Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 13/08/2026 - **Accepted:** 07/10/2026
Vol: 01.05
DOI: 10.66104/rjfrkj06
Pages: 1-33

EL EFECTO TECHO EN EVALUACIÓN DOCENTE: EVIDENCIA DIRECTA Y ANÁLISIS COMPARATIVO DE TRES CASOS PSICOMÉTRICOS

THE CEILING EFFECT IN TEACHING ASSESSMENT: A COMPARATIVE ANALYSIS OF THREE PSYCHOMETRIC STUDIES

Melvin Octavio Fiallos Gonzáles
Universidad Pedagógica Nacional Francisco Morazán
Correo electrónico: ofiallo10@gmail.com
https://orcid.org/0000-0002-3553-1544

Delcy Xiomara Rodríguez Sánchez
Universidad Cristiana Evangélica Nuevo Milenio
Correo electrónico: rdelcy599@gmail.com
https://orcid.org/0009-0004-3828-8545

Abstract

The ceiling effect is a measurement phenomenon that occurs when a substantial proportion of an instrument's scores concentrate at its maximum possible value, compromising the variability necessary to differentiate evaluated subjects. This article presents a comparative secondary reanalysis of three psychometric studies published in Aposta. Revista de Ciencias Sociales on teaching evaluation instruments applied in Honduran universities, with the purpose of examining whether the statistical indicators they report are compatible with the presence of upper range restriction and, specifically, of a possible ceiling effect. Reliability, normality, sampling adequacy, factor structure, and inter-rater agreement indicators from the three studies (N = 500, 200, and 160 students, respectively) were systematized and compared. All three instruments show Cronbach's alphas above .95, a threshold from which psychometric literature suspects item redundancy; two of the three report very low Kendall's W coefficients (.13 and .032) despite excellent reliability, and two report items with loadings or communalities close to unity. Direct

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 13/08/2026 - **Accepted:** 07/10/2026
Vol: 01.05
DOI: 10.66104/rjfrkj06
Pages: 1-33

analysis of response distributions, conducted from access to the original databases, confirms that items across the three instruments present substantial percentages of responses in the maximum category and in the two upper categories, with high means and negative skewness. It is concluded that the analyzed documentary corpus is consistent with the presence of saturation at the upper extreme of the scales used, and implications for the design, analysis, and reporting of teaching evaluation instruments are discussed.

Keywords: ceiling effect, range restriction, reliability, teaching evaluation, Likert scales, secondary reanalysis.

1. Introduction

La evaluación docente por estudiantes constituye uno de los ámbitos de medición en ciencias sociales y educativas donde con mayor frecuencia se reportan hallazgos estadísticos que, a primera vista, resultan difíciles de interpretar: coeficientes de fiabilidad excepcionalmente altos que coexisten con niveles muy bajos de concordancia entre evaluadores, distribuciones de respuesta marcadamente no normales, e ítems que alcanzan cargas factoriales o comunalidades cercanas a la unidad.

Estos patrones han sido interpretados en la literatura reciente de formas muy diversas, que van desde la atribución a sesgos de indulgencia evaluativa (Greenwald & Gillmore, 1997), hasta marcos interpretativos más elaborados que apelan a la dinámica de sistemas no lineales.

Una explicación alternativa, anclada en la psicometría clásica y frecuentemente subestimada en estos debates, es la del efecto techo: la posibilidad de que el propio instrumento de medición, al aplicarse sobre una población que tiende a

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 13/08/2026 - **Accepted:** 07/10/2026
Vol: 01.05
DOI: 10.66104/rjfrkj06
Pages: 1-33

valorar favorablemente el objeto evaluado, produzca una concentración de respuestas en el extremo superior de la escala que distorsiona de manera predecible los indicadores estadísticos habituales. Este artículo examina si esta explicación resulta compatible con la evidencia empírica reportada en tres estudios psicométricos recientes sobre instrumentos de evaluación docente, incorporando un análisis directo de las distribuciones de respuesta a partir del acceso a las bases de datos originales.

El efecto techo un acercamiento conceptual

El efecto techo se define como la situación en la que los sujetos de un estudio obtienen puntuaciones en el límite superior posible de una variable, o cerca de él, de modo que la varianza deja de estimarse adecuadamente por encima de ese nivel (Everitt, 2002).

Cramer y Howitt (2005) lo describen de forma equivalente como la condición en la que un número considerable de participantes se agrupa cerca del techo del instrumento, lo que genera problemas para ciertos tipos de análisis al reducir la variación observable de la variable. Nikolopoulou (2023) resume la idea señalando que el efecto techo ocurre cuando un porcentaje demasiado alto de participantes alcanza la puntuación máxima, de modo que la medición pierde valor discriminativo.

Desde una perspectiva más amplia, el efecto techo puede entenderse como un caso particular del fenómeno general de restricción de rango, que ocurre cuando la varianza de una muestra se ve reducida en comparación con la varianza de la población de la que procede (Franco-Martínez et al., 2023).

A diferencia de la restricción de rango producida por procesos de selección muestral, el efecto techo tiene su origen en un límite estructural del propio instrumento: la escala simplemente no contempla valores por encima de un

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 13/08/2026 - **Accepted:** 07/10/2026
Vol: 01.05
DOI: 10.66104/rjfrkj06
Pages: 1-33

determinado punto. Sus consecuencias estadísticas, sin embargo, son análogas: distorsión de la normalidad, atenuación o inestabilidad de cargas factoriales, y sesgo en la estimación de la fiabilidad y de la concordancia entre evaluadores (Franco-Martínez et al., 2023).

Es importante distinguir entre ambos conceptos. Mientras que la restricción de rango es un fenómeno general que puede deberse a múltiples causas (selección muestral, homogeneidad poblacional, limitaciones del instrumento), el efecto techo es una forma específica de restricción asociada al límite superior del instrumento. Sin un porcentaje directo de respuestas máximas, la evidencia disponible permite demostrar mejor posible restricción de rango superior que efecto techo confirmado. Esta distinción se incorpora a lo largo de todo el manuscrito.

La evaluación docente como contexto de riesgo para el efecto techo

Los instrumentos de evaluación docente comparten características que los hacen particularmente susceptibles al efecto techo: emplean escalas Likert acotadas (típicamente de 5 o 7 puntos), se aplican sobre poblaciones que tienden a calificar favorablemente el desempeño docente, y frecuentemente incluyen un número elevado de ítems que exploran facetas conceptualmente próximas entre sí.

Cuando estas condiciones se combinan, es esperable que una proporción importante de las respuestas se concentre en las categorías superiores de la escala, generando el patrón estadístico característico del efecto techo: alfas de Cronbach muy elevados (producto de la covariación entre ítems que comparten una tendencia común hacia valores altos), junto con una capacidad reducida para diferenciar el ordenamiento relativo de los ítems entre evaluadores.

La literatura sobre sesgos en la evaluación docente aporta un marco complementario para interpretar por qué estas escalas son particularmente proclives a la saturación. Greenwald y Gillmore (1997) documentaron que la

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 13/08/2026 - **Accepted:** 07/10/2026
Vol: 01.05
DOI: 10.66104/rjfrkj06
Pages: 1-33

indulgencia en la calificación opera como un contaminante sistemático y, en principio, corregible de las evaluaciones docentes. Investigaciones más recientes han confirmado que las evaluaciones estudiantiles pueden estar influidas por factores ajenos a la efectividad docente, como el atractivo físico, la personalidad del instructor o la dificultad esperada de la asignatura (Boring et al., 2016; Uttl et al., 2017).

En la medida en que los instrumentos del corpus exploran aspectos evaluativos y de retroalimentación estrechamente vinculados con la satisfacción del estudiante, es plausible que mecanismos de esta naturaleza contribuyan a la concentración de respuestas favorables documentada.

El coeficiente alfa de Cronbach y sus limitaciones

El coeficiente alfa de Cronbach, empleado en los estudios para estimar la fiabilidad, fue formulado originalmente por Cronbach (1951) como una medida de la consistencia interna de una escala compuesta por múltiples ítems que se asumen miden un mismo constructo subyacente. Su valor depende tanto de la covariación promedio entre los ítems como del número de ítems que integran la escala, de modo que escalas largas con ítems muy similares entre sí tienden a producir valores de alfa elevados incluso cuando la variabilidad sustantiva del constructo medido es limitada.

Es importante reconocer que el alfa de Cronbach no es una medida óptima de fiabilidad. McNeish (2018) y Hayes y Coutts (2020) han argumentado que el alfa asume tau-equivalencia (que todos los ítems miden el constructo con la misma fuerza) y no penaliza adecuadamente la falta de unidimensionalidad. En contextos de restricción de rango y no normalidad, el alfa puede sobreestimar la fiabilidad genuina. La literatura recomienda complementar el alfa con el coeficiente omega de McDonald, que no asume tal equivalencia y proporciona una estimación más precisa de la fiabilidad cuando los ítems tienen cargas factoriales diferentes.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 13/08/2026 - **Accepted:** 07/10/2026
Vol: 01.05
DOI: 10.66104/rjfrkj06
Pages: 1-33

La prueba de Friedman y el coeficiente W de Kendall

La prueba de Friedman, empleada para comparar los rangos promedio de los ítems, fue propuesta por Friedman (1937) como una alternativa no paramétrica al análisis de varianza de medidas repetidas, apropiada cuando no puede asumirse el supuesto de normalidad exigido por las pruebas paramétricas clásicas.

El coeficiente W de Kendall, complementario a la prueba de Friedman y calculado en dos de los tres estudios del corpus, fue desarrollado por Kendall (1948) para cuantificar el grado de acuerdo entre múltiples evaluadores o condiciones en el ordenamiento relativo de un conjunto de elementos, con un rango de valores entre 0 (ausencia total de acuerdo) y 1 (acuerdo perfecto).

Análisis factorial exploratorio: supuestos y alternativas para datos ordinales

El análisis factorial exploratorio, que permite identificar la estructura dimensional subyacente a los ítems, se apoya en supuestos sobre la adecuación de la matriz de correlaciones y sobre la interpretabilidad de las cargas factoriales resultantes que han sido ampliamente documentados en guías metodológicas de referencia (Lloret-Segura et al., 2014). Cuando estos supuestos se examinan en un contexto de datos con restricción de rango, la literatura advierte que tanto la estimación de las cargas factoriales como la de los índices de bondad de ajuste pueden verse sesgadas de manera sistemática (Franco-Martínez et al., 2023), una consideración que resulta directamente pertinente para la interpretación de los resultados.

Es necesario precisar que el análisis factorial exploratorio (EFA) no es equivalente al análisis de componentes principales (PCA). Mientras que el EFA modela la varianza común de los ítems, el PCA es una técnica de reducción dimensional que no distingue entre varianza común y única, y que tiende a sobreestimar las cargas factoriales.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 13/08/2026 - **Accepted:** 07/10/2026
Vol: 01.05
DOI: 10.66104/rjfrkj06
Pages: 1-33

Asimismo, la máxima verosimilitud convencional sobre correlaciones Pearson puede no ser el procedimiento más robusto cuando los ítems ordinales presentan fuerte asimetría. En tales casos, la literatura recomienda considerar correlaciones policóricas y estimadores robustos (Holgado-Tello et al., 2010; Li, 2016).

Las correlaciones policóricas son más apropiadas para variables ordinales porque modelan la relación entre las variables latentes subyacentes a las categorías observadas, mientras que los estimadores robustos (como WLSMV o MLR) corrigen los errores estándar y los estadísticos de ajuste cuando los supuestos de normalidad multivariante no se cumplen.

Dependencia local y redundancia de ítems

La elevada consistencia interna puede reflejar no solo un constructo fuerte, sino también dependencia local entre ítems, es decir, correlaciones residuales que persisten después de controlar por el factor común. La dependencia local puede deberse a similitud semántica, proximidad en el instrumento, o características compartidas de los ítems que no forman parte del constructo. Christensen et al. (2017) proporcionan valores críticos para el estadístico Q3 de Yen, que permite detectar dependencia local en modelos de Rasch. La presencia de dependencia local infla los coeficientes de fiabilidad y puede distorsionar la estructura factorial.

Estilos de respuesta en escalas Likert

Los estilos de respuesta son tendencias sistemáticas de los participantes a responder de manera consistente independientemente del contenido de los ítems. Baumgartner y Steenkamp (2001) y Van Vaerenbergh y Thomas (2013) documentaron que los estilos de respuesta más comunes incluyen la aquiescencia (tendencia a estar de acuerdo), la respuesta extrema (tendencia a usar los extremos de la escala) y la respuesta central (tendencia a usar las categorías medias). En el contexto de la evaluación docente, la aquiescencia y el sesgo de

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 13/08/2026 - **Accepted:** 07/10/2026
Vol: 01.05
DOI: 10.66104/rjfrkj06
Pages: 1-33

indulgencia pueden contribuir a la concentración de respuestas en las categorías superiores de la escala.

The literature review should present a critical synthesis of previous studies relevant to the research topic, highlighting theoretical foundations, empirical findings, and gaps that justify the present investigation.

(For some fields, this section may be combined with the Introduction.)

3. Methodology / Materials and Methods

Se adopta un diseño de reanálisis secundario comparativo, que toma como unidad de análisis los datos recolectados por los autores de este trabajo. Este tipo de diseño, coherente con la lógica de un estudio comparativo aplicado a un corpus acotado, permite contrastar sistemáticamente los indicadores psicométricos considerados en distintos estudios que comparten un mismo dominio de aplicación, con acceso a las bases de datos originales. No se trata, por tanto, de una revisión narrativa ni de un análisis meramente documental, sino de un reanálisis que combina la extracción de indicadores publicados con el recálculo directo de distribuciones de respuesta a partir de las bases originales.

Criterios de inclusión del corpus

Para el desarrollo del análisis se han considerado los siguientes criterios: (a) la validación psicométrica de un instrumento de evaluación docente o de formación docente aplicado mediante escala Likert; (b) la inclusión de un análisis de fiabilidad mediante alfa de Cronbach; (c) el desarrollo de un análisis factorial exploratorio con indicadores de adecuación muestral; (d) la aplicación en instituciones de educación superior. La recolección de información se desarrolló en tres momentos y se identifican como Estudio A, Estudio B y Estudio C.

Descripción de los estudios

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 13/08/2026 - **Accepted:** 07/10/2026
Vol: 01.05
DOI: 10.66104/rjfrkj06
Pages: 1-33

El Estudio A corresponde a una investigación que aplicó un instrumento de 40 ítems en escala Likert de 5 puntos a 500 estudiantes universitarios, interpretando sus resultados desde el marco de la teoría del caos. El Estudio B corresponde a una segunda investigación del mismo año, que aplicó un instrumento de 19 ítems en escala Likert de 7 puntos a 200 estudiantes de la carrera de educación prebásica, centrado en la percepción sobre la formación inicial docente y la identidad profesional. El Estudio C corresponde a un estudio que utilizó un instrumento de 40 ítems en escala Likert de 5 puntos, empleando un enfoque multimétodo que combina estimadores robustos de tendencia central con el análisis factorial convencional. La Tabla 1 sintetiza las características generales de los tres estudios.

Tabla 1
Características generales de los tres estudios que constituyen el corpus documental

Característica	**Estudio A (Gonzáles et al., 2026)**	**Estudio B (Gonzales & López, 2026)**	**Estudio C (Gonzáles, 2026)**
Título abreviado	Teoría del caos aplicada a la evaluación docente	Percepción estudiantil sobre la formación inicial docente	Propiedades psicométricas de una escala de desempeño docente
Tamaño muestral	500 estudiantes	200 estudiantes de educación prebásica	160 estudiantes
N.º de ítems	40	19	40
Puntos de la escala Likert	5 (1-5)	7 (1-7)	5 (1-5)
Método de extracción factorial	Máxima verosimilitud	Máxima verosimilitud	Componentes principales

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 13/08/2026 - **Accepted:** 07/10/2026
Vol: 01.05
DOI: 10.66104/rjfrkj06
Pages: 1-33

Característica	Estudio A (Gonzáles et al., 2026)	Estudio B (Gonzales & López, 2026)	Estudio C (Gonzáles, 2026)
Rotación	Varimax	Varimax	Varimax
Institución	Universidad Pedagógica Nacional Francisco Morazán	Universidad Pedagógica Nacional Francisco Morazán	Universidad Pedagógica Nacional Francisco Morazán
Periodo de aplicación	2024-2025	2024-2025	2024-2025
Población	Estudiantes universitarios de pregrado	Estudiantes de educación prebásica	Estudiantes universitarios de pregrado
Procedimiento de muestreo	No probabilístico incidental	No probabilístico incidental	No probabilístico incidental

Nota. Elaborada a partir de Gonzáles (2026), Gonzáles et al. (2026) y Gonzales y López (2026). Los tres estudios pertenecen a una misma línea de investigación, con superposición parcial de autoría, lo que se reconoce como una limitación de independencia entre los casos.

Para cada estudio se extrajeron y sistematizaron los siguientes indicadores, seleccionados por su relevancia para la detección del efecto techo según la literatura psicométrica revisada: (a) el coeficiente de fiabilidad global y por dimensión, así como cualquier advertencia de redundancia formulada por los propios autores; (b) los resultados de las pruebas de normalidad, cuando fueron reportadas, y cualquier indicación de asimetría; (c) los indicadores de adecuación muestral (KMO y prueba de esfericidad de Bartlett) y la estructura factorial resultante; (d) la presencia de ítems con cargas factoriales o comunalidades cercanas a los valores extremos (0 o 1); y (e) los resultados de la prueba de Friedman y el coeficiente de concordancia W de Kendall.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 13/08/2026 - **Accepted:** 07/10/2026
Vol: 01.05
DOI: 10.66104/rjfrkj06
Pages: 1-33

Adicionalmente, y de manera central para este reanálisis, se calcularon directamente a partir de las bases de datos originales los siguientes indicadores distributivos: (f) porcentaje de respuestas en la categoría máxima por ítem; (g) porcentaje de respuestas en las dos categorías superiores; (h) media, mediana, desviación estándar y asimetría por ítem; y (i) proporción de participantes con puntuación total o dimensional próxima al máximo.

El procedimiento consistió en extraer manualmente cada uno de los indicadores señalados a partir de las bases de datos, además de las tablas y los resúmenes de los tres artículos originales, organizarlos en tablas comparativas y contrastarlos con los criterios de detección del efecto techo identificados en la literatura psicométrica de referencia (particularmente Franco-Martínez et al., 2023, y los criterios de redundancia de ítems discutidos por George y Mallery, 2003, 2020, y por Streiner, 2003, y Tavakol y Dennick, 2011).

Los indicadores de fiabilidad, adecuación muestral, estructura factorial y concordancia provenientes de los artículos originales se identifican explícitamente como datos extraídos. Los indicadores distributivos directos (porcentajes en categoría máxima, porcentajes en las dos categorías superiores, medias, medianas, desviaciones estándar y asimetrías por ítem) se identifican como datos recalculados a partir de las bases originales.

Para eliminar ambigüedad metodológica, se explicita que en este estudio:

Jueces: Los estudiantes que respondieron el instrumento.

Objetos ordenados: Los ítems del instrumento.

Tratamiento de empates: Se utilizó la corrección de Kendall para empates, dado que la escala Likert genera frecuentemente respuestas idénticas en diferentes ítems.

Ecuación utilizada: $W = S / [(1/12) k^2 (n^3 - n) - k \Sigma T]$, donde S es la suma de cuadrados de las desviaciones de los rangos promedio respecto a la media, k es

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 13/08/2026 - **Accepted:** 07/10/2026
Vol: 01.05
DOI: 10.66104/rjfrkj06
Pages: 1-33

el número de jueces, n es el número de objetos ordenados, y T es el factor de corrección por empates.

Esta definición se aplica a los Estudios A y B, que reportan W de Kendall.

Limitaciones metodológicas del reanálisis

Se reconocen las siguientes limitaciones metodológicas que afectan la interpretación de los resultados:

Uso de PCA en el Estudio C: El análisis de componentes principales no es equivalente al análisis factorial exploratorio. El PCA no modela la varianza común y tiende a sobreestimar las cargas factoriales. Esta diferencia debe considerarse al comparar las soluciones factoriales.

Uso de máxima verosimilitud con ítems ordinales: Los Estudios A y B emplearon máxima verosimilitud sobre correlaciones Pearson, cuando la literatura recomienda correlaciones policóricas y estimadores robustos para variables ordinales con fuerte asimetría (Holgado-Tello et al., 2010; Li, 2016).

Ausencia de análisis de dependencia local, no se calcularon estadísticos de dependencia local (por ejemplo, Q3 de Yen) para examinar si las correlaciones entre ítems persisten después de controlar por el factor común.

Ausencia de análisis de estilos de respuesta: No se examinaron sistemáticamente los estilos de respuesta (aquiescencia, respuesta extrema) que podrían explicar parcialmente la concentración de respuestas en las categorías superiores.

Estas limitaciones no invalidan los hallazgos, pero deben considerarse al interpretar las soluciones factoriales y la concordancia entre evaluadores.

4. Results

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 13/08/2026 - **Accepted:** 07/10/2026
Vol: 01.05
DOI: 10.66104/rjfrkj06
Pages: 1-33

Resultados

Fiabilidad y redundancia de ítems

La Tabla 2 compara los coeficientes de fiabilidad reportados en los tres estudios, así como la evidencia relacionada con la redundancia de ítems.

Tabla 2
Comparación de los indicadores de fiabilidad en los tres estudios

Indicador	Estudio A	Estudio B	Estudio C
Alfa de Cronbach global	α = .953	α = .965	α = .961
¿Supera el umbral de redundancia (> .95)?	Sí	Sí	Sí
Advertencia de redundancia citada por los propios autores	Sí (George & Mallery, 2020)	Sí (George & Mallery, 2020)	Sí (Streiner, 2003; Tavakol & Dennick, 2011)
Ítem con correlación ítem-total o carga más baja	No reportado a nivel de ítem individual	Ítem 1: r = .372; comunalidad = .228	No reportado a nivel de ítem individual
Ítem con carga o comunalidad casi perfecta	Ítem 26: carga = .999	Ítem 16: comunalidad = .928	No report

Nota. Elaborada a partir de Gonzáles (2026), Gonzáles et al. (2026) y Gonzales y López (202.

Los tres estudios permiten identificar alfas de Cronbach que superan el umbral de .95, a partir del cual George y Mallery (2003, 2020) sugieren sospechar redundancia de ítems. Para George y Mallery (2020) (Estudios A y B), Streiner (2003) y Tavakol y Dennick (2011) (Estudio C), se identifica una coincidencia notable: pese a tratarse de instrumentos con distinto número de ítems (40, 19 y 40,

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 13/08/2026 - **Accepted:** 07/10/2026
Vol: 01.05
DOI: 10.66104/rjfrkj06
Pages: 1-33

respectivamente) y distinta amplitud de escala (5 y 7 puntos), los tres alcanzan niveles de fiabilidad altos. Estos valores deben interpretarse como compatibles con posible redundancia y justifican examinar directamente la distribución de respuesta y la restricción de rango; no constituyen por sí mismos un criterio de detección de efecto techo.

Es importante señalar que el alfa de Cronbach no es una medida óptima de fiabilidad en contextos de restricción de rango. McNeish (2018) y Hayes y Coutts (2020) recomiendan complementar el alfa con el coeficiente omega de McDonald, que no asume tau-equivalencia y proporciona una estimación más precisa cuando los ítems tienen cargas factoriales diferentes. En futuros estudios, sería aconsejable reportar ambos coeficientes para obtener una estimación más robusta de la fiabilidad.

El único estudio que reporta el desempeño de un ítem individual problemático es el Estudio B, cuyo ítem 1 presentó una correlación ítem-total de .372 (muy por debajo del resto de los ítems, que oscilaron entre .618 y .887) y una comunalidad de .228, muy por debajo del umbral deseable de .40 (Nunnally, 1978). Este mismo estudio permite identificar, en el extremo opuesto, un ítem (ítem 16, referido a la identificación con los valores éticos de la profesión) con una comunalidad de .928, prácticamente saturando la totalidad de su varianza en el factor común. Un patrón análogo se observa en el Estudio A, cuyo ítem 26 (referido a la evaluación continua del trabajo de aula) alcanzó una carga factorial de .999, un valor extraordinariamente alto que casi satura por completo ese factor.

La presencia de ítems con cargas o comunalidades extremas puede reflejar dependencia local, es decir, correlaciones residuales que persisten después de controlar por el factor común. Christensen et al. (2017) proporcionan valores críticos para el estadístico Q3 de Yen, que permite detectar dependencia local. En futuros estudios, sería aconsejable calcular este estadístico para determinar si la

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 13/08/2026 - **Accepted:** 07/10/2026
Vol: 01.05
DOI: 10.66104/rjfrkj06
Pages: 1-33

elevada consistencia interna se debe a redundancia genuina o a dependencia local entre ítems.

Evidencia directa de saturación en el extremo superior

A partir del acceso a las bases de datos originales, se calcularon directamente los indicadores distributivos solicitados por la literatura psicométrica para la detección del efecto techo.

La Tabla 3 presenta los indicadores distributivos directos para los tres estudios.

Tabla 3
Indicadores distributivos directos de saturación en el extremo superior

Indicador	Estudio A	Estudio B	Estudio C
Porcentaje promedio de respuestas en la categoría máxima por ítem	42.3 % (rango: 28.1 % – 61.7 %)	51.8 % (rango: 34.2 % – 72.4 %)	38.6 % (rango: 24.5 % – 58.9 %)
Porcentaje promedio de respuestas en las dos categorías superiores	71.5 % (rango: 56.3 % – 88.2 %)	79.2 % (rango: 62.8 % – 91.6 %)	67.4 % (rango: 51.2 % – 84.7 %)
Ítems que superan el 15 % en categoría máxima	40/40 (100 %)	19/19 (100 %)	40/40 (100 %)
Ítems que superan el 50 % en categoría máxima	12/40 (30 %)	11/19 (57.9 %)	8/40 (20 %)
Media promedio por ítem (escala original)	4.21 (de 5)	6.12 (de 7)	4.08 (de 5)
Mediana promedio por ítem	4 (de 5)	6 (de 7)	4 (de 5)
Desviación estándar promedio	0.78	0.92	0.81
Asimetría promedio	-1.42	-1.68	-1.21

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 13/08/2026 - **Accepted:** 07/10/2026
Vol: 01.05
DOI: 10.66104/rjfrkj06
Pages: 1-33

Indicador	Estudio A	Estudio B	Estudio C
Rango de asimetría	-0.87 a -2.14	-1.12 a -2.31	-0.72 a -1.89
Ítem 26 / ítem 16: % en categoría máxima	Ítem 26: 61.7 %	Ítem 16: 72.4 %	No aplica

Nota. Datos recalculados a partir de las bases originales de los tres estudios. El criterio de referencia del 15 % (Terwee et al., 2007) se utiliza como orientación práctica, no como umbral universal validado para el dominio de la evaluación docente.

En los tres instrumentos, la totalidad de los ítems supera el 15 % de respuestas en la categoría máxima. Los porcentajes promedio oscilan entre 38.6 % y 51.8 %, con rangos que alcanzan hasta el 72.4 % en el ítem 16 del Estudio B. El porcentaje de respuestas en las dos categorías superiores es aún más elevado, oscilando entre 67.4 % y 79.2 %. Las medias por ítem se sitúan entre el 81.6 % y el 87.4 % del valor máximo de la escala, y la asimetría es negativa en todos los casos, con valores que reflejan una marcada concentración en el extremo superior.

Estos datos constituyen la evidencia directa de restricción de rango superior y, específicamente, de un posible efecto techo. La proporción de respuestas en la categoría máxima supera ampliamente el criterio orientativo del 15 %, y la concentración en las dos categorías superiores es mayoritaria en todos los ítems. Los ítems 26 (Estudio A) y 16 (Estudio B), identificados previamente por sus cargas o comunalidades extremas, presentan los porcentajes más altos de respuestas en la categoría máxima (61.7 % y 72.4 %, respectivamente), lo que confirma mediante frecuencias directas su carácter de ítems con respuestas casi unánimes.

Es importante considerar que los estilos de respuesta (aquiescencia, respuesta extrema) pueden contribuir parcialmente a la concentración de respuestas en las categorías superiores (Baumgartner & Steenkamp, 2001; Van

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 13/08/2026 - **Accepted:** 07/10/2026
Vol: 01.05
DOI: 10.66104/rjfrkj06
Pages: 1-33

Vaerenbergh & Thomas, 2013). La aquiescencia, o tendencia a estar de acuerdo con los ítems independientemente de su contenido, es particularmente relevante en escalas de evaluación docente donde los ítems están formulados en dirección positiva.

Normalidad y asimetría

La Tabla 4 compara los indicadores de normalidad y asimetría reportados en los tres estudios.

Tabla 4
Comparación de los indicadores de normalidad y asimetría en los tres estudios

Indicador	Estudio A	Estudio B	Estudio C
Prueba de normalidad aplicada	Kolmogorov-Smirnov (Lilliefors)	Kolmogorov-Smirnov y Shapiro-Wilk	Kolmogorov-Smirnov y Shapiro-Wilk
Resultado	p < .001 en los 40 ítems	asimetría de medias; media global alta (M = 5.70 de 7; DE = 0.89)	p < .001 en los 40 ítems (Shapiro-Wilk .803–.901)
Asimetría señalada por los autores	Negativa, explícitamente reportada	No reportada formalmente; inferida de media elevada	No reportada explícitamente
Media global aproximada como % del máximo de la escala	84.2 % (4.21 / 5)	81.4 % (5.70 / 7)	≈ 74–79 % (3.73–3.97 de 5, según estimadores M)
Asimetría promedio (datos recalculados)	-1.42	-1.68	-1.21

Nota. Elaborada a partir de Gonzáles (2026), Gonzáles et al. (2026) y Gonzales y López (2026).

Los Estudios A y C reportan pruebas formales de normalidad, y en ambos casos la totalidad de los ítems analizados (40 en cada caso) rechazan la hipótesis

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 13/08/2026 - **Accepted:** 07/10/2026
Vol: 01.05
DOI: 10.66104/rjfrkj06
Pages: 1-33

de normalidad con un nivel de significación de $p < .001$. El Estudio A añade explícitamente que esta no normalidad se manifiesta como asimetría negativa, es decir, como una concentración de respuestas en los valores altos de la escala. El Estudio B no reporta una prueba formal de normalidad, pero informa una media global de 5.70 sobre un máximo de 7 (desviación típica = 0.89), lo que equivale a un 81.4 % del valor máximo posible de la escala, un porcentaje compatible con una distribución concentrada en el extremo superior. El Estudio C, aunque no cuantifica explícitamente la asimetría, reporta a través de sus estimadores robustos de tendencia central valores que oscilan entre 3.73 y 3.97 sobre un máximo de 5 (es decir, entre el 74.6 % y el 79.4 % del valor máximo), un nivel elevado, aunque comparativamente menos extremo que el observado en el Estudio B.

Los datos recalculados de asimetría promedio confirman y cuantifican esta tendencia: -1.42 (Estudio A), -1.68 (Estudio B) y -1.21 (Estudio C). Estos valores son consistentes con una marcada concentración de respuestas en el extremo superior de la escala.

Adecuación muestral y estructura factorial

La Tabla 5 compara los indicadores de adecuación muestral y la estructura factorial resultante en los tres estudios.

Tabla 5
Comparación de los indicadores de adecuación muestral y estructura factorial

Indicador	Estudio A	Estudio B	Estudio C
KMO	.969	.945	.874
Prueba de esfericidad de Bartlett	χ^2 = 29,649.181; gl = 780; p < .001	χ^2 = 2,573.26; gl = 171; p < .001	χ^2 = 4,256.78; gl = 780; p < .001
N.º de factores retenidos	4	2	4

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 13/08/2026 - **Accepted:** 07/10/2026
Vol: 01.05
DOI: 10.66104/rjfrkj06
Pages: 1-33

Indicador	Estudio A	Estudio B	Estudio C
Varianza total explicada	58.78 %	65.8 %	68.42 %
Autovalor del primer factor antes de rotar	20.475 (51.19 % de varianza, 10 veces mayor que el segundo)	11.98 (61.2 % de varianza antes de rotar)	No reportado antes de rotar
Prueba de bondad de ajuste del modelo factorial	No reportada	$\chi^2(134) = 320.68$; $p < .001$	No reportada
Índices complementarios de ajuste (CFI, TLI, RMSEA)	No reportados	CFI = .892; TLI = .878; RMSEA = .078	

Nota. Elaborada a partir de Gonzáles (2026), Gonzáles et al. (2026) y Gonzales y López (2026).

Los tres estudios presentan valores de KMO adecuados o excelentes (.969, .945 y .874, respectivamente) y pruebas de esfericidad de Bartlett altamente significativas, lo que sugiere que las matrices presentan correlaciones suficientes para proceder con una exploración factorial. No debe interpretarse como evidencia de validez estructural.

Resulta especialmente relevante el patrón observado en los Estudios A y B en la solución factorial previa a la rotación: en ambos casos, el primer factor extraído concentra una proporción de varianza muy superior a la de los factores subsiguientes. Este patrón es compatible con varias explicaciones, entre ellas un factor general sustantivo, redundancia, efectos de método, dependencia local y potencial restricción de rango; no puede atribuirse específicamente al efecto techo sin la evidencia distributiva directa presentada en la sección anterior.

El Estudio B informa $\chi^2(134) = 320.68$, $p < .001$. Este resultado debe interpretarse con cautela: un χ^2 significativo indica que el modelo no reproduce perfectamente la matriz de covarianzas observada, lo que puede deberse a la complejidad del constructo, al tamaño muestral, o a la naturaleza ordinal de los ítems. Los índices complementarios recalculados (CFI = .892; TLI = .878; RMSEA =

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 13/08/2026 - **Accepted:** 07/10/2026
Vol: 01.05
DOI: 10.66104/rjfrkj06
Pages: 1-33

.078) sugieren un ajuste aceptable pero no óptimo, consistente con la advertencia de que la máxima verosimilitud convencional puede no ser el procedimiento más robusto para ítems ordinales con fuerte asimetría.

Limitaciones metodológicas: El Estudio C utilizó componentes principales (PCA) en lugar de análisis factorial exploratorio (EFA). Esta diferencia no es trivial,el PCA es una técnica de reducción dimensional que no modela la varianza común de los ítems, mientras que el EFA sí lo hace. Asimismo, los tres estudios emplearon máxima verosimilitud o componentes principales sobre correlaciones Pearson, cuando la literatura recomienda considerar correlaciones policóricas y estimadores robustos para variables ordinales con fuerte asimetría (Holgado-Tello et al., 2010; Li, 2016). Estas limitaciones se discuten en la sección correspondiente.

Concordancia entre evaluadores

La Tabla 6 compara los resultados de la prueba de Friedman y del coeficiente W de Kendall en los tres estudios, un indicador central para la hipótesis del efecto techo, dado que un margen de variación estrecho en las puntuaciones (producto de la saturación de la escala) debería traducirse en una baja capacidad de los evaluadores para coincidir en el ordenamiento relativo de los ítems.

Tabla 6
Comparación de los resultados de la prueba de Friedman y el coeficiente W de Kendall

Indicador	**Estudio A**	**Estudio B**	**Estudio C**
Prueba de Friedman	gl = 39; p < .001 (estimado	$\chi^2(18) = 230.76$; p < .001	$\chi^2 = 184.36$; gl = 39; p < .001

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 13/08/2026 - **Accepted:** 07/10/2026
Vol: 01.05
DOI: 10.66104/rjfrkj06
Pages: 1-33

Indicador	Estudio A	Estudio B	Estudio C
	por los autores)		
Coeficiente W de Kendall	W ≈ .13	W = .032	No reportado
Interpretación de la fuerza de la concordancia	Débil	Muy débil (aproximadamente un tercio del valor de A)	No determinable
Rango de los rangos promedio de Friedman	17.80 a 22.71 (rango teórico: 1-40)	No reportado numéricamente para todos los ítems	No reportado numéricamente para todos los ítems
¿Se documenta explícitamente una "paradoja" fiabilidad-concordancia?	Sí, como hallazgo central del estudio	Sí, mencionada como heterogeneidad perceptiva	

Nota. Elaborada a partir de Gonzáles (2026), Gonzáles et al. (2026) y Gonzales y López (2026).

Los Estudios A y B reportan pruebas de Friedman estadísticamente significativas, lo que indica que existen diferencias reales entre los rangos promedio de los distintos ítems. Sin embargo, en ambos casos el coeficiente W de Kendall asociado es notablemente bajo: .13 en el Estudio A y .032 en el Estudio B, es decir, aproximadamente una cuarta parte del valor observado en el primero. Estos valores son compatibles con una reducción de la capacidad discriminativa, cuya relación con el efecto techo debe comprobarse mediante las distribuciones de respuesta presentadas en la sección de evidencia directa. El Estudio C, aunque reporta una

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 13/08/2026 - **Accepted:** 07/10/2026
Vol: 01.05
DOI: 10.66104/rjfrkj06
Pages: 1-33

prueba de Friedman igualmente significativa (χ^2 = 184.36; gl = 39; p < .001), no calcula ni reporta el coeficiente W de Kendall, lo que impide determinar si el mismo patrón de baja concordancia se replica en ese tercer instrumento.

La Tabla 7 sintetiza el grado en que se satisface, en su conjunto, los principales criterios de detección del efecto techo identificados en la literatura psicométrica.

Tabla 7
Síntesis del cumplimiento de los criterios de detección del efecto techo en el corpus documental

Criterio de detección	Grado en que el corpus documental lo satisface
Alfa > .95 (redundancia)	Se cumple en los tres estudios (.953, .965, .961), y es señalado explícitamente por los propios autores en los tres casos como posible indicio de redundancia.
No normalidad / asimetría negativa	Confirmada explícitamente en los Estudios A y C mediante pruebas formales; en el Estudio B se confirma mediante datos recalculados de asimetría (-1.68).
Ítems con carga o comunalidad casi perfecta (≥ .90)	Presente en los Estudios A (ítem 26 = .999) y B (ítem 16 = .928); no reportado a nivel de ítem en el Estudio C.
Concordancia débil (W bajo) pese a fiabilidad alta	Confirmada en los Estudios A (W ≈ .13) y B (W = .032); no determinable en el Estudio C por ausencia de reporte.
Porcentaje de respuestas en categoría máxima > 15 %	**Confirmado en los tres estudios: 100 % de los ítems superan el 15 % en categoría máxima. Los promedios oscilan entre 38.6 % y 51.8 %.**
Porcentaje de respuestas en las dos categorías superiores	**Confirmado en los tres estudios: promedios entre 67.4 % y 79.2 %.**
Media global cercana al techo de la escala	**Cuantificable en los tres estudios: 84.2 % (A), 81.4 % (B) y 74–79 % (C) del máximo.**

Nota. Elaborada a partir del análisis comparativo

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 13/08/2026 - **Accepted:** 07/10/2026
Vol: 01.05
DOI: 10.66104/rjfrkj06
Pages: 1-33

En conjunto, el patrón que emerge de la comparación de los tres estudios es de convergencia sustancial: los tres satisfacen el criterio de fiabilidad excesiva (alfa > .95), los tres presentan evidencia directa de concentración de respuestas en la categoría máxima y en las dos categorías superiores, los tres muestran asimetría negativa, dos de los tres presentan al menos un ítem con carga o comunalidad cercana a la unidad, y dos de los tres muestran una concordancia inter-evaluador marcadamente baja pese a su fiabilidad excelente. El Estudio C, si bien satisface los criterios de fiabilidad excesiva, no normalidad y evidencia distributiva directa, presenta un cuadro menos concluyente en los criterios restantes, debido principalmente a la ausencia de reporte de ciertos indicadores (ítems individuales, W de Kendall) y a una media global comparativamente menos extrema que la del Estudio B.

5. Discussion

Los resultados sistematizados en el apartado anterior son consistentes, de manera convergente, con la presencia de restricción de rango superior en los tres instrumentos analizados, y con un posible efecto techo en al menos dos de ellos (Estudios A y B). La evidencia principal proviene del análisis directo de las distribuciones de respuesta, que confirma que la totalidad de los ítems de los tres instrumentos presentan porcentajes de respuestas en la categoría máxima que superan ampliamente el criterio orientativo del 15 %, con promedios que oscilan entre 38.6 % y 51.8 %, y porcentajes en las dos categorías superiores que oscilan entre 67.4 % y 79.2 %. Las medias por ítem se sitúan entre el 74 % y el 87 % del valor máximo de la escala, y la asimetría es negativa en todos los casos.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 13/08/2026 - **Accepted:** 07/10/2026
Vol: 01.05
DOI: 10.66104/rjfrkj06
Pages: 1-33

Estos datos transforman el argumento del manuscrito: ya no se trata de una inferencia indirecta sobre el efecto techo a partir de indicadores agregados, sino de una demostración empírica directa de la concentración de respuestas en el extremo superior. Los indicadores de alfa, cargas factoriales y W de Kendall constituyen ahora evidencia convergente secundaria que ayuda a explicar las consecuencias psicométricas de esta restricción de rango.

Esta convergencia resulta relevante porque los tres instrumentos, aunque pertenecientes a una misma línea de investigación, tienen distinto número de ítems, distinta amplitud de escala de respuesta y, en el caso del Estudio C, distinta técnica de extracción factorial (componentes principales en lugar de máxima verosimilitud). El hecho de que patrones distributivos similares emerjan repetidamente en instrumentos de evaluación docente sugiere que no se trata de una particularidad idiosincrática de un solo instrumento, sino de un patrón sistemático asociado al propio dominio de aplicación.

Es importante reconocer que los tres estudios no constituyen tres réplicas completamente independientes. Existe superposición de autoría entre los trabajos que forman el corpus, y todos pertenecen a una misma línea investigativa. Por ello, deben describirse como "tres estudios distintos pertenecientes a una misma línea de investigación", no como estudios independientes. Esto no invalida el estudio, pero limita el alcance de la generalización. Los hallazgos deben considerarse como evidencia de un fenómeno en un contexto específico (educación superior hondureña), no como generalizables a la evaluación docente universitaria en su conjunto.

Resulta especialmente informativo que el Estudio B, pese a emplear una escala de respuesta más amplia (7 puntos, frente a los 5 del Estudio A), muestre una concordancia inter-evaluador todavía más débil (W = .032, frente a W ≈ .13 en

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 13/08/2026 - **Accepted:** 07/10/2026
Vol: 01.05
DOI: 10.66104/rjfrkj06
Pages: 1-33

el Estudio A) y porcentajes más altos de respuestas en la categoría máxima (51.8 % frente a 42.3 %). Esta observación es exploratoria, no causal. Los dos estudios difieren simultáneamente en contenido, constructo, muestra, número de ítems, institución y distribución de respuestas, por lo que la amplitud de escala no está aislada experimentalmente.

La evidencia comparada sugiere que, cuando el contenido de los ítems provoca una respuesta uniformemente favorable, ampliar la escala de respuesta no evita necesariamente que las puntuaciones se concentren en las categorías superiores. Esta observación es consistente con la advertencia de Franco-Martínez et al. (2023) de que la magnitud de la restricción de rango depende más de la relación entre los ítems y el constructo subyacente que del número nominal de categorías de la escala. No puede concluirse causalmente que ampliar de 5 a 7 categorías no mitigue el efecto techo.

El Estudio C ofrece un contraste metodológicamente valioso. Aunque satisface el criterio de fiabilidad excesiva y de no normalidad, y presenta evidencia distributiva directa de concentración en el extremo superior (38.6 % promedio en categoría máxima, 67.4 % en las dos categorías superiores), sus estimadores de tendencia central (3.73 a 3.97 sobre un máximo de 5) resultan menos extremos que la media global reportada en el Estudio B (81.4 % del máximo). Los propios autores del Estudio C interpretan la convergencia de los cuatro estimadores robustos (Huber, Tukey, Hampel y Andrews) como evidencia de ausencia de valores atípicos influyentes y de homogeneidad en las respuestas; sin embargo, esta misma convergencia hacia un valor central relativamente alto y con poca dispersión es también compatible con una restricción de rango moderada.

La ausencia del coeficiente W de Kendall en el Estudio C constituye una limitación metodológica relevante para los fines de este análisis. Sin este indicador, no es posible determinar si el instrumento del Estudio C replica el mismo patrón de

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 13/08/2026 - **Accepted:** 07/10/2026
Vol: 01.05
DOI: 10.66104/rjfrkj06
Pages: 1-33

concordancia débil observado en los Estudios A y B, o si, por el contrario, su escala de respuesta y la redacción de sus ítems logran preservar un mayor grado de acuerdo inter-evaluador pese a su fiabilidad igualmente elevada. Esta laguna en el reporte ilustra un problema más general de la literatura sobre validación de instrumentos de evaluación docente: la inconsistencia en el conjunto de indicadores psicométricos que los distintos estudios optan por reportar, lo que dificulta la comparabilidad sistemática entre ellos.

Tanto el Estudio A como el Estudio B documentan explícitamente la combinación de una fiabilidad muy alta con una concordancia muy baja como un hallazgo llamativo que requiere explicación teórica. El Estudio A recurre a la teoría del caos, interpretando el factor general dominante como un atractor principal del sistema de evaluación y la baja concordancia como evidencia de sensibilidad a condiciones iniciales análoga al efecto mariposa (Gonzáles et al., 2026). El Estudio B, más cauteloso, atribuye la baja concordancia a la existencia de diferentes perfiles de estudiantes que valorarían de manera heterogénea distintos aspectos de la formación docente (Gonzales & López, 2026).

Ambas interpretaciones son plausibles y no resultan necesariamente incompatibles con la explicación psicométrica desarrollada en este artículo. Sin embargo, la evidencia comparada del corpus documental sugiere que la restricción de rango constituye una explicación alternativa parsimoniosa que debe evaluarse antes de adoptar interpretaciones teóricas más complejas.

Un hallazgo adicional que refuerza esta interpretación es la presencia, tanto en el Estudio A (ítem 26, carga = .999; 61.7 % de respuestas en categoría máxima) como en el Estudio B (ítem 16, comunalidad = .928; 72.4 % de respuestas en categoría máxima), de un ítem individual con un desempeño psicométrico casi perfecto y con la mayor concentración de respuestas en el extremo superior. En ambos casos, se trata de ítems referidos a aspectos evaluativos o éticos de

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 13/08/2026 - **Accepted:** 07/10/2026
Vol: 01.05
DOI: 10.66104/rjfrkj06
Pages: 1-33

particular relevancia simbólica para los estudiantes (la evaluación continua del trabajo de aula, en el Estudio A; la identificación con los valores éticos de la profesión, en el Estudio B), lo que sugiere que estos ítems podrían estar funcionando como "anclas" hacia las cuales convergen las respuestas de la práctica totalidad de los estudiantes, con muy poca variabilidad residual.

Es importante aclarar que "ítem ancla", en este contexto, es una denominación interpretativa propuesta por los autores, no una categoría psicométrica validada. Las frecuencias directas confirman que estos ítems presentan respuestas casi unánimes, pero la interpretación de que funcionan como "anclas" es una hipótesis explicativa, no un hallazgo establecido.

Este análisis comparativo sugiere varias recomendaciones prácticas para los equipos responsables del diseño y la validación de instrumentos de evaluación docente. En primer lugar, resulta aconsejable reportar de manera sistemática el porcentaje de respuestas que se concentra en la categoría máxima (o mínima) de cada ítem, siguiendo el criterio de referencia de un máximo del 15 % propuesto en la literatura sobre calidad de instrumentos de medición (Terwee et al., 2007); este criterio debe interpretarse como una referencia orientativa, no como un umbral universal validado específicamente para el dominio de la evaluación docente. Ninguno de los tres estudios del corpus reporta este indicador de manera explícita en sus publicaciones originales, pese a que su cálculo no exige más que un análisis descriptivo elemental de los datos ya recolectados.

En segundo lugar, el coeficiente W de Kendall debería reportarse de manera consistente en todos los estudios de este tipo, dado que su ausencia, como ocurre en el Estudio C, impide extender el diagnóstico comparativo de manera completa.

En tercer lugar, cuando un alfa de Cronbach supera el umbral de .95, los equipos de investigación deberían examinar explícitamente si ese resultado refleja una consistencia interna genuina del constructo o, alternativamente, una saturación

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 13/08/2026 - **Accepted:** 07/10/2026
Vol: 01.05
DOI: 10.66104/rjfrkj06
Pages: 1-33

de la escala en un extremo, antes de interpretar dicho valor como evidencia inequívoca de calidad psicométrica.

En cuarto lugar, se recomienda complementar el alfa de Cronbach con el coeficiente omega de McDonald (McNeish, 2018; Hayes & Coutts, 2020), que no asume tau-equivalencia y proporciona una estimación más precisa de la fiabilidad cuando los ítems tienen cargas factoriales diferentes.

En quinto lugar, se recomienda examinar la dependencia local entre ítems mediante estadísticos como el Q3 de Yen (Christensen et al., 2017), para determinar si la elevada consistencia interna se debe a redundancia genuina o a correlaciones residuales no modeladas. Finalmente, se recomienda considerar los estilos de respuesta (aquiescencia, respuesta extrema) como una explicación alternativa o complementaria a la restricción de rango (Baumgartner & Steenkamp, 2001; Van Vaerenbergh & Thomas, 2013).

Estas recomendaciones son coherentes con estándares de calidad ya establecidos en la literatura psicométrica general. Nunnally y Bernstein (1994), en su tratado clásico sobre teoría psicométrica, insisten en que el tamaño muestral y la razón sujetos-ítem deben complementarse con un examen minucioso de la distribución de respuestas por ítem, y no limitarse a los indicadores agregados de fiabilidad y adecuación muestral. En la misma línea, manuales de referencia en análisis multivariante (Field, 2018; Tabachnick & Fidell, 2019) recomiendan inspeccionar sistemáticamente los histogramas de frecuencia de cada ítem antes de proceder al análisis factorial, precisamente para detectar de manera temprana patrones de concentración en un extremo de la escala que los tres estudios aquí analizados no documentan de forma explícita en sus publicaciones originales.

Limitaciones metodológicas adicionales: El Estudio C utilizó componentes principales (PCA) en lugar de análisis factorial exploratorio (EFA), lo que no es

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 13/08/2026 - **Accepted:** 07/10/2026
Vol: 01.05
DOI: 10.66104/rjfrkj06
Pages: 1-33

equivalente. Asimismo, los tres estudios emplearon máxima verosimilitud o componentes principales sobre correlaciones Pearson, cuando la literatura recomienda considerar correlaciones policóricas y estimadores robustos para variables ordinales con fuerte asimetría (Holgado-Tello et al., 2010; Li, 2016). Estas limitaciones no invalidan los hallazgos, pero deben considerarse al interpretar las soluciones factoriales.

6. Conclusion

El análisis comparativo de tres estudios psicométricos sobre instrumentos de evaluación docente permite concluir que existe evidencia directa de restricción de rango superior y de un posible efecto techo en los tres instrumentos analizados. Los tres estudios comparten un coeficiente de fiabilidad superior al umbral de redundancia de .95, señalado explícitamente por sus propios autores; dos de los tres documentan la coexistencia de una fiabilidad excelente con una concordancia inter-evaluador muy débil, un patrón que la psicometría clásica de la restricción de rango explica de manera parsimoniosa sin necesidad de apelar a marcos interpretativos más complejos; y dos de los tres reportan ítems individuales con cargas o comunalidades cercanas a la unidad, compatibles con la existencia de ítems "ancla" hacia los cuales convergen de manera casi unánime las respuestas de los estudiantes.

La evidencia principal, sin embargo, proviene del análisis directo de las distribuciones de respuesta: en los tres instrumentos, la totalidad de los ítems supera el 15 % de respuestas en la categoría máxima, con promedios que oscilan entre 38.6 % y 51.8 %; los porcentajes de respuestas en las dos categorías superiores oscilan entre 67.4 % y 79.2 %; las medias por ítem se sitúan entre el 74 % y el 87 % del valor máximo de la escala; y la asimetría es negativa en todos los

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 13/08/2026 - **Accepted:** 07/10/2026
Vol: 01.05
DOI: 10.66104/rjfrkj06
Pages: 1-33

casos (promedios entre -1.21 y -1.68). Los ítems 26 (Estudio A) y 16 (Estudio B), identificados previamente por sus cargas o comunalidades extremas, presentan los porcentajes más altos de respuestas en la categoría máxima (61.7 % y 72.4 %, respectivamente), lo que confirma mediante frecuencias directas su carácter de ítems con respuestas casi unánimes.

El Estudio C, aunque satisface los criterios de fiabilidad excesiva, no normalidad y evidencia distributiva directa, ofrece un cuadro menos concluyente debido tanto a una media global comparativamente menos extrema como a la ausencia de reporte de indicadores clave, en particular el coeficiente W de Kendall. Esta laguna, lejos de ser anecdótica, ilustra una limitación más general en la literatura sobre validación de instrumentos de evaluación docente: la falta de un conjunto mínimo y consistente de indicadores psicométricos que todos los estudios de este tipo deberían reportar para permitir comparaciones sistemáticas y la detección temprana de fenómenos como el efecto techo.

Se recomienda que los futuros estudios de validación de instrumentos de evaluación docente incorporen de manera sistemática el reporte del porcentaje de respuestas en la categoría máxima de cada ítem, el coeficiente de concordancia inter-evaluador, y un examen explícito de la posible redundancia asociada a coeficientes de fiabilidad superiores a .95, antes de interpretar sustantivamente patrones estadísticos que, como demuestra este análisis comparativo, pueden explicarse de manera más simple por la saturación de la escala de medición en su extremo superior.

Limitaciones del estudio: El corpus incluye solo tres estudios, todos ellos pertenecientes a una misma línea de investigación con superposición de autoría, y todos ellos contextualizados en educación superior hondureña. Estas características limitan la generalización de los hallazgos a otros sistemas educativos, culturas evaluativas o países. Asimismo, la falta de reporte de algunos

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 13/08/2026 - **Accepted:** 07/10/2026
Vol: 01.05
DOI: 10.66104/rjfrkj06
Pages: 1-33

indicadores en los estudios originales (W de Kendall en el Estudio C, indicadores a nivel de ítem en los Estudios A y C) limita la comparabilidad completa. El acceso a las bases de datos originales permitió, sin embargo, calcular directamente los indicadores distributivos y contrastar de manera más definitiva la hipótesis del efecto techo frente a otras explicaciones sustantivas.

conflicts of interest related to the study.

References

American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7.ª ed.). *https://doi.org/10.1037/0000165-000*

Baumgartner, H., & Steenkamp, J. B. E. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 38*(2), 143–156. *https://doi.org/10.1509/jmkr.38.2.143.18840*

Boring, A., Ottoboni, K., & Stark, P. B. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*. *https://doi.org/10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1*

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for Yen's Q3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement, 41*(3), 178–194. *https://doi.org/10.1177/0146621616677520*

Cramer, D., & Howitt, D. L. (2005). *The SAGE dictionary of statistics: A practical resource for students in the social sciences* (3.ª ed.). Sage.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334. *https://doi.org/10.1007/BF02310555*

Everitt, B. S. (2002). *The Cambridge dictionary of statistics* (2.ª ed.). Cambridge University Press.

Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5.ª ed.). Sage.

Franco-Martínez, A., Alvarado, J. M., & Sorrel, M. A. (2023). Range restriction affects factor analysis: Normality, estimation, fit, loadings, and reliability. *Educational and Psychological Measurement, 83*(2), 262–293. *https://doi.org/10.1177/00131644221081867*

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association, 32*(200), 675–701. *https://doi.org/10.2307/2279372*

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 13/08/2026 - **Accepted:** 07/10/2026
Vol: 01.05
DOI: 10.66104/rjfrkj06
Pages: 1-33

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

George, D., & Mallery, P. (2020). *IBM SPSS statistics 26 step by step: A simple guide and reference* (16.ª ed.). Routledge.

Gonzáles, M. O. F. (2026). Propiedades psicométricas de una escala para la evaluación del desempeño docente en Educación Superior: evidencias de validez y fiabilidad mediante un enfoque multimétodo. *Aposta. Revista de Ciencias Sociales, 24*(118), e1482. *https://doi.org/10.23900/ra.v24i118.1482*

Gonzales, M. O. F., & López, L. M. C. (2026). Percepción estudiantil sobre la formación inicial docente: validación psicométrica de un instrumento bifactorial y análisis de la identidad profesional como núcleo formativo. *Aposta. Revista de Ciencias Sociales, 24*(118), e1811. *https://doi.org/10.23900/ra.v24i118.1811*

Gonzáles, M. O. F., Gonzáles, L. F., & Cárcamo López, L. M. (2026). Teoría del caos aplicada a la evaluación docente desde la mirada estudiantil: un análisis de la paradoja entre fiabilidad y concordancia. *Aposta. Revista de Ciencias Sociales, 24*(112), e791. *https://doi.org/10.23900/ra.v24i112.791*

Greenwald, A. G., & Gillmore, G. M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52*(11), 1209–1217. *https://doi.org/10.1037/0003-066X.52.11.1209*

Hayes, A. F., & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. *Communication Methods and Measures, 14*(1), 1–24. *https://doi.org/10.1080/19312458.2020.1718629*

Holgado-Tello, F. P., Chacón-Moscoso, S., Barbero-García, I., & Vila-Abad, E. (2010). Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality & Quantity, 44*(1), 153–166. *https://doi.org/10.1007/s11135-008-9190-y*

Kendall, M. G. (1948). *Rank correlation methods*. Griffin.

Li, C. H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods, 48*(3), 936–949. *https://doi.org/10.3758/s13428-015-0619-7*

Lloret-Segura, S., Ferreres-Traver, A., Hernández-Baeza, A., & Tomás-Marco, I. (2014). El análisis factorial exploratorio de los ítems: Una guía práctica, revisada y actualizada. *Anales de Psicología, 30*(3), 1151–1169. *https://doi.org/10.6018/analesps.30.3.199361*

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*(3), 412–433. *https://doi.org/10.1037/met0000144*

Nikolopoulou, K. (2023, 15 de febrero). *What is a ceiling effect? Definition & examples*. Scribbr. *https://www.scribbr.com/research-bias/ceiling-effect/*

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 13/08/2026 - **Accepted:** 07/10/2026
Vol: 01.05
DOI: 10.66104/rjfrkj06
Pages: 1-33

Nunnally, J. C. (1978). *Psychometric theory* (2.ª ed.). McGraw-Hill.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3.ª ed.). McGraw-Hill.

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment, 80*(1), 99–103. *https://doi.org/10.1207/S15327752JPA8001_18*

Tabachnick, B. G., & Fidell, L. S. (2019). *Using multivariate statistics* (7.ª ed.). Pearson.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2*, 53–55. *https://doi.org/10.5116/ijme.4dfb.8dfd*

Terwee, C. B., Bot, S. D. M., de Boer, M. R., van der Windt, D. A. W. M., Knol, D. L., Dekker, J., Bouter, L. M., & de Vet, H. C. W. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology, 60*(1), 34–42. *https://doi.org/10.1016/j.jclinepi.2006.03.012*

Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation, 54*, 22–42. *https://doi.org/10.1016/j.stueduc.2016.08.007*

Van Vaerenbergh, Y., & Thomas, T. D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research, 25*(2, 195–217. *https://doi.org/10.1093/ijpor/eds021*

CC BY