ARTÍCULOS

Volumen XXVIII N° 1
2009 – Segunda Época

Error de Medición

Amaia Urdanibia
aurdanibia@gmail.com

Escuela de Psicología, Universidad Central de Venezuela

Resumen

Dentro de la Psicología, así como en el resto de las ciencias, siempre ha existido la necesidad de contar con medidas precisas de los fenómenos estudiados. Sin embargo, la dificultad en esta área de estudio, radica en la complejidad de las variables investigadas. En este sentido, dado que no se puede contar con medidas exentas de error, es importante determinar y comunicar la información respecto al error de medición; para que la misma sea considerada en las decisiones y conclusiones resultantes de los procesos de evaluación. A continuación se hará un recorrido que inicia en la Teoría Clásica de los Test y que muestra al error de medición como una manera alternativa y necesaria, en la determinación de la precisión de las medidas de las pruebas.

Palabras clave: *Error de medición, Teoría Clásica de los Test, Fiabilidad*

Measurement Error

Abstract

In Psychology, like in any other science, there has always been the need to have a precise measure of the phenomenon being studied. Nevertheless, the difficulty in this area of research lies in the complexity of the studied variables. In this sense, given the fact that there is no error free measure, it is important to determine and communicate the information on regard of the measurement error; so that it can be used in the decisions and conclusions resulting from evaluation processes. In this article there will be a revision starting from Classical Test Theory and that will show the measurement error as an alternative yet necessary element, in the determination of the precision of test measures.

Key words: *Measurement error, Classical Test Theory, Reliability*

> "Cuando medimos algo, bien sea en el campo de la física, de la biología o de las ciencias sociales, esa medición contiene una cierta cantidad de error aleatorio. La cantidad de error puede ser grande o pequeña, pero está siempre presente en cierto grado" (Thorndike, 1951, pág. 560).

El interés por computar la precisión de las mediciones, al analizar e interpretar los resultados obtenidos de las pruebas psicológicas; radica en la necesidad de controlar y reducir los errores asociados a la medición de constructos psicológicos.

Considerando que en la psicología no puede eliminarse del todo dicho error; es relevante, cuantificar e informar a los usuarios de pruebas, acerca del error de medición; para que éstos lo tomen en cuenta en sus decisiones sobre los puntajes obtenidos por los evaluados. Ahora bien, ¿cómo estiman los psicólogos el grado de error que hay en sus mediciones?

Los índices de fiabilidad informan sobre la calidad de las calificaciones arrojadas por un instrumento. Estos índices permiten conocer cuan exenta de error se encuentra una medida; por ello, conociendo la fiabilidad, es posible calcular el error de medición y emplearlo para hacer estimaciones de los puntajes de los sujetos. Sin embargo, en la mayoría de las pruebas, no se hace un reporte claro del error de medida, e incluso usuarios inexpertos desconocen que la información sobre el error debe encontrarse junto al índice de fiabilidad.

De acuerdo a lo que señalan Anastasi y Urbina (1998) la fiabilidad se refiere a la consistencia de los puntajes obtenidos por las mismas personas cuando se les evalúa en distintas ocasiones con la misma prueba, con conjuntos equivalente de reactivos o en otras condiciones de exanimación. Este concepto en su acepción más amplia, se refiere a la medida en que las diferencias individuales en los resultados pueden atribuirse a "verdaderas diferencias" en la variable considerada y el grado en que pueden deberse a errores azarosos.

Para determinar si un instrumento es confiable es preciso calcular el índice de fiabilidad, el cual se obtiene de la correlación entre un conjunto de puntuaciones; la magnitud de la correlación muestra qué porcentaje de la varianza total se debe a la varianza verdadera, y la magnitud restante es producto de la varianza de error o fuentes de error que se pueden presentar en

la construcción, administración y calificación del instrumento. Lo expuesto anteriormente está basado en que el puntaje obtenido por un examinado en una prueba está formado por dos elementos: un puntaje verdadero y puntaje de error (Magnusson, 1975).

Tomando en consideración el planteamiento de Magnuson, es notable que la precisión de una prueba puede ser vista de dos maneras; a través de los índices de fiabilidad, y mediante el error de medición. Ahora bien, es importante aclarar que la conceptualización anterior se fundamenta en el Modelo Lineal propuesto por Spearman en 1910. Según Gempp (2006), el paradigma de la puntación verdadera, expresado mediante la Teoría Clásica de los Test (TCT), constituye el fundamento técnico de la gran mayoría de las pruebas psicológicas; y permite explicar los elementos necesarios para estimar la fiabilidad de un instrumento y los errores de medida.

La propuesta de Spearman plantea como primer supuesto que, la puntuación empírica (observada u obtenida) de un sujeto en una prueba (que denominaremos *X)*, consta de dos componentes, la puntuación que verdaderamente le corresponde en esa prueba (que designaremos *V)*, y un cierto margen de error *(e)* (Muñiz, 1998). Este principio puede expresarse a través de la siguiente ecuación:

$$\boldsymbol{X = V + e} \qquad (1)$$

Esta ecuación indica claramente que cualquier medición realizada contendrá error. Por lo tanto, el interés de la TCT es determinar en qué medida la puntación observada es puntación verdadera y en qué medida es error. Es importante destacar que el puntaje de error puede resultar de errores constantes o de eventos aleatorios; los primeros, al afectar a todas las observaciones por igual no influyen en las comparaciones, por lo que no suelen considerarse; sin embargo, este tipo de errores deben ser tomados en cuenta en el caso de la determinación de la validez. Los errores producto de procesos aleatorios afectan a las observaciones de manera distinta, pudiendo provenir de variadas fuentes y constituye lo que se conoce como error de medición; el cual informa acerca de la discrepancia que existe entre el puntaje obtenido y el puntaje verdadero de una persona, pudiendo afirmarse que suministra una información complementaria al coeficiente de confiabilidad.

En este sentido, el error de medición introduce, como señala Gempp (2006), la noción de replicabilidad; ya que equivaldría a cualquier diferencia

observada entre los resultados obtenidos en la administración repetida de un mismo instrumento. Sin embargo, luego de múltiples medidas con una misma prueba cómo se sabe ¿cuál de todos los puntajes del sujeto es el verdadero? Para resolver esta incógnita, Spearman añadió al modelo el supuesto de que la puntuación verdadera de una persona en una prueba sería la que obtendría como promedio si se le aplicase infinitas veces el instrumento (Muñiz, 1998).

De este modo, se está considerando que las discrepancias observadas en cada situación de medición fueron producidas por procesos aleatorios, no controlables que afectaron a los resultados de manera imprevista; en ocasiones produciendo fluctuaciones superiores al puntaje verdadero y otras veces inferiores. Al asumir este supuesto, se están asumiendo simultáneamente otros de los planteamientos de Spearman, que señalan que no existe relación entre la puntuación verdadera de las personas y los errores de medida (r_{ve}= 0), y los errores de medida de los tests no están relacionados ($r_{ej,ek}$= 0) (Muñiz, 1998). Esto quiere decir, que al considerar a las medidas repetidas como independientes entre sí, el resultado de una medición no afecta el resultado de la siguiente medición; lo que confirma que el error de medida se comporta de forma aleatoria.

Gempp (2006) plantea que si el resultado obtenido en cada momento, no influye ni es influido por las aplicaciones anteriores o posteriores, se puede asumir al error como una variable aleatoria continua y esperar que luego de varias mediciones en un mismo sujeto, tenga un promedio igual a cero. Igualmente, se puede definir a la puntuación verdadera como el resultado promedio de una serie "hipotética" de puntajes observados replicados infinitamente (Gempp, 2006), por lo cual, no necesariamente representa al valor real del constructo. Este mismo autor expresa que la ecuación principal de la TCT no asume que exista alguna variable subyacente a los resultados de las pruebas; lo que implica que el puntaje verdadero (*V*) no está relacionado con el concepto de validez. Lo que se plantea desde la TCT es que cualquier prueba tiene asociada una puntuación verdadera, independientemente de que los reactivos que la componen representen adecuadamente el constructo. En conclusión una prueba puede ser confiable sin ser válida.

Ahora bien, es imposible lograr en la práctica, que un evaluado pueda responder a una prueba infinitas veces, en igualdad de condiciones y sin que las aplicaciones previas afecten las siguientes aplicaciones. Por ello, Gempp (2006) señala que dentro de la TCT se derivan algunos corolarios que permiten extender el primer supuesto, a una situación en la que se analizan las respuestas de un grupo de evaluados, en lugar de un caso individual.

Uno de las extensiones de la teoría plantea que, si los errores son aleatorios y no correlacionan con el puntaje verdadero ni entre sí, la varianza de los puntajes observados a través de los evaluados correspondería a la suma de las varianzas verdadera y de error, como se expresa a través de la siguiente ecuación:

$$\boldsymbol{var\ (x) = var\ (v) + var\ (e)} \qquad (2)$$

Esto indica que no se analiza entonces la variabilidad intrasujeto sino que se analizan las variaciones de las puntuaciones interindividuales. Y que la ecuación antes descrita representa lo plasmado en el primer supuesto; pero en este caso, descrito en términos de varianza. Asimismo, Gempp (2006) expone que a esta ecuación subyace el mismo problema que a la ecuación inicial, sólo hay un valor conocido (la varianza del puntaje observado) y existen dos incógnitas (la varianza de los puntajes verdadero y de error); por lo que no hay forma de resolver la ecuación y determinar el valor de la varianza de error.

Al respecto, Muñiz (1998) plantea que a partir del desarrollo del modelo propuesto por la TCT, es posible llegar a fórmulas operativas para la estimación de los errores *(e)*, y por lo tanto de los puntajes verdaderos *(V)* de los sujetos. Esto conduce a la conceptualización de la confiabilidad como la razón entre la varianza de los puntajes verdaderos y la varianza de los puntajes observados; es decir, la confiabilidad estima cuánto de verdadero hay en las puntuaciones observadas.

$$r_{xx'} = \frac{\boldsymbol{var\ (v)}}{\boldsymbol{var\ (x)}} \qquad (3)$$

De esta ecuación se deduce que si la puntuación de cada evaluado en la distribución observada, es la misma que su puntuación verdadera, la varianza observada será la misma que la de los puntajes verdaderos, y la varianza de error será cero, por la cual la confiabilidad resultaría igual 1.

En este sentido, cuando la confiabilidad aumenta, la varianza de error disminuye, lo que equivale a decir que la puntuación observada en un evaluado, se aproxima a su puntuación verdadera (Gempp, 2006). Asimismo, si la varianza de error aumenta, los puntajes observados constituyen una mala estimación de los puntajes verdaderos. Cuando todo el puntaje observado de un sujeto es puntaje de error, la varianza de los puntajes verdaderos será igual a cero y la varianza de los puntajes de error será la máxima e igual a la

varianza observada, con lo cual la confiabilidad resultaría equivalente a cero. Estos planteamientos conllevan a que la confiabilidad también pueda expresarse como:

$$r_{xx'} = 1 - \frac{var\ (e)}{var\ (x)} \qquad (4)$$

Nuevamente, a medida que la varianza de los puntajes de error disminuya la confiabilidad aumentará y viceversa. Es importante destacar que, como señala Muñiz (1998), el cálculo empírico del valor del coeficiente de fiabilidad no puede llevarse a cabo mediante las fórmulas antes descritas, ya que son meramente conceptuales; la estimación empírica puede obtenerse a través de varias estrategias, dentro de las que destacan: a) la correlación entre dos formas paralelas de la prueba, b) la correlación entre dos mitades aleatorias de la prueba, que luego es corregida mediante la fórmula de Spearman-Brown, y c) la correlación entre dos aplicaciones de una misma prueba a una muestra de personas. Cada uno de estos procedimientos tiene ventajas y desventajas y se ajustan mejor a unas situaciones que a otras. De igual modo, en todas las situaciones, el valor obtenido es un valor numérico entre 0 y 1, indicando a medida que se acerca a 1 que la prueba está siendo precisa en la medición.

De acuerdo con Gempp (2006) algunas de las propiedades del coeficiente de confiabilidad, expresado en las dos ecuaciones anteriores son las siguientes:

- Se deriva de la ecuación planteada en el primer supuesto de la TCT, por lo cual no es estrictamente calculable a partir de los datos.
- Su máximo valor es 1, cuando toda la varianza observada es atribuible a la varianza verdadera.
- Su mínimo valor es 0, cando no existe varianza verdadera y toda la varianza observada es debida a la varianza de error.

Ahora bien, puede demostrarse que la confiabilidad también puede ser definida como el coeficiente de determinación de la correlación entre los puntajes verdaderos y observados (Gempp, 2006), si lograra controlarse artificialmente todas las fuentes de varianza de error; en otras palabras como la proporción de varianza de los puntajes verdaderos explicada por la varianza de los puntajes observados. Este valor puede obtenerse a través de la siguiente ecuación:

$$r_{xv} = \sqrt{r_{xx'}} \quad (5)$$

Se observa que el coeficiente de determinación, al ser el cociente entre la desviación típica verdadera y la observada, puede escribirse como la raíz cuadrada de la confiabilidad. Tal índice informa la medida en que los puntajes observados representan a los puntajes verdaderos. Cuando la confiabilidad de la prueba resulta baja también lo será su índice de confiabilidad y los puntajes verdaderos estarán pobremente representados por los puntajes observados; pero si la confiabilidad es igual a 1, la puntuación verdadera se puede predecir con exactitud desde las puntuaciones observadas.

Muñiz (1998) señala que para cualquier valor de confiabilidad, la correlación entre los puntajes observados y verdaderos siempre será mayor que esta; por ello dentro de la TCT se plantea que la puntación observada de cualquier prueba siempre correlacionará más alto con su puntuación verdadera que con el puntaje observado, de otras pruebas.

Como se mencionó anteriormente, el problema con las ecuaciones de confiabilidad presentadas es que no permiten estimar el índice de confiabilidad, dado que requieren conocer previamente, la varianza de los puntajes verdaderos o de error, los cuales son las cantidades que se desean estimar a partir de los puntajes observados. Gempp (2006) plantea al respecto, que es un error común creer, que la confiabilidad puede calcularse; ya que la TCT define a la confiabilidad de forma tautológica, nunca puede calcularse directamente sino únicamente estimarse. Otro malentendido, es la falsa creencia de que existen varios tipos de confiabilidad; mientras lo que en realidad sucede es que hay distintas estrategias empíricas para estimarla. De este modo, dado que la confiabilidad no puede calcularse directamente, los distintos métodos y coeficientes sólo constituyen aproximaciones al índice.

Por otra parte, tampoco es adecuado interpretar la confiabilidad como una propiedad intrínseca de los instrumentos; ya que la TCT plantea que ésta es una propiedad de los puntajes arrojados por las pruebas. Esta distinción implica que al ser una propiedad de las calificaciones y no de los instrumentos, la confiabilidad depende de las características de la población en la que fue aplicada la prueba y sólo es generalizable a sujetos que pertenezcan a esa población. En este sentido, Gempp (2006) explica que dada la dependencia muestral de los estadísticos de la teoría clásica, la consecuencia es que los coeficientes de confiabilidad obtenidos en cualquier estudio representan simplemente la confiabilidad de los puntajes de los evaluados que participaron en el estudio y no son representativas del

instrumento sino de los individuos. Aunado a esto, se tiene que las muestras en Psicología pocas veces son demasiado grandes y con menor frecuencia son probabilísticas; por lo cual, resulta atrevido asumir que la confiabilidad para una muestra, constituya una propiedad del instrumento, que se mantendrá invariable cada vez que éste sea aplicado. Por esta razón, es imprescindible que los investigadores y usuarios calculen la confiabilidad de las puntuaciones en sus propias muestras.

Tal y como se planteó al inicio de este material, una forma alternativa pero equivalente de expresar la confiabilidad de las pruebas es a través del Error Típico de Medida. No importa el índice de confiabilidad que se emplee, (en cada caso hay razones prácticas para utilizar uno u otro), lo fundamental es que toda medición lleva asociado un grado de precisión que es calculable empíricamente. Esto quiere decir que, aun cuando el coeficiente de confiabilidad indica la consistencia, estabilidad o equivalencia de los puntajes observados en una prueba, no informa directamente la cantidad de discrepancias (error) que pueden esperarse entre unas mediciones y otras; lo cual se obtiene a partir del error de medición, el cual señala el tamaño del error que se acepta cometer cuando se asume la puntuación observada como la verdadera, resultando un indicador preciso de cuán cierta es la estimación del puntaje verdadero.

Según Magnusson (1975), para conocer el tamaño del error de medición se puede partir de una de dos hipótesis:

- *Errores diferentes para diferentes puntajes verdaderos*: Este supuesto plantea que los errores se encuentran representados por los desvíos típicos de la distribución de los puntajes verdaderos de un solo individuo alrededor de su puntaje verdadero.
- *Errores iguales para diferentes puntajes verdaderos:* Los errores se encuentran expresados por la dispersión de los puntajes obtenidos en torno a una calificación verdadera fija (la misma para todos los test paralelos).

De acuerdo con Magnusson (1975), partiendo del primer supuesto, Lord derivó una ecuación para determinar la incertidumbre de un puntaje obtenido; y supuso que un test puede considerarse una muestra de ítems seleccionados al azar de una población, de la cual pueden derivarse infinitos test paralelos al azar. Asimismo, un sujeto puede resolver un número de ítems de la población de donde se escogió la muestra. La proporción que ellos representan para ese individuo *"j"* es la probabilidad de que los reactivos que puede resolver sean

incluidos en la prueba, lo cual se denota como p_j. De este modo, el puntaje verdadero a estimar es:

$$T_j = n(p_j) \quad (6)$$

Donde *n* es el número de reactivos de la muestra. Para estimar el tamaño del error cometido cuando se acepta que un puntaje obtenido representa al verdadero, se calcula la desviación estándar mediante la ecuación:

$$Se = \sqrt{\frac{t_j\,(n - t_j)}{(n - 1)}} \quad (7)$$

Esta ecuación da una estimación de la desviación estándar de la distribución de los puntajes que resultan de examinar a un sujeto con un número infinito de test paralelos al azar (Magnusson, 1975).

La magnitud del error de medición obtenido bajo esta hipótesis, es una función del número de ítems del test y del número de ítems que el sujeto puede resolver, lo cual tiene como consecuencia que la desviación típica (error) es independiente del contenido de los ítems, de la frecuencia de las soluciones correctas y de las intercorrelaciones de los reactivos. Esto significa que pruebas con el mismo número de ítems deberán tener un error estándar promedio, sin importar que sean test de diferente naturaleza.

De igual modo, el tamaño del error estándar de medida, estimado con el modelo propuesto por Lord, será máximo cuando el individuo resuelva bien la mitad de los ítems, aumentará al aumentar o disminuir p_j (hacia los extremos) y finalmente será cero cuando p_j sea 1 ó 0. Es decir, mientras más extremo sea el valor del puntaje verdadero más pequeño será este error. Cuando este índice sea empleado para establecer intervalos de confianza, se debe tener en cuenta que está basado en las suposiciones de que los test paralelos son extraídos de una población infinita de ítems paralelos, que el número de reactivos que un sujeto puede contestar varía de un test paralelo a otro, aún cuando provengan de la misma población y que no se toman en cuenta las fuentes de varianza de error. No debe calcularse el error de medida a partir de este modelo para tests con límites de tiempo, ya que se basa en pruebas paralelas extraídas de una población de reactivos paralelos (Magnusson, 1975).

Por otro lado, partiendo del segundo supuesto, si se somete a una persona a infinitas pruebas paralelas se tendrá una distribución de infinitas

puntuaciones obtenidas, que asumirá una forma normal con una media aritmética que representará la puntuación verdadera del sujeto y una desviación típica que representa al error de medición, el cual indicará la variabilidad esperada de las calificaciones obtenidas en torno a las puntuaciones verdaderas.

De acuerdo con Magnusson (1975), el tamaño del error de medición será el mismo para todas las pruebas paralelas y por ende, para todos los individuos que las presentan; será independiente de la magnitud del puntaje verdadero y es mayor mientras más varíen los puntajes obtenidos respecto a los verdaderos.

Ahora bien, como es imposible administrar infinitas pruebas paralelas para determinar la varianza y con ello el error de medición, lo que sucede en la práctica es la administración de la prueba o una paralela, dos veces a un grupo de sujetos y la asunción de que las diferencias de las calificaciones individuales entre las dos aplicaciones, son el resultado de puntajes de error y que la varianza generada por tal distribución es el error estándar de medida.

Magnusson (1975) indica que de la ecuación de la confiabilidad se puede derivar una ecuación para computar el tamaño del error. De este modo, en la TCT el error de medición se define como la desviación estándar de la distribución de los puntajes de error (Gempp, 2006), y se expresa de la siguiente manera:

$$\boldsymbol{Em = Sx\sqrt{1 - r_{xx}}}, \qquad (8)$$

En la ecuación, *Sx* representa el valor de la desviación típica de los puntajes observados para la misma muestra de sujetos en los cuales se estimó la confiabilidad. De este modo, cuando la confiabilidad es igual a cero, el error asumirá su máximo valor; lo que implica que toda la variabilidad observada será explicada por el error. Pero si la confiabilidad es uno, significa que el error de medición asume su mínimo valor y la varianza observada será explicada en su totalidad por la varianza verdadera.

De acuerdo con Gempp (2006) una propiedad importante del error de medición, es que se expresa en las mismas unidades de escala que las puntuaciones de la prueba, lo que facilita su interpretación directa. Sin embargo, resulta incorrecto hacer comparaciones entre los errores de pruebas cuyos puntajes se expresan en unidades diferentes. Si se requiriese comparar la precisión de diferentes instrumentos, se deberá recurrir al coeficiente de

confiabilidad, ya que independientemente del instrumento este coeficiente siempre se expresa en las mismas unidades.

Por otra parte, el error de medición se relaciona de forma particular con la longitud de la prueba y la homogeneidad de la muestra. En este sentido, el error se ve afectado por variaciones en la longitud de la prueba; cuando se aumenta la cantidad de reactivos que conforman un instrumento, se incrementa la probabilidad de que los errores aleatorios se cancelen entre sí, reduciéndose con ello el error de medición. Respecto a la homogeneidad, el error resulta independiente de la variabilidad de la muestra para el cual se calculó, pues como es una característica del instrumento es el mismo independientemente de si el grupo es heterogéneo u homogéneo.

Ahora bien, cabe preguntarse cuáles son las fuentes del error más habituales en la medición psicológica. Este es un asunto que ha sido exhaustivamente estudiado por especialistas (Cronbach, 1947; Thorndike, 1951; Stanley, 1971; Schmidt y Hunter, 1996), que han llegado a clasificar con todo detalle las posibles fuentes de error; su trabajo puede reducirse a tres grandes categorías: a) la propia persona evaluada; b) el instrumento de medida utilizado y c) la aplicación, corrección e interpretación hecha por los profesionales. A continuación se describe cada una.

Las fuentes de varianza de error debida a las personas evaluadas vienen determinadas por múltiples elementos que Brown (1980) sistematiza de la siguiente forma: *Memoria, Ansiedad, Variables Fisiológicas, Motivación*; también pueden incluirse el *Aprendizaje* y el *Desarrollo-Maduración;* cualquier evento de este tipo que ocurra previo a la evaluación o durante la misma, puede influir en la cuantía de los errores.

Los errores provenientes del instrumento de medida, aluden a cualquier aspecto que produzca que los evaluados respondan a los reactivos sobre bases diferentes a la posesión de un rasgo o el conocimiento de la información preguntada; dentro de estos destaca el muestreo de contenido, el cual según Cohen y Swerdlik (2001) hace referencia a la variación entre los reactivos de una prueba, así como la variación entre reactivos de diferentes pruebas. En este sentido, de acuerdo con Spearman, las pruebas paralelas son aquellas conformadas por reactivos extraídos azarosamente de un banco infinito de ítems paralelos, lo que avala la ausencia de error por muestreo de contenido; sin embargo, en la práctica sólo pueden elaborarse pruebas equivalentes. Las cuales pretenden garantizar que el puntaje verdadero de la persona que será medido por una de las pruebas, también sea medido por la otra prueba; a

pesar de los esfuerzos esto no siempre se logra, produciendo lo que se ha denominado error por muestreo de contenido.

En cuanto a los errores provenientes de los procesos de administración, calificación e interpretación de las pruebas, se tiene que la ambigüedad en las instrucciones, los límites de tiempo, el marcaje erróneo en la hoja de respuestas, un ambiente inadecuado, el uso inapropiado de las plantillas de corrección, la falta de acuerdo entre calificadores, entre otros; pueden conducir a resultados alejados de los puntajes verdaderos.

Como pudo observarse, existen múltiples factores que afectan la confiabilidad de un instrumento causando error de medida; por esta razón, el constructor de pruebas, está en el deber de generar las estrategias necesarias para controlar esas diversas fuentes de varianza de error. Seguidamente se exponen algunos procedimientos para garantizar dicho control propuestos por McGuigan (1996):

- *Eliminación:* Consiste en evitar el posible efecto de variables extrañas eliminándolas de la situación de evaluación; por ejemplo, procurando un ambiente exento de ruidos e interrupciones, permite descartar el posible efecto de esas variables.
- *Constancia:* Se refiere a que cuando la variable extraña no puede eliminarse, debe mantenerse constante para cada una de las condiciones de evaluación, de modo que afecte por igual a todos los participantes. Esto es lo que se pretende con la estandarización, garantizar la uniformidad de procedimientos de aplicación y calificación del instrumento; de manera que todos los sujetos sean evaluados en igualdad de condiciones (Mismos materiales, instrucciones, entre otros).
- *Balanceo:* Consiste en equiparar el efecto de las variables extrañas en las distintas condiciones de evaluación, se lleva a cabo cuando no se pueden mantener constantes las condiciones. Esto podría emplearse si se quiere conocer el posible efecto de la práctica sobre los resultados de las pruebas.
- *Contrabalanceo*: Se refiere a que en situaciones de evaluación en las que se requieren repeticiones, cada condición debe presentarse a cada participante la misma cantidad de veces y cada condición tiene que ocurrir igual cantidad de veces en cada sesión de práctica. De igual modo, cada condición debe preceder y seguir a todas las condiciones, la misma cantidad de veces.
- *Aleatorización:* Hace referencia a la posibilidad de que cada uno de los participante tenga la misma probabilidad de ser elegido, se utiliza la

mayoría de las veces cuando no es factible aplicar alguna de las anteriores técnicas de control. Por ello se hace especial énfasis en que la elección de las muestras se lleve a cabo de forma probabilística.

Por otra parte, según Hernández, Fernández y Baptista (2006), el control se logra mediante dos formas:

1. *Varios grupos de comparación:* Es necesario que en las situaciones de evaluación que así lo requieran se tengan por lo menos dos grupos a comparar.
2. *Equivalencia de los grupos:* Además de tener más de un grupo es necesario que los grupos sean similares en todos los aspectos. Los grupos deben ser inicialmente equivalentes y equivalentes durante la situación de evaluación. De igual modo los instrumentos de medición deben ser iguales y aplicados de la misma manera. Para lograr la equivalencia de los grupos podría utilizarse la técnica del emparejamiento, este proceso consiste en igualar los grupos en torno a una variable específica, que puede generar influencia en la medición.

Es importante destacar que estas formas de control se utilizan para el análisis de los ítems y la validación de algunos tipos de pruebas. Específicamente, los diseños experimentales empleados, son los *grupos contrastados y diseños pre-postest.* En el primer caso, se seleccionan intencionalmente los miembros de los grupos de modo que los integrantes de uno de ellos posean la variable estimada por la prueba y los participantes del otro no la posean. Bajo igualdad de condiciones de evaluación se aplica la prueba a los dos grupos y se somete los datos a un análisis estadístico, esperando que la mayoría de las personas que poseen la variable se comporten de una manera y quienes no poseen tal variable lo hagan distinto. En el segundo caso, se selecciona un solo grupo de participantes al cual le administra la prueba, luego lo somete a alguna manipulación efectiva de la variable medida por el instrumento (tratamiento o condición) y posteriormente se aplica la misma prueba o una forma paralela. Se espera que si la prueba evalúa la variable estimada, los evaluados obtengan una ejecución distinta en el pretest que contrasta con su ejecución en el postest.

Tomando en consideración las estrategias descritas anteriormente, si todo se hace con rigor se minimizarán los errores en todo el proceso, y es precisamente de eso lo que informa la confiabilidad de la prueba, de los errores cometidos (Muñiz, 1998).

Ahora bien, más allá de su control, una de las cosas más importantes del error de medición es que previene el énfasis inadecuado en una sola puntuación; permitiendo considerar los resultados de una prueba psicológica como bandas de calificaciones y no como puntos exactos; de manera que, cuanto más bajo resulte el error de medición más precisa resultará la estimación del puntaje verdadero, ya que el rango donde se encontrará el mismo será corto. De acuerdo con Gempp (2006) existen varias estrategias para estimar el puntaje verdadero a partir del observado; sin embargo, sólo dos se consideran legítimas: la aproximación tradicional y la aproximación basada en regresión.

En este sentido, y desde el punto de vista tradicional, Anastasi y Urbina (1998) plantean que el error de medición se analiza como cualquier otra desviación estándar, de manera que puede emplearse para interpretar las calificaciones individuales. De este modo, conociendo el error de medición con un grado conocido de confianza, se pueden determinar los límites dentro de los cuales se encontrará el puntaje verdadero de un sujeto que obtuvo un puntaje obtenido determinado, lo cual ha sido denominado intervalo de confianza.

Partiendo de las suposiciones clásicas acerca de los errores, es posible derivar una ecuación que permita computar el intervalo de la escala dentro del cual se puede hallar la puntuación verdadera del sujeto con algún grado conocido de exactitud, cuando se tiene la puntuación obtenida del sujeto. Dado que se ha supuesto que los errores de medida son independientes de las calificaciones verdaderas que representan, independientes entre sí, y distribuidos normalmente, la desviación estándar puede usarse para determinar los intervalos de confianza y puede interpretarse de la misma forma que cualquier otro error estándar.

Gempp (2006) plantea que se trata de un procedimiento que consiste en utilizar el error de medición para construir un intervalo de confianza en torno al puntaje observado; teóricamente, el procedimiento se basa en asumir que los errores de medida se distribuyen normalmente. Las ecuaciones para obtener los límites inferior y superior del intervalo de confianza son:

$$L_i = x - Z(Em) \quad (9a) \qquad L_s = x + Z(Em) \quad (9b)$$

En ambas ecuaciones, x representa al puntaje observado de un sujeto, Em al error de medición y Z al valor de la distribución normal asociado a la magnitud del intervalo de confianza que se desea construir. En Psicología es

práctica común emplear un intervalo no direccional de un 95% de confianza, al que corresponde un valor *Z*= 1,96.

Como ilustración, se va a considerar a una persona que obtuvo 60 puntos en una escala de ansiedad, cuya media es 40 puntos, desviación 10 puntos y confiabilidad 0,84. ¿Cuál será su puntuación verdadera en la escala? Para determinar lo límites considerar un 95% de confianza.

$$\boldsymbol{Em} = 10\sqrt{1 - 0{,}84} = 4$$

$\boldsymbol{L_i} = 60 - 1{,}96(4) = 52{,}16$ $\boldsymbol{L_s} = 60 + 1{,}96(4) = 67{,}84$

Como se expuso previamente la puntuación obtenida representa la media de las puntuaciones verdaderas en infinitas aplicaciones del instrumento y dado que el error de medición es el tamaño del error que se asume cometer cuando se acepta tal supuesto, podría esperarse que el puntaje del sujeto estuviera, con un 95% de confianza, 7,84 puntos por encima y por debajo de su puntaje obtenido (60 puntos) específicamente en el intervalo de confianza 52 – 68 y fuera de ese rango solamente en el 5% de los casos. Es notable que, aun cuando los puntajes arrojados por la escala parecen confiables, el intervalo de confianza que contiene a la puntuación verdadera, con un 95% de confianza, es bastante amplio (15,68 puntos de magnitud).

Ahora bien, desde la aproximación basada en la regresión lineal es posible, igualmente, hacer una estimación puntual de la calificación verdadera y además construir un intervalo de confianza. De acuerdo con Gempp (2006) para emplear este método es necesario conocer, además de la confiabilidad y la desviación típica de los puntajes observados, el promedio obtenido por el grupo de referencia en el cual se calculó la confiabilidad de las calificaciones.

Según este mismo autor, esta aproximación está basada en un principio simple: el puntaje observado guarda una relación lineal con el puntaje verdadero. De este modo, si la confiabilidad fuera igual a 1 (lo que equivaldría a la ausencia de error) los puntajes observados se corresponderían perfectamente con los verdaderos. En dicha situación habría una relación de identidad entre ambos puntajes (*X=V)* para cualquier valor de la distribución y la recta de regresión tendría una pendiente igual a 1. A medida que la confiabilidad disminuye, la pendiente de la recta de regresión también disminuye; tiende a ser más horizontal.

De este modo, partiendo de tal relación es posible estimar la puntuación verdadera a partir de la observada empleando la ecuación de regresión:

$$\boldsymbol{v = r_{xx'}\,(x - \bar{X}) + \bar{X}} \qquad (10)$$

En la ecuación, *v* representa el puntaje verdadero, *x* el puntaje observado, $\bar{X}$ la media de los puntajes observados y $r_{xx'}$ la confiabilidad. La discrepancia entre los puntajes observados y verdaderos será mayor mientras disminuya la confiabilidad de la medición. Asimismo, en la TCT, las puntuaciones extremas en una prueba son, por definición, estimaciones menos confiables del puntaje verdadero que las calificaciones cercanas a la media grupal (Gempp, 2006).

El autor igualmente señala que como en todo modelo de regresión lineal, la predicción del puntaje verdadero a partir del observado no es perfecta y tiene asociada un error de predicción, conocido como Error Estándar de Estimación, y se obtiene a través de la siguiente ecuación:

$$\boldsymbol{Ee = Sx\sqrt{r_{xx'}\,(1 - r_{xx'})}} \qquad (11)$$

Una fórmula alternativa para el error estándar de estimación es a partir del error estándar de medida:

$$\boldsymbol{Ee = Em\sqrt{r_{xx'}}} \qquad (12)$$

Esta ecuación permite estimar el error de estimación cuando se desconoce la desviación típica de las puntuaciones y se cuenta con una buena estimación del error de medida (Gempp, 2006).

Del mismo modo que el error de medición permite obtener un intervalo de confianza para el puntaje verdadero en torno al puntaje observado, el error de estimación facilita la construcción de un intervalo de confianza para el puntaje verdadero estimado. Las ecuaciones para obtener los límites inferior y superior del intervalo de confianza son:

$$\boldsymbol{L_i = v - Z(Ee)} \quad (13a) \qquad \boldsymbol{L_s = v + Z(Ee)} \quad (13b)$$

Nótese que son las mismas ecuaciones presentadas para la construcción de intervalos con el error de medición; a diferencia que en este caso, es el error de estimación el que se emplea, y que se trabaja con los puntajes verdaderos (estimados) y no con los observados. Empleando el mismo ejemplo se obtendría lo siguiente:

$$\boldsymbol{v} = 0{,}84\boldsymbol{(}60 - 40\boldsymbol{)} + 40 = 56{,}8$$

$$\boldsymbol{Ee} = 10\sqrt{0{,}84\boldsymbol{(}1 - 0{,}84\boldsymbol{)}} = 3{,}67 \qquad \text{ó} \qquad \boldsymbol{Ee} = 4\sqrt{0{,}84} = 3{,}67$$

$$\boldsymbol{L_i} = 56{,}8 - 1{,}96\boldsymbol{(}3{,}67\boldsymbol{)} = 49{,}6 \qquad \boldsymbol{L_s} = 56{,}8 + 1{,}96\boldsymbol{(}3{,}67\boldsymbol{)} = 63{,}9$$

Se puede concluir que con un 95% de confianza el puntaje verdadero del sujeto se encuentra en un intervalo comprendido entre 50 y 64 puntos. Es notable que este intervalo (14,3 puntos de magnitud) es más pequeño que el obtenido por el método tradicional, aunque esta diferencia se debe como señala Gempp (2006) a que ambos métodos apuntan a objetivos distintos; el método tradicional construye un intervalo centrado en la puntuación observada (entre qué valores de la puntuación observada se encuentra el puntaje verdadero, con una probabilidad conocida) y el método de regresión construye un intervalo centrado en la estimación de la calificación verdadera (entre qué valores de la puntuación verdadera se encuentra el puntaje verdadero, con una probabilidad conocida).

Ahora bien, cuando se desea conocer si es significativa la diferencia existente entre los puntajes de un individuo en dos pruebas, o de distintos sujetos en un mismo test, resulta necesario disponer de un indicador de tal diferencia. (Magnusson, 1975). En cualquiera de los casos (diferencias intraindividuales o interindividuales) conociendo el tamaño de la diferencia entre los puntajes en cierta dirección, es posible garantizar que su probabilidad de aparición como resultado del error de medida, sea tan pequeña que pueda emplearse para hacer predicciones.

En este sentido, cuando se desea comparar a dos personas en una misma prueba, se parte del supuesto de que existe una distribución de diferencias de puntajes para infinitos sujetos en una misma prueba. Tales diferencias, son consideradas producto de la incapacidad de la prueba para medir los puntajes verdaderos de los dos individuos, por lo cual se afirma que es una distribución de errores cuya media es cero y su desviación viene dada por el error estándar de las diferencias.

El error estándar de las diferencias intraprueba resulta del error de medición que conllevan los dos puntajes involucrados. La ecuación para estimar el error estándar de las diferencias intraprueba se obtiene de la siguiente ecuación:

$$\boldsymbol{Sed = Sx \sqrt{2(1 - r_{xx'})}} \qquad (14)$$

Por su parte, cuando se requieren tomar decisiones que implican diferencias entre pruebas, es decir, intraindividuales, se asume que se administraron infinitos test paralelos *g y h,* a una misma persona y que no hay diferencias entre los puntajes *g y h* (hipótesis nula); seguidamente, se calcula la diferencia para cada par de puntajes en los test paralelos. Las diferencias de puntajes se distribuirán al azar generando diferencias que en unas circunstancias serán negativas (bajo la media), en otras positivas (sobre la media) y presentarán una media de cero. La varianza de esas diferencias constituye la varianza de error, de donde se desprende la ecuación que permitirá estimar el error estándar de las diferencias intertest:

$$\boldsymbol{Sed = Sx \sqrt{2 - r_{gg'} - r_{hh'}}} \qquad (15)$$

Ahora es posible calcular qué tan grande debe ser la diferencia en una dirección para que se le considere significativa con un nivel de confianza dado. Como señala Gempp (2006), el error estándar de medida de la diferencia entre dos calificaciones individuales indica la contribución del error estándar de medida de la puntuación de cada persona a la diferencia entre sus puntuaciones. En este sentido, una derivación de la aplicación anterior es la interpretación de la diferencia entre los puntajes obtenidos por dos evaluados en una misma prueba.

Por ejemplo, si una persona calificó en una escala de estrés con un puntaje menor que otra persona en esa misma escala, qué magnitud debe tener la diferencia entre ambos puntajes para asumir que refleja una diferencia "verdadera" y no que se debe a error de medida. Para contrastar los puntajes de ambas personas la misma prueba se debe construir un intervalo de confianza para ambos resultados, con cualquiera de los métodos descritos anteriormente, y como indica Gempp (2006), exigir que la diferencia entre dos puntuaciones sea igual o mayor que dos errores estándar de las diferencias para ser considerada significativa; esto está determinado en función de un intervalo de confianza del 95% cuyo puntaje típico z asociado

es de 1,96 es decir, aproximadamente las dos veces que plantea Gempp. En este sentido, si el primero sujeto calificó con 66 puntos y el segundo con 74 puntos y el error es de 3 puntos, se puede considerar que los resultados reflejan una diferencia sustantiva, ya que la diferencia empírica (74-66) entre los puntajes es de 8 puntos; mientras que la diferencia teórica (3x2) es igual a 6 puntos. También es posible realizar este análisis, calculando el error estándar de medida de la diferencia entre las puntuaciones obtenidas por una sola persona en dos test.

En resumen, estos son, a grandes rasgos, algunos de los aspectos más importantes acerca del error de medición. Su consideración conducirá a un uso más apropiado de los instrumentos de medida.

Referencias

Anastasi, A. y Urbina, S. (1998) *Test Psicológicos.* (7ma Ed.). México: Prentice Hall.

Brown, F. (1980) *Medición en Psicología y Educación*. México: El Manual Moderno.

Cohen, R. y Swerdlik, M. (2000). *Pruebas y Evaluación Psicológicas.* México: McGraw Hill.

Cronbach, L. J. (1947). Test reliability: its meaning and determination. *Psychometrika, 12*, 1-16.

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.

Hernández, R., Fernández-Collado, C. y Baptista, L. (2006). *Metodología de la Investigación* (4ta Ed.). DF, México: McGraw Hill.

Magnusson, D. (1975) *Teoría de los Test.* DF, México: Trillas.

McGuigan, F. (1996). Psicología experimental. DF, México: Prentice Hall.

Muñiz, J. (1998). La Medición de lo Psicológico. *Psicothema, 10* (1), 1-21.

Schmidt, F. L. y Hunter, J. E. (1996). Measurement error in psychological research: lessons from 26 research scenarios. *Psychological Methods, 1*(2), 199-223.

Stanley, J. C. (1971). Reliability. En R. L. Thorndike (Ed.), *Educational Measurement*. Washington. DC: American Council on Education.

Thorndike, R. L. (1951). Reliability. En E. L. Lindquist (Ed.). *Educational Measurement*. (pp. 560-620) Washington, DC, EEUU: American Council on Education.

(Recibido el 12 de Marzo de 2009 – Aceptado el 15 de Junio de 2009)