Avances en Medición, 3, 9-28
2005

INTERPRETACIONES DEL COEFICIENTE ALPHA DE CRONBACH

Víctor H. Cervantes[*]
Universidad Nacional de Colombia, Colombia

Resumen

El coeficiente α posee una larga historia dentro del desarrollo de la Teoría Clásica en psicometría, y como tal, ha recibido la atención de una amplia variedad de investigadores con el fin de comprender mejor su funcionamiento. Las diferentes interpretaciones que ha recibido dentro de la teoría de la confiabilidad son discutidas; así como diversos estudios que pretenden dar cuenta de las condiciones necesarias para su buen uso. Este trabajo resalta las limitaciones que se encuentran en el empleo de este estadístico en tanto estimador de la confiabilidad de una prueba.

Palabras clave: *Alpha de Cronbach, Teoría de la Confiabilidad, Teoría Clásica de los Tests*

Abstract

Within the history of Classical Test Theory, coefficient alpha has had a large role; thus, several researchers have investigated it in order to better understand its behaviour. The present article discusses different ways in which this coefficient has been interpreted in psychometrical reliability theory, as well as several studies which provide some guidelines on its proper use. The limitations encountered on using alpha as a reliability estimator under different actual test circumstances are highlighted.

Key words: *Cronbach's alpha, reliability theory, Classical Test Theory*

Para todo psicólogo, tanto investigador como profesional, resulta de vital importancia contar con instrumentos válidos y confiables. Por esta razón, cuando un investigador desarrolla una prueba psicológica lleva a cabo una evaluación rigurosa de las propiedades psicométricas de dicho instrumento. En primer lugar, evalúa la presencia de un nivel apropiado de confiabilidad, condición *sine qua non* puede obtenerse un instrumento válido y útil. Así, en el proceso de construcción de una prueba psicológica, la evaluación de la confiabilidad de la misma es un paso imprescindible para que pueda ser utilizada en la medición del atributo de interés. En este sentido, la estimación de la confiabilidad ha recibido un amplio interés desde la Teoría Clásica de los Tests (en adelante TCT). En la medida en que la confiabilidad de una prueba se halla definida como la precisión de los resultados obtenidos por medio de su aplicación, es decir, el grado en que la prueba se ve o no afectada por los diferentes errores aleatorios de medición, es necesario estimar el tamaño de estos errores. Desde la TCT se han trabajado varias fuentes de error de medición. La estimación del error asociado con alguna de ellas permite evaluar un aspecto diferente de la confiabilidad de una prueba.

[*] Departamento de Psicología. Universidad Nacional de Colombia. Bogotá, Colombia.
e-mail: vhcervantesb@unal.edu.co

Entre las principales fuentes de error de medición que se encuentran reseñadas en los libros clásicos de psicometría (v.g. Anastasi, 1954,1990; Nunnally, 1970; Brown, 1980; Thorndike, 1989,1996) es posible encontrar las siguientes: (a) los sujetos, (b) los ítems, (c) los evaluadores, (d) la situación de aplicación, (e) las interacciones entre estas fuentes, (f) el error aleatorio "puro". Las estimaciones tradicionales de la confiabilidad se hallan ligadas con los tres primeros tipos de fuentes de error, mientras se procura que los errores introducidos por la situación de aplicación sean minimizados mediante una estricta estandarización de la misma. Ligado con la estimación del error producido por los evaluadores se encuentra el coeficiente de confiabilidad interevaluadores; con la estimación del error originado por el conjunto de ítems empleados, contrastado con otros conjuntos de ítems posibles, se encuentra el coeficiente de equivalencia; y con el error debido a fluctuaciones de los individuos en el tiempo, se encuentra el coeficiente de estabilidad. El cálculo de estos coeficientes requiere que la prueba sea aplicada dos o más veces al mismo grupo de personas. Así, la aplicación de la prueba por dos o más evaluadores permite obtener una estimación de la confiabilidad interevaluadores (nótese que los observadores pueden, también, producir la evaluación a partir de una sola aplicación en la que califiquen de forma independiente a los sujetos, por ejemplo, empleando un video de las conductas que conforman los ítems); la aplicación de dos pruebas alternativas, compuestas por ítems construidos a partir de la misma definición de un atributo, permite la estimación del coeficiente de equivalencia de ambas pruebas; y la aplicación en dos momentos diferentes, de la misma prueba, permite estimar su estabilidad.

Junto con estas formas de evaluar la confiabilidad se encuentra, también, la evaluación por la consistencia interna de la prueba. Esta estimación indica la intercorrelación entre los distintos componentes de la prueba y, en este sentido, separa del conjunto la variación que corresponde a factores comunes de los ítems y la que corresponde a factores únicos de cada uno de ellos. Así entendida, la confiabilidad por consistencia interna puede tomarse como una forma de estimación de la equivalencia de los componentes entre sí y su estimación será entonces un coeficiente de equivalencia calculado a partir de una sola aplicación de la prueba (ver por ej. Gerbing & Anderson, 1988; Schmidt, Le & Ilies, 2003). En 1951, Cronbach propuso el coeficiente α como un estimador de este índice de equivalencia, con el cual generalizó un conjunto de diferentes métodos que se empleaban en la época para tal fin (Muñiz, 1996). Desde entonces, el uso del α se ha venido generalizando no sólo en la Psicología, sino también en gran cantidad de áreas afines dentro de las ciencias sociales, de la salud y de la estadística, como el estimador por excelencia de la confiabilidad de un compuesto de otras mediciones. Muestra de ello es que, tan sólo el artículo de Cronbach (1951) fue citado un promedio de 60 veces anuales entre 1966 y 1990 (ver Cortina, 1993), más la cantidad de estudios en los que se emplea sin llegar a citar la literatura para justificar su utilización, por tratarse de un estadístico de amplio uso (ver p. ej. Peterson, 1994). Junto con la generalización del empleo del α se ha producido, asimismo, la propagación de diversos usos de este estadístico (Cortina, 1993; Schmidt et al., 2003; Streiner, 2003). Ellos incluyen su utilización como estimador de la consistencia interna, estimador de la homogeneidad de un conjunto de variables, indicador de unidimensionalidad, evidencia de la calidad de una prueba, índice de validez de medidas psicológicas, estimador de confiabilidad interobservadores, entre

otros (Webster, 1957; Green, Lissitz & Mulaik, 1977; Cortina, 1993; Bernardi, 1994; Peterson, 1994; Iacobucci, 2001; Schmidt et al., 2003; Streiner, 2003). Podrá verse que no todos estos usos del coeficiente α son afortunados o pueden no resultar efectivos mediante su uso exclusivo. Actualmente se recomienda que la utilización del coeficiente α sea llevada a cabo con precaución, pues no es tan versátil como la variedad de los usos que se le han dado podría indicar. Se resalta su incapacidad para estimar el "error temporal" (transient error o error debido a fluctuaciones temporales en los individuos), la poca información que provee para determinar la existencia de unidimensionalidad, su dependencia de la cantidad de componentes que conforman la prueba, la alta variabilidad de su estimación muestral, la poca robustez de los estimadores muestrales de la varianza a partir de los cuales se obtiene, el sesgo que lo afecta cuando hay desviaciones de alguno(s) de los supuestos en que se apoya, y así sucesivamente. Dadas estas dificultades, tanto el coeficiente α como su estimador muestral ($\hat{\alpha}$), se han estudiado bajo diferentes condiciones para evaluar su comportamiento y proveer guías de acción para su uso.

Así, el coeficiente α y/o el $\hat{\alpha}$ han sido estudiados en lo que respecta al grado de homogeneidad de los componentes o ítems que conforman la prueba (Green et al., 1977; Cortina, 1993; Zimmerman, Zumbo & Lalonde, 1993; Osburn, 2000), a la presencia de correlaciones entre los errores de los ítems (Zimmerman et al., 1993; Komaroff, 1997; Raykov, 2001), a la cantidad de componentes (Niemi, Carmines & McIver, 1986; Cortina, 1993; Zimmerman et al., 1993), a la presencia de error temporal "no trivial" (Schmidt et al., 2003), a la distribución de los componentes (Christmann & Van Aelst, 2002; Yuan & Bentler, 2002; Norris & Aroian, 2004), a la distribución del atributo (Zimmerman et al., 1993); así como sus relaciones con otros estadísticos propuestos para la estimación de la consistencia interna de una prueba (Osburn, 2000). Sin embargo, dos constantes son comúnes a todas estas aproximaciones al funcionamiento del coeficiente α: por una parte, la evaluación de los efectos de los distintos factores sólo llega a una descripción de la tendencia encontrada, sin analizar el tamaño del efecto asociado; y por otra, aún más importante y derivada de la primera, las conclusiones que se le ofrecen como resultado al usuario del estadístico no superan una recomendación de caución en su uso. Adicionalmente, algunas de estas investigaciones encuentran resultados opuestos, por ej. sobre la distribución de los componentes; pero, al tener metodologías bastante disímiles entre sí sus resultados tampoco resultan directamente comparables.

Conceptos básicos de la Teoría Clásica de los Tests

El objeto principal de la Psicometría se encuentra en el estudio de "los atributos de las personas, en las pruebas que intentan medir estos atributos y en los ítems particulares que [. . .] componen las pruebas" (Thorndike, 1989,1996, p. 16). De este modo, uno de los principales objetivos de la Psicometría es el desarrollo de modelos que describan el proceso de la medición de estos atributos de forma eficiente. La TCT es el modelo más antiguo y fructífero propuesto dentro de este campo. Así, hoy en día su utilización se halla difundida entre psicólogos, educadores y otros profesionales como la herramienta apropiada para el

análisis de las pruebas psicológicas. Su empleo y evolución a lo largo del s. XX han mostrado las fortalezas y debilidades que posee (Muñiz, 1996; Herrera, Sánchez & Jiménez-Ávila, 2000; Gómez Benito & Hidalgo Montesinos, 2003)

La perspectiva clásica parte de la consideración del resultado obtenido por un individuo en una ocasión específica como el resultado de la adición de dos componentes: la puntuación verdadera del sujeto en la prueba y el error aleatorio de medida (Gulliksen, 1950; Lord & Novick, 1968; Muñiz, 1996; Brennan, 2001; Arévalo, 2002). En psicología y en otras ciencias sociales es normal no poder contar con un gran número de mediciones de un mismo sujeto; es común, incluso, no poseer resultados para una persona mas que de una sola ocasión o de una sola prueba; y en algunos casos sólo se poseen mediciones para cada individuo en una sola prueba de una única aplicación de la misma (Kane, 1996). De este modo, para la mayor parte de los casos el modelo puede expresarse como:

$$X_i = T_i + \varepsilon_i , \qquad (1)$$

donde: X_i es el puntaje obtenido por un sujeto *i* en una prueba en una ocasión.

T_i es la puntuación verdadera del sujeto en la prueba, este valor se considera constante en diferentes ocasiones para un sujeto *i* y una prueba fijos.

ε_i es el error aleatorio de medición en la ocasión para un sujeto *i* en una prueba.

En esta relación, el único término directamente observable es X_i, por lo cual la estimación de T_i y de ε_i se torna en el problema central del trabajo con este modelo. Ahora bien, nótese que en tanto que X_i es conocido, basta con definir sólo uno de los dos términos restantes para describir completamente una medición. La TCT opta por describir más detalladamente el componente de error a partir de la inclusión de algunos supuestos sobre su comportamiento a lo largo de un conjunto de mediciones en una población (Gulliksen, 1950; Lord & Novick, 1968; Nunnally, 1987; Thorndike, 1989,1996; Muñiz, 1996; Martínez, 1996; Herrera et al., 2000). Estos son los siguientes: (a) el valor esperado del error aleatorio es igual a cero (Ecuación 2), (b) el error se distribuye normalmente con media cero y varianza σ_ε^2 (Ecuación 3), (c) el error aleatorio de medición en una prueba no se encuentra correlacionado con la puntuación verdadera en la prueba, con el error de medición en otra prueba ni con la puntuación verdadera en otra prueba (Ecuación 4), y (d) las varianzas de las puntuaciones observadas, las puntuaciones verdaderas y del error son finitas y mayores que cero (Ecuación 5). Este último es incluído para evitar la trivialidad del modelo y de los demás supuestos.

$$E(\varepsilon) = 0 \qquad (2)$$

$$\varepsilon \approx n(0, \sigma_\varepsilon^2) \qquad (3)$$

$$\rho_{T,\varepsilon} = \rho_{\varepsilon_1,\varepsilon_2} = \rho_{\varepsilon_1,T_2} = 0 \qquad (4)$$

$$0 < \sigma_h^2 < \infty \quad \text{para} \quad h = T, X, \varepsilon \qquad (5)$$

En Thorndike (1989,1996) podemos ver que estos desarrollos conforman la base para la teoría de la confiabilidad de la TCT --aspecto que se disipa en otros textos y se confunde con

la Teoría Clásica en su conjunto. En este sentido, el modelo en el que ésta se apoya es, principalmente, un modelo sobre cómo el error aleatorio afecta la medición psicológica; así, puede verse que, previa cualquier interpretación del puntaje obtenido por una persona en una prueba, se requiere de una estimación de dicho error para la prueba en cuestión. Ciertamente, ésta es una característica inherente a todo procedimiento de medición [el hallarse afectado por errores aleatorios] y, dadas las características de los objetos de estudio de la psicología, se modela de la forma hasta aquí presentada (Lord & Novick, 1968; Nunnally, 1987; Herrera et al., 2000). Ahora bien, dentro de la TCT se consideran dos tipos de errores que afectan la medición: el error aleatorio y el error sistemático. Lo que resta de este documento se concentra exclusivamente en la teoría de la confiabilidad, la cual provee la definición del primero de estos tipos de error, así como de sus fuentes. El segundo tipo de error es tratado por la TCT en sus desarrollos de la teoría de la validez de las mediciones psicológicas.

De los supuestos del modelo se sigue que la variación de los puntajes observados en una prueba fija para una población es expresable como la suma de las varianzas en las puntuaciones verdaderas y el error aleatorio:

$$\sigma_X^2 = \sigma_T^2 + \sigma_\varepsilon^2 \qquad (6)$$

De estos componentes, el que en la práctica nos interesa es el correspondiente a σ_T^2, mas a partir de la aplicación de una prueba conocemos únicamente σ_X^2. La TCT, con el concepto de confiabilidad, pone en relación estas dos cantidades para estimar σ_T^2; este concepto se define por la correlación cuadrática entre el puntaje observado y la puntuación verdadera ($\rho_{X,T}^2$). Dados los supuestos, este coeficiente es igual a:

$$\rho_{X,T}^2 = \frac{\sigma_T^2}{\sigma_X^2} = 1 - \frac{\sigma_\varepsilon^2}{\sigma_X^2} \qquad (7)$$

es decir, la confiabilidad de una prueba equivale a la razón entre la varianza de las puntuaciones verdaderas y la varianza de los puntajes observados, y en esta medida expresa la proporción de la varianza observada en un grupo de puntajes que corresponde o puede ser atribuída a las variaciones entre las puntuaciones libres de error (Lord & Novick, 1968; Nunnally, 1987; Thorndike, 1989,1996; Muñiz, 1996).

Ahora bien, una cosa es indicar cómo todo proceso de medida se ve afectado por errores; algo distinto es cuantificar el error y especificar las condiciones de medida que contribuyen a él. Lograr estas dos últimas requiere de la especificación de las condiciones reales en las que se obtienen los puntajes observados y de qué es lo que constituye una medida "ideal" de un sujeto (Brennan, 2001). Para la TCT la variable de error se considera conformada por un compuesto de factores no controlados en el procedimiento de medición o que no tienen que ver con el objetivo de la misma. De este modo, cuando algún factor se considerado irrelevante para una cierta medición es tomado como parte del error, de forma análoga, si para una medición u objetivo diferente, dicho factor resulta relevante, se le tomará como parte de la puntuación verdadera (Anastasi, 1954,1990; Thorndike, 1989,1996). Así, tanto el error como la

puntuación verdadera son determinados por la situación y objetivo específicos en los que se realiza la medición (Lord & Novick, 1968). Conviene entonces, revisar cuáles son los factores que desde la TCT se han considerado como fuentes relevantes de error `aleatorio' de medida. Se considera que un procedimiento de medición es confiable si para un individuo una medida puede ser replicada, bien sea en una ocasión diferente, o que la calificación sea realizada por otra persona, o que para obtener la medida se emplee una prueba paralela o similar; o que resultados similares puedan ser obtenidos para diferentes personas con la misma cantidad o presencia del atributo que se busca medir (Anastasi, 1954,1990; Nunnally, 1970, 1987; Brown, 1980; Thorndike, 1989,1996; Streiner, 1993). Bajo esta concepción es evidente que para la TCT las fuentes que pueden introducir error en el proceso de medición y que merecen ser considerados como relevantes en la utilización del modelo que desde ella se propone (ver Ecuación 1), son: (a) los sujetos, (b) los ítems, (c) los evaluadores, y (d) la situación de aplicación (Cortina, 1993).

Este modo de aproximarse a la confiabilidad implica también que para un cierto procedimiento de medición (v.g. una prueba) habrá un coeficiente de confiabilidad diferente para cada una de las fuentes de error relevantes para el mismo. Dentro de cada una de las fuentes de esta clasificación es posible ubicar diversos factores o facetas que, cuando se tiene la oportunidad de estimar, conducen a diferentes aspectos de la confiabilidad. Así, entre los errores cuya fuente está dada por los sujetos se encuentran: las variaciones momentáneas en los evaluados (como variaciones de su atención), aquellas fluctuaciones en el tiempo que no son debidas a la puntuación verdadera del sujeto pero que afectan su ejecución (como cambios en el estado de ánimo, fatiga, sueño, etc.), características específicas del individuo (como su actitud hacia el empleo de pruebas psicológicas). Dentro de los errores originados en los ítems están los efectos del muestreo de los mismos (i.e. errores que están asociados con factores específicos de los ítems seleccionados), características propias de cada ítem que hacen que algunas personas reaccionen ante ellos de formas particulares (por ejemplo, que se haya empleado una cierta palabra en la redacción y no otra de sentido similar). Entre los errores originados por los evaluadores se encuentran las diferencias en la comprensión que poseen del atributo, así como errores análogos a los que tienen por fuente a los sujetos. Dentro de los errores cuya fuente es la situación de aplicación se encuentran las condiciones ambientales (como el clima, la iluminación, el ruido del lugar de aplicación), las diferencias en las diferentes sesiones de aplicación (como variaciones en las instrucciones verbales, errores tipográficos en las pruebas).

Es importante anotar que no todas las fuentes ni tipos de error son relevantes para un uso específico de una prueba y, por lo tanto, de acuerdo con el uso que dicho procedimiento de medida vaya a tener será necesario evaluar sólo alguno(s) de estos. En este sentido, los diferentes aspectos que pueden estimarse dependen del error que el investigador o el usuario de la prueba consideren relevante (Streiner, 2003). A partir de lo reseñado hasta aquí podemos ver que un procedimiento de medición tendrá multiples coeficientes de confiabilidad asociados con los diferentes usos a los que se desee someter en una población. Adicionalmente, el procedimiento que se sigue para evaluar la confiabilidad incluye: "el razonamiento básico [que lo sostiene], los procedimientos para recabar información y los procedimientos estadísticos

para analizarla" (Thorndike, 1989,1996, p. 178). El razonamiento del cual se parte para la evaluación de cada aspecto de la confiabilidad debe identificar los factores que se asocian con esa fuente de error que pueden ser controlados experimentalmente; los factores que no pueden ser controlados pero sí pueden ser aislados por el experimento y los que se consideran parte de la puntuación verdadera para el uso específico de la prueba [nótese que aquellos factores que no logren ser identificados y asignados correctamente terminan subsumidos en esta última categoría] A partir de este punto habrán de especificarse los procedimientos mediante los cuales se puede obtener información sobre los factores "aislables" y, finalmente, qué procedimientos estadísticos permiten estimar la confiabilidad en cada caso.

Cuando se consideran los factores que dan origen a errores de medición debidos a los sujetos, aquel que recibe mayor atención, en tanto que no puede ser controlado y que no se considera parte de la puntuación verdadera, es aquel que incluye las fluctuaciones de los sujetos en el tiempo; por su parte, los errores debidos a distracciones, se considera que pueden ser eliminados empíricamente mediante la sumatoria de los puntajes de los ítems (ver p. ej. Anastasi, 1954,1990; Schmidt et al., 2003). Para obtener información sobre estas variaciones temporales en los sujetos se necesita de la aplicación de la prueba en [por lo menos] dos ocasiones. Los puntajes así obtenidos son correlacionados entre sí, usualmente con el uso de la correlación de Pearson, y su valor es tomado como el coeficiente de confiabilidad, también llamado coeficiente de estabilidad (Anastasi, 1954,1990; Nunnally, 1987; Brown, 1980; Thorndike, 1989,1996).

En cuanto a los errores originados en los ítems, aquellos cuya fuente tiene lugar en factores únicos de los mismos, tales como la redacción, suelen verse como controlables empíricamente mediante la utilización de una prueba únicamente con la población para la cual fue diseñada y su adaptación [y análisis respectivo] para otras poblaciones. El error debido al muestreo de los ítems es de especial importancia en aquellas pruebas diseñadas bajo el modelo de rasgo latente, en donde se considera que las conductas observadas por medio de los ítems seleccionados son un efecto de un rasgo no ebservable directamente en las personas (Thorndike, 1989,1996; Bollen & Lennox, 1991). La estimación de la confiabilidad asociada con este error requiere de la aplicación de dos formas paralelas o similares de la misma simultáneamente; la correlación entre los puntajes observados en estas dos formas de la prueba constituye un estimador de la confiabilidad de ambas y se conoce como coeficiente de equivalencia (Lord & Novick, 1968; Anastasi, 1954,1990; Nunnally, 1987).

Asociada con la estimación del error producido por los evaluadores está la necesidad de separar la varianza aportada por los examinadores a la calificación en aquellas situaciones donde existe evidencia de que ésta resulta ostensible. Casos que precisan de la obtención de este índice abarcan instumentos de evaluación clínica, pruebas de creatividad o pruebas proyectivas de personalidad, en las que el juicio del examinador interviene en gran medida en el momento de obtener la calificación o puntuación observada de una persona (Anastasi, 1954,1990). La estimación de este coeficiente requiere de la calificación de cada sujeto por dos o más examinadores y de la correlación entre estos puntajes; para esta correlación se

recomienda el uso de un coeficiente de correlación intraclase (Streiner, 1993; Brennan, 2001; Sánchez Pedraza & Rosero Villota, 2003).

Por su parte, se procura que los errores asociados con la situación de aplicación sean minimizados por medio de una estricta estandarización de la misma (Anastasi, 1954,1990; Lord & Novick, 1968). En la descripción anterior de diferentes coeficientes de confiabilidad puede verse que el procedimiento estadístico por excelencia para estimar la confiabilidad de una prueba [considerada como una totalidad] es el de la correlación. Anteriormente, se mostró que la confiabilidad de una prueba está dada por la relación entre la varianza verdadera y la varianza observada de un prueba para una población (cf. Ecuación 7). En diversos textos puede encontrarse la identidad existente entre esta relación y la correlación entre dos conjuntos de puntuaciones obtenidos por la misma muestra de personas en una prueba ($\rho_{X,X'}$) (ver p. ej. Lord & Novick, 1968; Nunnally, 1970; Brown, 1980).

Esta enumeración no es exhaustiva de los posibles coeficientes de confiabilidad que pueden ser calculados para una prueba; del mismo modo, la correlación producto-momento de Pearson no es el único procedimiento estadístico para obtener la confiabilidad de una prueba. Nótese, además, que los coeficientes hasta ahora mencionados pueden estimar la influencia de una sola fuente de error a la vez en el puntaje observado de la prueba total. La estimación de la confiabilidad teniendo en cuenta más de una fuente de error puede llevarse a cabo gracias a un análisis de varianza de los puntajes totales observados a lo largo de los diferentes factores; a partir de este análisis de varianza puede obtenerse un coeficiente de confiabilidad correspondiente a una correlación intraclase (para una descripción de estas estimaciones, ver Thorndike, 1989,1996; Brennan, 2001). En general, la TCT se aproxima al análisis de una prueba desde esta perspectiva de prueba total (Lord & Novick, 1968; Nunnally, 1987; Arévalo, 2002); sin embargo, uno de los métodos clásicos más empleado parte de una ruptura con esta tendencia y busca estimar la confiabilidad a partir del análisis de sus partes: esta es la confiabilidad por consistencia interna.

Para la estimación de la confiabilidad por consistencia interna existen una gran variedad de métodos. En este escrito se desarrollan sólo un par de ellos: la confiabilidad por mitades y el coeficiente α de Cronbach. Todos los métodos por consistencia interna tienen en común que permiten la estimación de la confiabilidad a partir de una sola aplicación de una prueba, siendo esta su mayor ventaja práctica sobre los demás métodos. Esta característica llevó a que estos métodos se convirtieran en los más empleados tanto por quienes aplican las pruebas como por quienes las desarrollan (Thorndike, 1989,1996). La forma en que se obtiene la información necesaria para los cálculos de los estadísticos en estos métodos consiste en separar la prueba en diferentes partes y calificar cada una de ellas; la cantidad de partes en que es dividida la prueba, así como los criterios específicos que guían este proceso dependen del método escogido. Al realizar esta división del puntaje observado en partes estamos asumiendo que el modelo de medida adoptado se ajusta también a las partes resultantes; ahora bien, es fácilmente demostrable que si el modelo se ajusta para varias mediciones, también se ajusta para una medición igual a la sumatoria de éstas.

El método por mitades consiste en la división de la prueba en dos partes paralelas de la misma longitud. Cada una de estas mitades es calificada y los puntajes observados así obtenidos son correlacionados entre sí. Hasta este punto, el procedimiento de las dos mitades puede verse como un caso especial del procedimiento con el cual se obtiene un coeficiente de equivalencia, en el que cada una de las dos mitades es una forma de la prueba. Con esta correlación lo que se obtiene es una estimación de la confiabilidad de cada una de las dos mitades y no de la prueba completa; para obtener la confiabilidad de la prueba total se hace uso de la fórmula de Spearman-Brown (Anastasi, 1954,1990; Lord & Novick, 1968; Brown, 1980; Thorndike, 1989,1996), la cual en el caso de dos partes, es igual a:

$$\rho_{X,X'} = \frac{2\rho_{M,M'}}{1+\rho_{M,M'}} \tag{8}$$

donde $\rho_{M,M'}$ es la confiabilidad de ambas mitades. Esta fórmula asume que las dos mitades en que ha sido dividida la prueba son paralelas. Asimismo, existe una forma general de la Ecuación 8 que permite estimar la confiabilidad por consistencia interna de *n* partes en que se divida una prueba; en la siguiente sección se retomará esta fórmula.

Coeficiente α de Cronbach

El coeficiente α fue propuesto en 1951 por Cronbach como un estadístico para estimar la confiabilidad de una prueba, o de cualquier compuesto obtenido a partir de la suma de varias mediciones. Este coeficiente estima el valor de $\rho^2_{X,T}$ (cf. Ecuación 7) al evaluar la consistencia interna del conjunto de ítems o partes del compuesto; en este sentido, se corresponde con un coeficiente de equivalencia (Lord, 1955) y, por lo tanto, estima la varianza que en los puntajes observados corresponde a factores comunes de los diferentes ítems (Cronbach, 1951; Cotton, Campbell & Malone, 1957; Streiner, 1993; Schmidt et al., 2003). En su momento, el coeficiente α entró al campo psicométrico como un método con el cuál se generalizaron varias propuestas alternativas de estimar la consistencia interna (Muñiz, 1996); el principal de estos predecesores es la fórmula de Kuder y Richardson número 20 (propuesto en 1939, cf. Cronbach, 1951; Cotton et al., 1957; Thorndike, 1989,1996, entre otros), la cual puede verse como un coeficiente α para el caso especial en que todos los ítems que conforman la prueba se califican de modo dicótomo (e.g. correcto / incorrecto). Otro método con el que está relacionado es el propuesto por Hoyt en 1941, en el cual se estima la confiabilidad a través del análisis de varianza (Thorndike, 1989,1996; Bravo & Potvin, 1991; McGraw & Wong, 1996; Muñiz, 1996; Brennan, 2001; Bonett, 2003).

Una fórmula con la cual se calcula el coeficiente α es la siguiente (Cronbach, 1951, p. 305):

$$\alpha = \frac{n}{n-1}\frac{\sum_{k=1}^{n}\sum_{h=1}^{n}\sigma_{k,h}}{\sigma_X^2}; \quad \forall h \neq k \tag{9}$$

donde *n* es el número de partes, *k* y *h* son partes sobre las que se calcula el estadístico. Una fórmula equivalente puede encontrarse en la derivación del coeficiente α de Lord y Novick (1968, p. 89):

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_{k=1}^{n}\sigma_k^2}{\sigma_X^2}\right) \qquad (10)$$

Para la aplicación de esta fórmula (sea en la forma dada en la Ecuación 9 o en la Ecuación 10) se asume que los supuestos generales del modelo propuesto por la TCT se cumplen para todas las partes del compuesto (y, por ende, también para la sumatoria o prueba total). Si esto es cierto, el valor del α es igual o menor que la confiabilidad real del compuesto ($\rho_{X,T}^2$) (Cronbach, 1951; Lord & Novick, 1968). En estas condiciones, el α también se relaciona con la varianza de los factores subyacentes al conjunto de partes de la prueba, siendo tal que la varianza del factor general es menor o igual al α, y este es menor o igual a la varianza de los factores comúnes del conjunto (Cronbach, 1951; Green et al., 1977). Adicionalmente, ocurre que si el conjunto de las partes son paralelas, el α es exactamente igual a $\rho_{X,T}^2$, así como al valor esperado de todos los posibles coeficientes de confiabilidad por mitades obtenidos por la aplicación de la fórmula de Spearman-Brown (Cortina, 1993). Gracias a estas propiedades, a la gran facilidad para obtener los datos necesarios (una sola aplicación) y a la sencillez de su cálculo, el coeficiente ganó rápidamente en aceptación y aplicación --tómese como ejemplo que, tan sólo tres años después, en el texto introductorio a las pruebas psicológicas escrito por Anastasi (1954,1990) ya se puede encontrar como uno de los principales métodos para la estimación de la confiabilidad de una prueba.

Interpretaciones del coeficiente α.

La extensión del uso del α fue seguida, sin embargo, de una multiplicidad de interpretaciones, muchas veces contradictorias entre sí (Cortina, 1993). Esta situación ha llevado a que este estadístico haya sido sobreutilizado a lo largo del tiempo, así como a que se hayan ignorado las condiciones para las que fue desarrollado. Para hacer un recorrido conjunto de las cualidades y limitaciones que estos esfuerzos han obtenido, se seguirá en primer lugar el orden de las principales interpretaciones dadas al α que se encuentra en Cortina (1993), posteriormente serán expuestos los aspectos que no hayan sido incluidos por este autor.

Cortina (1993) ubica cinco interpretaciones aceptadas generalmente en la literarura. Estas son: (a) el coeficiente α es la media de todos los coeficientes de confiabilidad por mitades, (b) es el límite inferior de la confiabilidad de una prueba, (c) es una medida de la saturación del primer factor, (d) es igual a la confiabilidad en condiciones de τ–equivalencia y (e) es una versión general del coeficiente de equivalencia Kuder-Richardson (K-R 20). De estas cinco afirmaciones, la segunda y la cuarta se relacionan estrechamente.

El uso de la primera afirmación implica que se ve al coeficiente α como un estimador estable de la confiabilidad calculada por mitades. Una consecuencia de esto es que se puede considerar al α como un estimador más robusto que el obtenido por el método de las dos mitades. Es importante notar que, la identidad entre el α y el valor esperado de todas las posibles mitades se da sólo cuando las partes son paralelas, es decir, cuando sus varianzas y covarianzas son todas iguales, y su valor esperado es el mismo (cf. Infra). Cuando las varianzas entre los ítems no son iguales, el coeficiente α es igual al valor esperado del α calculado sobre las mitades de la prueba como partes, en lugar de los ítems individuales (Cronbach, 1951; Lord & Novick, 1968; Cortina, 1993).

La segunda y cuarta afirmaciones ponen de relieve que el α es un estimador de la confiabilidad de una prueba (Cronbach, 1951; Lord & Novick, 1968; Bravo & Potvin, 1991). La demostración de las dos afirmaciones puede encontrarse en Cronbach (1951), Novick y Lewis (1967) y Lord y Novick (1968). Cabe recordar que estos resultados son ciertos en la medida en que α es un coeficiente de equivalencia, esto implica que no tiene en cuenta ciertas fuentes de error, como el error temporal (Gerbing & Anderson, 1988; Becker, 2000; Osburn, 2000; Schmidt et al., 2003), y por ello no puede tomarse como reemplazo de un coeficiente de estabilidad como estimador de la confiabilidad de una prueba.

Para la presente discusión es conveniente incluir en este punto algunas de las definiciones introducidas por Novick y Lewis (1967) y Lord y Novick (1968). Estas definiciones se refieren a la relación entre las puntuaciones verdaderas y los puntajes observados de diferentes pruebas, a partir de las cuales se definen ciertas parejas de mediciones. El caso más estricto se da cuando para un par de mediciones: (a) sus puntuajes observados tienen igual valor esperado y varianza ($E(X_1)=E(X_2)$ y $\sigma^2_{X_1}=\sigma^2_{X_2}$), (b) la media de las puntuaciones verdaderas de ambas es la misma, así como sus varianzas y covarianzas ($E(T_1)=E(T_2)$ y $\sigma^2_{T_1}=\sigma^2_{T_2}=\sigma_{T_1T_2}$), y (c) el error de ambas mediciones tiene la misma distribución $E(\varepsilon_1)=E(\varepsilon_2)$ y $\sigma^2_{\varepsilon_1}=\sigma^2_{\varepsilon_2}$; a este par de mediciones se les llama, entonces, paralelas. Este par mide, entonces, el mismo atributo, en las mismas unidades y con la misma precisión. Los otros casos se dan con la relajación de algunas de estas restricciones. Así, un par de mediciones son τ–equivalentes cuando dejan de ser sostenibles las restricciones sobre los errores de las mediciones, pero se siguen tomando las que aplican sobre las puntuaciones verdaderas; este par mide el mismo atributo en las mismas unidades aunque con precisiones diferentes. Cuando se relaja la restricción sobre la localización de las puntuaciones verdaderas (i.e. sobre sus valores esperados), las medidas son llamadas "esencialmente τ–equivalentes". En este caso, el par mide el mismo atributo en las mismas unidades aunque en ubicaciones diferentes de la escala, y lo hace con precisiones diferentes. Como los resultados que se obtienen respecto al α son idénticos cuando las medidas son τ–equivalentes o sólo esencialmente τ–equivalentes, no se les considera como casos diferentes cuando se estudia el coeficiente α. El último caso es el de los pares de medidas congenéricas. En éstas, las restricciones sobre la variabilidad de las puntuaciones verdaderas también son relajadas. En este caso, el par de medidas mide el mismo atributo en unidades diferentes y con precisiones diferentes. Estas precisiones resultan importantes en la

medida que la demostración de la segunda y cuarta afirmaciones requiere que se asuma el cumplimiento de, por lo menos, las restricciones correspondientes a variables τ–equivalentes entre las partes que componen la prueba (Novick & Lewis, 1967; Lord & Novick, 1968; Lucke, 2005)

Con la tercera afirmación se sugiere que el α es una medida del grado en que un factor general subyace a un conjunto de ítems; con esto se sugiere, además, que el α puede usarse como un índice de homogeneidad. Este uso puede verse ilustrado en Cronbach (1951), Anastasi (1954,1990), Webster (1957), Nunnally (1970), Brown (1980), entre muchos otros. Sin embargo, se ha demostrado que dicha afirmación es falsa y que dicho empleo del α es inapropiado (ver p. ej. Cotton et al., 1957; Green et al., 1977; Zimmerman et al., 1993; Muñiz, 1996; Lucke, 2005). Lo que sí es cierto es que entre más próximos estén los ítems a una estructura unidimensional (i.e. homogénea) mejor será la estimación de la confiabilidad por el α, por lo que el uso recomendado es determinar primero la estructura factorial, y calcular el α sobre los factores obtenidos en vez de sobre la totalidad de los ítems (Cotton et al., 1957; Gerbing & Anderson, 1988; Thorndike, 1989,1996; Cortina, 1993; Osburn, 2000; Kamata, Turhan & Darandari, 2003).

La quinta afirmación se refiere al hecho de que el coeficiente K-R 20 es un caso especial del α cuando los ítems son calificados de forma dicótoma, el cual es demostrado por Cronbach (1951). Es interesante anotar, sin embargo, algunas peculiaridades propias del α cuando se obtiene de este modo (v.g. cuando se calcula el *K-R 20*): (a) por una parte, aún en el caso de unidimensionalidad, nunca es igual al porcentaje de la varianza de los factores comúnes (Cotton et al., 1957) y (b) es siempre inferior al α que se obtendría si se calificara al ítem en una escala continua (Feldt, 1993; Stöber, Dette & Musch, 2002); esto es relevante si se asume que esta calificación no es más que una dicotomización artificial en un cierto nivel del atributo de interés. Ambas condiciones reflejan la pérdida de precisión inherente al empleo de este nivel de medición (v.g. dicótomo vs. numérico-continuo).

Factores que afectan al coeficiente α.

En los aspectos que esas cinco afirmaciones y sus consecuencias tienen respecto al α sobresalen los aciertos y desaciertos que han tenido las diferentes interpretaciones de este estadístico con respecto a la varianza correspondiente a las puntuaciones verdaderas de las personas. De su desarrollo puede concluirse que el coeficiente α requiere del supuesto de la τ–equivalencia de sus partes. Los estudios considerados que muestran el sesgo presentado por el α cuando este supuesto no se cumple son especialmente relevantes cuando se considera que en la práctica es bastante difícil obtener medidas a las que se ajuste tan bien dicho modelo (Feldt, 1993; Gómez Benito, 1996). También es notoria la falta de consideración por las implicaciones que tienen los términos que definen el cálculo del coeficiente, así como de los otros supuestos en que se basa y que pueden no verse cumplidos en la práctica. Estos son los que se presentan a continuación.

Por una parte, los términos que intervienen en las fórmulas de cálculo del α son la longitud de la prueba (*n*), la covarianza entre las partes ($\sigma_{k,h}$) y la varianza total de la prueba (σ_X^2). Varios autores (Cronbach, 1951; Green et al., 1977; Niemi et al., 1986; Cortina, 1993) han precisado que el valor del coeficiente se incrementa a medida que *n* aumenta, siempre que se conserven constantes los demás términos y se mantengan los supuestos del modelo. Esta característica fue modelada por Cronbach (1951) relacionando el valor del α con la correlación promedio entre los ítems y con la longitud de la prueba. Esta relación puede expresarse de la siguiente forma (Niemi et al., 1986, p. 371):

$$\alpha = \frac{n\bar{\rho}_{kh}}{1 + \bar{\rho}_{kh}(n-1)} \qquad (11)$$

donde $\bar{\rho}_{kh}$ representa la intercorrelación promedio entre los ítems.

De esta formulación resulta evidente la relación mencionada; adicionalmente, esta formulación es equivalente a la fórmula de Spearman-Brown generalizada a una prueba de longitud *n*. Esta relación ha sido estudiada, además, bajo situaciones en las que los supuestos del α no se cumplen; Green et al. (1977) y Cortina (1993) mostraron que aún en condiciones de multidimensionalidad (i.e. violación del supuesto de τ-equivalencia) el valor del α puede superar el nivel de 0.7. Por ejemplo, en el caso más extremo estudiado por Cortina (1993), el valor del α resulta igual a 0.64 cuando la prueba tiene 18 ítems, con una estructura de tres factores independientes y una interrcorrelación promedio de tan sólo 0.06 (0.3 entre los ítems del mismo factor); este es mayor que 0.7 cuando la intercorrelación promedio llega a 0.1 y los demás términos se mantienen constantes. Se ha encontrado que esta relación entre el valor del α y la longitud de la prueba es curvilínea y que empieza a estabilizarse en longitudes de prueba menores a 19 (Komorita & Graham, 1965; Cortina, 1993). Streiner (1993), por su parte, afirma que escalas de 20 ítems o más suelen obtener valores de α alrededor de 0.9. Otro aspecto importante a considerar se encuentra en el hecho que actualmente la mayoría de pruebas psicológicas que se desarrollan tienen una longitud inferior a 20 ítems (Cortina, 1993). La importancia observada en el valor obtenido en el α tiene su motivo en las guías que al respecto han sido ofrecidas con el tiempo en las cuales la recomendación de Nunnally (1987) (el valor del α debería ser igual o superior a 0.7) ha resultado la que mayor aceptación ha recibido en la comunidad (ver p. ej. Bernardi, 1994; Peterson, 1994). De estos resultados puede afirmarse que la influencia de la longitud de la prueba es relevante para muchos casos prácticos en que no es viable aplicar pruebas de una longitud de 20 o más ítems, para los cuales es posible ver una variabilidad considerable en el estadístico.

En la Ecuación 11 también se encuentra expresada la relación entre el α y la covarianza entre los ítems. Sin embargo, es mucho más notorio a partir de la Ecuación 9 que entre mayor sea la covarianza entre los ítems ($\sigma_{k,h}$) comparada con la varianza total de la prueba (σ_X^2), mayor será el valor del α. La relación entre la covarianza media de los ítems con el valor del α ha sido estudiada por Green et al. (1977), Niemi et al. (1986), Cortina (1993), entre otros.

Niemi et al. (1986) mostraron que una reducción considerable en la interrelación media de los ítems puede disminuir el valor del estadístico aún cuando se aumenta la longitud de la prueba.

En resúmen, las tres condiciones de las cuales depende el valor del coeficiente α que son discutidas por Cortina (1993) son: la dimensionalidad del conjunto de ítems, el nivel de covariación de los ítems entre sí y la cantidad de ítems o partes que conforman la prueba o compuesto. De estas tres, la primera tiene que ver con el cumplimiento o no de alguno de los supuestos en los que se apoya el α, específicamente con el supuesto de τ-equivalencia de las partes. La violación de este supuesto ha recibido la mayor atención de la literatura, al ser la condición necesaria y suficiente (bajo los supuestos de la TCT) para que el coeficiente sea igual a la confiabilidad del compuesto; sin embargo, no es el único supuesto que ha recibido atención en diferentes estudios. Asimismo, variables diferentes al valor que adquiere el α también han sido estudiadas. A continuación se presentan diferentes estudios sobre estos otros aspectos.

El abordaje del estudio de los supuestos sobre los que se apoya el coeficiente α de Cronbach puede dividirse en tres grupos. Por una parte, se pueden abordar los supuestos específicos del α; pueden, por otra, abordarse los supuestos de la TCT en los cuales se inscribe; o, finalmente, pueden abordarse los supuestos de los procedimientos estadísticos que emplea. En la primera categoría encontramos los estudios que han abordado el supuesto de la τ-equivalencia, y que fueron revisados previamente en este documento. Dentro del segundo grupo, aquel supuesto que mayor atención ha recibido es el de la independencia de los errores de diferentes mediciones entre sí (cf. Ecuación 5); más adelante se mostrarán algunos de estos estudios. En cuanto al tercer conjunto, ha sido estudiado el comportamiento del α con respecto a la violación de los supuestos sobre la distribución normal de las variables (en este caso del atributo medido) en la población sobre la cual descansa la precisión de la varianza y la covarianza. A continuación se presenta una serie de dichos estudios.

El interés por el efecto de la violación del supuesto de independencia de los errores de medida en el α de Cronbach es más bien reciente; sin embargo, es de considerable importancia, pues es claro que diferentes medidas aplicadas en la misma ocasión, como los ítems presentados en una sola prueba e incluso de diferentes pruebas aplicadas en una misma sesión, pueden ser influidas por fuentes de error comunes (Zimmerman et al., 1993; Raykov, 2001). Este problema ha sido abordado por medio de estudios de simulación (Zimmerman et al., 1993; Komaroff, 1997; Shevlin, Miles, Davies & Walker, 2000), así como por un análisis matemático (Raykov, 2001). El análisis de Raykov (2001) demuestra que cuando existe correlación entre los errores de los distintos ítems el coeficiente α pierde precisión y puede bien subestimar o sobreestimar la confiabilidad real del compuesto. Para ese autor la variable de interés fue el sesgo que se produce en el α cuando las partes son cuando menos congenéricas y se presenta esta interrelación entre los errores. La recomendación que ofrece ese autor es la verificación previa de los supuestos con el uso de modelos de estructuras de covarianza para evaluar cuán grande puede ser el sesgo en la estimación de la confiabilidad. Los estudios de Zimmerman et al. (1993) y Komaroff (1997) se preocupan especialmente del

efecto que la correlación entre los errores tiene en el valor calculado del α. Zimmerman et al. (1993) también estudiaron el sesgo del α para estimar la confiabilidad real; en este estudio encontraron que la presencia de correlación entre los errores conllevaba a una sobreestimación de la confiabilidad. La presencia y tamaño de esta sobreestimación fue evaluada de forma descriptiva (o "al ojo"), del mismo modo que las posibles interacciones con los otros factores que estudiaron: (a) la violación del supuesto de τ-equivalencia y (b) la falta de normalidad en la distribución poblacional del atributo. Komaroff (1997) por su parte se centra en la obtención de una correción de la fórmula del α que tenga en cuenta la relación entre los errores; la principal limitación de este estudio es que sólo fue considerado el valor obtenido del α y no su relación con la confiabilidad del compuesto, parámetro de interés al emplear este coeficiente. El estudio de Shevlin et al. (2000) examina la relación entre la correlación de los errores, el tamaño de muestra y el tamaño de la carga factorial de los ítems, asumiendo unidimensionalidad en los mismos, con el valor del α y con la diferencia entre la estimación muestral ($\hat{\alpha}$) y el α poblacional; también examina la distribución muestral empírica. Los resultados de este estudio son acordes con los de Zimmerman et al. (1993) y Komaroff (1997), e indican un aumento en el valor del $\hat{\alpha}$ cuando se presentan errores entre las partes; una ventaja que ofrece este estudio es que el efecto de estos factores no fue analizado únicamente de forma descriptiva, por lo que pudieron dar cuenta de las interacciones entre estos tres factores. Shevlin et al. (2000) encontraron que el error tiene un efecto significativo sobre el valor del $\hat{\alpha}$, y que además interactúa tanto con el tamaño de la muestra como con el peso factorial de las partes; el grado de correlación entre los errores presentó un efecto más grande entre menor fuese la muestra, así como entre menor fuese el peso factorial de los ítems. Este estudio tampoco relaciona el $\hat{\alpha}$ con la confiabilidad del compuesto.

En cuanto a los estudios relacionados con el tercer grupo de supuestos, los desarrollos más importantes se encuentran en el estudio de la distribución muestral del $\hat{\alpha}$ teniendo en cuenta la distribución de los puntajes de las partes sobre las que se calcula; también dentro de este grupo podemos encontrar estudios sobre el valor y la precisión del coeficiente bajo diferentes distribuciones del atributo en la población. A lo largo de la década de los 60, Kristof (Kristof, 1970; van Zyl, Neudecker, & Nel, 2000) y Feldt (1965) derivaron de forma independiente la teoría referente a la distribución muestral del $\hat{\alpha}$ bajo los supuestos de la TCT, de τ-equivalencia de las partes y de distribución normal multivariada de las partes. Más recientemente han aparecido desarrollos para emplear esta distribución en pruebas de hipótesis sobre el estimador del $\hat{\alpha}$ (Feldt, 1990), en obtención de intervalos de confianza del $\hat{\alpha}$ (Feldt, 1990; Bravo & Potvin, 1991; McGraw & Wong, 1996) y para hacer pruebas de hipótesis sobre diferencias de dos coeficientes α (e.g. Feldt & Ankenmann, 1999; Silver, 2001); adicionalmente, van Zyl et al. (2000) derivaron la distribución asintótica del $\hat{\alpha}$ sin restricciones sobre la existencia de τ-equivalencia entre las partes. La derivación de van Zyl et al. (2000) fue también estudiada por Yuan y Bentler (2002), quienes encontraron que dicha distribución asintótica parece ser robusta, además, ante violaciones del supuesto de normalidad multivariada de las partes. Bajo estos hallazgos, Duhachek y Iacobucci (2004) derivaron un método para obtener interválos de confianza bajo esta distribución muestral del $\hat{\alpha}$. Podemos encontrar estudios de simulación sobre la distribución muestral del $\hat{\alpha}$ en Zimmerman et al.

(1993), Shevlin et al. (2000) y Duhachek y Iacobucci (2004). En los primeros dos estudios la distribución muestral es analizada de forma descriptiva, y en general encuentran que la variabilidad del $\hat{\alpha}$ es menor cuando se tiene un mayor tamaño de muestra; Zimmerman et al. (1993) observan en sus datos una alta variabilidad del $\hat{\alpha}$ en la mayor parte de sus tratamientos, Shevlin et al. (2000) encuentra que dicha variabilidad, expresada respecto a su desviación estándar: (a) no lo hace ininterpretable, y (b) parece verse afectada tanto por la longitud de la prueba como por el peso factorial de los ítems. El estudio de Duhachek y Iacobucci (2004) analiza el efecto de la longitud de la prueba, del tamaño de la muestra, de la intercorrelación entre los ítems y de la falta de unidimensionalidad, en el error tipo I (contra el valor del α) y en la amplitud del interválo de confianza derivado; asimismo, compararon estos efectos entre diferentes propuestas de obtención de interválos de confianza. Con estos resultados, derivaron un interválo de confianza para la diferencia de dos coeficientes α.

Tabla 1
Cuadro comparativo de estudios sobre el α de Cronbach

Estudio	**Respuesta**[a]	**Factores**[b]
Cronbach (1951)	α	n, N, $\bar{\rho}_{kh}$, λ
Cotton et al. (1957)	α	λ, Dimensionalidad
Green et al. (1977)	α	n, $\bar{\rho}_{kh}$, λ, Dimensionalidad
Niemi et al. (1986)	α	n, $\bar{\rho}_{kh}$
Reuterberg y Gustafsson (1992)	$\hat{\alpha}$, $\hat{\rho}^2_{X,T}$	Dimensionalidad
Cortina (1993)	α	n, $\bar{\rho}_{kh}$, dimensionalidad
Feldt (1993)	α, $\hat{\alpha}$, *S.B.*	Dificultad de los ítems, $\bar{\rho}_{kh}$
Zimmerman et al. (1993)	$\hat{\alpha}$, $\rho^2_{X,T}$	n, N, ε_{kh}, Dimensionalidad
Komaroff (1997)	$\hat{\alpha}$	ε_{kh}, Dimensionalidad
Osburn (2000)	α	Dimensionalidad
Shevlin et al. (2000)	α, $\Delta\alpha$	N, λ, ε_{kh}, Interacciones
Raykov (2001)	$\Delta\rho^2_{X,T}$	ε_{kh}
Christmann y Van Aelst (2002)	$\hat{\alpha}$, $\Delta\rho^2_{X,T}$	Distribución de los ítems
Yuan y Bentler (2002)	α	Distribución de los ítems
Kamata et al. (2003)	$\hat{\alpha}$, $\Delta\rho^2_{X,T}$	N, λ, Dimensionalidad
Duhachek y Iacobucci (2004)	$\hat{\alpha}$	n, N, $\bar{\rho}_{kh}$, Dimensionalidad, Interacciones
Norris y Aroian (2004)	$\hat{\alpha}$	Distribución de los ítems
Lucke (2005)	α, ω	Dimensionalidad

Notas: [a] $\Delta\alpha$: Diferencia entre $\hat{\alpha}$ y el valor poblacional especificado en las condiciones del estudio, $\Delta\rho^2_{X,T}$: diferencia entre el $\hat{\alpha}$ y el valor real de la confiabilidad especificada en el estudio, S.B: Fórmula de Spearman-Brown (Ecuación 8), $\hat{\rho}^2_{X,T}$: estimación de la confiabilidad a partir de un análisis factorial confirmatorio.

[b] n: Número de partes-ítems sobre los que se calcula el coeficiente, N: Tamaño de la muestra, λ: cargas factoriales de los ítems en el(los) factor(es) subyacentes

Sobre la violación de los supuestos de normalidad multivariada de las partes Zimmerman et al. (1993) y Christmann y Van Aelst (2002) realizaron dos estudios de simulación. En ambos estudios se tiene en cuenta el sesgo de la estimación de la confiabilidad por parte del $\hat{\alpha}$; Zimmerman et al. (1993) analizan el efecto de la falta de normalidad en las partes de un modo descriptivo; en su estudio la forma de la distribución varía entre condiciones diferentes y es común a todas las partes de la misma condición. En el reporte de Christmann y Van Aelst (2002), la forma de la distribución también fue variada entre condiciones, pero emplearon distribuciones no normales diferentes; los análisis del sesgo del $\hat{\alpha}$ fueron asimismo realizados de forma descriptiva. Mientras que Zimmerman et al. (1993) encuentran que el $\hat{\alpha}$ parece ser robusto a las violaciones de normalidad en las partes que conforman la prueba, Christmann y Van Aelst (2002), por su parte, encuentran resultados opuestos.

En la tabla 1 puede verse resumido el conjunto de características de varios estudios realizados sobre el α de Cronbach. Puede observarse que la mayoría de los estudios se han concentrado únicamente en cómo se ve afectado el valor del $\hat{\alpha}$ o del α poblacional, también es evidente que son pocos los estudios que han analizado el sesgo del $\hat{\alpha}$ respecto a la confiabilidad real de la prueba. El principal factor analizado es claramente el que tiene que ver con la dimensionalidad del conjunto o prueba; esto es lo relativo al supuesto de τ-equivalencia entre las partes. Sólo en dos de los estudios que emplearon métodos de Monte Carlo se llevó a cabo un metamodelo; en ambos se trató de un modelo de ANOVA sobre el $\hat{\alpha}$ en el cual las interacciones de los factores considerados en el estudio fueron analizados; en el estudio de Shevlin et al. (2000) si bien se consideró la diferencia entre el $\hat{\alpha}$ y el α poblacional, esta relación no fue modelada. En cuanto a los estudios de Monte Carlo el número de réplicas varió entre 1 (Green et al., 1977) y 2000 (Zimmerman et al., 1993) por condición. Finalmente, ninguno de los estudios reportó el empleo de técnicas de reducción de varianza de los estimadores de Monte Carlo.

Discusión

Puede verse que el lugar que tiene actualmente el coeficiente α de Cronbach para la exploración de la confiabilidad de un compuesto de diferentes mediciones es realmente importante. El uso generalizado que ha tenido y la amplitud de las interpretaciones de las que ha sido objeto muestran, sin embargo, la falta de consenso en la comunidad sobre cómo y cuándo es empleado correctamente; así también, puede encontrarse una amplia variedad de estudios en los cuales se busca encontrar guías para este objetivo.

En este trabajo se han documentado los principales factores que en estos estudios han mostrado tener alguna influencia en la precisión del coeficiente α como estimador de la confiabilidad. El autor encuentra que los factores considerados en diferentes estudios no han sido interpretados de forma consistente de uno a otro; por esta razón, los resultados de los mismos muestran dificultades para ser considerados en su conjunto por los diferentes usuarios de pruebas y de otras formas de medición que obtienen información de mediciones compuestas. El principal factor analizado a lo largo de los diferentes estudios aquí

considerados, y sobre el cual existen las conclusiones más solidas para los desarrolladores y usuarios de pruebas, es la dimensionalidad del compuesto; en este sentido, puede recomendarse el uso del coeficiente únicamente en aquellos casos en los cuales se haya comprobado la unidimensionalidad del conjunto, entendida como mínimo como una condición de ítems congenéricos y preferiblemente entendida como τ-equivalencia, pues se ha demostrado que sólo en esta situación puede interpretarse el valor obtenido para el α como un estimador de la confiabilidad del compuesto.

Otros factores considerados demuestran afectar la precisión de la estimación de la confiabilidad que puede llevarse a cabo. Los resultados, sin embargo, no son lo suficientemente claros para estos factores, con la excepción de la presencia de correlación entre los errores de medida de los ítems o partes del conjunto. En el caso de la presencia de errores correlacionados es posible evaluar cuánto afectan los mismos la estimación; esta evaluación requiere del ajuste de un modelo de estructuras de covarianza a los datos obtenidos, método que puede resultar impracticable o altamente inestable con ciertos tamaños de muestra y especialmente bajo ciertas condiciones distribucionales de los ítems. Cuando dicho modelamiento es factible, puede obtenerse una estimación más directa de la confiabilidad que la obtenida con el uso del α de Cronbach. En esta medida, aún no es claro en qué casos podrá considerarse que el uso del coeficiente es apropiado, aunque sí es posible establecer casos en los que su obtención deja de ser imprescindible.

En general, las propiedades del α de Cronbach parecen no ser lo suficientemente deseables ante situaciones de aplicación adversas; sin embargo, la evidencia al respecto no es tampoco concluyente más que para un par de los posibles casos que es posible encontrar en la práctica. Actualmente se encuentran alternativas al α para la estimación de la confiabilidad, especialmente desde modelos factoriales de análisis de los datos, las cuales pueden en un futuro volverse mejores opciones; la popularidad que tiene el α podría, entonces, ser desafiada y en esta medida un uso más conciente y moderado de este estadístico sería factible. Para esto se necesita aún de mayor investigación sobre cuán preciso resulta el α, especialmente en las condiciones que poseen la mayor parte de las investigaciones que utilizan este estadístico, tales como pruebas cortas con tamaños de muestra reducidos.

Referencias

Anastasi, A. (1990). *Psychological testing* (6th ed.). New York: MacMillan Publishing Company. (Trabajo original publicado en 1954)

Arévalo, I. (2002). *Estudio comparativo del índice de dificultad en la teoría clásica de los tests, la teoría de respuesta al ítem y el análisis bayesiano*. Tesis de pregrado no publicada, Universidad Nacional de Colombia.

Becker, G. (2000). How important is transient error in estimating reliability? going beyond simulation studies. *Psychological Methods*, 5, 370-379.

Bernardi, R. (1994). Validating research results when Cronbach's alpha is below .70: A methodological procedure. *Educational and Psychological Measurement*, 54, 766-775.

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, 110, 305-314.

Bonett, D. G. (2003). Sample size requirements for comparing two alpha coefficients. *Applied Psychological Measurement*, 27, 72-74.

Bravo, G., & Potvin, L. (1991). Estimating the reliability of continuous measures with Cronbach's alpha or the intraclass correlation coefficient: Toward the integration of two traditions. *Journal of Clinical Epidemiology*, 44, 381-390.

Brennan, R. L. (2001). Generalizability theory. New York: Springer-Verlag.

Brown, F. G. (1980). *Principios de la medición en psicología y educación*. México D.F., México: Manual Moderno.

Christmann, A., & Van Aelst, S. (2002). *Robust estimation of Cronbach's alpha* (Tech. Rep. No. SFB475 42/02). Dortmund, Alemania: Universität Dortmund.

Cortina, J. M. (1993). What is coefficient alpha? an examination of theory and applications. *Journal of Applied Psychology*, 78, 98-104.

Cotton, J. W., Campbell, D. T., & Malone, R. D. (1957). The relationship between factorial composition of test items and measures of test reliability. *Psychometrika*, 22, 347-357.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 1-16.

Duhachek, A., & Iacobucci, D. (2004). Alpha's standard error (ASE): An accurate and precise confidence interval estimate. *Journal of Applied Psychology*, 89, 792-808.

Feldt, L. S. (1965). The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. *Psychometrika*, 30, 357-370.

Feldt, L. S. (1990). The sampling theory for the intraclass reliability coefficient. *Applied Measurement in Education*, 3, 361-367.

Feldt, L. S. (1993). The relationship between the distribution of item difficulties and test reliability. *Applied Measurement in Education*, 6, 37-48.

Feldt, L. S., & Ankenmann, R. D. (1999). Determining sample size for a test of the equality of alpha coefficients when the number of part-tests is small. *Psychological Methods*, 4, 366-377.

Gerbing, D. W., & Anderson, J. C. (1988). An updated paradigm for scale development incorporating unidimensionality and its assessment. *Journal of Marketing Research*, 25, 186-192.

Gómez Benito, J. (1996). Aportaciones de los modelos de estructuras de covarianza al análisis psicométrico. En J. Muñiz (Ed.), *Psicometría* (pp. 457-554). Madrid, España: Universitas.

Gómez Benito, J., & Hidalgo Montesinos, M. D. (2003). Desarrollos recientes en psicometría. *Avances en Medición*, 1, 17-36.

Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement*, 37, 827-838.

Gulliksen, H. (1950). *Theory of mental tests*. New York: John Wiley & Sons, Inc.

Herrera, A. N., Sánchez, N. R., & Jiménez Ávila, H. (2000). De la teoría clásica de los tests a la teoría de respuesta al ítem. *Aula Psicológica*, 3, 293-332.

Iacobucci, D. (Ed.). (2001). Methodological and statistical concerns of the experimental behavioral researcher [Número monográfico]. *Journal of Consumer Psychology*, 10(1 & 2).

Kamata, A., Turhan, A., & Darandari, E. (2003, April). *Estimating reliability for multidimensional composite scale scores*. Presentado en el encuentro anual de la American Educational Research Association, Chicago. Recuperado el 26 de Marzo de 2005, de http://garnet.acns.fsu.edu/~akamata/papers/

Kane, M. (1996). The precision the measurement. *Applied Measurement in Education*, 9, 355-379.

Komaroff, E. (1997). Effect of simultaneous violations of essential tau-equivalence and uncorrelated errors on coefficient alpha. *Applied Psychological Measurement*, 21, 337-348.

Komorita, S. S., & Graham, W. K. (1965). Number of scale points and the reliability of scales. *Educational and Psychological Measurement*, 25, 987-995.

Kristof, W. (1970). On the sampling theory of reliability estimation. *Journal of Mathematical Psychology*, 7, 371-377.

Lord, F. M. (1955). Sampling fluctuations resulting from the sampling of test items. *Psychometrika*, 20, 1-22.

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

Lucke, J. F. (2005). The α and the ω of congeneric test theory: An extension of reliability and internal consistency to heterogeneous tests. *Applied Psychological Measurement*, 29, 65-81.

Martínez, R. (1996). *Psicometría: teoría de los tests psicológicos y educativos*. Madrid, España: Síntesis.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1, 30-46.

Muñiz, J. (1996). Fiabilidad. En J. Muñiz (Ed.), *Psicometría* (pp. 1-47). Madrid, España: Universitas.

Niemi, R. G., Carmines, E. G., & McIver, J. P. (1986). The impact of scale lenght on reliability and validity. *Quality y Quantity*, 20, 371-376.

Norris, A. E., & Aroian, K. J. (2004). To transform or not transform skewed data for psychometric analysis. *Nursing Research*, 53, 67-71.

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. Psychometrika, 32, 1-13.

Nunnally, J. (1970). *Introducción a la medición psicológica*. Buenos Aires, Argentina: Centro Regional de Ayuda Técnica.

Nunnally, J. (1987). *Teoría psicométrica*. México D.F., México: Trillas.

Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods*, 5, 343-355.

Peterson, R. A. (1994). A meta-analysis of Cronbach's coefficient alpha. *Journal of Consumer Research*, 21, 381-391.

Raykov, T. (2001). Bias of coefficient α for fixed congeneric measures with correlated errors. *Applied Psychological Measurement*, 25, 69-76.

Reuterberg, S.-E., & Gustafsson, J.-E. (1992). Confirmatory factor analysis and reliability: Testing measurement model assumptions. *Educational and Psychological Measurement*, 52, 795-811.

Sánchez Pedraza, R., & Rosero Villota, N. E. (2003). Revisión crítica de las escalas de medición de manía. *Avances en Medición*, 1, 37-70.

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual differences constructs. *Psychological Methods*, 8, 206-224.

Shevlin, M., Miles, J. N. V., Davies, M. N. O., & Walker, S. (2000). Coefficient alpha: a useful indicator of reliability? *Personality and Individual Differences*, 28, 229-237.

Silver, N. C. (2001). DIFALPHA: A FORTRAN 77 program for testing the difference between independent α coefficients with different test lengths. *Applied Psychological Measurement*, 25, 68.

Stöber, J., Dette, D. E., & Musch, J. (2002). Comparing continuous and dichotomous scoring of the balanced inventory of desirable responding. *Journal of Personality Assessment*, 78, 370-389.

Streiner, D. L. (1993). A checklist for evaluating the usefulness of rating scales. *Canadian Journal of Psychiatry*, 38, 140-148.

Streiner, D. L. (2003). Being inconsistent about consistency: When coefficient alpha does and doesn't matter. *Journal of Personality Assessment*, 80, 217-222.

Thorndike, R. L. (1996). *Psicometría aplicada*. México D.F., México: Limusa. (Trabajo original publicado en 1989)

van Zyl, J. M., Neudecker, H., & Nel, D. G. (2000). On the distribution of the maximum likelihood estimator of Cronbach's alpha. *Psychometrika*, 65, 271-280.

Webster, H. (1957). Item selection methods for increasing test homogeneity. *Psychometrika*, 22, 395-403.

Yuan, K.-H., & Bentler, P. M. (2002). On robustness of the normal-theory based aymptotic distributions of three reliability coefficient estimates. *Psychometrika*, 67, 251-259.

Zimmerman, D. W., Zumbo, B. D., & Lalonde, C. (1993). Coefficient alpha as an estimate of test reliability under violation of two assumptions. *Educational and Psychological Measurement*, 53, 33-49.

Manuscrito recibido en Marzo de 2005
Aprobado para publicación en Abril de 2005