Coeficientes de confiabilidad de instrumentos escritos en el marco de la teoría clásica de los *tests*

Reliability coefficient of written tools in the frame of the classical theory of *tests*

Silvio F. Soler Cárdenas

Maestro en Ciencias en Educación Médica, Profesor Auxiliar, Investigador Agregado, Escuela Nacional de Salud Pública, La Habana, Cuba.

RESUMEN

OBJETIVO: evaluar la confiabilidad en el marco de la teoría clásica de los *tests* y precisar las condiciones bajo las cuales el coeficiente alfa de *Cronbach* constituye la mejor alternativa.
DESARROLLO: el coeficiente alfa de *Cronbach* es el recurso numérico más utilizado para evaluar la confiabilidad de instrumentos escritos. Como prueba de esto, baste decir que desde su publicación en 1951 hasta la fecha, ha sido citado más de 5 590 veces. Sin embargo, se ha comprobado que en no pocas situaciones este coeficiente no es la mejor alternativa para evaluar la confiabilidad. Se discutieron varias alternativas para la evaluación de la confiabilidad y se precisaron las condiciones bajo las cuales el coeficiente alfa de *Cronbach* constituye la mejor alternativa.
CONCLUSIONES: la teoría clásica de los *tests* constituye un modelo apropiado para desarrollar el concepto de confiabilidad y las diferentes fórmulas de cálculo de mayor utilidad en las aplicaciones, el coeficiente alfa de *Cronbach* es el indicador más utilizado para cuantificar la consistencia interna de los *tests* y es necesario tener en cuenta los supuestos que sustentan la correcta aplicación del coeficiente alfa para hacer una interpretación adecuada de su valor numérico.

Palabras clave: Coeficientes de confiabilidad, coeficiente alfa de *Cronbach*, teoría clásica de los *tests*.

ABSTRAC

AIM: to assess reliability in the frame of the classical theory of *tests* and to exactly specify the conditions under which *Cronbach's* alpha coefficient is the best option.
DEVELOPMENT: *Cronbach's* alpha coefficient is the most used numerical resort to evaluate the reliability of written tools. As a proof, it is enough to say that from its publication in 1951 up to now, it has been cited more than 5 590 times. However, it

has been proved that this coefficient is not always the best option to assess reliability. Various alternatives were discussed for the assessment of reliability, and the conditions under which *Cronbach's* alfa coefficient is the best choice were determined.
CONCLUSIONS: the classical theory of tests is an appropriate model to develop the concept of reliability and the calculation formulae more useful for the applications. This coefficient is the most used indicator to quantify the internal consistency of *tests*, and it is necessary to bear in mind the suppositions that support the correct application of alpha coefficient to make a proper interpretation of its numerical value.

Key words: Reliability coefficients, Cronbach's alpha coefficient, classical theory of *tests*.

INTRODUCCIÓN

En la actividad laboral e investigativa de los profesionales de la salud en ocasiones se presenta el problema de elaborar y aplicar instrumentos escritos con la finalidad de cuantificar determinados atributos personales de un grupo de individuos. Como ejemplo, se pueden poner los siguientes casos:

· Un investigador está interesado en conocer en qué medida ha aumentado la calidad de vida de los enfermos de la tercera edad después de someterlos a un determinado régimen de rehabilitación.

· Se quiere evaluar el nivel de desarrollo de habilidades psicomotoras de un grupo de obreros que manejan equipos de izaje en la construcción de puentes.

· Un docente necesita confeccionar un examen escrito para evaluar a un grupo de estudiantes en una asignatura dada.

· Se desea conocer la opinión que tienen los estudiantes de primer año de la carrera de Medicina con respecto a cambios en el plan de estudios.

En todos los casos anteriores, si se aplicara el mismo instrumento 2 o varias veces a las mismas personas, es muy probable que se obtengan resultados diferentes, es decir, de la aplicación reiterada de esos instrumentos se obtiene una serie de puntajes que poseen un determinado grado de variabilidad. Evidentemente, cuanto mayor sea esa variabilidad, menor será la precisión de las conclusiones y generalizaciones derivadas de los resultados del mencionado instrumento.

En general, cuando se aplica un *test* de cualquier tipo (ya sea un *test* de actitudes, un *test* de rendimiento en una tarea específica o simplemente un examen escrito para explorar conocimientos), el puntaje obtenido depende de un conjunto de condiciones internas (propias del examinado) y externas (el medio) y por tanto, el mismo *test* aplicado a la misma persona pero en momentos diferentes puede arrojar puntajes diferentes. En la práctica se presentan frecuentemente numerosos

factores difíciles de controlar y que a la larga determinan la inconsistencia de los puntajes de un *test*.

Cronbach[1] hace referencia a 4 grupos importantes de factores:

1. Características generales y duraderas del examinado.

2. Características duraderas y específicas del examinado.

3. Características generales y momentáneas del examinado.

4. Características temporales y no generales del examinado.

Variados han sido los enfoques presentados en la literatura para cuantificar la inconsistencia de puntajes de tests como consecuencia de la influencia de los factores mencionados anteriormente. En este sentido, *Guilbert*[2] presenta los conceptos de *objetividad, pertinencia, equilibrio, equidad, discriminación* y *eficacia*. No obstante, para este fin hay un concepto que, en esencia, desde que fue definido se mantiene vigente en la teoría y aplicaciones de *tests*: el concepto de *confiabilidad* presentado por el psicólogo británico *Charles Spearman*, en sus trabajos de los años 1904-1913.[3]

DESARROLLO

El problema de los errores de medición de instrumentos evaluativos de los *tests* de rendimiento y en general de los *tests* psicológicos, es un asunto que siempre ha ocupado a los investigadores dedicados a la exploración de atributos psicológicos. A partir de los trabajos desarrollados por el psicólogo británico *Charles Spearman* a principios del siglo xx (1904-1913), se cimentaron las bases de lo que hoy día se conoce como *teoría clásica de los tests* o también, *modelo clásico del puntaje verdadero* (*classical true score model*). Desde entonces ha aparecido en la psicología y en la educación una gran cantidad de literatura relacionada con los conceptos de confiabilidad, validez y otros afines. Varios autores han reformulado el planteamiento original de *Spearman*, permitiendo así que esa teoría se haya desarrollado tanto en los aspectos metodológicos como en las aplicaciones.

A partir de trabajos de *Allen* y *Yen*;[4] *Crocker* y *Algina*;[5] *Steyer*;[6] *Gruijter* y *Kamp*[7] y *Meliá,*[8] se presentan los fundamentos de la teoría clásica mediante los cuales se pueden deducir diferentes fórmulas de cálculo e interpretaciones de la confiabilidad.

Supuestos de la teoría clásica de los *tests*

En sus trabajos de principios del siglo xx, Spearman desarrolló la teoría de los *tests* partiendo de 6 supuestos que se explican a continuación:

La idea básica de su modelo consiste en establecer que cuando se aplica un *test* el *puntaje observado* ("X") se puede expresar como la suma de 2 componentes, uno representa el *puntaje verdadero* ("V") y el otro el *error de medición* ("E").

En símbolos este supuesto se expresa mediante:

$X_p = V_p + E_p$ (I)

En esta igualdad, el miembro izquierdo denotado con el símbolo "X_p", representa el puntaje que se obtiene al aplicar el *test* en cuestión a la persona "p" (puntaje observado); el símbolo "del miembro derecho de la expresión es el puntaje verdadero que posee la persona sometida al test y "E_p" es el error de medición.

Por ejemplo, a Juan, Sara y Ramón se les aplica un *test* de 10 preguntas de tipo verdadero-falso. Para estas personas, el puntaje del *test* se define como la cantidad de respuestas contestadas correctamente, de manera que los puntajes posibles son los números 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 y 10.

Si Juan conoce realmente las respuestas de 7 preguntas; pero por azar ofrece 2 incorrectas (de las 7 que conoce correctamente), su puntaje observado sería igual a "5". En términos del supuesto anterior (I) este hecho se expresa mediante $7 + (-2) = 5 = X_{Juan} = V_{Juan} + E_{Juan}$, es decir, el puntaje verdadero es 7 y el error es igual a -2.

Si Sara conoce solamente las respuestas de 4 *ítem*s pero adivina otros 3 entonces su puntaje sería $7 = 4 + 3 = X_{Sara} + V_{Sara} + E_{Sara}$

Si Ramón conoce las respuestas correctas de 8 *ítems*, pero olvida la de 1 de ellos y adivina otro. En este caso los errores positivos y negativos se cancelan y su puntaje es igual a $8 = 8 + 0 = X_{Ramón} + E_{Ramón}$

En la práctica no es posible conocer de antemano el puntaje real ni tampoco el error de medición. En realidad, el supuesto (I) establece que el puntaje observado (variable directamente observable) se expresa como la suma de 2 variables que no son observables directamente.

Mediante el segundo supuesto se establece la definición de valor verdadero ("V"):

El puntaje verdadero es igual al promedio de todos los puntajes observados que se obtendrían al aplicarle a una persona el mismo *test* una gran cantidad de veces (en teoría se supone que se realizan infinitas aplicaciones).

En símbolos:

$E(X_p) = V_p$ (II)

Si se le aplicara a Juan el *test* una cantidad infinita de veces, el promedio de los puntajes obtenidos sería 7. Para esta definición de V_p se asume que las aplicaciones del *test* son independientes, es decir, en cada una de ellas no influye el resultados de las aplicaciones anteriores. Debido a que en la práctica no es posible garantizar la independencia de los resultados de un *test* ni tampoco su aplicación una cantidad infinita de veces, el puntaje verdadero V_p siempre es una constante desconocida.

Nótese además, que el puntaje verdadero depende vitalmente de cuan fáciles o difíciles son las preguntas del *test* y por tanto, no necesariamente es un número que representa adecua damente la intensidad de un rasgo de la persona en cuestión.

Si por ejemplo, a un estudiante que tiene un bajo nivel de conocimientos en un determinado complejo de materias se le aplica un examen con preguntas muy fáciles, es muy probable que los puntajes observados obtenidos al aplicar muchas veces ese examen serían altos y por ende también su puntaje verdadero. Sucedería lo contrario si se tratase de un examen de preguntas difíciles.

El supuesto (I) establecido anteriormente, expresa claramente la presencia de errores en los puntajes cuando se aplica un *test*. En principio, no hay ninguna razón para suponer que esos errores son similares para diferentes *tests* ni tampoco para un mismo *test* cuando se aplica en diferentes ocasiones. De aquí la necesidad de evaluar la magnitud de esos errores. No obstante, para definir un indicador que exprese el nivel de precisión de un *test* es necesario formular un supuesto de carácter estadístico relativo a los puntajes verdaderos y a los errores de medición.

Se establece que no existe relación lineal entre los errores de medición y los puntajes verdaderos obtenidos en un conjunto de personas cuando se aplica un determinado *test*, es decir, individuos con altos puntajes verdaderos no necesariamente tienen altos valores para los errores de medición (ya sean positivos o negativos).

En términos estadísticos esto se expresa diciendo que *la correlación entre puntajes verdaderos y errores de medición es igual a cero*.

En símbolos:

Corr (V, E)= 0 (III)

Definición de confiabilidad

Cuando se aplica un *test* que cumple los supuestos (I), (II) y (III) a un conjunto de personas con la finalidad de medir determinado atributo, es deseable que los puntajes observados X_p no se aparten mucho de los puntajes verdaderos V_p, o sea que el error de medición sea pequeño. Esta propiedad asociada a la precisión de los puntajes de un *test* es lo que se llama *confiabilidad*.

Naturalmente, en las aplicaciones se exige que los *tests* que se aplican sean confiables pero no siempre es así, de modo que es necesario cuantificar de alguna manera esta propiedad para que en cada caso concreto puedan emitirse juicios con respecto a su magnitud.

En varias ramas de la ciencia que utilizan procedimientos de medición, suelen definirse indicadores basados en la varianza de los datos para cuantificar la confiabilidad. El caso de los *tests* no es una excepción en este sentido y por eso, la definición de confiabilidad requiere la determinación de las varianzas de los puntajes (observados y verdaderos) y del error de medición.

Si se aplica un *test* a un grupo de N personas entonces los puntajes observados y verdaderos, y los errores de medición cumplen las siguientes ecuaciones:

$X_1 = V_1 + E_1$

$X_2 = V_2 + E_2$

$X_N = V_N + E_N$

Si σ^2 (X), σ^2 (*V*) y σ^2 (*E*) denotan respectivamente las varianzas de los puntajes observados y verdaderos y de los errores de medición entonces se cumple que:

σ^2 (X) = σ^2 (*V*) + σ^2 (*E*)

Esta expresión significa que la varianza de los puntajes observados se puede expresar como la suma de las varianzas de puntajes verdaderos y de errores de medición.[8]

Ahora bien, cuanto menor sea la varianza del error σ^2 (*E*), ayor será la precisión de los puntajes del *test* subyacente. En particular si σ^2 (*E*) = 0 entonces no hay error de medición y la varianza de los puntajes observados coincide con la varianza de los puntajes verdaderos, es decir, σ^2 (X) = σ^2 (*V*). De aquí se deduce que un *test* tiene la máxima precisión cuando el cociente σ^2 (*V*)/σ^2 (X) es igual a la unidad, y en la medida en que este cociente sea menor que 1, menor será la precisión del correspondiente *test*. Estas reflexiones sustentan las siguientes definiciones:

Definición 1

Coeficiente de confiabilidad

Si a un grupo de personas se les aplica un *test* que cumple los supuestos (I), (II) y (III) y si σ^2 (X), σ^2 (V) y σ^2 (E) denotan respectivamente las varianzas de los puntajes observados y verdaderos, y de los errores de medición (se supone que σ^2 (X) > 0), entonces el coeficiente de confiabilidad de los puntajes observados de dicho *test* se denota por ρ y se define mediante:

$$\rho = \frac{\sigma^2(V)}{\sigma^2(X)} \quad (*)$$

Nótese que se habla de la confiabilidad de los puntajes observados y no del *test* subyacente. La confiabilidad es una propiedad de los puntajes y no del correspondiente *test*.[9]

Nótese además, que el coeficiente de confiabilidad es un número comprendido entre 0 y 1 y representa la proporción de la varianza de los puntajes observados que le corresponde a la varianza de los puntajes verdaderos.

Por ejemplo, si se sabe que la confiabilidad de un *test* es igual a 0,85, entonces se cumple que 85 % de la varianza de los puntajes observados se debe a la varianza de los puntajes verdaderos.

En las aplicaciones para evaluar la confiabilidad es necesario disponer de un valor estimado del parámetro definido en la igualdad (*). Este es uno de los problemas básicos de la teoría y práctica de los *test* y para su solución se ha propuesto varias alternativas que básicamente consisten en añadir otros supuestos al modelo que define la teoría clásica.

En algunas áreas (antropometría, laboratorios clínicos), es típico evaluar la confiabilidad a partir de repeticiones, en el mismo objeto de procedimientos de medición de interés, o sea, la confiabilidad se concibe como la consistencia de las mediciones cuando se hacen repeticiones. Según *Cronbach,*[10] este era el punto de vista sostenido por varios investigadores durante las primeras décadas del siglo xx. Durante ese tiempo, algunos aplicaban 2 veces el mismo *test* y evaluaban la consistencia de los puntajes obtenidos mediante el cálculo de un coeficiente de correlación; pero este proceder suscitaba dudas porque el resultado de la primera medición podía influir en el de la segunda. Debido a esto se hizo habitual la aplicación de 2 *tests* diferentes pero que, en cierto sentido, exploraban los mismos atributos. Estos son los llamados *tests paralelos*, cuyas variantes de diseño, conducen a diferentes fórmulas para evaluar la confiabilidad.

Definición 2

Tests paralelos

Se plantea que 2 *tests A* y *B* son paralelos (formas paralelas) si ambos cumplen los supuestos (I), (II) y (III) y además:

(IV): Los puntajes verdaderos son iguales.

(V): Las varianzas de los errores de medición son iguales.

(VI): La correlación entre los errores de medición es cero.

Las formas paralelas de un *test* deben ser construidas independientes, pero deben satisfacer ciertas especificaciones: deben contener el mismo número de preguntas, con idénticos formatos y abarcar el mismo tipo de contenido. Asimismo, el nivel de dificultad de cada pregunta debe ser el mismo. Hay que controlar para su comparabilidad, las instrucciones para la aplicación y los límites de tiempo para la aplicación de cada forma.

En términos de los supuestos discutidos anteriormente, se puede decir que la teoría clásica de los *tests* es el modelo que cumple los supuestos (I), (II), (III), (IV), (V) y (VI). A veces se le llama simplemente modelo de *tests paralelos*, para destacar el hecho de que los resultados se deducen partiendo de este tipo de *test*.

¿Cómo se evalúa la confiabilidad de un *test* en el marco de la teoría clásica?

Se supone que el *test* en cuestión cumple los supuestos (I), (II), (III).

Se diseña entonces otro *test* paralelo al inicial; ambos se aplican separadamente y después se calcula el coeficiente de correlación lineal entre los puntajes obtenidos. El número obtenido es el valor del coeficiente de confiabilidad para cada uno de los *tests*.[8]

Diseños para evaluar la confiabilidad de *tests* paralelos

Existen 3 diseños clásicos[5] que proporcionan formas paralelas de un *test*:

1. *Test* repetido.

2. Formas alternas.

3. División en mitades.

El coeficiente de confiabilidad tiene una interpretación diferente para cada una de esas situaciones. Los 2 primeros requieren la aplicación de 2 *tests* (1, 2 veces o 2, 1 vez). Para el tercero es suficiente una aplicación.

Test repetido (coeficiente de estabilidad)

Este diseño consiste en aplicar el mismo *test* en 2 ocasiones diferentes, separadas por cierto lapso de tiempo a un grupo de personas. El coeficiente de confiabilidad es igual al coeficiente de correlación lineal entre los puntajes observados en las 2 aplicaciones; se le llama *coeficiente de estabilidad.*

Formas alternas (coeficiente de equivalencia)

Este diseño consiste en construir 2 formas similares del *test* (forma 1 y forma 2) destinadas al mismo grupo de personas. Las formas deben ser aplicadas dentro de un período corto de tiempo. Para su aplicación se divide aleatoriamente el grupo de personas en 2 mitades; a la primera mitad se le aplica primero la forma 1 y después la forma 2, y para la segunda mitad se invierte el orden de aplicación de las formas.

En este caso el coeficiente de confiabilidad es igual al coeficiente de correlación lineal entre los puntajes observados en ambas formas del *test*. Este coeficiente recibe el nombre de *coeficiente de equivalencia.*

Para la aplicación de este diseño es necesario garantizar que las formas constituyan *tests* paralelos, de lo contrario no hay garantía de que el coeficiente de equivalencia obtenido sea una medida aceptable de la confiabilidad.

División en mitades (coeficiente de consistencia interna)

Para estos procedimientos se supone que el *test* se aplica una sola vez y está formado por varios *ítems* paralelos. Se supone además, que el puntaje observado de cada persona se obtiene sumando los puntajes observados de los *ítems*.

El *test* original se divide en 2 partes que se denominan *test A* y *test B*, de modo tal que cada una tenga la misma cantidad de *ítems* y constituyan formas paralelas. Para cada forma se calcula el puntaje sumando los puntajes de los correspondientes *ítems*. De esta manera, cada persona tiene 2 puntajes, el del *test A* y el del *test B*.

Hay 3 variantes clásicas para evaluar la confiabilidad de este diseño:

1. Fórmula de *Spearman-Brown*.

2. Fórmula de *Rulon*.

3. Fórmula de *Guttman*.

¿Qué relación existe entre los procedimientos de división en mitades?

Las fórmulas de *Rulon* y *Guttman* son numéricamente idénticas, es decir, cuando se aplican a un mismo *test* con una determinada división en 2 mitades, proporcionan el mismo valor para la confiabilidad.

Por otra parte, *Cronbach*[11] demuestra que cuando las varianzas de las 2 mitades son iguales, los procedimientos de *Rulon* y *Spearman-Brown* proporcionan idénticos resultados; aun si el cociente de las desviaciones típicas de los puntajes observados de las mitades se encuentra entre 0,9 y 1,1, ambos procedimientos proporcionan resultados muy parecidos.

No se presentan diferencias apreciables entre los métodos de división en mitades cuando las varianzas de las partes no difieren considerablemente.

Los procedimientos de división en mitades tienen el inconveniente de que proporcionan más de un valor para el coeficiente de confiabilidad y en la práctica resulta bastante difícil seleccionar uno de ellos para cuantificar la consistencia interna. Por ejemplo, un *test* de 10 preguntas se puede dividir en 2 mitades de 126 maneras diferentes, proporcionando otros tantos valores para la confiabilidad. Otro de 20 preguntas, de 4 862 maneras. Para un *test* de 50 preguntas el correspondiente número es 379 231 819 313 256 (una cantidad superior a 379 billones).

Coeficiente alfa de Cronbach

Con la finalidad de resolver el problema de la multiplicidad de valores que proporcionan los métodos de división en mitades, en 1951 *Cronbach*[11] propone el coeficiente alfa:

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum_{i=1}^{k} \sigma^2(i)}{\sigma^2(X)}\right]$$

En esta fórmula la constante *k* representa la cantidad de *ítems* del *test*; σ^2 (X) es la varianza de los puntajes observados y $\sigma^2(X_1)$; $\sigma^2(X_2)$; ... $\sigma^2(X_k)$ son las varianzas de los *ítems*.

Una fórmula numéricamente equivalente a la anterior es:

$$\alpha = \frac{2k}{(k-1)\sigma^2(X)} \sum_{i=1}^{k} \sum_{j \neq i}^{k} \sigma(i)\sigma(j)\rho(i;j)$$

El símbolo ρ (i;j) representa el coeficiente de correlación lineal entre el i-ésimo *ítem*

y el j-ésimo. De aquí que, cuanto mayor sean estas correlaciones, mayor será el valor de la confiabilidad y por tanto, se dice que el coeficiente alfa es una medida de la consistencia interna, la cual depende de las correlaciones entre todos los *ítems* a diferencia de los métodos de división en mitades, en donde solo tiene en cuenta la correlación entre las mitades consideradas. Es por esto que, a menos que se tenga un interés especial en la consistencia interna de 2 partes determinadas de un *test*, el coeficiente alfa de *Cronbach* es preferible a los métodos de división en mitades.

Puede apreciarse que existen varias alternativas para evaluar la confiabilidad de los *tests*. De los 3 diseños clásicos que proporcionan formas paralelas, 2 de ellos, el *test repetido* y el de *formas alternas,* requieren de la aplicación repetida del *test*.

El principal inconveniente de estas variantes reside en la determinación del tiempo que separa a las 2 aplicaciones del *test*. Si el intervalo es bastante corto, las personas recordarán muchas de sus primeras respuestas y entonces, el efecto memoria puede aumentar la magnitud del coeficiente de correlación.

Si el intervalo de tiempo de separación es amplio, entonces se corre el riesgo de que la propiedad que se está midiendo sufra un cambio ostensible y por tanto, la correlación calculada no refleje fielmente la confiabilidad del *test*.

La interpretación de la confiabilidad en estos casos plantea 2 interrogantes:

1. Cuando se obtiene un coeficiente pequeño, ¿significa esto que el *test* proporciona una medición no confiable de la propiedad en cuestión o que ésta ha cambiado durante el tiempo comprendido entre las 2 aplicaciones? Si se estima que lo que mide el *test* es un atributo cambiante, entonces es posible que no se cumpla el supuesto (IV) y no se justifica la evaluación de la confiabilidad mediante un coeficiente de correlación.

2. ¿Cómo interpretar la correlación entre los puntajes observados si en las respuestas de las personas en la segunda aplicación influyen efectos inducidos por la primera como son: memoria, práctica, aprendizaje y otros aspectos? En este caso no es razonable suponer que el coeficiente de correlación constituya una medida adecuada de la confiabilidad.

En algunas áreas (como psicología y educación), no es razonable aplicar un *test* 2 veces a un mismo grupo de personas y por tanto, se descartan los 2 procedimientos anteriores. Esta fue la situación que enfrentó *Spearman* en sus investigaciones de principios del siglo xx y por eso, inventó el procedimiento de división en mitades que proporciona 2 puntajes para un mismo *test*.[10] Posteriormente, aparecieron en la literatura otras fórmulas basadas en este procedimiento que, aunque distintas, generalmente proporcionaban resultados numéricos aproximadamente iguales.

El coeficiente alfa de *Cronbach* es otra de las fórmulas propuestas para cuando el *test* se aplica solamente una vez. Este coeficiente es indudablemente, el recurso numérico más utilizado para evaluar la consistencia interna: en primer lugar, el trabajo donde se presenta al coeficiente alfa (publicado en 1951) fue citado alrededor de 131 veces anualmente durante el quinquenio 1995-2000. Para tener una idea de cuan grande es ese número, baste tener en cuenta que en ese período el número promedio anual de citas de un artículo de ciencias sociales fue de 11.[10]

En segundo lugar, *Hogan*, *Benjamin*, y *Brezinski*,[12] reportan la frecuencia de aplicación de varios tipos de coeficientes de confiabilidad que aparecen en una muestra sistemática seleccionada de la *APA-Published Directory of Unpublished Experimental Mental Measures*. Este directorio comprende 37 revistas profesionales de educación, psicología y sociología. Al coeficiente alfa le correspondió el mayor porcentaje (66,5 %); el segundo lugar lo ocupó el *test-retest* (19,0 %) y todos los demás se aplicaron en menos del 5 % de los casos.

En tercer lugar, *Liu* y *Zumbo*[13] reportan que en una revisión del *Social Sciences Citations Index* del período 1966-1995, el artículo de *Cronbach* de 1951 había sido citado aproximadamente 60 veces por año en un total de 278 revistas que cubren varias áreas de investigación relacionadas con psicología, educación, sociología, estadística, medicina, enfermería, ciencias políticas y economía.

Es probable que la gran cantidad de aplicaciones del coeficiente alfa como medida de la consistencia interna se deba a la relación que tiene con los métodos de división en mitades y con varios de los coeficientes definidos en la primera mitad del siglo xx. En cuanto a esto, *Cronbach*[10] demostró que el coeficiente alfa:

· Es igual al promedio de todos los coeficientes que se obtienen al aplicar el procedimiento de *Rulon* a todas las divisiones posibles del *test* en 2 mitades. En otras palabras, es una medida resumen de los datos que se obtienen al aplicar la fórmula de *Rulon* a todas las divisiones de un *test*.

· Tiene como casos particulares a los coeficientes reportados por *Kuder* y *Richardson* (1937); *Dressel* (1940); *Hoyt* (1941); J*ackson* y *Ferguson* (1941) y *Guttman* (1945).

· Se puede aplicar a un *test* con varios tipos de formato (preguntas dicotómicas; de selección múltiple; de enlace y de respuestas cortas).

Supuestos para la aplicación del coeficiente alfa

En el anteriormente mencionado artículo de 1951, *Cronbach* no precisó los supuestos que garantizan la aplicación correcta de su coeficiente. Concretamente, expresó que "… en este artículo, suponemos dada la fórmula de alfa y no establecemos ningún supuesto con respecto a ella. A pesar de que diferentes autores han establecido sus propios conjuntos de axiomas, nosotros procedimos en la dirección opuesta, examinando las propiedades de a y arribando a una interpretación".[10]

No obstante, el alfa de *Cronbach* dista mucho de ser un coeficiente adecuado para medir la consistencia interna de todo tipo de *test*, *Gruijter*[7] demuestra que si se cumplen los supuestos (I), (II) y (III) de la teoría clásica, entonces el valor de a es menor o igual que el coeficiente de confiabilidad. Por otra parte *Cruz*[14] destaca que el coeficiente alfa puede tomar valores negativos, o sea, es un coeficiente que puede estimar por defecto la confiabilidad hasta tal punto que puede arrojar resultados absurdos.

Para la aplicación del coeficiente alfa es necesario que se cumplan las siguientes condiciones:

1. Se aplica un *test* integrado por al menos 2 *ítems* a un grupo de personas (como mínimo 2 personas).

2. El puntaje del *test* es igual a la suma de los puntajes de los *ítems*.

3. Se cumplen los supuestos (I), (II) y (III) de la teoría clásica de los *tests*.

4. Los puntajes verdaderos de los *ítems* difieren en una constante (que puede ser nula).

5. Los errores de medición de los *ítems* son incorrelacionados.

De las condiciones anteriores se deduce que si los *ítems* del *test* son paralelos entonces alfa coincide con el coeficiente de confiabilidad.

Si no se cumple la condición "1", es imposible calcular alfa.

Si falla la condición "2", no tiene sentido el cálculo de alfa.

Sobre la base de un estudio de simulación, *Zimmerman* y colaboradores[15] demuestran que el coeficiente alfa subestima la confiabilidad cuando no se cumple la condición "4" y la sobreestima cuando se incumple la condición "5".

CONCLUSIONES

· La teoría clásica de los *tests* constituye un modelo apropiado para desarrollar el concepto de confiabilidad y las diferentes fórmulas de cálculo de mayor utilidad en las aplicaciones.

· El coeficiente alfa de *Cronbach* es el indicador más utilizado para cuantificar la consistencia interna de los *tests*.

· Es necesario tener en cuenta los supuestos que sustentan la correcta aplicación del coeficiente alfa para hacer una interpretación adecuada de su valor numérico.

REFERENCIAS BIBLIOGRÁFICAS

1. Cronbach LC. Fundamentos de la exploración psicológica. Instituto Cubano del Libro. La Habana: Edición Revolucionaria; 1968.

2. Guilbert JJ. Guía pedagógica. Reimpreso por la Organización Panamericana de la Salud. Ginebra: Organización Mundial de la Salud; 1977.

3. Williams RH, Zimmerman DW, Zumbo BD, Ross D. Charles Spearman. British Behavioral Scientist. Human Nature Review. 2003;3:114-8 (12 march). [En línea]. [Acceso: 10 de junio de 2004]. URL disponible en: http://human-nature.com /nibbs/03/ spearman. html

4. Allen MJ, Yen WM. Introduction to measurement theory. California: Brooks/Cole Publishing Company; 1979.

5. Crocker L, Algina L. Introduction to classical and modern Test Theory. USA: Harcourt Brace Jovanovich College Publishers; 1986.

6. Steyer R. Classical (Psychometric) Test Theory. [En línea]. [Acceso: 12 de diciembre de 2003]. URL disponible en: http://www2.uni-jena.de/svw/metheval/materialien/publikationen/ctt.pdf

7. Gruijter DN, Kamp LJ. Statistical test theory for education and psychology. [En línea]. [Acceso: 3 de febrero de 2006]. URL disponible en: http://icloniis.fsw.leidenuniv.nl/gruijter/statistical%20test%20theory%20for%20education%20and%20psychology.pdf

8. Meliá JL. Teoría de la Fiabilidad y la Validez. Valencia: Cristóbal Serrano; 2001. www.uv.es/psicometria2

9. Brennan RL. An essay on the history and future of reliability from the perspective of replications. Journal of Educational Measurement. 2001;38(4):295-317.

10. Cronbach L. My current thoughts on coefficient alpha and successor procedures. [En línea]. [Acceso: 22 de mayo de 2004]. URL disponible en: http://www.stanford.edu/dept/SUSE/SEAL/Reports_Papers/My%20Current%20ThoughtsSubmit.doc

11. Cronbach L. Coefficient alpha and the internal structure of tests. [En línea]. [Acceso: 10 de junio de 2004]. URL disponible en: http://www.unc.edu/~rcm/psy330/cronbach.1951.pdf

12. Thomas P. Hogan, Amy Benjamin, Kristen L. Brezinski. Reliability methods: a note on the frequency of use of various types. [Acceso: 22 de septiembre de 2006]. URL disponible en: http://epm.sagepub.com/cgi/reprint/60/4/523.pdf

13. Liu Y, Zumbo BD. The impact of outliers on Cronbach's coefficient alpha estimate of reliability: visual analogue scales. 2007;67;620 Educational and psychological measurement. [Acceso: 14 de noviembre de 2007]. URL disponible en: http://epm.sagepub.com/cgi/reprint/67/4/620

14. Cruz D, Helmstadter G. The problem of negative reliabilities. [Acceso: 3 de noviembre de 2006]. URL disponible en: http://epm.sagepub.com/cgi/content/abstract/53/3/643

15. Zimmerman DW, Zumbo BD, Lalonde C. Coefficient alpha as an estimate of Test Reliability under violation of two assumptions. Educational and psychological measurement. 1993;53;33. [Acceso : 14 de noviembre de 2007]. URL disponible en: http://epm.sagepub.com/cgi/reprint/53/1/33

Recibido: 31 de marzo de 2008.
Aprobado: 23 de marzo de 2008.

Silvio F. Soler Cárdenas. Escuela Nacional de Salud Pública, Calle I esq. a Línea, La Habana, Cuba. E-mail: ssolercu@infomed.sld.cu