Psicología, tercera época ▪ Revista digital arbitrada
Vol. 35, N° 2-2016 | pp. 61-84 | ISSN: 1316- 0923

La confiabilidad en los instrumentos estandarizados de medición

Luisa E. Lezama
Escuela de Psicología, Universidad Central de Venezuela
luisalezam@yahoo.com

Resumen

Independientemente del tipo de instrumento estandarizado (normativo o criterial, de personalidad o aptitud, de calificación automatizada o semi-objetivo), éste requiere mostrar que las puntuaciones que arroja son confiables. Este artículo por medio de una investigación documental, expone la conceptualización de esta propiedad psicométrica,anclándola en el modelo lineal de Spearman y en su concepto de error de medición. Además, se detallan los procedimientos de test-retest, formas equivalentes, dos mitades y consistencia interna, especificando por qué se habla de coeficiente de estabilidad temporal y de equivalencia. Igualmente se identifican el modo cómo la homogeneidad de la muestra, la longitud de la prueba y la velocidad de respuesta afectan la confiabilidad. Finalmente, se alude a la confiabilidad de acuerdo entre jueces, rescatando que no está amparada dentro de la teoría clásica de los tests.

Palabras clave: Confiabilidad, Error de medición, Coeficientes de estabilidad temporal y de equivalencia, Tipos de confiabilidad.

Recibido: 20 de abril de 2016
Aceptado: 22 de junio de 2016
Publicado: 10 de enero de 2017

Psicología ▪ Refereed journal
Volume 35, Issue 2-2016 | Pages 61-84 | ISSN: 1316- 0923

Reliability in standardized measurement instruments

Luisa E. Lezama
Escuela de Psicología, Universidad Central de Venezuela
luisalezam@yahoo.com

Abstract

Indistinctly from the type of standardized instrument (norm or criterion referenced, personality or aptitude, with automated or semi-objective qualification), it is required from such tools to demonstrate reliability in the yielded scores. This paper, through a documentary research, exposes the conceptualization of this psychometric property, anchoring it in the Spearman's linear model and in its concept of measurement error. Also, the test-retest, equivalent forms, split-half and internal consistency methods are detailed, specifying why coefficients of temporal stability and equivalence are discussed. As well, the ways in which the homogeneity of the sample, the length of the test and the speed of response affect the reliability, are identified. Finally, it is discussed the inter-rater reliability, rescuing that it is not covered by the classical test theory.

Keywords: Reliability, Measurement error, Temporal stability and equivalence coefficients, Types of reliability

Received: Apr 20, 2016
Accepted: Jun 22, 2016
Published: Jan 10, 2017

Dentro de los múltiples campos aplicados de la Psicología es frecuente que los profesionales recurran al uso de instrumentos estandarizados de medición como parte de los recursos que emplean en sus evaluaciones. La calidad de tales evaluaciones se relacionan directamente con las características de esos instrumentos de recolección de información. En el caso de las pruebas estandarizadas tal calidad puede ser consultada en los manuales de prueba, donde se detallan los procedimientos de confiabilidad y validez que habilitan su uso.

La lectura acerca de las propiedades psicométricas de algunos manuales y artículos de revistas invitan a realizar precisiones en torno al tema, específicamente, de la propiedad psicométrica confiabilidad esperando que tales precisiones faciliten la comprensión de lo que se hace, el por qué se hace y promueva las estimaciones de confiabilidad con criterios más técnicos que logísticos.

Para lograr esos objetivos se hará un recorrido detallado por elementos teóricos específicos de la confiabilidad, que se expondrán a la luz de los asuntos prácticos en los que se comprometen los elaboradores de instrumentos, con lo cual la disertación está dirigida, básicamente, a diseñadores de pruebas, pero su naturaleza será de fácil comprensión para usuarios de tests, quienes dispondrán de información para juzgar la calidad de los instrumentos estandarizados que usan durante sus evaluaciones.

Conceptualización de Confiabilidad

Si se aplica una Prueba de Competencia Lingüística a un niño el lunes y el martes de la misma semana, bajo idénticas condiciones, y si la prueba no capta errores, el niño debería obtener más o menos el mismo puntaje en las dos mediciones, en cuyo caso se diría que el proceso de medición es confiable. Si las calificaciones del niño son distintas, esto se deberá a que la prueba es sensible a medir error, ya que sus competencias lingüísticas deberían ser las mismas tanto el lunes como el martes. En este ejemplo se diría que los resultados son no confiables.

El concepto de confiabilidad, justamente, hace alusión a esa repetibilidad o replicabilidad de los resultados en la medición de una variable. Un proceso de medición es confiable si las medidas que se hacen "carecen de errores", con lo cual los resultados obtenidos en una determinada ocasión con la misma

prueba o su equivalente, bajo las mismas condiciones, son reproducibles, es decir, son consistentes.

Se aprecia que un concepto central dentro del contexto de la confiabilidad es el de *error*. Los errores de medida de los que se ocupa la confiabilidad son aquéllos no sometidos a control, que son azarosos, que pueden estar presentes en un momento de la medición, pero no en otro y que resultan inevitables durante el proceso de medir. Tales fuentes de error provienen de la estabilidad temporal de la variable, de la equivalencia entre pruebas o de la consistencia entre los reactivos.

Así pues, la confiabilidad se puede definir como un término genérico que refiere a la repetibilidad de los resultados de la medición y que, específicamente, alude a la estabilidad o la equivalencia, según sean las fuentes de varianza de error a las que sea sensible el método para estimarla (Cureton, 1965). La confiabilidad, así entendida, tiene su anclaje en el conocido *modelo lineal de Spearman*.

Modelo lineal de Spearman: derivación estadística de la confiabilidad

La *teoría clásica de la confiabilidad* propuesta por *Spearman* en 1910 argumenta que ésta se establece tomando medidas repetidas de una variable con un instrumento, o su forma paralela, bajo idénticas condiciones. Mientras el resultado obtenido por diferentes personas evaluadas sea el mismo, se dice que la medida es repetible y el proceso es confiable.

El *modelo lineal*, como también se conoce a la *teoría clásica de la confiabilidad*, parte de una serie de supuestos e implicaciones, que permiten llegar a la derivación estadística de la confiabilidad. Según Santisteban (1990) el primero de ellos es conocido como *supuesto fundamental* y postula que el *puntaje obtenido t* por un individuo *j* en un test t_j, está formado por dos elementos: un puntaje verdadero T_j y un puntaje error e_j. Esto es: $t_j = T_j + e_j$.

El puntaje verdadero T_j es el mismo para cada individuo en infinitos tests paralelos, de modo que cualquier diferencia a nivel de su puntaje observado, será debida al puntaje de error captado por la prueba y no al puntaje verdadero del individuo; el *puntaje de error* e_j, alude a los llamados puntajes azarosos que son adjudicables a la sensibilidad del instrumento para captar aquellos eventos cuyo efecto varía de una ocasión a otra. Quedan excluidos

de esta categoría los denominados *errores constantes*. El puntaje de error es el resultado de la diferencia entre el puntaje verdadero y el observado, de tal modo que: $e_j = t_j - T_j$.

El *supuesto de nulidad de los errores*, es el segundo propuesto dentro de la teoría, y postula que los errores se anulan entre sí, por lo que debe suponerse que la media aritmética de los errores es cero ($M_e = 0$) para un número infinito de sujetos en un mismo test, y para un solo sujeto en infinitos tests paralelos.

El tercero, es el *supuesto de correlación cero entre los errores*, y asume que la correlación entre los puntajes de error es cero ($r_{ee} = 0$) para un número infinito de personas en un mismo test; y, para un solo sujeto en infinitos tests paralelos, debido a que los errores son azarosos o "no sistemáticos", con lo cual no producen varianza conjunta y la correlación es cero.

El último, es el *supuesto de correlación cero entre los puntajes verdaderos y de error* ($r_{eT} = 0$) para un número infinito de sujetos en un mismo test; y, para un solo sujeto en infinitos tests paralelos, debido a que el error es totalmente independiente del puntaje verdadero, con lo cual no producen varianza conjunta resultando cero su correlación.

De los supuestos anteriormente expuestos, se desprenden una serie de implicaciones, que permiten relacionar distintas variables e inferir conceptos de variables inobservables (T_j, por ejemplo) a partir de variables observables (t_j, por ejemplo). Según Santisteban (1990) y Muñiz (1992), las más relevantes de tales implicaciones son:

a. El valor esperado de la puntuación verdadera (T_j) será igual a la media de las calificaciones observadas (t_j) en infinitos tests paralelos (con lo cual se le acepta como un concepto matemático). Ello es completamente razonable si se recuerda que: 1) $t_j = T_j + e_j$. y que 2) la media de los errores es cero ($M_e = 0$), en infinitos test paralelos; de manera que, la puntuación verdadera será igual a la media de las puntuaciones observadas, siempre que hayan sido obtenidas de pruebas repetidas de manera independiente.

b. La ecuación de regresión de las puntuaciones obtenidas sobre las observadas, es la de la línea recta que pasa por el origen y tiene como pendiente la unidad.

c. La varianza de la distribución *t* resulta de la suma de la varianza de los puntajes verdaderos, más la varianza de los puntajes de error, más las

covarianzas entre ellos $S^2_t = S^2_T + S^2_e + cov_{Te}$, pero dado que la correlación entre los puntajes verdaderos y de error es cero, la covarianza será cero y la ecuación quedará como: $S^2_t = S^2_T + S^2_e$, de donde se desprende que $S^2_e = S^2_t - S^2_T$.

d. La varianza total del test S^2_t es igual para los test paralelos; la varianza de los puntajes verdaderos S^2_T es la misma para todos los test paralelos, puesto que los sujetos contribuyen con el mismo puntaje a las diversas distribuciones de puntajes verdaderos; la varianza de error S^2_e, es la misma ya que es el producto de la sensibilidad del instrumento a captar errores azarosos. Pero mientras las personas tienen el mismo puntaje verdadero T_j en infinitos test paralelos, el tamaño y la dirección de los puntajes de error ej variarán azarosamente de un test paralelo a otro.

A partir de las consideraciones previas, es posible derivar de la ecuación general para calcular una correlación $\left(r_{t_1 t_2} = \frac{\sum(t \cdot t)}{N \cdot S_{t_1} \cdot S_{t_2}}\right)$, la fórmula especial para el cálculo del coeficiente de confiabilidad. Supóngase que se obtienen los puntajes de un individuo *j* en dos test paralelos (t_{j1} y t_{j2}), por lo que se estableció previamente, ellos se representan como $t_{j1} = T_{j1} + e_{j1}$ y $t_{j2} = T_{i2} + e_{i2}$, si se sustituye con estos símbolos la ecuación anterior se tendrá que: $r_{t_1 t_2} = \frac{\sum(T_j + e_{j_1})(T_j + e_{j_2})}{N S_{t_1} S_{t_2}}$, que al ser desarrollada queda de este modo: $\frac{\sum T_j^2}{N S_{t_1} S_{t_2}} + \frac{\sum T_j e_{j_2}}{N S_{t_1} S_{t_2}} + \frac{\sum e_{j_1} T_j}{N S_{t_1} S_{t_2}} + \frac{\sum e_{j_1} e_{j_2}}{N S_{t_1} S_{t_2}}$

Como derivación de los supuestos del *modelo lineal de Spearman*, el segundo y tercer término se hacen cero, debido a que la correlación entre los puntajes verdaderos y de error es cero ($r_{eT} = 0$); el cuarto término también se hace cero, toda vez que no existe correlación entre los puntajes de error ($r_{ee} = 0$), quedando solo el primer término de la ecuación $\frac{\sum T_j^2}{N S_{t_1} S_{t_2}}$.

En vista de que las desviaciones estándares son iguales para los tests paralelos, es decir $S_{t1} = S_{t2}$, ese primer término de la ecuación queda $\sum T_j^2 / N s_j^2$ donde T_j representa la desviación de los puntajes de los sujetos respecto a la media de las distribuciones de los puntajes verdaderos, que partido entre *N*, es igual a la varianza de los puntajes verdaderos $\sum T_j^2 / N = s_T^2$. Dicho término queda ubicado en el numerador de la expresión que es dividida entre S^2_t por lo tanto la correlación es $r_{tt} = \frac{s_T^2}{s_t^2}$ adicionalmente, si se sustituye en el numerador $S^2_T = S^2_t - S^2_e$, la expresión también puede escribirse como $r_{tt} = 1 - \frac{s_e^2}{s_t^2}$, lo cual implica que la confiabilidad es la proporción de varianza total que es varianza verdadera, el resto del porcentaje de varianza total, es explicada por la varianza de error. Es decir, que mientras la varianza verdadera explique más la varianza total, menor será la varianza de error y mayor será la confiabilidad de la medida.

Como el coeficiente de confiabilidad se calcula para determinar qué tan similares son grupos de datos, su valor oscila entre cero (absolutamente no semejante) y uno (absolutamente semejante). En términos generales se puede decir que, si el puntaje de cada individuo en la distribución total tj es el mismo que su puntaje verdadero T_j, la varianza total será la misma que la de los puntajes verdaderos S^2_T y la varianza de error S^2_e será cero, con lo cual, la confiabilidad será 1. Si todo el puntaje obtenido t_j por un individuo es un puntaje de error e_j, la varianza verdadera S^2_T será cero, la varianza de error S^2_e será máxima y será igual a la varianza total S^2_t , con lo cual la confiabilidad será cero.

Si un coeficiente de confiabilidad es por ejemplo de .90 (cercana a 1), el 90% de la varianza total estará conformada por varianza verdadera y 10% por varianza de error, con lo cual la medición es bastante confiable. Por el contrario, si el coeficiente de confiabilidad es de .20 (cercano a 0), solo el 20% de la varianza total está conformada por varianza verdadera y 80% por varianza de error, con lo cual la medición será poco confiable.

La información provista por el coeficiente de confiabilidad se puede complementar con la del *índice de confiabilidad* $r_{tT}= \sqrt{r_{tt}}$ que provee la correlación entre los puntajes verdaderos y los puntajes obtenidos, en tanto se logre controlar artificialmente todas las fuentes de varianza de error (suposición teórica).

Error de medición

Como se indicó anteriormente, una calificación está constituida por la suma de un puntaje verdadero más un puntaje de error $t_j=T_j+e_j$; éste último es el responsable de que la calificación de un individuo sea distinta de un test paralelo a otro, ya que, según se supuso, el puntaje verdadero siempre es el mismo.

El puntaje de error puede ser el resultado de: a) procesos sistemáticos o errores constantes que afectan a todas las observaciones por igual, por lo que no influyen en las comparaciones, siendo normalmente desconsiderados, ó b) procesos aleatorios que afectan a ciertas observaciones de manera diferente que a otras, en cuyo caso se le denomina *sesgo*, el cual puede provenir de las más variadas fuentes.

De esos eventos aleatorios, que producen *sesgo* en las calificaciones, se ocupa la confiabilidad. Ellos pueden provenir de muy variadas fuentes, las cuales

son sistematizadas por Cohen y Swerdlik (2001) y Brown (1980) en tres grandes rubros, a saber, error durante la *administración* y *calificación* del instrumento, dentro de la *prueba* y debido al *examinado*.

La fuente de varianza de error, que tiene que ver con la *administración* y *corrección* de la prueba, incluye factores como la modificación de las instrucciones de administración del instrumento, el marcaje erróneo en la hoja de respuestas, el empleo de un ambiente con ruidos, la mala ventilación e iluminación, las fallas en el registro del tiempo, el empleo inapropiado de las plantillas de corrección, etc. Esta fuente de varianza de error se controla fácilmente especificando en el manual del test los pasos para su empleo, a objeto de que los procedimientos se sigan estrictamente (Cohen y Swerdlik, 2001; Brown, 1980).

Los *errores dentro de la prueba*, aluden a cualquier aspecto que haga que un individuo responda los reactivos sobre bases distintas de su conocimiento de la respuesta o posesión del rasgo, con lo cual, el asunto recae directamente sobre la muestra de items que constituyen la prueba, lo que es de especial atención cuando se trate de pruebas paralelas al azar o equivalentes.

Para el modelo lineal de *Spearman*, las pruebas paralelas al azar son aquellas conformadas por reactivos extraídos azarosamente de un banco infinitos de items paralelos, lo que garantiza la ausencia de error por muestreo de contenido. Ello es un asunto eminentemente teórico, en la práctica sólo se pueden elaborar pruebas equivalentes. Éstas, son instrumentos construidos con la misma fundamentación teórica y psicométrica; es decir, se diseñan de manera tal que midan la variable con el mismo formato; se derivan de la misma tabla de especificaciones y se ajustan para arrojar similares estadísticos (tendencia central y dispersión), tanto a nivel de los ítems, como de la prueba total; además para para arrrojar una correlación entre las formas, similar a la que se presentaría si se reaplicara cada forma (Nunnally y Bernstein,1999).

Aún cuando se construyan las pruebas apegadas a condiciones de estricta equivalencia, como las señaladas anteriormente, siempre habrá una parte del puntaje verdadero de la persona que será medida por una de las pruebas, pero no por la otra, produciendo lo que se denomina *error por muestreo de contenido* (una prueba muestrea un contenido no muestreado por la otra). Al procedimiento de confiabilidad que recoja tal varianza de error, se le denomina *coeficiente de equivalencia*.

La tercera fuente de varianza de error contemplada es la debida a las *personas evaluadas*, y es la que resulta más difícil de manejar, aún cuando se mantengan bajo estricto control el resto de las condiciones de administración.

Los elementos que aportan varianza de error por parte del examinado son sistematizados por Brown (1980) en los siguientes términos: *motivación*, que produce sus efectos si los sujetos están diferencialmente dispuestos para la prueba; *ansiedad*, que tiene efectos perjudiciales sobre la ejecución y produce comportamientos erráticos, por lo que es recomendable otorgar información para bajar los niveles de temor explicando el objetivo de la prueba, haciendo que las personas se familiaricen con la situación, etc.; *variables fisiológicas*, ya que si una persona toma un test, enferma o fatigada probablemente obtendrá calificaciones diferentes, en comparación a cuando está sana y vigil; *memoria*, dado a que los efectos del recuerdo de la primera aplicación de la prueba, pueden afectar la segunda, si dicha administración se produce en ocasiones sucesivas o muy cercanas en el tiempo.

Dentro de esta tercera fuente de varianza de error, requieren especial atención los procesos de aprendizaje, desarrollo y educación, ya que ellos causan que el puntaje verdadero cambie genuinamente. Por ejemplo, si algunas personas reciben un curso de instrucción, su conocimiento acerca de ese tópico mejorará, por lo que el instrumento administrado antes y después de tal instrucción revelará una diferencia que, si bien se tratará como error, será producto de la fluctuación genuina del puntaje verdadero. Este tipo de error es un buen ejemplo de lo que en la literatura psicométrica se denomina *muestreo de tiempo*, ya que ocurren fluctuaciones del puntaje verdadero como producto del lapso transcurrido entre la administración de los tests; al procedimiento de confiabilidad que recoge tal varianza de error se le denomina *coeficiente de estabilidad temporal*.

Magnusson (1975) indica que de la ecuación de confiabilidad $r_{tt} = 1 - \frac{s_e^2}{s_t^2}$ se deriva la fórmula para computar el *error de medición*, ya que al despejar de la misma S_e^2 quedará como $S_e^2 = S_t^2(1-r_{tt})$, que al expresarse como desviación típica, queda como $S_e = S_t\sqrt{1-r_{tt}}$. El error de medición así determinado, se analiza según Anastasi y Urbina (1998), como cualquier otra desviación estándar, razón por la cual puede emplearse para interpretar las calificaciones. Así, por ejemplo, conociendo el error de medición, con un grado particular de exactitud, se puede determinar los límites dentro de los cuales está el puntaje verdadero de una persona con un puntaje obtenido dado, lo cual se conoce en la literatura como *intervalo de confianza*.

El coeficiente de confiabilidad y el error de medición, tal como puede haberse apreciado, son dos formas de expresar la confiabilidad de una prueba, pero cada uno de ellos informa aspectos particulares. El primero, indica la consistencia de las puntuaciones obtenidas en una prueba, pero no informa directamente la cantidad de variabilidad (error) que puede esperarse al hacer las mediciones, la cual se obtiene a partir del error de medición, que señala el tamaño del error que se acepta cometer cuando se asume la calificación obtenida como si fuera la verdadera. De modo que resulta un indicador más preciso de cuán cierta es la estimación del puntaje verdadero, aún con coeficientes de confiabilidad relativamente altos.

Si bien el coeficiente de confiabilidad y el error de medición suministran información complementaria en cuanto a la calidad de un test, para comparar entre pruebas diferentes, sólo se puede recurrir a la confiabilidad ya que, independientemente del instrumento, este coeficiente siempre se expresa en las mismas unidades. Recuérdese que la correlación se obtiene a través de la ecuación *producto momento de Pearson*, la cual informa sobre la correspondencia entre el lugar (en desviaciones típicas) ocupado por la persona en las distribuciones, de modo que, independientemente de los instrumentos, siempre se expresa en las mismas unidades. En cambio, el error de medición se expresa en los mismos términos que las puntuaciones de la prueba. Así, si las calificaciones están dadas como respuestas correctas, número de errores, etc., el error se expresará en esas mismas unidades, resultando incorrecto metodológicamente las comparaciones entre tests expresados en unidades diferentes.

Procedimientos para determinar la confiabilidad de las medidas

Si bien la confiabilidad se expresa como un coeficiente de determinación, los arreglos experimentales para obtener las calificaciones objeto de tal estudio, conlleva a distintos tipos de confiabilidad, a saber: test-retest, formas alternas, dos mitades y consistencia interna.

Según Cohen y Swerdlik (2001) corresponde a quien conduce el estudio identificar el método que más convenga a sus fines, considerando entre otros aspectos, el objetivo de la prueba, la homogeneidad del contenido y la variabilidad de la medida. A continuación se describen los métodos para estimar la confiabilidad partiendo de Brown (1980), Cohen y Swerdlik (2001), Magnusson (1975), Anastasi y Urbina (1998) y Aiken (1996)

haciendo especial énfasis en las fuentes de varianza de error a las que es sensible la medición, según el método de confiabilidad empleado.

El procedimiento de test-retest es el método más obvio. Consiste en administrar el mismo instrumento en una segunda ocasión, a una misma muestra de personas, bajo idénticas circunstancias, para luego correlacionar las calificaciones recogidas, a través de la ecuación *producto momento de Pearson*.

La varianza de error a la que es sensible este método procede de varias fuentes: cambios climáticos importantes, interrupciones durante la administración, ruidos, una punta rota, enfermedad, fatiga, ansiedad, entre otras, susceptibles de control haciendo los arreglos experimentales pertinentes. No obstante, la varianza de error más relevante que recoge este método deviene del paso del tiempo, no por el tiempo en sí mismo, sino porque en su devenir pueden ocurrir cambios importantes en los puntajes verdaderos de los sujetos producto del aprendizaje, la maduración o el desarrollo. A esa fuente de varianza se le conoce como varianza de error por *muestreo de tiempo* y el resultado del estudio de retest provee un coeficiente de *estabilidad temporal de las medidas*.

Para decidir el lapso que se dejará transcurrir entre el test y el retest, es indispensable disponer de un claro conocimiento teórico acerca de la variable medida; e impedir que el tiempo, por exceso o déficit, pueda modificar el puntaje verdadero genuino, con lo cual la inconsistencia entre las calificaciones test-retest subestimaría la confiabilidad; poco tiempo entre las medidas facilitaría el recuerdo de las respuestas, con lo que la consistencia entre las calificaciones sobreestimará esta propiedad psicométrica. Todo ello deja clara la importancia que tiene para los constructores disponer de una clara conciencia de las experiencias relevantes ocurridas entre el test y su reaplicación.

Cuando se informa, por ejemplo, que el resultado de un estudio de retest es de ,97, se está diciendo que esa medida goza de una importante estabilidad temporal, ya que ese valor da cuenta de que el 97% de la varianza total es debida a la varianza de los puntajes verdaderos y solo un 3% es debida a varianza de error, específicamente devenida del muestro de tiempo; es decir, que los resultados obtenidos en la medición serán poco sensibles a los cambios fortuitos de la cotidianidad de los evaluados o del entorno en

que se administra la prueba, al menos para el plazo de tiempo que se dejó transcurrir para ejecutar el estudio de retest.

Un indicador de estabilidad temporal, para Cohen y Swerdlik (2001) es indispensable cuando las calificaciones obtenidas con una prueba van a ser empleadas para tomar decisiones a largo plazo como ocurre con los instrumentos de aptitudes, habilidades, intereses y de personalidad. Resulta menos útil cuando el rasgo medido es poco estable en el tiempo.

El segundo tipo de confiabilidad que se debe mencionar es el de *formas paralelas al azar*, *equivalentes* o *alternas*. Es pertinente acotar que, si bien los términos pruebas paralelas al azar y pruebas equivalentes o alternas son empleados indistintamente, existe una diferencia entre ellas. Cuando se habla de las primeras, se hace referencia a tests que se han conformado al seleccionar azarosamente los items de un banco infinito de reactivos paralelos; en cambio, se trata de pruebas equivalentes o alternas cuando ambas se han construido a propósito para que sean equivalentes teórica y psicométricamente. Esto último es lo que se hace en la práctica.

Para establecer la confiabilidad a través de formas equivalentes, se construyen dos tests independientes para que sirvan a las mismas especificaciones, garantizando que tienen el mismo número de reactivos, que éstos están expresados de la misma forma, que usan las mismas instrucciones y que arrojan los mismos estadísticos, tanto para los items como para las pruebas.

Luego de construir los tests con las especificaciones ya señaladas, se les administran contrabalanceando el orden (control de variable extraña) a un mismo grupo de personas, para finalmente establecer la correlación entre las calificaciones.

No obstante, aún cumpliéndose las más estrictas condiciones de paralelismo, siempre habrá un puntaje verdadero medido por un instrumento y no por el otro, esto es lo que Magnusson (1975) denomina varianza de error por muestreo de contenido, caso en el cual, el estudio de confiabilidad genera un coeficiente de equivalencia. Así, si la correlación arroja, por ejemplo, un coeficiente de ,97, se sabrá que el 97% de la varianza total es debida a varianza verdadera; y el 3% a la varianza de error que deviene de la falta de equivalencia entre el contenido de las dos pruebas.

También se pueden administrar las formas equivalentes con un lapso entre ellas (formas paralelas demoradas) en cuyo caso, la correlación obtenida

supondrá tanto un *coeficiente de equivalencia* como de *estabilidad temporal*, dado que todos los factores que producen inconsistencia en el paradigma de formas alternas (muestreo del contenido) como en el de retest (muestreo del tiempo) están presentes. Al aplicar este diseño se pueden esperar los coeficientes más bajos de confiabilidad que puedan obtenerse en un método para estimar esta propiedad psicométrica.

En algunas ocasiones, administrar una prueba en más de una ocasión resulta logísticamente difícil o muy costoso; en otras oportunidades la naturaleza de la prueba dificulta elaborar una forma equivalente. En tales circunstancias, se puede recurrir al procedimiento de *división por mitades* o *split-half* que separa una misma prueba en dos partes y la correlación entre esos pares de puntajes suministra el coeficiente de confiabilidad.

Se puede dividir la prueba en dos subpruebas, si a una aportan puntaje los items pares y a la otra aportan puntajes los items impares; también se pueden conformar dos grupos de calificaciones si a un grupo aportan calificaciones la primera y la cuarta parte de los items y al otro aportan puntos la segunda y la tercera parte de los items. En cualquier combinación, lo relevante es procurar garantizar la equivalencia entre las mitades ya que es fácil que tal paralelismo se pierda, si se desatiende el contenido de los reactivos, el nivel de dificultad de los items, el efecto de la práctica, la fatiga, etc.

Después de obtener las calificaciones mitades, independientemente del modo en que se decida dividir el instrumento, se procede a establecer la confiabilidad de las medidas a través de la *ecuación producto momento de Pearson*, pero como se dispone de la información para "media prueba" se procede a aplicar la ecuación de la *profecía de Spearman-Brown* $r_{nn} = \frac{n r_{tt}}{1+(n-1)\ r_{tt}}$ que es una ecuación general empleada cuando el instrumento ha variado su longitud n veces. De esta ecuación general, se obtiene una particular, que se aplica cuando la prueba es doblada en longitud, que es lo que ocurre en el caso de la confiabilidad de dos mitades $r_{nn} = \frac{2\ r_{tt}}{1+ r_{tt}}$, en el que ambas pruebas tienen medias y varianzas iguales.

A partir de la ecuación general de S*pearman- Brown*, no solamente es posible saber cómo varía la confiabilidad tras cambios en la longitud en la prueba, sino también es posible saber cuánto es necesario variar esa longitud para lograr coeficientes de confiabilidad particulares. Para ello se aplica la fórmula $n = \frac{r_{tt_n}(1-r_{tt})}{r_{tt}(1-r_{tt_n})}$, donde la *n* obtenida se transforma a número de items, e indica el número de reactivos paralelos, que deben agregarse o reducirse de

la prueba, para lograr valores de la confiabilidad deseada (debe recordarse que ésto es en papel y aún se requiere examinarlo en la práctica).

El coeficiente de confiabilidad obtenido a través de la división en mitades, es un *coeficiente de equivalencia*, ya que recoge información acerca de fallas en el muestreo de contenido entre las dos formas, es decir, da a conocer en qué medida cada mitad muestrea apropiadamente el contenido que se está midiendo, con lo cual, si el estudio dio .95 se sabe que el 95% de la varianza total es varianza verdadera, y un 5% es varianza de error, debido a la falta de equivalencia entre las dos mitades de la prueba.

Según Anastasi y Urbina (1998), *Rulon y Guttman* presentan una ecuación alternativa para determinar la confiabilidad de par-impar la cual únicamente requiere la varianza de las diferencias entre las calificaciones de cada persona en ambas mitades de la prueba (DE^2_d) y la varianza de las puntuaciones totales (DE^2_x). De manera que la confiabilidad de dos mitades de *Rulon y Guttman*, viene dada por la ecuación $r_{tt} = 1 - \frac{DE_d^2}{DE_x^2}$. Al aplicar esta fórmula, no se hacen arreglos especiales para ordenar los items, dado que no exige igual variabilidad en las dos mitades de la prueba.

La estrategia de división por mitades es inapropiada cuando se emplean items encadenados, en cuyo caso, la cadena entera debe asignarse a una sola mitad ya que, de lo contrario, la similitud de los resultados puede sobreestimar la confiabilidad. Esta estrategia también resulta inapropiada para pruebas de velocidad que, como requieren una atención especial, serán detalladas más adelante.

Otro modo de establecer la confiabilidad y quizás el más empleado, es el *análisis de consistencia interna*, que se refiere a la administración del instrumento una sola vez a un grupo de estandarización, para luego determinar el grado en que los items están correlacionados, mediante la corrida estadística correspondiente. Si esas correlaciones son altas y positivas se dice que hay consistencia entre los reactivos de una prueba, lo cual implica que el saber cómo se desempeña una persona en un item, permite predecir cómo lo hará en los otros. Si los items guardan pocas correlaciones entre ellos, se dice que la medida es inconsistente (Brown, 1980).

Anastasi y Urbina (1998) señalan que la estrategia de consistencia interna se hace acompañar de dos fuentes de varianza de error: por un lado, el muestreo del contenido del instrumento, por lo que cabe hablar de un *coeficiente de equivalencia* y por otro, la heterogeneidad del área de conducta

muestreada, en atención a lo cual se le alude a un coeficiente de consistencia interna. Mientras más homogénea sea el área de conducta muestreada, más consistencia habrá entre los reactivos, mayor será la posibilidad de que la gente repita su patrón de respuesta y mayor será la confiabilidad de la medida.

La homogeneidad no solo se refiere a la consistencia de todos los items de una prueba psicológica, sino que se puede aplicar a las subpruebas o agrupaciones de reactivos dentro de ella. El nivel apropiado de análisis dependerá de la estructura de la prueba y de la finalidad de dicho análisis. Así, es posible que una prueba heterogénea se pueda componer de cierto número de subpruebas homogéneas.

Para obtener la consistencia interna de una prueba se puede recurrir a los coeficientes *alfa de Cronbach, Kuder Richardson 20 ó 21*, o el coeficiente de *Hoyt*. La elección del estadístico apropiado tiene que ver con el número que genera la medida, con la dificultad de los items o con el análisis de la varianza.

Se emplea *coeficiente alfa de Cronbach* cuando se requiere conocer la consistencia interna de un instrumento compuesto por items tipo Likert, como ocurre en los inventarios de interés y en pruebas de personalidad (Aiken, 1996; Brown, 1980). En estos casos se efectúa una sola administración del instrumento y se procede a emplear esta fórmula general $r_{tt} = \left(\frac{n}{n-1}\right)\frac{s_t^2 - \sum s_i^2}{s_t^2}$, donde $\sum S_i^2$ es la sumatoria de la varianza de los items y S_t^2 la varianza total del instrumento. El *coeficiente alfa* se puede interpretar como la correlación promedio entre pruebas psicológicas de la misma longitud tomadas del mismo dominio. Esto es fundamentalmente teórico; en la práctica sería la correlación promedio entre las posibles divisiones en mitades que se puedan generar del instrumento que se está estudiando.

El coeficiente de confiabilidad de *Kuder-Richardson*, se emplea cuando se requiere conocer la consistencia interna en instrumentos cuyos reactivos son binarios, pero en los que las divisiones en mitades planificadas resultan inconvenientes, como cuando se emplean items encadenados (Brown, 1980). El procedimiento consiste en administrar la prueba en una sola oportunidad y luego, haciendo uso de las ecuaciones de *Kuder y Richardson*, derivadas a partir del *alfa de Cronbach*, se obtiene el coeficiente respectivo.

La fórmula 20 de *Kuder-Richardson* (*KR-20*) $r_{tt} = \left(\frac{n}{n-1}\right)\frac{s_t^2 - \sum pq}{s_t^2}$ (*n* es el número de items) parte del supuesto de que los items poseen diferentes niveles

de dificultad; mientras que la fórmula 21 de *Kuder-Richardson* (*KR-21*) $r_{tt} = \left(\frac{n}{n-1}\right)\frac{n s_t^2 - \bar{X}\left(n - \bar{X}\right)}{n s_t^2}$ parte del supuesto de que todos los reactivos tienen iguales niveles de dificultad y arroja niveles más conservadores de este coeficiente. Tal supuesto raramente se cumple, por lo que esta ecuación se usa poco, siendo más empleada la *KR-20*.

Finalmente, el *coeficiente de Hoyt* se emplea cuando la medida se expresa en una escala continua y se basa en obtener estimaciones de la varianza verdadera y de error a partir de las Medias Cuadradas del Análisis de Varianza. Aquí la Media Cuadrada Residual (MCR) sería el indicador de la varianza de error y la diferencia entre la Media Cuadrada de Las Personas (MCP) y la media cuadrada residual dividida por el número de medidas de los examinados (*k*) será un indicador de la varianza verdadera. La varianza total se estima al sumar los estimadores de las otras dos varianzas. La expresión se puede escribir así: $r_{xx} = \frac{MCP - MCR}{MCP + (k-1)MCR}$.

Antes de cerrar el tema de los tipos de confiabilidad, es menester recordar que los cuatro estudios referidos hasta acá se fundamentan en la teoría clásica de las pruebas, con lo cual se expresan como un coeficiente de confiabilidad, con valores que oscilan entre cero y uno, e informan el porcentaje de varianza de error debido a muestreo de tiempo (test-retest) o a muestreo de contenido (formas alternas, dos mitades o consistencia interna); proveyendo en el primer caso, un coeficiente de estabilidad temporal y en el segundo, un coeficiente de equivalencia.

Especificidad y complementariedad de los métodos de confiabilidad

Cualquier método de confiabilidad no es adecuado para todo proceso de medición, sino que se requiere considerar el objetivo de la prueba, el modo de contestar, la variabilidad de la medida, la homogeneidad del contenido y el tiempo concedido para contestar; de ahí que se hable de la especificidad de los métodos de confiabilidad.

Si el instrumento va a contribuir a la toma de decisiones que deberán ser estables en el tiempo (contratación de personal, por ejemplo), es apropiado recurrir al método de retest o de formas equivalentes demoradas, ya que ellas informan cuán estable será en el tiempo la medida. Si el contenido examinado es homogéneo, se puede recurrir a un estudio de consistencia interna, pero si es heterogéneo, tal estudio puede subestimar la confiabilidad, por lo que quizás lo apropiado sería recurrir a un estudio de retest o a uno de formas

equivalentes sucesivas. Cuando el rasgo medido cambia rápidamente (estrés, por ejemplo) una estrategia de retest arrojará una confiabilidad muy baja, es decir, subestimará al coeficiente, resultando apropiado emplear alguna estrategia de consistencia interna. No así ocurre con variables estables en el tiempo (inteligencia, por ejemplo), en cuyo caso es pertinente emplear formas alternas demoradas o retest.

Si el test es de velocidad, es recomendable emplear cualquier estrategia a excepción de aquéllas que informen acerca de consistencia interna, ya que se sobreestima la confiabilidad debido a que la variable es tan homogénea que la repetibilidad de las respuestas estará garantizada. Si la prueba es de poder, la decisión versará sobre aspectos como, fines de la prueba y homogeneidad del contenido.

Dada la especificidad de la confiabilidad, lo que corresponde al constructor de tests es indicar aspectos como la técnica que empleó para tomar las mediciones, el lapso transcurrido entre las administraciones, referir las posibles experiencias de los evaluados (educativas, laborales, etc.) y describir en detalle la muestra a la que se recurrió para realizar el estudio. El valor del coeficiente de confiabilidad obtenido para el mismo instrumento, pero con otra técnica, situación o muestra, proporcionará información completamente diferente. En tal sentido, es imprescindible hacer las precisiones antes descritas en el manual de prueba (Brown ,1980).

Como se ha dicho, cada método para determinar la confiabilidad capta determinada varianza de error. Así, los métodos de test-retest y formas alternas demoradas, son sensibles al muestreo de tiempo; los métodos de formas equivalentes sucesivas o demoradas, dos mitades y consistencia interna, captan la varianza de error por muestreo de contenido. Entonces, un informe adecuado de la confiabilidad deberá hacer un ejercicio de complementariedad para mostrar todas las evidencias que sean adecuadas para el instrumento particular.

Si por ejemplo se va a llevar a cabo el estudio de un Test de Creatividad que se empleará para hacer selección, lo apropiado sería verificar su confiabilidad a través de un procedimiento de retest para dar cuenta de que arroja medidas repetibles después de un tiempo; y, quizás un estudio de consistencia interna que informe la equivalencia entre los reactivos de la medida.

Factores que influyen en los coeficientes de confiabilidad

Es claro que, si se implementa un estudio de *formas paralelas demoradas*, se obtendrán coeficientes de confiabilidad mucho más bajos que si se ejecuta un estudio de test-retest, ya que el primero es sensible a dos fuentes de varianza de error (muestreo de tiempo y contenido) en tanto el segundo solo recoge una fuente de varianza de error (muestreo de tiempo). Adicional al método mismo, hay otros aspectos que afectan el coeficiente de confiabilidad, ellos son la *homogeneidad de la muestra de personas y la longitud de la prueba*, lo cual es de indispensable conocimiento para los diseñadores, que podrán plantearse distintos escenarios de construcción considerando esos aspectos.

Como se indicó en las derivaciones del modelo de *Spearman*, la varianza verdadera es distinta de una muestra de personas a otra, ya que deviene de las personas examinadas, pero la varianza de error, en tanto depende de la sensibilidad del instrumento para captar fuentes de error, siempre es la misma para una prueba particular independientemente de la muestra a quien se administre. De lo anterior, se desprende que la confiabilidad depende de la heterogeneidad de los puntajes verdaderos (Magnusson, 1975).

Supóngase el ejemplo de que se examina la confiabilidad de una prueba cuya varianza de error es 2 (no varía, pues es propia del instrumento), para una muestra heterogénea cuya varianza total es de 8, y luego para una muestra homogénea cuya varianza obtenida es de 4, se puede observar fácilmente, a partir de la ecuación $r_{tt} = 1 - \frac{s_e^2}{s_t^2}$, que en el caso de la prueba heterogénea la confiabilidad es de ,75 y en el caso de la prueba homogénea es de ,50. Es decir, es más confiable la prueba en el caso de la muestra heterogénea, ya que la varianza de error explica en menor medida la varianza total (25%) en tanto que en la muestra homogénea la explica en mayor medida (50%).

Es posible computar la confiabilidad nueva (r_{uu}) de una prueba que se administra a un grupo cuya homogeneidad varía con relación a la homogeneidad de la muestra original recurriendo a la ecuación $r_{uu} = 1 - \frac{s_t^2(1-r_{tt})}{s_u^2}$ donde el único término nuevo es S^2_u que es la varianza de la nueva muestra. Con esta ecuación, es posible acceder directamente a la información del valor de confiabilidad ante cambios en la muestra de examinados, lo cual es de especial relevancia de tomarse en cuenta la especificidad de la medida.

La confiabilidad de una prueba, también es una función del *número de reactivos que la conforman*. Mientras mayor es el número de items incluidos en un test (supóngase infinitos), mayor es la probabilidad de que los puntajes

de error se cancelen entre sí; de que el puntaje total se acerque al puntaje verdadero de la persona y de que la confiabilidad sea uno (Magnusson, 1975). Distintas ecuaciones permiten conocer qué pasa con cada una de las varianzas antes esos cambios de longitud de la prueba.

El cambio en la varianza total de un test, cuya longitud varía viene dada por la ecuación $S^2_t=nS^2_t[1+(n-1)]r_{tt}$, donde n es la nueva longitud de la prueba y los restantes datos son los estadísticos iniciales del instrumento. Al duplicar el número de items, la n es igual a 2 y la ecuación se expresa así $S^2_{2t}=2S^2_t(1+r_{tt})$.

Cuando la longitud de un instrumento cambia, la varianza de los puntajes verdaderos aumenta en el cuadrado del número de veces que se incrementa esa longitud, como se aprecia en esta ecuación $S^2_{nT}=n^2S^2_T$; y si el número de items se duplica, la ecuación puede expresarse como $S^2_{2T}=4S^2_T$, de donde puede afirmarse que, la varianza verdadera nueva, aumenta cuatro veces con relación a la varianza verdadera del test original.

La varianza de error, al aumentar n veces una prueba, se calcula a partir de la ecuación $S^2_{ne}=nS^2_e$; y cuando la longitud del test es duplicada, la fórmula queda $S^2_{2e}=2S^2_e$, lo que revela que la varianza de error, aumenta proporcionalmente a la longitud del test.

A partir de lo argumentado previamente, es posible concluir que cuando a un test se le agregan items, la varianza verdadera se incrementa como el cuadrado del número de veces que se alarga el test, en tanto que la varianza de error aumenta proporcionalmente a tal incremento de longitud. Como la varianza verdadera aumenta con mayor "rapidez" que la varianza de error, representará en mayor proporción a la varianza total y las medidas del test se harán más confiables.

Se puede dar cuenta del cambio en el valor de la confiabilidad, a través de los cambios a nivel de las varianzas total, verdadera y de error, pero también se puede acceder a esa cuantía directamente a través de la ecuación de *Spearman-Brown* $r_{tt_n}=\frac{n\,r_{tt}}{[1+(n-1)]\,r_{tt}}$, donde n es la nueva longitud de la prueba cuyo impacto se está indagando y rtt es la confiabilidad inicial. De esa misma ecuación, se puede derivar otra fórmula $n=\frac{r_{tt_n}(1-r_{tt})}{r_{tt}(1-r_{tt_n})}$ para conocer qué longitud (n) debe tener un instrumento con una confiabilidad determinada (,80 p.e.). para que tenga un valor de confiabilidad nuevo deseable (,90 p.e.). La n obtenida indicará la longitud nueva para ese valor de confiabilidad buscado. Luego corresponderá verificarlo en la práctica.

Ghiselli (1964) acota que tales ecuaciones no sólo se emplean cuando se incrementa la longitud del instrumento, sino también cuando ella se reduce, con lo cual permite hacer todos los estudios de este tipo que se requieran. Tal información permite a los elaboradores de pruebas plantearse escenarios de construcción y explorar cuántos items agregar o eliminar de la prueba que diseña, para propiciar determinados valores de confiabilidad. Luego corresponderá verificarlo en la práctica.

Velocidad de la prueba y confiabilidad

El estudio de confiabilidad de las medidas tomadas con una prueba de velocidad requiere una atención especial. Como señalan Anastasi y Urbina (1998) una prueba es de velocidad cuando las calificaciones dependen exclusivamente de la rapidez de ejecución de las personas. La prueba está conformada por reactivos de dificultad considerablemente baja para el nivel de habilidad de la muestra a la cual se dirige, pero el límite de tiempo que se concede para responderla es tan reducido que nadie puede terminar todos los items (ninguna persona obtiene calificaciones perfectas). De manera que el resultado de cada individuo sólo revela su velocidad de trabajo.

A los fines de disponer de un indicador de si la prueba es o no de velocidad, Anastasi y Drake (1954) recomiendan el uso de la ecuación de *Gulliksen* $H^2 = \frac{s_c^2}{s_t^2}$. Donde el numerador S^2_c, es la varianza de los reactivos completados por las personas, es decir, número de reactivos contestados por cada persona, menos el promedio de los reactivos contestados, dividido entre el número total de reactivos contestados. El denominador S^2_t, es la varianza total de la prueba. Si tales varianzas son iguales, la razón será 1, concluyéndose que se trata de una prueba de velocidad. También refieren la fórmula *Cronbach y Warrington* $H^2_{(a)} = \frac{s_c^2 \sum x_{\lambda p}^2}{N s_t^2}$ aplicable cuando los items que componen la prueba varían en su nivel de dificultad; donde N es el número de casos y $\sum x_{\lambda p}^2$ es una medida de exactitud de los últimos dos reactivos intentados. El índice obtenido a través de cualquiera de estas ecuaciones revela el porcentaje de la varianza total que es debida a varianza de velocidad.

En dichas pruebas, los coeficientes de confiabilidad obtenidos por los métodos de dos mitades corregidos, o cualquier otra evaluación de consistencia interna (*alfa de Cronbach, Kuder-Richardson o coeficiente de Hoyt*) resultan inapropiados, ya que éstos se fundamentan en la consistencia del número de errores cometidos por el examinado, en tanto las pruebas de velocidad se basan en la rapidez de ejecución de esas personas, ya que nadie

se equivoca. Luego, es improcedente explicar la varianza de velocidad desde la consistencia del número de errores, por lo que siempre es aconsejable emplear retest o formas paralelas para evaluar la confiabilidad de los instrumentos de velocidad.

Cuando sea absolutamente necesario, confiabilizar una prueba de velocidad a través de una sola administración, se puede recurrir a una modalidad de división por mitades, pero en función del tiempo y no de los reactivos. Por esta vía, las puntuaciones se basan en partes de la prueba cronometradas por separado. Una forma de efectuar tal división consiste en aplicar dos mitades equivalentes con distintos límites de tiempo, por ejemplo, se imprimen los reactivos pares e impares por separado y se le asigna a cada conjunto la mitad del tiempo previsto para toda la prueba y, finalmente, debe aplicarse la fórmula de Spearman-Brown para obtener la confiabilidad de la prueba entera. No obstante, más que recurrir a estos artificios, se recomienda usar el método de retest o formas alternas a la hora de trabajar la confiabilidad en pruebas de velocidad.

Confiabilidad por acuerdo entre jueces

En la literatura más heterodoxa, se alude al acuerdo entre jueces entre los estudios de confiabilidad. A pesar de su gran relevancia dentro de la Psicología, donde el uso de instrumentos semi-objetivos es cotidiana, la exposición acerca de este estudio se ha dejado para el final no por infrecuente o irrelevante, sino porque no se ancla dentro del modelo lineal de *Spearman*.

En el andamiaje teórico que se ha expuesto hasta este momento de la presente exposición, se ha conceptualizado la confiabilidad como la estabilidad o exactitud con que se miden los puntajes verdaderos de las personas, a la que se accede por medio de un estudio correlacional que a la postre informa la proporción de varianza de error debida a las fluctuaciones temporales (muestreo de tiempo) o a las diferencias entre conjuntos de reactivos paralelos (muestreo de contenido).

No obstante, dentro de esa propuesta teórica no se considera el aporte de error debida al examinador, que es especialmente relevante en los instrumentos semi-objetivos (subescala de WISC, p. e.), en las pruebas proyectivas, en los registros y en las observaciones conductuales. Para esos instrumentos, un indicador de confiabilidad de sus mediciones es tan necesario como los estudios de confiabilidad tradicionales anteriormente expuestos; con lo cual,

conocer y hacer uso del método de confiabilidad de *acuerdo entre jueces* dentro de la disciplina, es determinante.

Cuando se trata de instrumentos de observación, de registros conductuales o de escalas semi-objetivas, se alude a la confiabilidad como el grado en que dos o más observadores concuerdan en el registro de la ocurrencia del asunto observado.

Son múltiples los aspectos que pueden afectar el registro de las puntuaciones verdaderas de los sujetos y con ello disminuir la confiabilidad. Lacasella (2000) hace referencia a: el sesgo del observador, entendido como su tendencia a cambiar el asunto que inicialmente se decidió atender, la reactividad del que registra, que alude a sus reacciones cuando sabe que está siendo evaluado a través de la confiabilidad de sus registros, las expectativas del observador, que consiste en hacer registros en función de lo que espera o desea como resultado, la habilidad del observador para registrar simultáneamente los distintos aspectos observados. Éstos disminuirán la confiabilidad del instrumento siempre que afecten a uno de los observadores y no al otro, razón por la cual, un buen entrenamiento y ensayos para poder ejecutar el registro es indispensable, con miras a controlar esas fuentes de varianza de error.

Para establecer la confiabilidad entre calificadores, se verifica que éstos entiendan cómo se define el asunto a examinar, se les entrena en el sistema de anotaciones, se les solicita a dos o más de ellos que registren simultáneamente el mismo asunto, y se procesan los datos obtenidos respetando la dimensión relevante de la conducta examinada. Si bien cada investigador puede darse su propio esquema de análisis, ayuda mucho considerar los índices propuestos por Lacasella (2000):

Tabla 1
Dimensión de la conducta e identificación del porcentaje de acuerdos.

Dimensión	Formulas
Frecuencia	$\%\ de\ acuerdos = \frac{Número\ menor\ de\ observaciones}{Número\ mayor\ de\ observaciones} \cdot 100$
Frecuencia y Duración	$\%\ de\ acuerdos = \frac{acuerdos}{acuerdos + desacuerdos} \cdot 100$
Duración	$\%\ de\ acuerdos = \frac{Duración\ acumulada\ menor}{Duración\ acumulada\ mayor} \cdot 100$
Actividades planificadas	$\%\ de\ acuerdos = 1 - \frac{\sum de\ las\ diferencias\ obtenidas\ de\ las\ personas\ en\ la\ tarea}{100\ número\ menor\ de\ presentes}$

Nota. De "*Metodología para el desarrollo del estudio infantil desde la perspectiva conductual*", por R. Lacasella, 2000.

Los criterios anteriores, si bien han sido propuestos en el marco del análisis experimental de la conducta, sirven muy bien a quienes trabajan con pruebas semi-objetivas, quienes también pueden recurrir a otros índices como el de Hambleton y Novick o el coeficiente *Kappa*, que, usados regularmente en el contexto de test criteriales, pueden ser empleados también acá en tanto suponen identificar acuerdo entre calificadores.

El índice de Hambleton y Novick informa la proporción de sujetos consistentemente clasificados por dos evaluadores en las categorías consideradas a través de la ecuación $rtt = \sum_{j=1}^{m} pjj = \frac{N_{11}}{N} + \cdots + \frac{N_{mm}}{N}$. El coeficiente *Kappa*, que también es un indicador de concordancia, informa la proporción de clasificaciones consistentes entre dos evaluadores, más allá que las esperables por el azar $k = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$, donde *Pr(a)* es la proporción de concordancias observadas y *Pr(e)* es la proporción esperada de que K evaluadores coincidan (Martínez, 1996).

En cualquiera de los casos, lo relevante es destacar que este método de confiabilidad, al informar los errores debido al examinador, da cuenta de la repetibilidad de las calificaciones de las personas, independientemente de quien lo examine, lo cual lo hace irrelevante si la prueba es de calificación objetiva, pero indispensable cuando se trate de instrumentos semiobjetivos, proyectivos o registros conductuales.

De lo expuesto en este trabajo se desprende que cada método para establecer confiabilidad es sensible a determinadas fuentes de error, quedando a criterio de quien construye un instrumento seleccionar el método pertinente según la utilidad del mismo.

Referencias bibliográficas

Aiken, L. (2003). *Test psicológicos y evaluación*. (11ra Ed.). México: Prentice Hall.

Anastasi, A. & Drake, J. (1954). An Empirical Comparison of Certain Techniques for Estimating the Reliability of Speeded Tests. *Educational and Psychological Measurement, 14* (3), 529- 540.

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

Brown, F. (1980). *Principios de la Medición en Psicología y Educación*. (4ta Ed.). México: El Manual Moderno.

Cohen, R. & Swerdlik, M. (2001). *Pruebas y Evaluación Psicológica. Introducción a las pruebas y a la Medición*. (4ta Ed.). México: McGraw Hill.

Cureton, E. (1965). Reliability and Validity: Basic Assumptions and Experimental Designs. *Educational and Psychological Measurement, 25*(2), 327-346.

Ghiselli, E. (1964). *Theory of Psychological Measurement*. New York. McGraw Hill.

Lacasella, R. (2000). *Metodología para el desarrollo del estudio infantil desde la perspectiva conductual*. Caracas: Fondo Editorial de la Facultad de Humanidades y Educación. Universidad Central de Venezuela.

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

Martínez, R. (1996). *Psicometría: Teoría de los test psicológicos y educativos*. España: Síntesis

Muñiz, J. (1992) Teoría Clásica de los tests. Madrid: Pirámide.

Nunnally, J. & Bernstein, I. (1996). *Teoría Psicométrica*. (3ra Ed.). México: McGraw Hill.

Santisteban, C. (1990). *Psicometría: Teoría y práctica en la construcción de tests*. Madrid: Ediciones Norma, S.A.