Ciencia Latina
Revista Multidisciplinar

¿Qué tan apropiadamente reportaron los autores el Coeficiente del Alfa de Cronbach?

Héctor Francisco Ponce Renova
Hector.ponce@uacj.mx

Diana Irasema Cervantes Arreola
Diana.cervantes@uacj.mx

Alejandro Jesús Robles Ramírez
Jesus.robles@uacj.mx
Universidad Autónoma de Ciudad Juárez
Ciudad Juárez- México

RESUMEN

Se evaluó el uso del Coeficiente del Alfa de Cronbach en una muestra de artículos publicados desde el año 2000-2017 (n = 111) por cuatro revistas científicas mexicanas enfocadas en la educación. La metodología consistió en extraer información de la muestra para compararla con ciertos criterios sugeridos en la psicometría. Los resultados mostraron que la mayoría de los autores no usaron apropiadamente este coeficiente. De lo cual se dedujo que la falta de una interpretación adecuada, puede llevar a conclusiones erróneas acerca del concepto de la confiabilidad.

Palabras clave: Alfa de cronbach; confiabilidad; psicometría; educación

¿How well did authors report the Cronbach's Coefficient Alpha?

ABSTRACT

The use of Cronbach's Alpha Coefficient was evaluated in a sample of articles published since 2000-2017 (n = 111) by four Mexican journals focused on education. The methodology consisted of extracting information from the sample to compare it with certain criteria suggested by psychometricians. The results showed that most authors did not use this coefficient properly. The lack of adequate interpretation can lead to erroneous conclusions about the concept of reliability.

Keywords: Cronbach's alpha; psychometrics; reliability; education.

Artículo recibido: 05 de Abril 2021
Aceptado para publicación: 28 de Mayo 2021
Correspondencia: hector.ponce@uacj.mx
Conflictos de Interés: Ninguna que declarar

1. INTRODUCCIÓN

Se usan instrumentos de medición como test y encuestas en muchas áreas de la educación (e.g., educación especial para evaluar el coeficiente intelectual de estudiantes con algún tipo de discapacidad). La medición de algún constructo involucra muchas veces al Coeficiente del Alfa de Cronbach (i.e, alfa o α). Por otro lado, la interpretación y uso del alfa ha sido un problema debido a muchos errores, omisiones e imprecisiones al igual que con otras propiedades psicométricas en la literatura en el idioma ingles (Henson y Roberts, 2006; Hogan, Benjamin, y Brezinski, 2000; Taber, 2017; Whittington, 1998; Worthington y Whittaker, 2006; entre otros). Más al respecto, el alfa ha sido llamado confiabilidad en la literatura, pero no son sinónimos (ver a Henson, 2001). Sin embargo, Taber (2017) encontró que en general los autores no han distinguido entre estos dos conceptos y los usan intercambiablemente. Para efectos del presente manuscrito, el alfa será considerado, como uno de los procesos de confiabilidad o consistencia interna. Dados estos problemas, el *objetivo* del presente manuscrito fue evaluar el uso del alfa en artículos publicados (n = 111; durante 2000-2017) en cuatro revistas científicas mexicanas en español (Tabla 1) que aparecieron en el portal del Consejo Nacional de la Ciencia y Tecnología (CONACYT) y en el ranking de *Scimago Journal & Country Rank* (SJCR).

La pregunta de investigación para este escrito se derivó del objetivo: ¿Qué tan apropiadamente reportaron los autores el Coeficiente del Alfa de Cronbach? Para contestar esta pregunta, se abordaron cinco *aspectos* que han aparecido recurrentemente en la literatura del alfa los cuales también fungieron como los ***Criterios de Evaluación de los Artículos*** de la muestra (ver la metodología):

- Uso de la definición moderna del concepto de confiabilidad a través del α.
- Información sobre el *tamaño del α* estimado.
- Calculo de un *intervalo de confianza* del α.
- Estimación de un α global y por constructo.
- Descripción de las *características del estudio* como uso de datos empíricos y procesos de validación de los puntajes.

1.1 Uso de la Definición Moderna del concepto del α

La conceptualización de la confiabilidad en psicometría ha ido cambiando a través del tiempo (Sawilowsky, 2000). Sin embargo, algunos autores en la literatura educativa en

inglés *no* han usado una conceptualización e interpretación moderna de confiabilidad (Thompson, 2003). Este último autor calificó como desafortunada la interpretación de algunos autores de educación y psicología cuando declaran que un *instrumento es confiable*. Previamente, Thompson (1992) explicó que este tipo de vocabulario ha evidenciado un descuido y una conceptualización errónea porque la confiabilidad o consistencia interna ha sido una propiedad de los puntajes. La definición moderna de confiabilidad fue dada por Feldt y Brennan (1989), y fue retomada y recomendada en 1999 por Wilkingson y el Grupo de Trabajo de la Asociación Americana de Psicología [APA]:

> Es importante recordar que un test no es confiable o no confiable. De este modo, los autores deben de proveer coeficientes de confiabilidad de los puntajes de los datos que están siendo analizados, aunque el foco de la investigación no sea la psicométrica. (p. 597)

En otras palabras y bajo la Teoría Clásica del Puntaje Verdadero (TCPV; *Classical True-Score Theory*), la confiabilidad o consistencia interna *no* radica en el instrumento en sí mismo sino en los puntajes obtenidos a través de este (Feldt y Brennan, 1989; Gronlund y Linn, 1990; Thompson, 1994; Thompson, 1992; Wilkinson y el Grupo de Trabajo de la APA, 1999). Según Thompson (2003), este problema de mala interpretación de la confiabilidad o consistencia interna puede llevar tanto a los investigadores como a sus lectores a conclusiones inexactas. Una conclusión inexacta seria que el instrumento no es confiable, cuando sus puntajes serían los no confiables en una situación.

1.2 Información sobre el Tamaño del α Estimado

Según George y Mallery (2003), los coeficientes del α tienen ciertos mínimos para considerarlos desde *inaceptable*s hasta *excelentes* y son:

- $\alpha \geq .90$ es *excelente*;
- $\alpha \geq .80$ es *bueno*;
- $\alpha \geq .70$ es *aceptable*;
- $\alpha \geq .60$ es *cuestionable*;
- $\alpha \geq .50$ es *pobre*;
- y $\alpha < .50$ es *inaceptable*.

Cabe la posibilidad de que un alfa sea negativo, y Thompson (2003) recomendó que se debe de revisar el significado del ítem para ver que está pasando o reportar el alfa como

cero. Además, Streiner (2003) argumento que un alfa > .90 es demasiado alto y podría sugerir que los ítems son muy redundantes entre sí. En contraparte, Taber (2017) discutió que hay que considerar el contexto de donde se obtuvo el alfa antes de calificarlo de algún modo. Otra crítica, Sijtsma (2009) dijo que el alfa *no* provee un valor preciso de confiabilidad de los puntajes sino proporciona un límite bajo. Sin embargo, el mismo Cronbach (1951) ya había declarado que el alfa era una aproximación y que la consistencia interna era una noción que no tenía un significado claro y unánimemente acordado.

Otro aspecto del tamaño del α es la longitud del instrumento. Es decir, y sin considerar otras variables, si se incrementa el número de ítems de un instrumento, el alfa tiende a incrementar también (cf. Schultz, Whitney y Zickar, 2014). De hecho, se puede calcular el cambio en el α de acuerdo a la *Fórmula de la Profecía de Spearman-Brown*. Esta es una ecuación que estima la confiabilidad de los puntajes de un instrumento si el número de ítems es modificado (cf. Schultz et al., 2014). Al respecto de la longitud del instrumento, Herman (2015, p. 8) explico que el alfa tiende a subestimar el nivel de consistencia interna de los puntajes cuando se tienen menos de 10 ítems. Este autor continúo diciendo que el alfa ofrece un promedio de las correlaciones inter-ítems como un valor alternativo a la consistencia interna.

1.3 Cálculo un Intervalo de Confianza del α

De las poblaciones se obtienen parámetros y de las muestras de estas se estiman estadísticas (Schumacker y Tomek, 2013). Las muestras son una parte y representan a la población y se usan para poder generalizar los resultados de algún estudio. Las estadísticas representan los valores de los parámetros (e.g., promedio, desviación estándar, alfa, etc.), pero se debe de asumir que hay un error al obtener la estadística. Este error de muestreo es la diferencia que hay entre un parámetro y una estadística. Cuando se reporta un intervalo de confianza (*IC*), se considera este error para informar al lector si hubo un 95% o 99% de confianza. Con un 95% de confianza, se esperaría que de 100 muestras tomadas al azar de la población 95 de ellas contuviera el parámetro de interés (Cumming, 2013). La precisión de los *IC* suele ser sensible al tamaño de la muestra (i.e., entre más grande la muestra menos error).

Fan y Thompson (2001) elaboraron una serie de guías para instruir y pedir a los autores que reporten *IC* junto con su respectiva metodología cuando usen el alfa o alguna otra

estadística de confiabilidad. Estos últimos autores declararon que su propósito es ayudar al lector a entender que todas las estimaciones estadísticas, incluyendo aquellas de confiabilidad, son afectadas por un error de muestreo. Según ellos, estos requisitos podrán facilitar el entendimiento de que los instrumentos *no* están impregnados con confiabilidad invariante.

Sobre el tamaño de la muestra para generalizar un coeficiente de α, Churchill y Peter (1984) encontraron en su meta-análisis que el alfa tenía una relación negativa con el tamaño de la muestra. Mientras en otro meta-análisis, Petterson (1994) *no* encontró relación entre el tamaño de la muestra y el α. Más al respecto, Fleiss (1986) sugirió que una muestra de un tamaño de 15-20 participantes sería suficiente para estimar un α. Por el otro lado, Nunnally y Bernstein (1994) recomendaron una muestra de 300 o más. Más al respecto, Bonnett (2002) explico que para el tamaño de una muestra se usa un test para estimar el *IC* del alfa. Este último autor dijo que el tamaño óptimo de la muestra debe de basarse en criterios como el poder estadístico (i.e., probabilidad de rechazar una hipótesis nula cuando es falsa; ver a Cohen, 1988), tamaño del efecto o precisión deseada del alfa. Bonnett (2002) declaro que las sugerencias basadas en simples recomendaciones muchas veces son engañosas y mostro una serie de fórmulas para estimar el tamaño de la muestra basadas en los criterios antes mencionados.

1.4 Estimación de un α global y por constructo

Gardner (1995) enfatizó la importancia para el cálculo del α para los puntajes de todos los ítems que subyacen en el mismo constructo: si un instrumento contiene cinco constructos, deben de calcularse cinco α. Además, si el instrumento contiene un constructo de segundo nivel en el cual subyacen los demás constructos, DeVellis (2016) sugirió estimar un alfa global también. Un coeficiente alfa global es estimado de los puntajes de todos los ítems del instrumento. Para el ejemplo de los cinco constructos, si hubiera un alfa global, se calcularían estos cinco coeficientes alfas más el global, dando seis.

La condición para usar el alfa es que todos los ítems (por lo menos dos) deben de pertenecer a un solo constructo (unidimensional que es un concepto que se sustenta matemáticamente en el análisis exploratorio de factores y en una teoría psicológica en específico; ver a Thompson, 2004). Otra condición es que se usa el α en un test cuando una respuesta involucra ítems con puntajes (e.g., 1-10) o encuestas que tienen una escala

tipo Likert (e.g. 1-5; ver a Likert, 1932). Para escalas dicotómicas (e.g., acierto vs. error), se usa los coeficientes de *KR-20* y *KR-21* (ver a Kuder y Richardson, 1937).

1.5 Descripción de las Características del Estudio como uso de Datos Empíricos y Procesos de Validación de los Puntajes

La ventaja del α es que solo se necesita administrar una vez a los participantes. En este presente manuscrito, se describen algunas características de los artículos de la muestra. Algunos de estos son: el uso de datos empíricos, procesos de validación de los puntajes, tamaño de sus muestras de participantes/observaciones, sus coeficientes de desviación estándar, creación, adaptación o replicación de un instrumento. Esto con el propósito de ilustrar que elementos contuvieron estos artículos publicados.

Se seleccionó el portal del CONACYT porque este organismo público y parte del Gobierno Federal Mexicano declaro que es responsable de la elaboración de políticas de ciencia y tecnología para solucionar problemas y bienestar de la población (Galindo, 2017). La misión de este organismo (Galindo, 2017, párr. 1) es "… impulsar y fortalecer el desarrollo científico y la modernización tecnológica de México, mediante la formación de recursos humanos de alto nivel, la promoción y el sostenimiento de proyectos específicos de investigación y la difusión de la información científica y tecnológica." Como parte de esta tarea de traer el bienestar a la población mexicana de más de 120 millones por medio de la ciencia y tecnología, el CONACYT mantiene un portal de revistas científicas. Dada la misión del CONACYT, se asumió que las revistas contienen las mejores prácticas en el quehacer científico en general. En lo particular, se asumió que estas mejores prácticas incluirían uso apropiado del α. Se encontraron nueve revistas en su portal con la palabra clave: *educación*.

Además de aparecer en el portal del CONACYT, la fuente debía de aparecer en algún ranking internacional. Al estar ranqueada una revista, se asume que estas fuentes pueden ser leídas más allá de México. Se usó el ranking de SJCR porque tiene fuentes en español. Solo cuatro de las revistas de las nueve del portal del CONACYT fueron ranqueadas por SJCR, así que solo estas cuatro fueron parte de la muestra. Según ranking de SJCR, las revistas publicadas en México ubican a este país en el lugar 29º de la lista de países. De los países donde el español probablemente predomina, solo España se ubica antes que México en el lugar 10º.

1.6 Vacío en la literatura

En las cuatro revistas de la muestra (Tabla 1), se llevó una serie de búsquedas con las palabras clave: confiabilidad y Coeficiente del Alfa de Cronbach. Se identificó un vacío en la literatura del alfa porque ninguno de los artículos (n = 111) realizo una evaluación de la literatura de este coeficiente de consistencia interna como las de Henson y Roberts (2006); Hogan et al. (2000); Taber (2017); Whittington (1998); Worthington y Whittaker (2006). Como parte de la justificación del presente manuscrito es llenar este vacío, y la otra parte es servir como una guía para los autores que usen el alfa. Por lo tanto, el alcance es para todas las investigaciones educativas y psicológicas, entre otras, que usen algún tipo de medición y, por ende, un proceso de confiabilidad con el α.

El coeficiente *alfa* ha sido clasificado como un coeficiente de *consistencia interna*. El alfa depende de la homogeneidad de un grupo de ítems que en conjunto miden un constructo (Henson, 2001). Un constructo o atributo psicológico puede ser: desarrollo del lenguaje, madurez social o conocimiento de matemáticas, entre muchos otros más. Crocker y Algina (2006) definieron a un constructo:

> Los atributos psicológicos son constructos. Ellos son conceptos hipotéticos – productos de una imaginación científica informada de científicos de las ciencias sociales quienes intentan desarrollar teorías para explicar el comportamiento humano. La existencia de tales constructos nunca puede ser absolutamente confirmada. De este modo el grado al cual los constructos psicológicos caracterizan a un individuo pueden ser solamente una inferencia de los comportamientos observados. (p. 4)

Los ítems fueron previamente derivados de la definición operacional de un constructo (cf. Crocker y Algina, 2006). En otras palabras, el contenido de los ítems debe de abarcar íntegramente la definición de un constructo. Es importante resaltar que un coeficiente de consistencia interna *no* es una medida directa de *confiabilidad*, sino una estimación teórica derivada de la TCPV (Henson, 2001). Una medida directa seria el peso de una persona en una báscula y la confiabilidad de los puntajes radicaría en la consistencia en la que este aparato midiera los kilogramos en una y otra ocasión. En cambio (i.e., en un modelo teórico), las variables (constructos) no se pueden observar directamente sino a través de una forma indirecta (i.e., instrumento). Se supone que se puede medir este

constructo, pero nunca se podría confirmar su existencia (ver a Crocker y Algina (2006). Haciendo su mejor esfuerzo y en varias ocasiones, cuando una persona corre, canta, cocina, etc., se da cuenta que difícilmente puede tener un resultado idéntico al anterior. Según Feldt y Brennan (1989), esta inconsistencia emana de una variedad de factores, dependiendo en la naturaleza de la medición. La naturaleza puede ser una medición física (e.g., metros, peso, etc.) o psicométrica (i.e., un instrumento). Entre estos factores existen sutiles variaciones en eficiencias físicas y mentales, fluctuaciones incontrolables de las condiciones externas, e inconsistencias en la parte de aquellos que evalúan la actuación de la otra persona (Feldt y Brennan, 1989). Con esta explicación de la variación, se puede esperar que una persona no obtenga el mismo resultado cuando contesta un instrumento. En la TCPV, se usan varios análisis de confiabilidad (e.g., consistencia interna, pre y post test, y formas equivalentes del instrumento). Entre los análisis de consistencia interna están el coeficiente del α (Cronbach, 1951; ecuación 1), y el *KR-20* y *KR-21* (Kuder y Richardson, 1937). Para este manuscrito, se adoptó la confiabilidad de la TCPV porque ha sido muy usada por investigadores con el α (Hogan et al., 2000; Taber, 2017; Thompson, 2002). Según la TCPV, el puntaje verdadero es el nivel que una persona posee de un constructo (e.g. inteligencia, autoestima, etc.). Sin embargo, al medir estos niveles de los constructos, que es lo se podría observar a través de un instrumento, se asume un error de medición (el puntaje que se observa = el puntaje verdadero + un error de medición). El error se define como la discrepancia entre el puntaje verdadero y el observado (Crocker y Algina, 2006). En la fórmula del alfa (ecuación 1) se captura la TCPV:

$$\alpha = (k / k - 1) \text{ X } (1 - \Sigma^{k}_{i=1} \alpha_{yi}^2 / \alpha_{xi}^2) \text{ (ecuación 1)}$$

k = número de ítems.

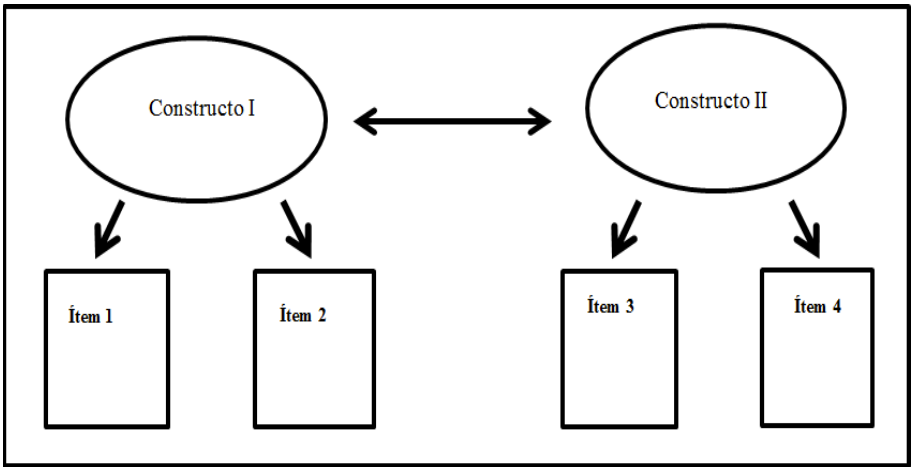
α_{yi}^2 = Varianza dentro de cada ítem; Cuando es $\Sigma^{k}_{i=1} \alpha_{yi}^2$, implica que cada uno de las varianzas se sumara (*suma de varianzas dentro del ítem*).

α_x^2 = Varianza que se aplica a la suma de los puntajes de los ítems a través de cada persona (*varianza entre personas*).

Retomando el tema, la cuantificación de las consistencias e inconsistencias en la persona evaluada constituye la esencia del análisis de confiabilidad (Feldt y Brennan, 1989). En resumen, la confiabilidad puede ser expresada como uno menos la varianza del error para expresarlo con un decimal donde obtener un coeficiente de 1 significaría cero errores de

medición. La Gráfica 1 contiene un ejemplo de un modelo teórico para ser analizado psicométricamente en la TCPV con dos constructos correlacionados y dos ítems por constructo (cf. Crocker y Algina, 2006). La flecha de los constructos hacia los ítems indica una relación de causa y efecto (ver a Byrne, 2016 para modelos causales). En otras palabras, los constructos son variables no observables, pero se suponen que son la causa de los niveles de un participante en una escala (cf. Thompson, 2004). Del nivel de esta asociación se apoya la existencia y medición de un constructo que causa ciertos puntajes alcanzados en los ítems (DeVellis, 2016).

Gráfica 1: Modelo Teórico con Dos Constructos y Dos Ítems por Constructo



Fuente: *el presente manuscrito*

2. ESTRATEGIAS METODOLÓGICAS

La metodología usada fue similar a la de los siguientes autores: Henson y Roberts (2006); Hogan et al. (2000); Taber (2017); Whittington (1998); Worthington y Whittaker (2006). Estos autores identificaron una serie de revistas científicas mediante palabras claves en torno a procesos psicométricos de validación y confiabilidad. Luego, evaluaron estas publicaciones bajo ciertos criterios y dieron recomendaciones. De una manera similar a la de estos autores, en el presente manuscrito se seleccionaron algunas revistas y artículos para compararlas con ciertos criterios y dar recomendaciones.

2.1 Selección de las revistas

Se usaron el portal del CONACYT y la página de SJCR para seleccionar las cuatro revistas (para la justificación de la selección de estas revistas ver la sección: ***Fuente de la Muestra de Artículos***). El CONACYT tenía listadas 137 revistas científicas en su portal en el verano del 2017. Se hizo una búsqueda con la palabra *educación* y se

encontraron nueve fuentes de las cuales solo cuatro (Tabla 1) aparecen en la página de SJCR (ver a SCImago, 2007). SJCR estimó el peso de las cuatro revistas que se refiere al número de citas que los artículos han tenido. SCImago (2007, párr. 1) definió su índice para ranquear las revistas: "Es una medición del impacto de una revista. Influencia o prestigio. Expresa el numero promedio de citas recibidas en el año seleccionado por los documentos publicados en la revista en los 3 años previos."

Tabla 1. Revistas de la muestra.

Fuente	**Ranking según SJCR para el 2015:** Lugar en el ranking**; *índice*; cuarto en el que se ubican**
Revista Electrónica de Investigación Educativa (REDIE)	**862**; *0.105*; Cuarto: 4
Revista Latinoamericana de Investigación en Matemática Educativa (RLIME)	**731**; *0.16*; Cuarto: 4
Revista Mexicana de Investigación Educativa (RMIE)	**465**; *0.348*; Cuarto: 2
Perfiles Educativos (PE)	**875**; *0.103*; Cuarto: 4

Nota: El CONACYT las clasifico con el número IV y bajo Humanidades y Ciencias de la Conducta.
Fuente*: El CONACYT y SJCR.*

2.2 Selección de los artículos

Para este propósito, se buscaron artículos dentro de las cuatro revistas del 2000-2017 que contuvieran las palabras clave: confiabilidad o Coeficiente del Alfa de Cronbach. Se consideró que a partir del 2000 porque es después de la definición moderna de la confiabilidad de Wilkinson y el Grupo de Trabajo de la APA (1999).

En cuanto a los criterios de evaluación de los artículos, tomando en cuenta la definición moderna del concepto de confiabilidad a través del α y para volver operacional la definición moderna y compararla con la muestra, se construyó una definición operacional de la confiabilidad o consistencia interna basada: Feldt y Brennan (1989); Gronlund y Linn (1990); Thompson (1994); Thompson (1992); Wilkinson y el Grupo de Trabajo de la APA (1999). Esta definición fue: *la confiabilidad o consistencia interna radica en los puntajes del instrumento*. Si el autor se acercaba a esta definición al hablar del α en algún fragmento de cualquier sección de su artículo, se marcaria: *uso apropiado del concepto*.

Por el contrario, si el autor declaraba que el instrumento era confiable en cualquier sección del artículo, se marcaria como: *sin uso apropiado*. Entonces, el autor estaba usando una conceptualización errónea. También, había la posibilidad de que la atribución al alfa no fuera clara (al instrumento o a los puntajes), y, entonces, se marcaria: *Ambiguo*. La última posibilidad era que el artículo no cayera dentro de estas tres categorías antes mencionadas (e.g., sin referencia al alfa) y, entonces, seria: *otro tipo*.

2.3 Información sobre el tamaño del α estimado

Se promediaron los valores del alfa reportados por revista con su *SD* y rango, así como un promedio para las cuatro revistas con *SD y rango*. Para volver operacional el tamaño del α, se adoptaron los mínimos de George y Mallery (2003) así como la posible redundancia explicada por Streiner (2003) cuando se rebasa el α de .90. También, se consideró la relación entre el α y el número de ítems.

Referente al cálculo de un intervalo de confianza del y para volver operacional los *IC* del alfa, se adoptó la sugerencia de Fan y Thompson (2001): era suficiente el reportar que se estimó un intervalo de confianza y como se hizo. También, se describió el tamaño de las muestras de los artículos. Y de la operacionalidad de los α por constructo y global, se buscó en el artículo los reportes de los autores al respecto. Entonces, se extrajo la información al respecto.

2.4 Descripción de las características del estudio como uso de datos empíricos y procesos de validación de los puntajes

Se describen algunas características: el uso de datos empíricos, procesos de validación de los puntajes, tamaño de sus muestras de participantes/observaciones, sus coeficientes de *SD*, creación, adaptación o replicación de un instrumento. Esto sirvió para ilustrar los elementos de la muestra. En calcular algunas estadísticas (i.e., promedio), se utilizó la mediana debido a que es menos sensible a valores extremos que pueden inflar el promedio (ver a Ponce, 2016). Un mayor extremo o atípico puede estar de 2 o más *SD* del promedio.

3. RESULTADOS Y DISCUSIÓN

3.1 Uso de la definición moderna del concepto de confiabilidad a través del α

Según el análisis del presente estudio (Tabla 2), los resultados del total de artículos (*n* = 111) mostraron que el *uso apropiado del concepto* fue de 4.85%, *sin uso apropiado* de 85.44%, *ambiguo* de 9.71%, y ocho artículos (*otro tipo*) no pudieron ser clasificados: Las razones fueron publicación antes de 1999, ensayos sin tratar del alfa, o usaban la

confiabilidad desde la teoría de respuesta al ítem (IRT) que es diferente al concepto de confiabilidad de la TCPV. La revista con el mayor porcentaje del *uso apropiado* fue la RLIME con 20%, pero con un número de artículos pequeño (5) relativo a las otras fuentes. Por otro lado, los autores de RMIE y PE no usaron el concepto apropiadamente.

Tabla 2: Frecuencias y porcentajes del uso del α

Revista	Frecuencia	Años de Publicación	Uso Apropiado (%)	Sin Uso Apropiado (%)	Ambiguo (%)	Otro Tipo
REDIE	60	2000-2017(60)	4(7.14)	49(87.50)	3(5.36)	4
RMIE	38	2002-2016(36*)	0(0)	31(86.11)	5(13.89)	2
PE	8	2010-2017(8)	0(0)	5(83.33)	1(16.67)	2
RLIME	5	2007-2013(5)	1(20)	3(60)	1(20)	0
Total	111	109	5(4.85)	88(85.44)	10(9.71)	8

Nota: *dos artículos de esta revista fueron publicados antes del 1999.

Fuente*: Elaboración propia.*

En la Tabla 3, se encuentra una muestra de algunos de los fragmentos de los artículos por revista que fueron clasificados: *uso apropiado*, *sin uso apropiado*, *ambiguo* u *otro tipo*. Solo se tomaron pequeños fragmentos de los artículos para evidenciar la clasificación, pero también se consideró el contexto (i.e., todo el artículo) para ubicarlo bajo cierta categoría. Por cuestión de espacio, no fue posible colocar todos los fragmentos, pero se dio una muestra de ellos. Por ejemplo (REDIE), el siguiente fragmento se clasifico como *uso apropiado* de la definición: "Los datos psicométricos obtenidos en este estudio sobre la Escala-C proporcionan una fiabilidad y una validez adecuadas para el uso de la escala, tanto a nivel de investigación como a nivel psicopedagógico" (p. 11). Se clasifico de esta manera porque se interpretó que el fragmento se acerca a la definición del concepto al atribuirle las propiedades psicométricas a los datos. Por otro lado, el siguiente fragmento se clasifico como *sin uso apropiado* de la definición: "La confiabilidad del instrumento CPIE, se midió con el alfa de Cronbach…" (p. 5). La razón fue que se interpretó que en este fragmento se estaba atribuyendo la confiabilidad al instrumento en sí mismo lo cual

es erróneo. Otro ejemplo fue el siguiente fragmento que se clasifico como ambiguo: "Estos valores dan cuenta de una adecuada consistencia de las mediciones hechas" (pp. 146-147). La razón para clasificarlo como ambiguo fue que no era claro si la confiabilidad fue atribuida a los puntajes (esto sería lo correcto) o al instrumento (esto no sería lo apropiado).

Tabla 3: Resultados del análisis de contenido de los artículos de la revista

REDIE
Uso Apropiado. **Dentro de los paréntesis se encuentra el número de página del fragmento en el artículo.**
"Los datos psicométricos obtenidos en este estudio sobre la Escala-C proporcionan una fiabilidad y una validez adecuadas para el uso de la escala, tanto a nivel de investigación como a nivel psicopedagógico." (11)
"…uno no valida un instrumento de medición, sino el uso específico que se le da a las puntuaciones o resultados obtenidos." (2)
"…a través de la discusión de sus propósitos, validez y confiabilidad." (37)
"…las puntuaciones tienen un error muy pequeño…" (153)
Sin Uso Apropiado
"La confiabilidad del instrumento CPIE, se midió con el alfa…" (5)
"…high interrater reliabilities can be attained with shell-generated items." (7)
"...en la psicología cognitiva estructural con un interesante apoyo en nuevos usos de herramientas típicas de la psicometría." (1)
"...se trata de un instrumento bien construido..." (10)
"…el CAPIC ofrece un nivel alto de fiabilidad y una consistencia interna adecuada." (12)
Fue Ambiguo
"Estos valores dan cuenta de una adecuada consistencia de las mediciones hechas." (146-147)
"...la consistencia interna mediante el [α];" (151)
"…se procedió a evaluar la consistencia interna…" (195)
Otro Tipo
Un artículo era una entrevista donde no se definió al α. En dos artículos se trató la confiabilidad desde la perspectiva de la Teoría de la Respuesta al Ítem, pero esta es

diferente de la α. En un artículo que fue un ensayo se habló de Cronbach, pero no desde la perspectiva de sus coeficientes de consistencia interna.

RMIE

Uso Apropiado

No hubo un artículo en esta revista donde se reportara el uso apropiado de la definición.

Sin Uso Apropiado

"…dedicada a determinar la validez y confiabilidad del Ceda…" (1278)

"…verificar la validez y confiabilidad del instrumento…" (401)

"La confiabilidad del instrumento se verifico mediante el coeficiente [α]…" (1233)

"El instrumento, validado y confiabilizado mediante procedimiento de jueceo y [α]…" (1233)

"La consistencia interna de los puntajes medida con el [α] fue de .81." (1099)

Ambiguo

"…se estableció la homogeneidad entre ítems para cada una de las tres dimensiones del instrumento mediante análisis de consistencia interna de [α]." (722)

"…los análisis de confiabilidad de Conciencia Social…" (140)

"…el índice de confiabilidad fue de .90 según él [α]." (199)

*En este artículo se calculó el [α], pero no se interpretó este coeficiente.

"…más adelante se detienen en los aspectos de la validez y confiabilidad de las evaluaciones…" (815)

Otro Tipo

Dos artículos fueron publicados antes de 1999 así que no fueron analizados porque este año se tomó como el inicio de la definición moderna.

PE

Uso Apropiado

No hubo un artículo en esta revista donde se reportara el uso apropiado de la definición.

Sin Uso Apropiado

"…para determinar la validez y confiabilidad del instrumento." (89)

"…líneas de investigación que garanticen la confiabilidad y validez de estas pruebas…" (131)

"…una alta correlación que sustenta la consistencia del instrumento." (116)

"…por medio de la elaboración de instrumentos de medición válidos y confiables…" (84)

"Se utilizó un cuestionario con una escala tipo Likert, con una confiabilidad de .92…" (103)
Ambiguo
"…se siguió un procedimiento sistemático que permitió obtener buenas características de validez y confiabilidad." (98)
Otro Tipo
Los otros dos artículos fueron ensayos donde se abordó el concepto de confiabilidad para analizar los procedimientos para el ranking de universidades y de la deserción escolar. Sin embargo, no se trató desde la psicometría del α.
RLIME
Uso Apropiado
"Las respuestas de los participantes presentan un [α] …"(347)
Sin Uso Apropiado
"…para estimar la fiabilidad del instrumento…" (107)
"…para la fiabilidad del instrumento…" (151)
"The reliability of this instrument…" (283)
Ambiguo
"Las propiedades psicométricas de confiabilidad…mostraron que la escala AMMEC es adecuada para los fines que fue diseñada…" (306)

Fuente*: Elaboración propia.*

3.2 Información sobre el tamaño del α estimado

Para el tamaño del α por revista (Tabla 4), el promedio del α fue de .84 (*SD* = .09) y cada una de las fuentes estuvo en un rango promedio de .83-.87. De acuerdo a George y Mallery (2003) estos serían *buenos*. Los valores mínimos estuvieron por abajo del .70 en tres fuentes. La REDIE tuvo un mínimo *pobre* con .55. El resto se situó entre .60 y .70: *cuestionables*. La única fuente con un mínimo *aceptable* fue la PE con .79. Para el máximo, los artículos por publicación estuvieron en rango de *excelente*. Por otro lado, Streiner (2003) diría que estos máximos podrían ser demasiado redundantes porque están arriba de .90. Para incrementar los α menores a .70, se podría estimar el número de ítems a agregar de acuerdo a la fórmula de Spearman-Brown. Taber (2017) diría que habría que considerar el contexto de cada instrumento y muestra para entender mejor la magnitud de estos α.

Tabla 4 Tamaño del α y su reporte por constructo y global

Revista	Promedio del α *y* (*SD*)	(Mínimo-Máximo)	Un α por Constructo (frecuencia de artículos por revista con datos empíricos)	% de α por Constructo entre datos empíricos	Reporte de un α Global (frecuencia de artículos por revista con datos empíricos)	% de Reporte de un α Global entre datos empíricos
REDIE	.85(.11)	(.55-.99)	28(55)	50.90	45(55)	81.82
RMIE	.83(.09)	(.62-.99)	24(34)	70.59	32(34)	94.12
PE	.87(.06)	(.79-.93)	2(4)	50	4(4)	100
RLIME	.83(.11)	(.68-.92)	1(5)	20	4(5)	80
Total	.84(.09)		55(98)	56.12	85(98)	86.73

Fuente: Elaboración propia.

3.3 Cálculo de un intervalo de confianza del α

Se buscó en todos los artículos de la muestra para ver cuáles eran los *IC* reportados. En ninguno de los artículos se reportó un *IC* para el alfa. Al parecer el trabajo de Fan y Thompson (2001) no ha sido tomado en cuenta en lo que concierne a los autores de estos artículos.

3.4 Estimación de un α global y por constructo

En la Tabla 4, se muestran los artículos con datos empíricos en los cuales se calcularon el α por constructo (55 de un total de 98 con datos empíricos: i.e., 56.12%). En estos artículos, se siguió la recomendación de Gardner (1995) de reportar un alfa por constructo que es lo que correspondería. Por revista, la RLIME fue la que menos reporto: con 20%. La RMIE fue la que más reporto: 70.59%. En 85 artículos de 98 se reportó un α global que represento el 86.73% el cual, según DeVellis (2016), sería un *constructo* de segundo nivel donde subyacerían los demás. El porcentaje mayor del reporte de un α Global fue para PE (100%) y la que menos fue RLIME (80%).

3.5 Descripción de las características del estudio como uso de datos empíricos y procesos de validación de los puntajes

Los artículos contenían en su mayoría datos empíricos (Tabla 5): el 89.90% (98 artículos de 109) con datos que permitieron a sus respectivos autores estimar los coeficientes del α. Asimismo, el 66.33% de los artículos mostro algún procedimiento relacionado con la

validación de los datos como el Análisis Exploratorio o Confirmatorio de Factores, así como el análisis de contenido por expertos. En la RLIME fue donde más se usaron los procesos de validación (100%) y donde menos fue la RMIE con un 64.71%.

En estos 98 artículos con datos empíricos, se utilizaron 156 instrumentos. Como había 23 artículos con más de un instrumento y 75 con un solo instrumento, se usó la mediana para evitar la influencia del menor número de artículos sobre le mayor (mediana = 1; *SD* = 1.43). Para más información sobre cómo usar la mediana para evitar valores extremos se recomienda a Ponce (2016).

Tabla 5 Artículos con datos empíricos, procedimientos de validación y tamaño de la muestra

Revista	Frecuencia de Artículos con datos empíricos (frecuencia de artículos por revista)	% de Artículos con datos empíricos por revista	Frecuencia de Procedimiento de Validación (frecuencia de artículos por revista con datos empíricos)	% de Artículos con Procedimiento de Validación	Mediana	*SD*	(Mínimo-Máximo)
REDIE	55(60)	91.67	36(55)	65.45	282.5	74,976.15	(7-548,756)
RMIE	34(36)	94.44	22(34)	64.71	233	25,703.73	(30-15,404)
PE	4(8)	50	3(4)	75	127	268.72	(54-625)
RLIME	5(5)	100	4(5)	80	186	208.02	(63-573)
Total	98(109)	89.90	65(98)	66.33	211.50	58,192.96	

Nota. Entre las cuatro revistas contaron con un total de 156 instrumentos; la mediana = 1 (*SD* = 1.43). *Nota:* Los artículos de REDIE tuvieron 77 instrumentos; mediana = 1 (*SD* = 1.43); RMIE con 63 instrumentos y mediana = 1 (*SD* = 1.85); PE con 11 instrumentos y mediana = 1 (*SD* = 3.50); y RLIME con 5 instrumentos y mediana = 1 (*SD* = 1.81).

Fuente*: Elaboración propia.*

En la Tabla 5, la mediana del tamaño de las muestras de los artículos fue de 211.50 (*SD* = 58,192.96). La REDIE tuvo la mediana más alta con 282.5 (*SD* = 79,976.15), y la con menos fue la PE (mediana = 127; *SD* = 268.72).

La mediana de constructos de los instrumentos fue de 4 (*SD* = 3.90; Tabla 6). La RLIME

tuvo la mediana más alta con 9 constructos (*SD* = 6.68) y la más baja fue PE con 1 constructo (*SD* = 1). También, se muestran el mínimo y máximo número de constructos y sus ítems en los instrumentos por revista.

Tabla 6 Estadísticas descriptivas del número de constructos por instrumento y sus ítems

Revista	Mediana del número de Constructos y (*SD*)	(Mínimo-Máximo)	Mediana del número de Ítems y (*SD*)	(Mínimo-Máximo)	Mediana de los Puntos de la Escala y (*SD*)	(Mínimo-Máximo)
REDIE	4(3.22)	(1-15)	26.50(24.42)	(10-128)	5(1.77)	(2-10)
RMIE	4(4.07)	(1-22)	29.50(35.81)	(10-175)	5(1.90)	(2-10)
PE	1(0)	(1-1)	16(4.12)	(10-20)	5(2.06)	(2-7)
RLIME	9(6.68)	(4-18)	49(25.91)	(20-84)	5(0.45)	(4-5)
Total	4(3.90)		26.50(28.49)		5(1.54)	

Nota. Ninguno de los autores de la muestra reporto un intervalo de confianza para los coeficientes alfa calculados. *Fuente.* Análisis de la presente investigación.

Fuente*: Elaboración propia.*

La mediada del número de ítems fue de 26.50 (*SD* = 28.49) con las cuatro revistas (Tabla 6). La revista con la mediana más alta fue la RLIME con 49 (*SD* = 28.91) y la que tuvo la menor fue PE con 16 (*SD* = 4.12). Asimismo, aparecen el mínimo y máximo número de ítems en los instrumentos por revista.

La mediana de los puntos de la escala fue de 5 tanto para el total como para cada revista (Tabla 6). En lo que difirieron hasta cierto punto las revistas entre sí, fue en sus valores de desviación estándar y en sus mínimos y máximos. El mínimo fue de dos puntos lo cual implicaría que se usó el *KR-20* o *KR-21* para escalas dicotómicas, pero como SPSS calcula el alfa y estos coeficientes sin hacer la distinción en la hoja de resultados, la diferencia debió de haber pasado desapercibida. Por ello, muchos autores no se dan cuanta que usaron algo diferente al alfa.

En el 75.51% de los artículos, se presentó por lo menos algún instrumento de nueva creación o su adaptación (Tabla 7). Se identificó como de nueva creación cuando los autores lo declararon de alguna manera (e.g., se desarrolló un nuevo test, etc.). Asimismo, se identificó como una adaptación de un instrumento cuando los autores declararon que se había modificado un instrumento previo para los propósitos de su propia investigación. En el 12.24% de los artículos clasificados con un instrumento de replicación, los autores

declararon que por lo menos estaban usando un instrumento previamente desarrollado por alguien más en otra ocasión. En un mismo porcentaje de artículos 12.24%, sus autores no declararon el origen de su instrumento.

Tabla 7 Estadísticas descriptivas de los instrumentos de nueva creación o replicación

Revista	Frecuencia de Artículos con datos empíricos	Instrumentos de Nueva Creación o Adaptación	Instrumentos de Replicación	Instrumentos sin Información de su Origen
REDIE	55	42(76.36%)	4(7.27%)	9(16.36%)
RMIE	34	23(67.65%)	8(23.53%)	3(8.82%)
PE	4	4(100%)	0(0%)	0(0%)
RLIME	5	5(100%)	0(0%)	0(0%)
Total	98	74(75.51%)	12(12.24%)	12(12.24%)

Fuente*: Elaboración propia.*

Dados estos resultados, el presente manuscrito fue una evaluación de la literatura en español y una propuesta metodológica en psicometría para el uso del alfa, donde puede visualizarse que en dichos artículos es necesario usar apropiadamente el coeficiente.

La pregunta de investigación planteada sobre: ¿Qué tan apropiadamente reportaron los autores el Coeficiente del Alfa de Cronbach? Puede ser contestada con base a los cinco criterios que se usaron para evaluar la muestra de artículos. Estos criterios se explican en forma jerárquica a continuación desde donde se necesita poner más atención al que menos para futuras publicaciones de artículos.

La mayoría de los autores de la muestra han insistido erróneamente en que la confiabilidad radica en el instrumento per se, y no han seguido las recomendaciones de Feldt y Brennan (1989), Gronlund y Linn (1990,) Thompson (1994), Thompson (1992), Taber (2017) y Wilkinson y el Grupo de Trabajo de la APA (1999). La interpretación moderna y operacional del alfa fue: *la confiabilidad o consistencia interna radica en los puntajes del instrumento*.

Las implicaciones de no interpretar la confiabilidad correctamente es que algunas de las conclusiones que se saquen de los estudios también serán incorrectas (Thompson, (2003). Una conclusión errónea seria que el test es confiable, así que no habría necesidad de calcular el coeficiente del α. Se podría usar tal y como está. Esto iría en contra de lo que dijeron Wilkinson y el Grupo de Trabajo de la APA, 1999.

El tamaño de alfa fue *bueno* en promedio, según dirían George y Mallery (2003). Por otro lado, hubo ciertos coeficientes alfa que fueron posiblemente redundantes (>.90), y esto sería según Streiner (2003): i.e., muy altos. Asimismo, otros coeficientes alfa fueron menores a .70 lo cual los coloca de *pobres* a *cuestionables* de acuerdo a George y Mallery (2003). Al contrario, para el cálculo de cada alfa habría que estudiar a cierta profundidad para entender el contexto (Taber, 2017).

Por otro lado, se podría mejorar bastante el cálculo de *IC* del alfa porque nadie de la muestra lo hizo. Al *no* tomar en cuenta la recomendación de Fan y Thompson (2001) de calcular un *IC* para el alfa, se trata al α como si fuera un parámetro (i.e., invariante), pero es en realidad una estimación (estadística) que contiene un error de muestreo. Esta idea de que un alfa es un parámetro más de una población podría llevar a concluir que no es necesario calcular uno para una muestra de estudio. Sin tener evidencia de la consistencia interna de los puntajes de un instrumento, se podrían realizar relaciones entre variables o comparaciones de grupos que no tendrían evidencia.

Para estimar el nivel de confianza de un *IC* es necesario calcular el tamaño de una muestra. Para el tamaño de la muestra, Bonnett (2002) explicó que el tamaño óptimo de la muestra debe de basarse en criterios la precisión deseada del alfa, entre otros.

4. CONCLUSIÓN O CONSIDERACIONES FINALES

Se concluye que un poco más de la mitad de los autores calculó un alfa por constructo, pero todavía se podría mejorar bastante este aspecto y seguir las recomendaciones de Gardner (1995). La mayoría de los autores reporto un alfa global lo cual implicaría un constructo de segundo nivel según DeVellis (2016).

Los artículos contenían: la mayoría datos empíricos, poco más de la mitad contenían algún procedimiento de validación; la mediana de observaciones estuvo por encima de 200, la mediana de constructos de los instrumentos fue de 4; la mediana del número de ítems fue de 26.50; la mediana de los puntos de la escala fue de 5; y el 75.51% presentó por lo menos algún instrumento de nueva creación o su adaptación.

Entre las limitaciones más importantes de este presente artículo esta que solo se cubrieron cuatro revistas de nueve del CONACYT. Por otro lado, se cubrieron las revistas que si están ranquedas por SJCR y tienen probablemente más peso internacional. Otra limitante fue es que no se explicó como estimar intervalos de confianza para el α (ver a Fan y Thompson, 2001). Estos intervalos de confianza permiten darse una idea de cuando error

hay en una medición y de estimar un rango donde se podría encontrar el α de la población de interés.

La recomendación para los investigadores, editores, y el CONACYT es que sigan los cinco criterios usados para evaluar el alfa, especialmente al CONACYT por su posible peso en la elección de revistas científicas para publicar. El énfasis de la recomendación seria para usar la definición moderna del alfa y de estimación de *IC* de los puntajes del α para darse una idea de dónde estaría el nivel del alfa de la población. Para futuras investigaciones, sería recomendable evaluar más revistas publicadas en español para evaluar sus prácticas referentes al coeficiente alfa.

5. LISTA DE REFERENCIAS

Bonnet, D. (2002). Sample size requirements for testing and estimating Coefficient Alpha. *Journal of Educational and Behavioral Statistics*, *27*(4), 335-340. Recuperado de https://journals.sagepub.com/doi/10.3102/10769986027004335

Byrne, B. (2016). *Structural equation modeling with Amos: Basic concepts, applications, and programming* (3ª Ed.). Nueva York: Routledge.

Churchill, G. y Peter, P. (1984). Research design effects on the reliability of rating scales: A meta-analysis. *Journal of Marketing Research*, 31, 360-375. Recuperado de https://www.jstor.org/stable/3151463?seq=1

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2ª Ed.). Hillsdale, Nueva Jersey: Psychology Press.

Cronbach, L. (1951). Coefficient Alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334. Recuperado de https://link.springer.com/article/10.1007/BF02310555

Crocker, L., y Algina, J. (2006). *Introduction to classical and modern test theory*. Nueva York: Rinehart and Winston.

Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Nueva York: Routledge.

DeVellis, R. (2016). *Scale development: Theory and Applications* (4ª. Ed.). Los Angeles: Sage.

Fan, X. y Thompson, B. (2001). Confidence intervals about score reliability coefficients. *Educational and Psychological Measurement, 61*(4), 517-531. Recuperado de https://psycnet.apa.org/record/2001-01813-008

Feldt, L. y Brennan, R. (1989). Reliability. En *Educational measurement* (3ª Ed.), editado por Robert Linn, 105-146. Nueva York: McMillan.

Fleiss, J. (1986). *Design and analysis of clinical experiments*. Nueva York: Wiley.

Galindo, E.(10 de julio de 2017) El Conacyt. Recuperado de http://www.conacyt.gob.mx/index.php/el-conacyt.

Gardner, P.L. (1995). Measuring attitudes to science: unidimensionality and internal consistency revisited. *Research in Science Education, 25*(3), 283-289. Recuperado de https://link.springer.com/article/10.1007/BF02357402

George, D. y Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference. 11.0 update* (4ª Ed.). Boston: Allyn & Bacon.

Gronlund, N. y Linn, R. (1990). *Measurement and evaluation in teaching* (6ª Ed.). Nueva York: Macmillan.

Henson, R. (2001). Understanding internal consistency reliability estimates: A conceptual primer on Coefficient Alpha. *Measurement and Evaluation in Counseling and Development, 34* (3), 177-188. Recuperado de https://www.tandfonline.com/doi/abs/10.1080/07481756.2002.12069034

Henson, R. y Roberts, K. (2006). Exploratory factor analysis reporting practices in published research. *Advances in social science methodology*, *66*(3), 393-416. Recuperado de https://www.researchgate.net/publication/247728606_Use_of_Exploratory_Factor_Analysis_in_Published_ResearchCommon_Errors_and_Some_Comment_on_Improved_Practice

Herman, B. (2015). The influence of global warning science views and sociocultural factors on willingness to mitigate global warning. *Science Education, 99*(1), 1-38. Recuperado de https://onlinelibrary.wiley.com/doi/abs/10.1002/sce.21136

Hogan, T., Benjamin, A. y Brezinki, K. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement*, 60: 523-531. Recuperado de https://journals.sagepub.com/doi/10.1177/00131640021970691

Kuder, F. y Richardson, M. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*(3), 151-160. Recuperado de https://link.springer.com/article/10.1007/BF02288391

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology.* Nueva York: Columbia University Press.

Messick, Samuel. (1989). Validity. En *Educational measurement* (3ª Ed.), editado por Robert Linn, 13-103. Nueva York: Mcmillan.

Nunnally, J. y Bernstein, I. (1994). The assessment of reliability. *Psychometric theory*, *3*(1), 248-292. Recuperado de https://www.scirp.org/(S(i43dyn45teexjx455qlt3d2q))/reference/ReferencesPapers.aspx?ReferenceID=1960143

Petterson, R. (1994). A meta-analysis of Cronbach's Coefficient Alpha. *Journal of Consumer Research,* 21, 381-391. Recuperado de https://academic.oup.com/jcr/article-abstract/21/2/381/1799516

Ponce, H. (2016). Evaluación de los índices de reprobación de la Universidad usando intervalos de confianza. III Congreso de Investigación Educativa en El Estado de Chihuahua en Ciudad Juárez, Chihuahua, noviembre 7 del 2016.

Sawilowsky, S. (2000). Psychometrics versus datametrics: Comment on Vacha-Haase's reliability generalization method and some EPM editorial policies. *Educational and Psychological Measurement*, 60, 157-173. Recuperado de https://journals.sagepub.com/doi/10.1177/00131640021970439

Schultz, K., Whitney, D. y Zickar, M. (2014). *Measurement Theory in Action: Case Studies and Exercises* (2ª Ed.). Nueva York: Routledge.

Schumacker, R. y Tomek, S. (2013). *Understanding statistics using R*. Nueva York: Springer Science & Business Media.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's Alpha. *Psychometrika*, *74*, 107-120. Recuperado de https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2792363/

Streiner, D. (2003). Starting at the beginning: an introduction to Coefficient Alpha and internal consistency. *Journal of personality assessment*, *80*(1), 99-103. Recuperado de https://www.tandfonline.com/doi/abs/10.1207/S15327752JPA8001_18

Taber, K. (2017). The use of Cronbach's alpha when developing research instruments in science education. *Res Sci Educ*, 1-24. Recuperado de https://www.repository.cam.ac.uk/handle/1810/262956

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association.

Thompson, B. (2003). Understanding reliability and Coefficient Alpha, Really. En *Score reliability: Contemporary thinking on reliability issues*, editado por Bruce Thompson, 3-23. Thousand Oaks, California: Sage.

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement*, 54, 837-847. Recuperado de https://www.karger.com/Article/PDF/337710

Thompson, B. (1992). Two and one-half decades of leadership in measurement and evaluation. *Journal of Counseling and Development*, 70, 434-438. Recuperado de https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1556-6676.1992.tb01631.x

Whittington, D. (1998). How well do researchers report their measures?: An evaluation of measurement in published educational research. *Educational and Psychological Measurement*, *58*(1): 21-37. Recuperado de https://journals.sagepub.com/doi/10.1177/0013164498058001003

Wilkinson, L. y el Grupo de Trabajo de la Asociación Americana de Psicología. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54 (8), 594-604. Recuperado de https://www.apa.org/pubs/journals/releases/amp-54-8-594.pdf

Worthington, R. y Whittaker, T. (2006). Scale development research a content analysis and recommendations for best practices. *The Counseling Psychologist*, *34*(6), 806-838. Recuperado de https://journals.sagepub.com/doi/10.1177/0011000006288127