AERA, APA, NCME, 1999, p. 35)*

Tal como sucede con el *EEM* tradicional, la sugerencia de los *Standards* respecto a los  $EEM_{COND}$  ha tenido cierto impacto en la medición de rendimiento máximo, pero un efecto nulo entre quienes desarrollan, adaptan o emplean tests clínicos o de personalidad. Como prueba de ello, una búsqueda reciente en la literatura especializada arroja apenas dos estudios que examinan los errores de medida condicionales en pruebas de personalidad, específicamente en el MMPI (Saltstone, Skinner & Tremblay, 2001) y en el EPQ-R (Ferrando, 2003). ¿A qué puede atribuirse que el  $EEM_{COND}$  haya merecido tan escasa atención, pese a su innegable utilidad?

Sin pretender una respuesta definitiva, es probable que los múltiples problemas prácticos para calcular los  $EEM_{COND}$  sean los que hayan impedido su popularización en la comunidad de usuarios de tests psicológicos. Entre ellos, el primero y más concreto es que los  $EEM_{COND}$  deben calcularse para cada posible puntaje de la prueba, lo que supone contar con más datos que sólo la fiabilidad y la desviación estándar de la muestra normativa, como ocurría con el  $EEM$  convencional. A ello se agrega que la estimación de los  $EEM_{COND}$  no aparece como una opción disponible en las rutinas de análisis psicométrico de los paquetes estadísticos más populares (e.g. SAS, SPSS, STATISTICA) sino únicamente en programas especializados de análisis psicométrico, de uso frecuente en el campo de la medición educativa pero casi desconocidos para investigadores clínicos o de la personalidad. Un tercer problema es la falta de claridad sobre el método más apropiado para estimar los  $EEM_{COND}$ , lo que dificulta enormemente optar por una técnica concreta y más aún localizar y programar las ecuaciones y algoritmos necesarios para hacer los análisis.

Entre los métodos disponibles para estimar los  $EEM_{COND}$  se pueden distinguir dos grandes aproximaciones. Una alternativa bastante potente es utilizar análisis basados en Teoría de Respuesta al Ítem (TRI), aunque se debe pagar el costo de trabajar con modelos que necesitan muestras más grandes y supuestos más estrictos que los habituales en la investigación psicológica. Como consecuencia, muchos de los modelos de TRI rara vez ajustan a datos de pruebas tradicionales de personalidad (e.g. MMPI, CPI, NEO-FFI, EPQ, etc), lo cual hace que las estimaciones de  $EEM_{COND}$  obtenidas sean poco confiables. Otro problema práctico de la TRI es que requiere operar software y modelos de análisis altamente sofisticados, especialmente cuando los instrumentos tienen ítems politómicos (e.g. escalas tipo Likert), lo que nuevamente escapa a las expectativas y formación de los investigadores en psicología de la personalidad o clínica. Además, el error de medida condicional estimado con TRI no es exactamente equivalente al  $EEM_{COND}$  propiamente tal, lo que añade dificultades adicionales a la tarea<sup>1</sup>. Aunque recientemente han aparecido varios estudios que intentan aplicar modelos de TRI a pruebas de personalidad o psicopatología, el único trabajo dirigido específicamente a estimar errores de medida condicionales en una prueba de personalidad a partir de TRI es el estudio de Ferrando (2003), referido al EPQ-R. Este autor, sin embargo, inves-

tigó los errores de medida basados en TRI y no derivó los correspondientes  $EEM_{COND}$ .

La segunda aproximación reúne a varios procedimientos basados en la Teoría Clásica de los Test o sus derivaciones. En este marco se pueden diferenciar tres grupos de métodos. Los más tradicionales (Mollenkopf, 1949; Thorndike, 1951; Livingston, 1982) requieren muestras relativamente grandes y exigen supuestos poco realistas (e.g. paralelismo estricto entre medidas), además de estar diseñados originalmente para ítems dicotómicos (que son los habituales en pruebas de rendimiento máximo, pero no en tests de personalidad). Sin embargo, bajo ciertas condiciones entregan una buena aproximación a los  $EEM_{COND}$ . Otro grupo de métodos (Jajoura, 1986; Brennan, 1998) estiman el  $EEM_{COND}$  utilizando *Teoría de la Generalizabilidad*, lo que tiene el inconveniente de requerir un modelo de análisis escasamente conocido, que además es difícil de comprender y manejar apropiadamente fuera del ámbito de la medición educativa. Por último, el tercer grupo de métodos se basa en la aplicación de los modelos binomiales propuestos por Lord (1964, 1965) para el análisis de pruebas de rendimiento máximo. Aunque estos métodos funcionan sólo con ítems dicotómicos, ofrecen al menos dos grandes ventajas respecto a otras alternativas.

La primera es que se trata de modelos con supuestos empíricamente testeables, lo cual representa un enorme progreso respecto a los restantes modelos clásicos. Como se recordará, la TCT se construye sobre una serie de supuestos (ver Muñiz, 2001, para revisión técnica, o Gempp, 2006, para una presentación didáctica) que *no son comprobables*; simplemente se *asumen* sin que exista modo alguno de verificar si son correctos. Los modelos binomiales, en cambio, parten desde una serie de supuestos que, de ser correctos, permiten formular ciertas *hipótesis* sobre los datos empíricos. Concretamente, permiten predecir la forma que tendrá la distribución de frecuencias de las puntuaciones observadas en el test. Posteriormente estas hipótesis pueden contrastarse directamente con los datos, comparando la distribución de frecuencias *hipotetizada* por el modelo con la distribución de frecuencias *observada* en los datos. Si ambas exhiben un grado aceptable de concordancia se puede concluir que las hipótesis del modelo son plausibles en los datos y que, por tanto, los supuestos del modelo son aceptables para ese test en particular. Ello significa que el modelo psicométrico está “ajustado” a los datos y existe aval empírico para las inferencias psicométricas que se hagan del test<sup>2</sup>.

<sup>1</sup> El llamado error de medida en el contexto de la TRI es verdaderamente un error de estimación del rasgo latente que subyace al modelo. Para fines prácticos, esta estimación de error sólo es útil cuando las puntuaciones del test se calculan a partir de los patrones de respuesta a los ítems mediante algún procedimiento inherente a la TRI como la máxima verosimilitud o la estimación bayesiana. Cuando la puntuación en el test se calcula simplemente sumando las respuestas a los ítems (es decir, el método habitual), los  $EEM_{COND}$  son la estimación de error más apropiada.

<sup>2</sup> La capacidad de los modelos binomiales para comprobar empíricamente los supuestos en que se basan es la causa de que sean denominados genéricamente “Teoría Fuerte” (o Robusta) de la Puntuación Verdadera, en comparación con la TCT, que al no contar con mecanismos para comprobar sus propios supuestos es catalogada como “Teoría Débil” de la Puntuación Verdadera. Nótese que la incapacidad de la TCT para comprobar sus propios supuestos la convierte en una tautología y no en una Teoría científica propiamente tal.

La segunda ventaja, de tipo eminentemente práctico, es que los modelos binomiales requieren de poca información para estimar los  $EEM_{COND}$ : basta con conocer la distribución de frecuencias de las puntuaciones totales, la confiabilidad y el número de ítems del test, sin que sea necesario conocer el patrón de respuestas o información estadística de cada uno de los ítems. Esta característica facilita estimar los  $EEM_{COND}$  a partir de datos secundarios (e.g. los datos contenidos en las normas del test) sin ninguna necesidad de contar con los datos originales de los ítems. Esto permite, a su vez, que cualquier investigador o usuario pueda estimar errores de medida condicionales a partir de la información contenida en el manual del test o en datos extraídos desde algún estudio de estandarización.

Por las dos razones anteriores no es sorprendente que las propuestas teóricas más fructíferas para estimar los  $EEM_{COND}$  se desarrollaran en el marco de los modelos binomiales (e.g. Lord, 1955, 1984; Keats, 1957) y que en los trabajos comparados ésta sea la aproximación más recomendada (Feldt, et al., 1985; Qualls-Payne, 1992). Sin embargo la casi totalidad de los estudios realizados con este tipo de modelos se circunscriben a problemas de medición educativa o a pruebas de rendimiento máximo. Aparentemente, la única excepción es el trabajo de Saltstone et al. (2001) en que se utilizó un método binomial *simplificado* para estimar los  $EEM_{COND}$  de las escalas del MMPI.

El objetivo del presente trabajo es demostrar la aplicación de un tipo particular de modelo psicométrico de la familia de modelos binomiales, denominado *Modelo Beta Binomial Compuesto de 4 Parámetros* (Lord, 1964, 1965) para estimar los  $EEM_{COND}$  de las escalas del Eysenck Personality Questionnaire Revised Version [EPQ-R] adaptado para Chile por Bustos y Meneses (1991). Específicamente, este trabajo se propuso: (1) evaluar el ajuste de un *Modelo Beta Binomial Compuesto* a las escalas del EPQ-R; (2) estimar los  $EEM_{COND}$  para cada una de las escalas, y (3) comparar los  $EEM_{COND}$  estimados con el  $EEM$  tradicional. Todos los análisis del estudio se basan en las normas del EPQ-R para la Región Metropolitana de Chile, producidas por Kaplan y Liberman (1992), que se encuentran todavía en uso. La decisión de utilizar datos secundarios, así como la justificación de los objetivos específicos del estudio ameritan algunas explicaciones adicionales.

Respecto a la relevancia del primer objetivo específico, una revisión de las investigaciones en el área indica que, a la fecha, no se han realizado estudios que apliquen modelos binomiales a pruebas clínicas o de personalidad. Esto se explica, en parte, porque los modelos binomiales originales funcionan mejor cuando los ítems tienen una dificultad homogénea y las puntuaciones del test tienen una distribución aproximadamente normal o, al menos, simétrica. Estos requisitos son, obviamente, irreales para pruebas de psicopatología o personalidad, en que las distribuciones altamente asimétricas son la norma (ver Micceri, 1989, para

una revisión empírica de la “no normalidad” de la mayoría de las distribuciones de datos en Psicología). Sin embargo, modelos binomiales más complejos y flexibles, como el *Modelo Beta Binomial Compuesto* (Lord, 1964, 1965) permiten relajar muchos de estos supuestos y deberían ser capaces de ajustarse sin problemas a test de personalidad. El primer objetivo de este estudio apunta en esa dirección y pretende explorar empíricamente si este tipo de modelo psicométrico funciona apropiadamente con datos de un test de personalidad.

En línea con el razonamiento anterior, el segundo objetivo específico se justifica en la medida que pretende aportar resultados empíricos sobre los  $EEM_{COND}$  de un cuestionario de personalidad ampliamente utilizado, como el EPQ-R, respecto del cual se carece actualmente de esa información. El tercer objetivo surge de la necesidad de poner en perspectiva los  $EEM_{COND}$  y comprobar si la información psicométrica que aportan para un test de personalidad es lo suficientemente valiosa en comparación con el  $EEM$  tradicional, como para justificar el esfuerzo técnico que significa su cálculo.

Por último, consideramos preferible utilizar datos secundarios provenientes del estudio de estandarización del EPQ-R en la Región Metropolitana, en lugar de recoger una nueva muestra, fundamentalmente por dos razones. La primera es de tipo pragmático: al trabajar con los datos de Kaplan y Liberman (1992), los presentes resultados permiten *complementar* las tablas normativas vigentes para la Región Metropolitana, lo que esperamos contribuya a un uso más informado de ellas. La segunda razón es que la finalidad última que anima este trabajo es mostrar que resulta perfectamente factible estimar los  $EEM_{COND}$  aun cuando se carezca de información sobre las respuestas a cada uno de los ítems. En este sentido, tal como se demostrará enseguida, los datos disponibles en las tablas normativas son suficientes para hacer las estimaciones necesarias.

## Método

### *Instrumento*

El Eysenck Personality Questionnaire es un test de personalidad ampliamente conocido y validado a través del mundo (Barrett, Petrides, Eysenck & Eysenck, 1998). La versión revisada (EPQ-R) fue adaptada para Chile por Bustos y Meneses (1991), mediante la traducción de la mayoría de los ítems y la sustitución de otros. El cuestionario adaptado está compuesto por 100 ítems presentados como preguntas (e.g. “¿Cuando usted sale, le gusta encontrarse con gente conocida?”) que deben responderse utilizando dos alternativas (“Sí” y “No”). La puntuación de cada ítem es, por tanto, dicotómica.

Los ítems se agrupan en cuatro escalas, de las cuales tres corresponden a las dimensiones básicas de la personalidad propuestas en la teoría del autor: *Neuroticismo* (24 ítems), *Extraversión* (23 ítems) y *Psicoticismo* (32 ítems). En general, la *Extraversión* [E] es la menos clínica de las dimensiones y se concibe como un continuo de inhibición – excitación temperamental, que se expresa conductualmente en comportamientos, emociones y cogniciones de tipo introvertido, en un polo, y extravertido, en el otro polo. El *Neuroticismo* [N] es entendido como una predisposición a la irritabilidad e inestabilidad emocional, cuyo polo opuesto es la estabilidad emocional. Puntuaciones altas en esta escala se interpretan como evidencia de malestar psicológico (distress) y sintomatología neurótica clásica. La dimensión de *Psicoticismo* [P] tiene una definición menos nítida en la medida que involucra rasgos tradicionalmente atribuidos a las personalidades esquizoides y psicopáticas, además de alteraciones conductuales. Aquellos individuos con puntuaciones elevadas en esta escala se caracterizan por un patrón comportamental y afectivo muy cercano al concepto clásico de psicopatía.

A estas tres dimensiones el EQP-R agrega una escala de *Veracidad*, compuesta por 21 ítems que evalúan deseabilidad social y manejo de impresión, cuya finalidad es aportar un criterio cualitativo a la interpretación de los resultados observados en las tres dimensiones básicas. Debido a su nombre en inglés (Lie) comúnmente se la denomina escala L.

#### *Datos utilizados en el estudio*

Los datos utilizados en este trabajo provienen directamente de la Tesis de Licenciatura de Kaplan y Liberman (1992). Estos autores estandarizaron el EPQ-R en la Región Metropolitana de Chile utilizando una muestra no probabilística por cuotas, compuesta por 1666 adultos de ambos sexos (798 hombres y 868 mujeres), de más de 20 años de edad, pertenecientes a distintos niveles educacionales y socioeconómicos. Más detalles de la muestra pueden consultarse en el estudio de estandarización.

A partir de esta muestra, Kaplan y Liberman (1992) produjeron varios juegos de normas, en puntuaciones T y Percentiles, diferenciados por sexo, grupo de edad y nivel educacional, todas las cuales son presentadas en el Anexo X de su Tesis. En el presente trabajo se utilizan únicamente los baremos en Puntuaciones T para hombres y mujeres, para las cuatro escalas del EPQ-R adaptado, que corresponden a las ocho tablas que se encuentran al inicio del Anexo X. Debido a los fines del estudio no se consideró necesario analizar las normas para distinto grupo de edad y nivel educacional. Además, hacerlo supondría subdividir excesivamente la muestra.

Una característica particularmente atractiva de los resultados de Kaplan y Liberman (1992) es que son muy

detallados e incluyen, además de la equivalencia entre la Puntuación Bruta, Puntuación T y percentil correspondiente, la distribución de frecuencias absolutas y relativas de las puntuaciones totales de cada escala. La disponibilidad de las distribuciones de frecuencia originales evita inferirlas desde las normas, aumentando la precisión de los análisis. Desde el punto de vista de los análisis que se realizan en este artículo, la distribución de frecuencias permite calcular la media y desviación estándar para cada escala y contar con un criterio para evaluar el ajuste del modelo.

Además, se obtuvieron las medias y desviaciones típicas de cada escala, diferenciadas por género, desde la *Tabla 15* de la Tesis (Kaplan & Liberman, 1992, p. 75) y se compararon con las calculadas a partir de la distribución de frecuencias, obteniéndose un resultado consistente<sup>3</sup>. Las medias y desviaciones típicas para hombres y mujeres en cada escala son presentadas en la *Tabla 1*.

El último dato necesario para los presentes análisis es una estimación de la fiabilidad por consistencia interna para cada escala, diferenciada por género, que fue extraída desde la *Tabla 6* de Kaplan y Liberman (1992, p. 63). Los coeficientes alfa de Cronbach obtenidos son presentados en la *Tabla 1*.

#### *Modelo de análisis*

Los análisis se llevaron a cabo utilizando el *Modelo Beta Binomial Compuesto de 4 Parámetros* (Lord, 1964, 1965), una versión generalizada de los modelos binomiales. La descripción técnica del modelo, sus supuestos y derivación, excede con creces a los objetivos de este trabajo, pero afortunadamente no es imprescindible para comprender su funcionamiento. Básicamente los modelos binomiales asumen que la Puntuación Verdadera del test, expresada como proporción de respuestas correctas (en el rango 0 a 1), tiene una distribución de tipo Beta. Ésta es una distribución de probabilidad para variables continuas cuya forma es descrita por dos parámetros denominados Alfa y Beta. Simultáneamente se asume que para todos los evaluados con una misma Puntuación Verdadera la distribución de sus respectivas Puntuaciones Observadas puede describirse por una distribución de tipo binomial cuyos parámetros corresponderán al número de ítems del test y a la Puntuación Verdadera expresada como proporción. Al combinar ambas distribuciones (beta y binomial) el modelo predice que la distribución de las Puntuaciones Observadas en el test tendrá la forma de una distribución conocida como hipergeométrica negativa. Trabajos clásicos de Keats y Lord (1962) y de Lord y Novick (1968) demostraron empírica-

<sup>3</sup> Sin embargo, los valores calculados a partir de la distribución de frecuencias y los presentados en la *Tabla 15* de Kaplan y Liberman (1992, p. 75) no coinciden con los reportados en el Anexo X de la misma Tesis. Suponemos que podría tratarse de un error tipográfico de esa sección del informe.

mente que esta distribución de probabilidad es bastante flexible y útil para describir apropiadamente la distribución de Puntuaciones Observadas de la mayoría de los tests.

Existen varias versiones del modelo binomial básico recién comentado. La versión *Beta Binomial Compuesta de 4 Parámetros* utilizada en este trabajo agrega dos parámetros adicionales a la distribución hipotética de las Puntuaciones Verdaderas del test, para acotar los límites inferior y superior de la distribución. De esta manera se logra aumentar la flexibilidad del modelo y mejorar el ajuste a la distribución de Puntuaciones Observadas. Como es fácil deducir, el nombre del modelo se debe al número de parámetros que emplea (Alfa, Beta, Límite Inferior y Límite Superior) para describir la distribución Beta de Puntuaciones Verdaderas, aunque en realidad el modelo completo tiene seis parámetros en total. Los dos restantes corresponden al número de ítems del test (que no necesita ser estimado) y al parámetro  $k$  (ver Lord, 1964, 1965) que depende directamente de la fiabilidad del test. En este trabajo, la estimación del modelo se realizó mediante el método desarrollado por Hanson (1991). Para derivar los  $EEM_{COND}$  en la misma métrica en que están expresadas las normas (Puntuaciones T) se utilizaron las ecuaciones propuestas por Kolen, Hanson & Brennan (1992). Como insumos básicos para el análisis se emplearon la distribución de Puntuaciones Observadas, el número de ítems y la fiabilidad del test, para cada escala, en las muestras de hombres y mujeres, respectivamente.

## Resultados

El análisis comenzó estimando los parámetros del modelo *Beta Binomial Compuesto de 4 Parámetros* y evaluando su ajuste empírico a las distribuciones de frecuencia de cada una de las cuatro escalas del EPQ-R, en las muestras de hombres y de mujeres. Para ello se comparó la distribución predicha por el modelo y la distribución real de puntuaciones observadas. En la Figura 1 se puede observar que en todas las escalas la distribución de frecuencias real (línea delgada segmentada) es muy consistente con la distribución de frecuencias ajustada (línea gruesa continua), lo que indica un buen ajuste del modelo. Como criterio estadístico de evaluación del ajuste se compararon las distribuciones con una prueba de Chi Cuadrado. Los resultados, que se presentan en las Tablas 2 y 3 para la submuestras de cada género, confirman que no hay diferencia significativa ( $p > 0.05$ ) entre las distribuciones empírica y ajustada en todas las escalas, con la excepción de *Psicoticismo* en la muestra de mujeres, donde se observa una discrepancia significativa ( $p = 0.03$ ). No obstante, en términos descriptivos y considerando la razón *Chi Cuadrado/g.l.* el ajuste puede considerarse satisfactorio (Burnham & Anderson, 2002). En las Tablas 2 y 3 también se presentan los parámetros estimados para el modelo. Es interesante notar que en varios casos el límite inferior

de la distribución fue igual cero, lo que hizo innecesario ajustar los cuatro parámetros del modelo, ocasionando un incremento en los grados de libertad del modelo.

También puede observarse que las distintas escalas muestran distribuciones muy diferentes entre sí. Mientras *Veracidad* y *Extraversión* exhiben distribuciones relativamente simétricas, *Psicoticismo* presenta una distribución positivamente asimétrica con un rango de variación muy restringido respecto al número máximo de ítems de la escala (35 ítems). *Neuroticismo*, por su parte, muestra una distribución de puntajes claramente diferente para los hombres que para las mujeres. Para complementar la inspección visual de las distribuciones, en la Tabla 4 se presentan los momentos de la distribución de puntajes normativos expresados en Puntuaciones T, para la muestra de hombres y de mujeres<sup>4</sup>.

Una vez satisfecho el primer objetivo específico del estudio, se procedió a realizar los análisis necesarios para cumplir con los dos objetivos restantes. Concretamente, se estimaron los  $EEM$  tradicionales tanto para los puntajes brutos como para los puntajes normativos en métrica T. Los resultados se presentan en la Tabla 5. Además, se estimaron los  $EEM_{COND}$  para cada puntaje T, en cada una de las escalas. La distribución de los  $EEM_{COND}$  es graficada en la Figura 2 (distribución en forma de U invertida) y se la compara con el  $EEM$  tradicional estimado para los puntajes T (línea continua).

Puede apreciarse cómo el  $EEM$  tradicional, al ser constante para todos los puntajes, no refleja apropiadamente el nivel de error de cada escala. Por ejemplo, en el caso de *Veracidad* el error condicional es casi equivalente al  $EEM$  en el centro de la escala, pero para las puntuaciones bajo  $T=45$  o sobre  $T=55$  es dramáticamente más bajo.

En *Psicoticismo*, por su parte, el  $EEM$  proporciona una estimación demasiado conservadora del error de medida (3.22 y 3.18 puntos), mientras los  $EEM_{COND}$  advierten que el error de medida en torno al centro de la escala supera los 4 puntos en métrica T y es sistemáticamente superior al  $EEM$  para casi todo el rango de las puntuaciones.

En *Neuroticismo* podemos observar que el error de medida condicional es realmente superior al  $EEM$  estimado en la zona central de la escala, pero decrece significativamente para puntuaciones menores a  $T=42$  o mayores a  $T=58$ .

Otros efectos interesantes son algunas asimetrías entre géneros. Por ejemplo, en *Extraversión* tanto el  $EEM$

<sup>4</sup> Puede observarse que, a diferencia de lo que cabría teóricamente esperar para las Puntuaciones T, la media y desviación típica no corresponden a 50 y 10 puntos, respectivamente. Aparentemente las normas están construidas centrando la distribución en las medias reportadas en el Anexo X de la Tesis que, sin embargo, discrepan de las medias empíricas de la distribución y de las exhibidas en otra sección del mismo estudio (ver Nota 3). Ignoramos a qué se debe esta decisión que, en todo caso, no afecta los resultados del presente trabajo toda vez que la transformación entre las puntuaciones y puntuaciones T es perfectamente lineal ( $r = 1$ ) en todos los casos.

Figura 1: Distribuciones de frecuencia empíricas y estimadas para cada una de las escalas del EPQ-R, por género.

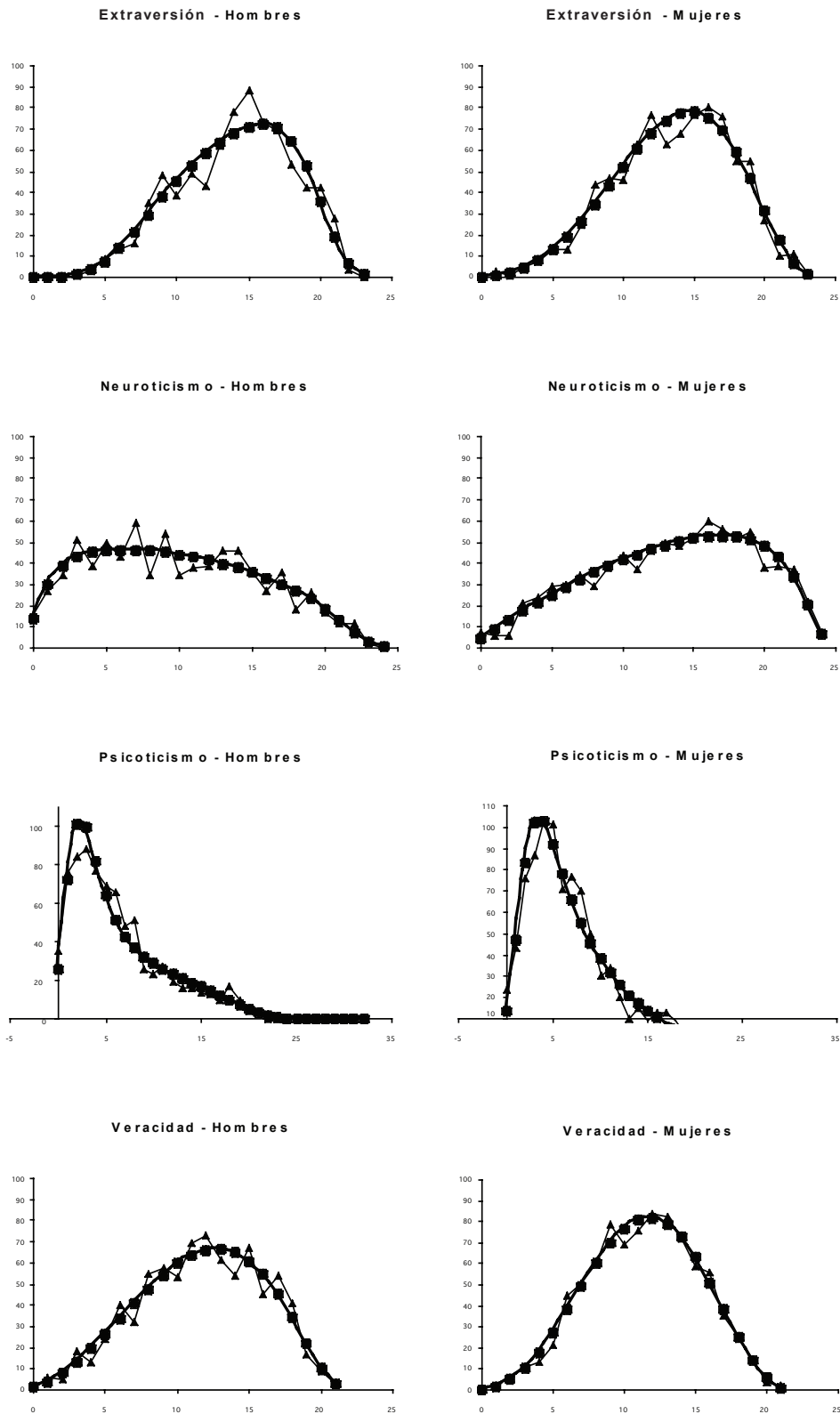


Figura 2: Errores Estándar de Medida Condicionales versus tradicionales, estimados para las normas en Puntuaciones T de cada escala del EPQ-R, en ambos géneros.

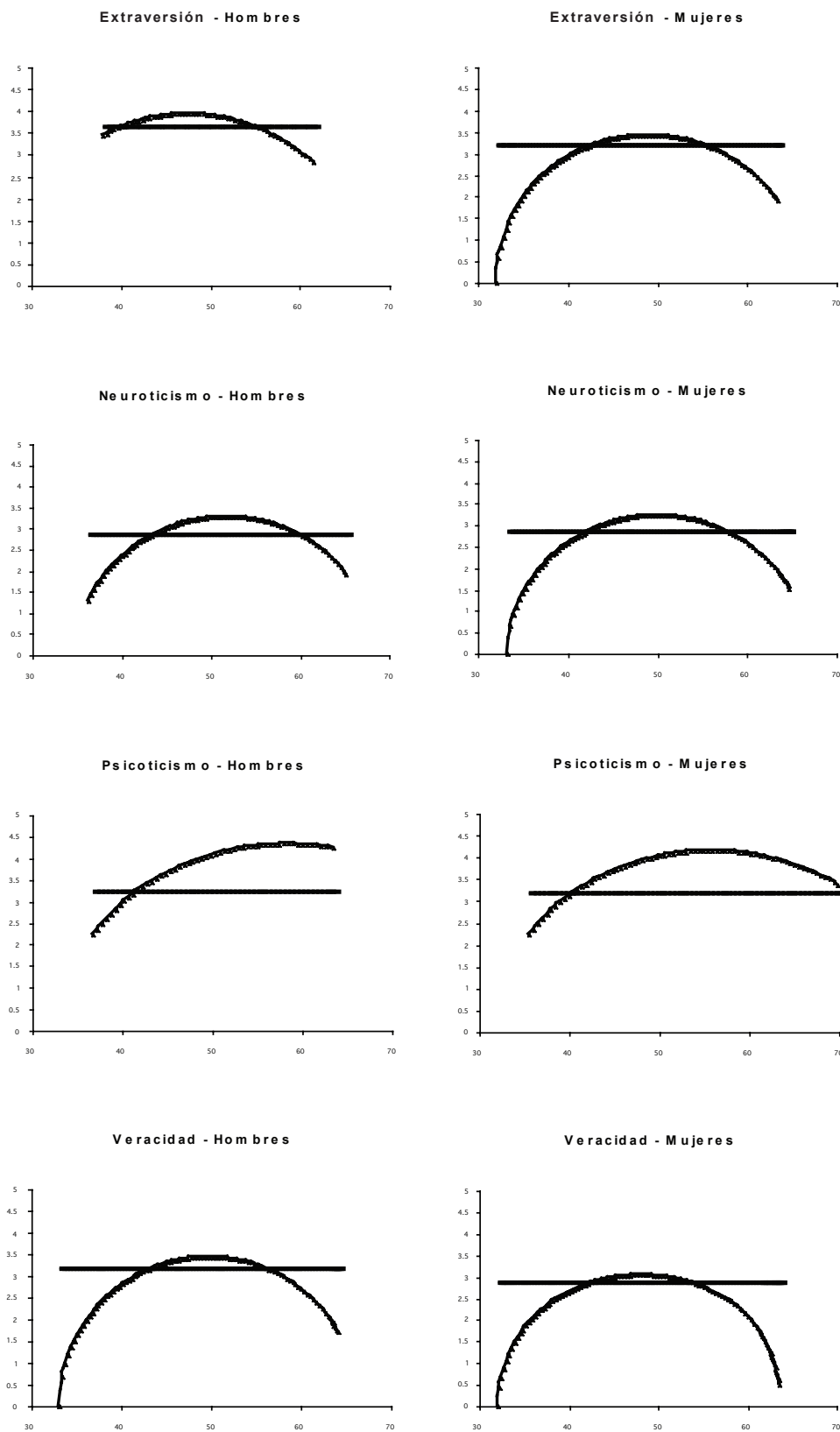


Tabla 1: Número de ítems, media, desviación típica y alfa de Cronbach para cada escala del EPQ-R en la muestra de estandarización de Kaplan y Liberman (1992)

Escala	N. Ítems	Hombres (N=798)			Mujeres (N=868)		
		Media	DT	Alfa	Media	DT	Alfa
Extraversión	23	14.14	4.01	.72	13.58	4.16	.74
Neuroticismo	24	9.90	5.75	.87	13.58	5.71	.87
Psicoticismo	32	6.37	5.05	.83	6.44	4.31	.76
Veracidad	21	11.69	4.30	.78	11.36	3.87	.78

Tabla 2: Parámetros del Modelo Beta Binomial Compuesto y prueba de ajuste para la muestra de hombres (N=798)

Escala	k	Alfa	Beta	Lim. Inf.	Lim. Sup.	Chi2	g.l.
Extraversión	1.17	1.77	1.22	0.26	0.86	28.43	19
Neuroticismo	1.00	1.20	1.56	0.03	0.90	27.69	20
Psicoticismo	0.94	0.43	1.31	0.06	0.61	36.37	28
Veracidad	0.99	3.24	2.21	0.00	0.93	21.61	18

Tabla 3: Parámetros del Modelo Beta Binomial Compuesto y prueba de ajuste para la muestra de mujeres (N=868)

Escala	k	Alfa	Beta	Lim. Inf.	Lim. Sup.	Chi2	g.l.
Extraversión	1.17	4.51	2.51	0.00	0.92	25.92	20
Neuroticismo	1.18	1.99	1.33	0.00	0.94	15.88	21
Psicoticismo	0.85	0.81	3.74	0.07	0.80	43.39	28
Veracidad	3.05	4.41	3.61	0.00	0.98	10.84	18

Tabla 4: Media, desviación típica, asimetría y curtosis de la distribución de normas en Puntuaciones T para la muestra de hombres y de mujeres

Escala	Hombres (N=798)				Mujeres (N=868)			
	Media	DT	Asim.	Curt.	Media	DT	Asim.	Curt.
Extraversión	52.10	6.90	-0.28	2.41	52.22	6.28	-0.30	2.61
Neuroticismo	48.82	7.99	0.22	2.09	52.24	7.95	-0.29	2.19
Psicoticismo	43.14	7.82	1.02	3.32	41.44	6.49	1.07	3.98
Veracidad	51.65	6.79	-0.21	2.40	49.55	6.12	-0.11	2.50

Tabla 5: Error Estándar de Medida tradicional, estimado para las puntuaciones brutas y para las puntuaciones T, en la muestra de hombres y de mujeres

Escala	Hombres (N=798)		Mujeres (N=868)	
	EEMbruto	EEMT	EEMbruto	EEMT
Extraversión	2.12	3.65	2.12	3.20
Neuroticismo	2.07	2.88	2.06	2.87
Psicoticismo	2.08	3.22	2.11	3.18
Veracidad	2.02	3.18	1.82	2.87

convencional como los  $EEM_{COND}$  son más altos para los hombres que para las mujeres, indicando que la escala tiene un desempeño psicométrico diferencial por género. Concretamente en estas normas el EPQ-R entrega una medida de *Extraversión* más precisa para las mujeres que para los hombres.

### Discusión

Este trabajo se propuso tres objetivos: (1) evaluar el ajuste de un *Modelo Beta Binomial Compuesto* a las escalas del EPQ-R adaptado en Chile, (2) estimar los  $EEM_{COND}$  de cada escala y (3) comparar el análisis del error medida basado en el  $EEM$  tradicional con el análisis que se desprende de los  $EEM_{COND}$ , para evaluar si estos últimos hacen un aporte significativo que justifique su cálculo.

Respecto al primer objetivo específico los resultados demuestran claramente que el modelo psicométrico utilizado fue capaz de ajustar apropiadamente los datos de las cuatro escalas (N, E, P y L) del EPQ-R. Este hallazgo por sí solo es muy promisorio considerando que hasta la fecha no existen estudios publicados que documenten la factibilidad de usar modelos binomiales en test de personalidad. Además, se observa (ver Figura 1) que el *Modelo Beta Binomial Compuesto* fue capaz de adaptarse a distribuciones muy disímiles; incluso a aquellas con niveles extremos de asimetría (*Psicoticismo*). Este resultado sugiere la conveniencia de seguir explorando la aplicación de modelos psicométricos de la familia binomial en test de personalidad, toda vez que el cálculo de  $EEM_{COND}$  es apenas una de las muchas ventajas que ofrecen sobre el modelo clásico (ver Lord, 1964, 1965, para una discusión de algunas aplicaciones).

Al mismo tiempo, el resultado observado en *Psicoticismo*, típico de las escalas que evalúan rasgos psicopatológicos (la mayoría de los casos se concentran bajo la media), anticipa un buen funcionamiento de estos modelos en escalas clínicas o sintomatológicas, aconsejando la realización de futuros estudios en esta área.

Una limitación para emprender esta tarea es que los modelos binomiales han sido diseñados originalmente para ítems dicotómicos, mientras en evaluación clínica y de la personalidad prima el uso de ítems con formato graduado. Hay dos soluciones al problema. Una es dicotomizar los ítems, dado que en muchos casos no implicará una pérdida significativa de la calidad métrica de la escala (ver López, 2005, para un ejemplo elocuente en el caso de una escala de depresión). Además, desde un punto de vista psicométrico, la puntuación total calculada como la suma de las respuestas a los ítems es siempre un estimador eficiente del rasgo subyacente en el caso de ítems dicotómicos, mientras que tratándose de ítems graduados la sumatoria de los ítems no siempre representa apropiadamente al rasgo latente medido (van der Ark, 2005). Otra alternativa es recurrir a

extensiones recientes de los modelos binomiales diseñados explícitamente para ítems politómicos (Lee, 2001). Aunque es legítimo dudar del valor práctico de modelos tan sofisticados frente a las técnicas tradicionales, una buena razón para recomendar el uso de modelos binomiales es que sirven de puente conceptual y práctico entre el modelo clásico de análisis y construcción de tests y los desarrollos más modernos en el área, íntimamente ligados a la TRI. Dado el rápido avance de estos últimos modelos, es probable que en pocos años los tests convencionales que no incorporen los nuevos avances en psicometría caigan rápidamente en la obsolescencia.

En relación con el segundo y tercer objetivo del estudio, los resultados muestran que los  $EEM_{COND}$  permiten hacer inferencias más precisas sobre la calidad métrica de las escalas que el mero uso del  $EEM$  convencional.

En el caso de las escalas sin un componente clínico como *Extraversión* y *Veracidad*, la simple inspección de la Figura 2 revela que el  $EEM$  subestima levemente el error de medida en el centro de la escala y sobreestima groseramente el error en los puntajes altos o bajos (i.e. 10 puntos de distancia a  $T=50$ ). ¿Qué consecuencias prácticas tienen estos  $EEM_{COND}$  sobre las decisiones diagnósticas tomadas a partir de estas escalas? Como anticipamos en la introducción, los  $EEM_{COND}$  son más consistentes con las interpretaciones típicas que los usuarios formulan a partir de los tests. Concretamente, que las puntuaciones extremas son más fiables y entregan resultados más consistentes que las puntuaciones en torno al centro de la escala. Por ejemplo, una puntuación  $T=35$  en *Extraversión*, obtenida por una mujer, tiene un error de medida de apenas  $T=2.2$  puntos, y no de  $T=3.2$  puntos como indica el  $EEM$  convencional. Para puntuaciones aun más bajas en esta escala, el error de medida es prácticamente nulo, de manera que si concluimos que esa persona es introvertida, la certidumbre del resultado es altísima.

Esta típica distribución en forma U invertida de los  $EEM_{COND}$  tiene otras consecuencias prácticas sobre la interpretación de las puntuaciones de los tests. La primera concierne a los errores de clasificación, que surgen cuando se usan las puntuaciones de los tests para asignar a las personas evaluadas a categorías discretas, mutuamente excluyentes. Un ejemplo típico es el uso de puntos de corte en medidas de screening psicopatológico, para lo cual los investigadores intentan encontrar un valor de la escala que maximice simultáneamente la sensibilidad y especificidad de la clasificación diagnóstica. En este caso, la pertinencia de los  $EEM_{COND}$  es mencionada en el *Standard 2.14* previamente citado (AERA, APA, NCME, 1999, p. 35). Brevemente explicado, el problema es que la metodología habitual para identificar puntos de corte (Curvas ROC) no incorpora el dato del error de medida de la escala. Sin embargo, al tomar en cuenta esa información sabemos que mientras más lejano se encuentre el punto de corte del centro de la escala, menor error de medida tendrá, logrando una

clasificación más consistente, Si se considera la distribución de los  $EEM_{COND}$  junto a la sensibilidad y especificidad, es posible estimar el error de clasificación debido a errores de medida e incorporarlo formalmente al análisis y establecimiento de los puntos de corte (e.g. Hanson, 1991).

Otra implicancia concreta de la forma de U invertida que suele tener la distribución de los  $EEM_{COND}$  es arrojar una nueva perspectiva sobre la importancia del coeficiente de fiabilidad en la interpretación de los resultados del test. El coeficiente de fiabilidad es un indicador del grado en que los resultados de un test pueden generalizarse a los resultados que una persona podría obtener si fuera evaluada con una medida equivalente del mismo constructo. Tal como hemos planteado en un trabajo anterior (Gempp, 2006), el coeficiente de fiabilidad es un medio para cuantificar el error de medida y no un fin en sí mismo. Si consideramos la distribución de los  $EEM_{COND}$ , debemos concluir que la fiabilidad *no es la misma* para todo el rango de puntuaciones del test y que las puntuaciones más extremas son más fiables que las puntuaciones cercanas a la media de puntuaciones. Ello significa, entre otras cosas, que un test con una fiabilidad moderada podría, no obstante, entregar decisiones altamente fiables cuando éstas se basen en puntuaciones extremadamente altas o bajas.

Por otro lado, en las escalas que evalúan características clínicas como *Psicoticismo*, la discrepancia entre el  $EEM$  y los  $EEM_{COND}$  es más dramática, mostrando que el primer indicador subestima sistemáticamente el error de medida de la escala y que sólo los  $EEM_{COND}$  son sensibles a la asimetría en la distribución de puntuaciones. Ello vuelve a alertar sobre la importancia de explorar el análisis de los errores condicionales de medida en escalas diseñadas para evaluar psicopatología<sup>5</sup>.

En cuanto a la aplicación práctica de los resultados obtenidos, anunciada en la introducción del artículo, éstos permiten complementar las normas elaboradas por Kaplan y Liberman (1992). Por ejemplo, si un varón obtiene 18 puntos en la escala de *Neuroticismo* las normas indican que le corresponde una puntuación  $T=60$ , y los presentes resultados añaden que esa puntuación tiene un error asociado de  $T=2,8$  puntos. Algunas guías para asistir la interpretación de los tests utilizando el error de medida han sido discutidas en Gempp (2006).

Para finalizar, creemos pertinente concluir este artículo con una breve reflexión acerca de la escasa disponibilidad de datos públicos sobre los tests adaptados y construidos en Chile y en Latinoamérica. Una mirada imparcial a la historia de la Psicología muestra que los tests psicológicos son uno de los desarrollos tecnológicos más exitosos de la disciplina. Además, los tests son empleados continuamente

en el trabajo académico y profesional de los psicólogos. Por ello, sorprende la ausencia de una base de datos organizada que compile y sistematice en forma periódica los instrumentos disponibles, sus fundamentos técnicos y baremos interpretativos, como mínimo. Creemos que la inexistencia de herramientas de este tipo es indirectamente responsable del uso poco informado y riguroso del instrumental psicométrico, con las múltiples consecuencias negativas que esto genera. A ello se agrega que los artículos publicados en revistas especializadas rara vez proporcionan suficiente información técnica para que los lectores puedan formar sus propias conclusiones o realizar estudios independientes sobre los resultados reportados, lo que evidentemente retrasa el desarrollo científico de la disciplina. Como ejemplo de estas dificultades, la elaboración de esta investigación debió dedicar muchos meses de trabajo sólo a la tarea de localizar datos normativos sobre el EPQ-R con la información necesaria para hacer las estimaciones (que, recordemos, se reduce simplemente a la distribución de frecuencias, medias, desviaciones típicas y fiabilidades de cada escala). Lamentablemente, muchos autores son reticentes a compartir sus datos y/o plantean abiertamente su aprensión de que un re-análisis detecte algún error en sus propios resultados.

En este sentido, creemos que propuestas como la de Wicherts, Borsboom, Kats & Molenaar (2006) referidas a que las revistas especializadas soliciten a los autores un respaldo electrónico de sus datos en la Web, es una iniciativa constructiva que bien podría resultar un aporte en el medio latinoamericano. Por último, sentimos un deber felicitar la honestidad intelectual de tantos autores de Tesis de Licenciatura quienes, como Kaplan & Liberman (1992), exponen abiertamente sus resultados para ser utilizados en la comunidad científica, contribuyendo así al desarrollo colectivo de la Psicología.

## Referencias

- American Educational Research Association, American Psychological Association & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Anastasi, A. & Urbina, S. (1998). *Test psicológicos*. México: Prentice-Hall.
- Barrett, P.T., Petrides, K.V., Eysenck, S.B.G. & Eysenck, H.J. (1998). The Eysenck Personality Questionnaire: an examination of the factorial similarity of P. E. N. and L. across 34 countries. *Personality and Individual Differences*, 25, 805-819.
- Blixt, S.L. & Shama, D.D. (1986). An empirical investigation of the standard error of measurement at different ability levels. *Educational and Psychological Measurement*, 46, 545-550.
- Brennan, R.L. (1998). Raw-score conditional standard errors of measurement in Generalizability Theory. *Applied Psychological Measurement*, 22(4), 307-331.
- Burnham, K.P. & Anderson, D.R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. New York: Springer.
- Bustos, M.T. & Meneses, C. (1991). *Adaptación del Eysenck Personality*

<sup>5</sup> No obstante, debe tomarse en cuenta que este resultado también es producto de que las puntuaciones T no estén centradas en la media empírica de la distribución (ver Nota 4).

- Questionnaire Revised Version a la población adulta del área metropolitana de Chile.* Tesis de Licenciatura en Psicología, no publicada, Escuela de Psicología, Universidad Diego Portales.
- Educational Testing Service. (2002). *ETS Standards for Quality and Fairness*. Princeton, NJ: Educational Testing Service.
- Feldt, L. S. & Brennan, R. L. (1989). Reliability. En: R. L. Linn (Ed.), *Educational Measurement* (3rd ed.) (pp. 105-146). New York: American Council on Education and MacMillan.
- Feldt, L. S., Steffen, M. & Gupta, N. C. (1985). A comparison of five methods for estimating the standard error of measurement at specific score levels. *Applied psychological measurement*, 9, 351-361.
- Ferrando, P.J. (2003). The accuracy of the E, N and P trait estimates: an empirical study using the EPQ-R. *Personality and Individual Differences*, 34, 665-679.
- Gempp, R. (2006). El Error Estándar de Medida y la Puntuación Verdadera de los tests psicológicos: Algunas recomendaciones prácticas. *Terapia Psicológica*, 24(2), 117-130.
- Hanson, B.A. (1991). *Method of Moments Estimates for the Four-Parameter Beta Compound Binomial Model and the Calculation of Classification Consistency Indexes*. ACT Research Report Series 91-5. Iowa City, IA: American College Testing Program.
- International Test Commission. (2000). *International Guidelines for Test use*. Stockholm: International Test Commission.
- Jarjoura, D. (1986). An estimator of examinee-level measurement error variance that considers test form difficulty adjustments. *Applied Psychological Measurement*, 10, 175-186.
- Kaplan, M. & Liberman, L. (1992). *Estandarización del Test EPQ-R (Eysenck Personality Questionnaire, Revised version) adaptación Bustos-Meneses 1991, a la población urbana adulta del área metropolitana de Chile*. Tesis de Licenciatura en Psicología, no publicada, Escuela de Psicología, Universidad Diego Portales.
- Keats, J.A. (1957). Estimation of error variances of test scores. *Psychometrika*, 22, 29-41.
- Keats, J.A. & Lord, F.M. (1962). A theoretical distribution for mental test scores. *Psychometrika*, 27(1), 59-72.
- Kolen, M., Hanson, B.A. & Brennan, R.L. (1992). Conditional standard errors of measurement for scale scores. *Journal of Educational Measurement*, 29, 285-307.
- Lee, W. (2001). *A multinomial error model for tests with polytomous items*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Seattle.
- Livingston, S.A. (1982). Estimation of the conditional standard error of measurement for stratified tests. *Journal of Educational Measurement*, 19, 135-138.
- López, J.A. (2005). Ítems politómicos vs. dicotómicos: Un estudio metodológico. *Anales de Psicología*, 21(2), 339-344.
- Lord, F.M. (1955). Estimating test reliability. *Educational and Psychological Measurement*, 15, 325-336.
- Lord, F.M. (1957). Do tests of the same length have the same standard error of measurement? *Educational and Psychological Measurement*, 17, 510-521.
- Lord, F.M. (1959). Tests of the same length do have the same standard error of measurement. *Educational and Psychological Measurement*, 19, 233-239.
- Lord, F.M. (1964). *A strong true score theory, with applications*. Educational Testing Service Research Bulletin 64-19. Princeton, NJ: Educational Testing Service.
- Lord, F.M. (1965). A strong true score theory, with applications. *Psychometrika*, 30, 239-270.
- Lord, F.M. (1984). Standard errors of measurement at different ability levels. *Journal of Educational Measurement*, 21, 239-243.
- Lord, F.M. & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison Wesley.
- Micceri, T. (1989). The unicorn, the normal curve and other improbable creatures. *Psychological Bulletin*, 105(1), 156-166.
- Mollenkopf, W.G. (1949). Variation of the standard error of measurement. *Psychometrika*, 14, 189-229.
- Muñiz, J. (2001). *Teoría Clásica de los Test*. Madrid: Pirámide.
- Qualls-Payne, A.L. (1992). A comparison of score level estimates of the standard error of measurement. *Journal of Educational Measurement*, 29(3), 213-225.
- Saltstone, R., Skinner, C. & Tremblay, P. (2001). Conditional standard error of measurement and personality scale scores: an investigation of classical test theory estimates with four MMPI scales. *Personality and Individual Differences*, 30, 691-698.
- Thorndike, R.L. (1951). Reliability. En E. F. Lindquist (Ed.), *Educational measurement* (pp. 560-620). Washington, DC: American Council on Education.
- van der Ark, A. (2005). Stochastic ordering of the latent trait by the sum score under various polytomous IRT models. *Psychometrika*, 70(2), 283-304.
- Wicherts, J.M., Borsboom, D., Kats, J. & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, 61, 726-728.