



Universidad Pedagógica Experimental Libertador
Vicerrectorado de Investigación y Postgrado
Instituto Pedagógico “Rafael Alberto Escobar Lara”
Subdirección de Investigación y Postgrado

HACIA UNA EVALUACIÓN DOCENTE MÁS SÓLIDA ARTICULANDO ESTIMADORES M, CONTRASTE DE FRIEDMAN, PRUEBAS DE ADITIVIDAD Y MÉTRICAS DE FIABILIDAD

Autor: Melvin Octavio Fiallos Gonzáles

ofiallos10@gmail.com

<https://orcid.org/0000-0002-3553-1544>

Universidad Pedagógica Nacional Francisco Morazán
Tegucigalpa, D.C. - Honduras

PP. 747-769



HACIA UNA EVALUACIÓN DOCENTE MÁS SÓLIDA ARTICULANDO ESTIMADORES M, CONTRASTE DE FRIEDMAN, PRUEBAS DE ADITIVIDAD Y MÉTRICAS DE FIABILIDAD

Autor: Melvin Octavio Fiallos Gonzáles

ofiallos10@gmail.com

<https://orcid.org/0000-0002-3553-1544>

Universidad Pedagógica Nacional Francisco Morazán
Tegucigalpa, D.C. - Honduras

Recibido: Mayo 2026

Aceptado: Julio 2026

Resumen

Esta investigación propone un enfoque estadístico integral para evaluar el desempeño docente mediante estimadores robustos, pruebas de normalidad, análisis de fiabilidad (alfa de Cronbach), contraste de Hotelling y prueba de Friedman. Con una metodología cuantitativa y escala Likert aplicada a 160 estudiantes, se busca validar el instrumento e identificar fortalezas y áreas de mejora. Los resultados muestran consistencia interna adecuada y ausencia de valores atípicos significativos, destacando como fortalezas la organización y retroalimentación, y como debilidades la aplicabilidad del aprendizaje y el dinamismo participativo. El enfoque multimétodo demuestra ser un modelo robusto, evidenciando que la integración de perspectivas estadísticas complementarias ofrece conclusiones más sólidas y matizadas que cualquier técnica aislada.

Palabras clave: inferencia estadística, análisis estadístico, evaluación docente.

TOWARDS A MORE ROBUST TEACHER EVALUATION BY ARTICULATING M- STIMATORS, FRIEDMAN CONTRAST, ADDITIVITY TESTS AND RELIABILITY METRICS

Abstract

This research proposes a comprehensive statistical approach to evaluate teaching performance through robust estimators, normality tests, reliability analysis (Cronbach's alpha), Hotelling's contrast and Friedman's test. Using a quantitative methodology with a Likert scale applied to 160 students, the study aims to validate the instrument and identify



strengths and areas for improvement. The results show adequate internal consistency and the absence of significant outliers, highlighting organization and feedback as strengths, while the applicability of learning and participative dynamism is identified as weaknesses. The multimethod approach proves to be a robust model, demonstrating that the integration of complementary statistical perspectives offers more solid and nuanced conclusions than any isolated technique.

Key words: statistical inference, statistical analysis, teacher assessment.

Introducción

La evaluación docente constituye un componente relevante de los sistemas de aseguramiento de la calidad en la educación superior, esto, debido a que proporciona información que puede utilizarse para la retroalimentación del profesorado, la identificación de áreas de mejora y la toma de decisiones académicas. Sin embargo, la utilidad de estas evaluaciones depende no solo de la calidad del instrumento empleado, sino también de la rigurosidad con que se analizan e interpretan los datos obtenidos. La literatura especializada ha advertido que las puntuaciones de evaluación docente pueden estar condicionadas por múltiples fuentes de variabilidad, entre ellas las características de los estudiantes, los cursos, los instrumentos de medición y el contexto en el que se desarrolla la evaluación, lo que plantea desafíos para la interpretación de los resultados como indicadores directos de la efectividad docente (Abrami et al., 2007; Quansah et al., 2024). De igual manera, la evidencia empírica ha cuestionado la utilización de las evaluaciones estudiantiles como medida única de la efectividad del profesorado, particularmente cuando se pretende establecer una relación directa entre las puntuaciones obtenidas y los resultados de aprendizaje de los estudiantes (Uttl et al., 2017).

A esta problemática de medición se añade un desafío de naturaleza estadística. Los datos provenientes de escalas de evaluación pueden presentar distribuciones asimétricas, concentración de respuestas, valores atípicos, heterogeneidad y otras características que deben ser examinadas antes de seleccionar los procedimientos de análisis. En consecuencia, la aplicación automática de técnicas estadísticas sustentadas en



determinados supuestos puede conducir a interpretaciones poco sensibles a la estructura real de los datos. No obstante, la presencia de desviaciones respecto de la normalidad no implica, por sí misma, que los métodos paramétricos deban ser descartados; la pertinencia de cada procedimiento depende también del tamaño muestral, el diseño del estudio, la escala de medición, la independencia de las observaciones, la homogeneidad de las varianzas y el objetivo específico de la inferencia (Field, 2018; Wilcox, 2017). Por ello, el problema metodológico fundamental consiste en determinar hasta qué punto las conclusiones sobre la evaluación docente permanecen estables cuando los mismos datos son examinados mediante procedimientos estadísticos con diferentes supuestos y niveles de sensibilidad frente a las características de la distribución.

En este sentido, limitar el análisis a un único indicador o procedimiento estadístico puede proporcionar una representación parcial del fenómeno estudiado. La media, por ejemplo, puede ser sensible a valores extremos, mientras que la mediana ofrece una medida de tendencia central más resistente; de manera análoga, las pruebas paramétricas y no paramétricas responden a supuestos y objetivos inferenciales diferentes. La estadística robusta surge precisamente como una alternativa para producir estimaciones menos sensibles a observaciones atípicas y a determinadas desviaciones de los modelos convencionales (Huber, 1981; Maronna et al., 2019). Así, más que asumir que un método es universalmente superior a otro, resulta pertinente examinar la convergencia o divergencia de los resultados obtenidos mediante estrategias analíticas complementarias.

La necesidad de realizar este estudio se fundamenta, por tanto, en la importancia de fortalecer la calidad de las inferencias derivadas de los datos de evaluación docente. Analizar estos datos desde una única perspectiva puede ocultar patrones relevantes o generar conclusiones dependientes de los supuestos particulares del procedimiento utilizado. En contraste, la integración de enfoques descriptivos, paramétricos, no paramétricos y robustos permite valorar el comportamiento de los datos desde diferentes perspectivas y determinar si los principales hallazgos son consistentes ante distintas estrategias analíticas. Este enfoque no pretende sustituir un método estadístico por otro,

sino generar evidencia sobre la estabilidad de las conclusiones y proporcionar una interpretación más completa del fenómeno evaluado.

En consecuencia, el presente estudio tuvo como objetivo analizar en qué medida la integración de múltiples perspectivas estadísticas contribuye a una comprensión más profunda, rigurosa y confiable de los datos de evaluación docente. Para ello, se propuso contrastar los resultados obtenidos mediante diferentes procedimientos estadísticos, considerando las características empíricas de los datos y los supuestos asociados a cada técnica. De esta manera, se busca aportar evidencia metodológica que permita determinar cuándo los resultados son convergentes, cuándo presentan diferencias y cómo dichas diferencias pueden modificar la interpretación de la evaluación docente. Esta perspectiva resulta particularmente pertinente en contextos educativos en los que las decisiones académicas pueden apoyarse en los resultados de estas evaluaciones, pues una interpretación estadísticamente más sólida puede contribuir a reducir conclusiones simplistas y favorecer decisiones sustentadas en evidencia (Maronna et al., 2019; Quansah et al., 2024).

Marco Teórico

Desarrollo de los estimadores robustos

La conceptualización de estimadores robustos nace como respuesta a los estimadores de máxima verosimilitud, ante la presencia de valores que se consideran atípicos. Huber (1964) propuso una clase de estimadores que permiten disminuir la sensibilidad a las observaciones. Los estimadores M (máxima verosimilitud generalizada) que representan una generalización del método de máxima verosimilitud, donde la función de pérdida cuadrática es reemplazada por una función ρ que crece más lentamente para valores grandes de los residuos (Huber, 1981).

Los tipos de estimadores como el M de Huber, buscan la representación de un equilibrio entre eficiencia y robustez Huber (1981), esto permite la asignación de pesos

decrecientes a las observaciones extremas, sin rechazarlas completamente. El estimador bponderado de Tukey permite el rechazo de valores atípicos, este estimador es más robusto que el de Huber y tiene un punto de ruptura del 50% (Beaton y Tukey, 1974; Tukey, 1960). Andrews et al. (1972) proporcionan una evaluación exhaustiva de este y otros estimadores robustos. El estimador de Hampel, permite definir tres regiones de crecimiento lineal, una región intermedia y una extrema. Este estimador es particularmente flexible porque permite un control independiente sobre diferentes rangos de valores extremos. La elección de estas constantes específicas ha sido validada mediante estudios de simulación que demuestran su efectividad en diversas condiciones (Hampel, 1968).

Además, se encuentra el estimador de Onda de Andrews, que produce una transición suave hacia el rechazo completo de valores atípicos. En Andrews (1974) y Andrews et al. (1972) se demostró que este estimador tiene propiedades óptimas bajo ciertas condiciones de contaminación.

Evaluación del desempeño docente

La evaluación del desempeño docente constituye un componente estratégico de los sistemas de aseguramiento de la calidad educativa, en tanto permite generar evidencia sobre las prácticas profesionales que contribuyen al desarrollo de los procesos de enseñanza y aprendizaje. Sin embargo, su significado trasciende la asignación de una puntuación o la clasificación de un docente como satisfactorio o no satisfactorio. El desempeño docente constituye un constructo complejo y multidimensional que involucra conocimientos profesionales, planificación, organización de las experiencias de aprendizaje, interacción pedagógica, estrategias de enseñanza, evaluación del aprendizaje, gestión del ambiente educativo y compromiso con el desarrollo profesional. En consecuencia, su evaluación requiere una definición conceptual previa del constructo, la identificación de dimensiones observables y la selección de procedimientos de medición capaces de representar adecuadamente dicha complejidad (Danielson, 2007; Organization for Economic Cooperation and Development, 2013).

Desde esta perspectiva, evaluar el desempeño docente supone establecer una correspondencia entre lo que se entiende teóricamente por buena enseñanza y las evidencias empíricas utilizadas para valorarla. Por su parte, Danielson (2007) concibe la práctica docente como un conjunto articulado de responsabilidades profesionales agrupadas en dominios relacionados con la planificación y preparación, el ambiente de aprendizaje, la instrucción y las responsabilidades profesionales. Este enfoque resulta particularmente relevante porque desplaza la evaluación desde una concepción centrada exclusivamente en resultados hacia una comprensión más amplia de las prácticas que hacen posible una enseñanza eficaz. La evaluación, bajo esta perspectiva, puede convertirse en un instrumento para promover la reflexión profesional, orientar la retroalimentación y favorecer procesos de desarrollo docente, en lugar de limitarse a funcionar como mecanismo de clasificación.

La dimensión formativa de la evaluación adquiere especial importancia cuando su finalidad es el mejoramiento de la práctica. La Organization for Economic Cooperation and Development (OECD, 2013) sostiene que los sistemas de evaluación docente pueden constituir un mecanismo para fortalecer la calidad de la enseñanza y el desarrollo profesional cuando los procedimientos, criterios y usos de los resultados se encuentran articulados con procesos de retroalimentación y mejora. Esto implica distinguir entre la evaluación concebida con fines formativos y aquella utilizada con propósitos sumativos o de rendición de cuentas, puesto que ambas finalidades pueden requerir diferentes fuentes de evidencia, criterios de interpretación y consecuencias institucionales. En este sentido, la evaluación adquiere valor cuando sus resultados permiten identificar fortalezas y áreas susceptibles de desarrollo, y no únicamente cuando producen una puntuación susceptible de comparación.

La consideración del desempeño docente como un constructo multidimensional tiene consecuencias directas sobre la manera en que deben interpretarse sus resultados. Una puntuación global puede ofrecer una síntesis útil del comportamiento de los datos, pero también puede ocultar diferencias sustantivas entre las dimensiones que integran el constructo. Dos docentes podrían obtener promedios globales similares y,



simultáneamente, presentar perfiles diferentes en aspectos como organización de la enseñanza, claridad de las explicaciones, interacción con los estudiantes, estrategias didácticas o evaluación del aprendizaje. Por esta razón, la evaluación del desempeño no debería depender exclusivamente de un indicador agregado, sino considerar la estructura dimensional del instrumento y el comportamiento específico de sus componentes.

Esta consideración resulta particularmente importante en la educación superior, donde las prácticas docentes se desarrollan en contextos disciplinares, institucionales y pedagógicos diversos. En este ámbito, las evaluaciones realizadas por estudiantes constituyen una de las fuentes más utilizadas para obtener información sobre la experiencia de enseñanza. Sin embargo, una evaluación de este tipo representa una fuente específica de evidencia y no necesariamente el constructo completo de desempeño docente. Una revisión sistemática reciente encontró que la validez de las evaluaciones estudiantiles está condicionada por diferentes fuentes de error y por factores relacionados con el contexto, el diseño del instrumento y las condiciones en que se realiza la evaluación (Quansah et al., 2024).

La literatura también ha documentado que las puntuaciones obtenidas mediante evaluaciones estudiantiles pueden verse afectadas por variables que no necesariamente representan la calidad de la enseñanza. Estudios sobre la validez de estas evaluaciones han identificado efectos asociados con la percepción de agrado o simpatía hacia el docente y con el interés previo de los estudiantes por la asignatura, factores que pueden introducir distorsiones en las puntuaciones (Spooren et al., 2013; Zabaleta, 2007). De manera más reciente, Cook et al. (2024) desarrollaron un modelo para examinar diferentes fuentes de sesgo en las evaluaciones estudiantiles y señalaron que factores estadísticos, cognitivos, contextuales y temporales pueden contribuir a generar representaciones potencialmente inexactas del desempeño docente.

Por tanto, una cuestión consiste en distinguir entre medir una percepción sobre la enseñanza y evaluar integralmente el desempeño docente. Las evaluaciones estudiantiles pueden proporcionar información valiosa sobre determinadas dimensiones de la



experiencia educativa, pero no deberían interpretarse automáticamente como una medida exhaustiva de la efectividad profesional del docente. Esta distinción es fundamental para evitar inferencias que excedan aquello que realmente permiten sostener los datos.

La discusión sobre la validez de las evaluaciones docentes ha adquirido particular relevancia debido al uso creciente de sus resultados en decisiones institucionales. Las evaluaciones estudiantiles pueden utilizarse para proporcionar retroalimentación, modificar estrategias de enseñanza e informar procesos de promoción o desarrollo profesional; sin embargo, la literatura muestra resultados heterogéneos respecto de su relación con otros indicadores de efectividad docente y aprendizaje. Uttl et al. (2017), por ejemplo, encontraron que las evaluaciones estudiantiles no presentaban una relación consistente con el aprendizaje de los estudiantes.

La convergencia de estos hallazgos permite sostener que la evaluación docente debe abordarse desde una perspectiva de múltiples evidencias. La cuestión no consiste en determinar si una fuente de información es válida o inválida de manera absoluta, sino en establecer qué dimensión del desempeño representa, cuáles son sus fuentes potenciales de error y cómo puede complementarse con otras evidencias. Esta perspectiva es coherente con los enfoques contemporáneos de evaluación, en los cuales la interpretación de un constructo complejo requiere considerar diferentes indicadores y fuentes de información antes de formular conclusiones sobre el desempeño profesional (OECD, 2013; Quansah et al., 2024).

En consecuencia, la selección de los procedimientos estadísticos debe responder a las características de las variables, al nivel de medición, a la distribución de las observaciones, a la presencia de valores atípicos, a la estructura de las dimensiones y, fundamentalmente, a las preguntas que orientan la investigación.

Esta consideración resulta especialmente relevante cuando los datos proceden de escalas de respuesta ordinales y presentan distribuciones asimétricas, concentración de



respuestas o valores extremos. En tales circunstancias, la utilización exclusiva de estadísticas basadas en supuestos específicos puede proporcionar una representación incompleta del fenómeno. Sin embargo, metodológicamente tampoco resulta adecuado asumir que toda desviación de la normalidad obliga automáticamente a sustituir los procedimientos paramétricos por pruebas no paramétricas. La decisión debe considerar conjuntamente los supuestos del método, las características de los datos, el diseño de investigación y el objetivo inferencial. Desde esta perspectiva, la estadística robusta proporciona herramientas para reducir la sensibilidad de las estimaciones frente a observaciones influyentes y determinadas desviaciones de los supuestos convencionales (Huber, 1981; Maronna et al., 2019; Wilcox, 2017).

Por consiguiente, el análisis estadístico de la evaluación docente debe orientarse a determinar no solo los niveles promedio observados, sino también la estabilidad de los resultados bajo diferentes procedimientos analíticos. El análisis descriptivo permite caracterizar la distribución de las puntuaciones y establecer patrones generales; las técnicas inferenciales permiten examinar diferencias y asociaciones; los procedimientos no paramétricos permiten contrastar determinadas hipótesis sin depender de supuestos distribucionales tan restrictivos; y los métodos robustos permiten valorar la influencia de observaciones potencialmente atípicas sobre las estimaciones y conclusiones. La complementariedad entre estos enfoques permite determinar si los hallazgos son consistentes o si dependen de las características particulares de un procedimiento estadístico.

Marco metodológico

El enfoque del estudio fue cuantitativo, con modalidad descriptiva siguiendo a Creswell (2009) y Cardeña Ojeda (2024). Se ha considerado una muestra para el pilotaje (Bernal, 2010; Sepúlveda García et al., 2018) de 160 estudiantes en Honduras, con el objetivo de verificar la fiabilidad, previo se ha generado la revisión por cinco jueces expertos para garantizar la validez del contenido, para facilitar la eliminación de ítems poco relevantes (Pulido y Rodríguez, 2014). Además, se ha considerado la revisión teórica

de la matriz de variables para el desarrollo lógico de construcción del cuestionario.(Fiallos y Fiallos, 2024; Pineda García et al., 2025). Las variables que se ha considerado se muestran en la tabla 1.

Tabla 1

Evaluación teórica de los constructos en el contexto educativo

Variables	Indicador	Sub indicador
Diseño y desarrollo micro currículo	Dominio metodológico	Estrategias de enseñanza
		Estrategias de aprendizaje
		Manejo de metodologías didácticas
	Evaluación del aprendizaje	Estrategias de evaluación
		Diseño evaluativo
		Planificación de la evaluación
	Retroalimentación	Evaluación de proceso y producto
		Devolución con observaciones
	Dinámica de aula	Interacción pedagógica
		Actividad pedagógica

El dominio metodológico Beltrán (2017); Bermejo y Ballesteros (2014), se refieren al dominio en la organización y selección de las estrategias relacionadas a la enseñanza, aprendizaje con relación al tiempo y los contenidos. La evaluación del aprendizaje nos acerca al proceso de planificación de la evaluación, el diseño de instrumentos y las estrategias que se integran para la verificación de aprendizajes (Olivos, 2016). La retroalimentación está considerada como medio de reflexión de lo que se aprende apoyada de la dinámica de aula para la interacción y las actividades de aula (Gimeno Sacristán, 1998).

Resultados, análisis y discusión

Los datos en la Tabla 2, muestran una notable consistencia entre los cuatro estimadores-M para todos los ítems evaluados. Para el ítem *El docente utiliza diferentes*

recursos didácticos, los estimadores producen valores de 4.03 (Huber), 4.03 (Tukey), 4.00 (Hampel) y 4.03 (Andrews). Esta consistencia tiene implicaciones importantes.

En primer lugar, la ausencia de diferencias significativas entre los estimadores sugiere que no hay valores atípicos influyentes en los datos. Como señalan Maronna et al. (2019), si existieran valores atípicos severos, los estimadores más robustos (Tukey y Hampel) producirían estimaciones más bajas que el estimador de Huber, que es menos resistente.

La similitud observada indica que la distribución de respuestas es homogénea y que los promedios reportados son representativos de la tendencia central. En segundo lugar, la consistencia entre métodos robustos valida el uso de estadísticos descriptivos convencionales, aunque con la precaución de que los datos no son normales, como se demostrará más adelante.

Tabla 2
Estimadores M

Ítems	Estimador- M de Huber ^a	Biponderado de Tukey ^b	Estimador- M de Hampel ^c	Onda de Andrews ^d
El docente utiliza diferentes formas de explicar los contenidos de la clase.	3.52	3.51	3.49	3.51
Las explicaciones del docente facilitan la comprensión de los temas.	3.90	3.89	3.84	3.89
El docente utiliza ejemplos prácticos durante la clase.	3.74	3.87	3.74	3.87
El docente promueve la participación de los estudiantes.	3.98	3.98	3.93	3.98
Las actividades de clase ayudan a comprender mejor los contenidos.	3.73	3.72	3.67	3.72
En las clases se participa dinámicamente en las actividades.	3.49	3.48	3.49	3.48
Las actividades permiten construir nuestro propio aprendizaje.	3.88	3.85	3.82	3.85

Cont. Tabla 2.

La evaluación considera diferentes aspectos del aprendizaje.	3.85	3.83	3.81	3.83
El docente explica los resultados de las evaluaciones.	4.08	4.09	4.05	4.09
Los comentarios recibidos permiten mejorar nuestro trabajo.	3.96	3.95	3.92	3.95
El docente orienta como mejorar el aprendizaje	4.09	4.08	4.02	4.08
La retroalimentación se realiza a tiempo.	3.76	3.74	3.70	3.74
Las observaciones ayudan a mejorar el aprendizaje.	3.97	3.97	3.91	3.97
En clase existe participación activa.	3.92	3.90	3.87	3.90
Los estudiantes interactúan entre sí.	3.91	3.89	3.87	3.89
Se promueve el diálogo durante la clase.	3.97	3.96	3.91	3.96
Las clases son dinámicas.	3.77	3.73	3.73	3.73
Las actividades de clase mantienen el interés de los estudiantes.	4.09	4.09	4.07	4.09

a. La constante de ponderación es 1.339.

b. La constante de ponderación es 4.685.

c. Las constantes de ponderación son 1.700, 3.400 y 8.500.

d. La constante de ponderación es $1.340 \cdot \pi$.

La verificación del supuesto de normalidad constituye un paso crítico en el análisis estadístico, particularmente cuando se planea utilizar pruebas paramétricas. Las pruebas de Kolmogorov-Smirnov (con corrección de Lilliefors) y Shapiro-Wilk, mostradas en la tabla 4 son dos de los métodos más utilizados para evaluar si una muestra proviene de una población con distribución normal (Razali & Wah, 2011).

La prueba de Shapiro-Wilk es generalmente considerada más potente que la de Kolmogorov-Smirnov, especialmente para muestras pequeñas y moderadas ($n \leq 50$). Su estadístico W varía entre 0 y 1, donde valores cercanos a 1 indican normalidad. Por su parte, la prueba de Kolmogorov-Smirnov compara la función de distribución empírica con la función de distribución normal teórica.

Tabla 3

Pruebas de normalidad del cuestionario

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
El docente utiliza diferentes formas de explicar los contenidos de la clase.	.237	40	.000	.898	40	.002
Las explicaciones del docente facilitan la comprensión de los temas.	.230	40	.000	.865	40	.000
El docente utiliza ejemplos prácticos durante la clase.	.294	40	.000	.863	40	.000
El docente promueve la participación de los estudiantes.	.305	40	.000	.822	40	.000
Las actividades de clase ayudan a comprender mejor los contenidos.	.264	40	.000	.865	40	.000
En las clases se participas dinámicamente en las actividades.	.189	40	.001	.901	40	.002
Las actividades permiten construir nuestro propio aprendizaje.	.214	40	.000	.846	40	.000
Se trabaja en equipo para realizar actividades.	.260	40	.000	.862	40	.000
Es aplicable a la realidad lo aprendido en clase	.195	40	.001	.886	40	.001
Se puede aplicar lo aprendido en diferentes situaciones.	.219	40	.000	.901	40	.002
Reflexionamos sobre lo que aprendemos en clase.	.191	40	.001	.900	40	.002
El docente utiliza diferentes métodos para enseñar.	.225	40	.000	.892	40	.001
Las actividades de clase están bien organizadas.	.244	40	.000	.864	40	.000
Las metodologías utilizadas facilitan el aprendizaje.	.190	40	.001	.864	40	.000
El docente utiliza diferentes recursos didácticos.	.271	40	.000	.840	40	.000
Las actividades ayudan a comprender mejor los temas.	.227	40	.000	.844	40	.000
Las evaluaciones están relacionadas con lo visto en clase.	.246	40	.000	.883	40	.001
Se utilizan diferentes formas de evaluación.	.194	40	.001	.859	40	.000
Con las evaluaciones se reflexiona sobre lo aprendido	.213	40	.000	.889	40	.001
Se socializan los criterios y formas de evaluación	.248	40	.000	.875	40	.000
Las evaluaciones reflejan el aprendizaje.	.228	40	.000	.888	40	.001
Las evaluaciones están organizadas durante el curso.	.235	40	.000	.829	40	.000
Los instrumentos de evaluación son claros	.280	40	.000	.849	40	.000
Las instrucciones de evaluación son explicadas claramente	.222	40	.000	.838	40	.000
Los procesos de evaluación son importantes para conocer nuestro progreso	.228	40	.000	.876	40	.000

Cont. Tabla 3.

El docente evalúa el trabajo de aula durante el desarrollo de las actividades.	.200	40	.000	.870	40	.000
Se evalúan tanto los procesos como los resultados de las actividades de aula.	.261	40	.000	.868	40	.000
Las actividades de clase forman parte de la evaluación.	.235	40	.000	.859	40	.000
Los trabajos finales son evaluados con criterios claros.	.233	40	.000	.867	40	.000
La evaluación considera diferentes aspectos del aprendizaje.	.234	40	.000	.870	40	.000
El docente explica los resultados de las evaluaciones.	.285	40	.000	.803	40	.000
Los comentarios recibidos permiten mejorar nuestro trabajo.	.249	40	.000	.860	40	.000
El docente orienta como mejorar el aprendizaje	.268	40	.000	.820	40	.000
La retroalimentación se realiza a tiempo.	.199	40	.000	.877	40	.000
Las observaciones ayudan a mejorar el aprendizaje.	.292	40	.000	.823	40	.000
En clase existe participación activa.	.208	40	.000	.855	40	.000
Los estudiantes interactúan entre sí.	.265	40	.000	.857	40	.000
Se promueve el diálogo durante la clase.	.287	40	.000	.836	40	.000
Las clases son dinámicas.	.219	40	.000	.865	40	.000
Las actividades de clase mantienen el interés de los estudiantes.	.246	40	.000	.830	40	.000

a. Corrección de la significación de Lilliefors

Los resultados de ambas pruebas son concluyentes: para todos los 40 ítems, el valor de significación (Sig.) es .000, lo que indica que se rechaza la hipótesis nula de normalidad ($p < .001$). Esto tiene implicaciones metodológicas para el análisis de la información.

La primera implicación se dirige hacia que los datos de evaluación docente no cumplen el supuesto de normalidad requerido para pruebas paramétricas. Como señalan Frumento y Bottai (2017), la violación de este supuesto puede llevar a inferencias incorrectas si se utilizan métodos paramétricos sin las debidas correcciones. Una segunda implicación apunta a que la no normalidad no se debe necesariamente a la presencia de valores atípicos, como ya demostró la consistencia de los estimadores-M. Esta puede estar ligada a respuesta a la naturaleza de las escalas Likert que producen distribuciones asimétricas y discretas (Rhemtulla et al., 2012). Una tercera implicación es que la violación

Cont. Tabla 4.

Los procesos de evaluación son importantes para conocer nuestro progreso	20.93
El docente evalúa el trabajo de aula durante el desarrollo de las actividades.	20.11
Se evalúan tanto los procesos como los resultados de las actividades de aula.	20.10
Las actividades de clase forman parte de la evaluación.	22.50
Los trabajos finales son evaluados con criterios claros.	21.46
La evaluación considera diferentes aspectos del aprendizaje.	20.90
El docente explica los resultados de las evaluaciones.	23.20
Los comentarios recibidos permiten mejorar nuestro trabajo.	22.03
El docente orienta como mejorar el aprendizaje	23.15
La retroalimentación se realiza a tiempo.	19.05
Las observaciones ayudan a mejorar el aprendizaje.	21.34
En clase existe participación activa.	21.70
Los estudiantes interactúan entre sí.	21.78
Se promueve el diálogo durante la clase.	22.08
Las clases son dinámicas.	20.38
Las actividades de clase mantienen el interés de los estudiantes.	23.66

El análisis de los rangos promedio más altos o mejor evaluados por cada estudiante consultado son: Las actividades de clase mantienen el interés de los estudiantes (23.66), Las evaluaciones están organizadas durante el curso (23.34), El docente explica los resultados de las evaluaciones (23.20), El docente orienta cómo mejorar el aprendizaje (23.15), Las instrucciones de evaluación son explicadas claramente (23.03). Esto sugiere que el docente es particularmente efectivo en aspectos relacionados con organización, claridad en la evaluación, y retroalimentación. Los ítems más bajos, Se puede aplicar lo aprendido en diferentes situaciones (15.53), En las clases se participa dinámicamente en las actividades (16.64), Reflexionamos sobre lo que aprendemos en clase (17.66), El docente utiliza diferentes formas de explicar los contenidos (17.38), con ello se identifican áreas de mejora, particularmente en lo que respecta a transferencia del aprendizaje, dinamismo y reflexión metacognitiva.

A lo anterior se adiciona la fiabilidad del cuestionario, el coeficiente Alfa de Cronbach (α) es la medida más utilizada para evaluar la consistencia interna de un instrumento de medición. Desarrollado por Cronbach (1951), se basa en el promedio de las correlaciones entre los ítems que componen una escala, Los criterios comúnmente

aceptados para interpretar el alfa de Cronbach son (George & Mallery, 2003): $\alpha \geq 0.9$: Excelente, $0.8 \leq \alpha < 0.9$: Bueno, $0.7 \leq \alpha < 0.8$: Aceptable, $0.6 \leq \alpha < 0.7$: Cuestionable, $\alpha < 0.6$: Inaceptable.

El coeficiente Alfa de Cronbach obtenido es .961, clasificado como *excelente* según todos los criterios estándar. Indica que los 40 ítems del cuestionario miden un mismo constructo subyacente de manera consistente. En términos psicométricos, esto sugiere que el instrumento tiene alta unidimensionalidad, es decir, todos los ítems contribuyen a medir la misma variable latente. El valor idéntico del alfa basado en elementos tipificados (.961) indica que no hay problemas de escalamiento. Esto significa que las varianzas de los ítems son homogéneas, lo cual es un requisito para la interpretación del alfa como medida de consistencia interna. Un alfa tan alto sugiere que los ítems podrían ser redundantes. Los valores de α superiores a 0.95 pueden indicar que algunos ítems están midiendo exactamente lo mismo, lo que podría justificar una reducción de la escala. De manera complementaria se considera la prueba de Hotelling es el vector hipotético de medias poblacionales, y la matriz de covarianzas muestral (Anderson, 2003).

Perlman (2019) señala que la prueba T^2 de Hotelling, es uno de los triunfos del análisis multivariante clásico, siendo uniformemente más potente entre las pruebas invariantes, admisible, Bayes propia, y local y asintóticamente minimax entre todas las pruebas. Los resultados de la tabla 5, muestran un T-cuadrado de Hotelling de 480.468, que se transforma en un estadístico F de .316 con 39 y 1 grados de libertad, produciendo una significación de .917 ($p > .05$). Esto indica que no se rechaza la hipótesis nula de que el vector de medias de los ítems es igual a cero.

Tabla 5
Prueba T cuadrado de Hotelling

T-cuadrado de Hotelling	F	gl1	gl2	Sig.
480.468	.316	39	1	.917



Conclusiones

El análisis integrado de los datos de evaluación docente mediante cinco aproximaciones metodológicas complementarias ha permitido obtener conclusiones robustas y consistentes. La convergencia de resultados entre métodos paramétricos y no paramétricos, robustos y clásicos, fortalece la validez de las inferencias realizadas.

Se ha identificado que los datos no son normales, lo que justifica el uso de métodos robustos y no paramétricos además de la no hay presencia de valores atípicos influyentes, lo que valida los promedios reportados.

Con los valores encontrados el instrumento tiene fiabilidad excelente, aunque potencialmente con redundancia, es de considerar que existen diferencias significativas entre los ítems, identificándose claras fortalezas y áreas de mejora, además, la ausencia de interacción significativa indica consistencia en los patrones de respuesta.

Se debe tomar en cuenta que el enfoque multimétodo constituye un modelo para el análisis riguroso de datos de evaluación educativa, demostrando cómo la integración de perspectivas estadísticas complementarias puede producir conclusiones más sólidas y matizadas que cualquier método único.

Referencias

- Abrami, P.C., d'Apollonia, S., Rosenfield, S. (2007). The dimensionality of student ratings of instruction: What we know and what we do not. En Perry, R.P., Smart, J.C. (eds.) *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective*. Springer. https://doi.org/10.1007/1-4020-5742-3_10
- Anderson, T. W. (2003). *An introduction to multivariate statistical analysis* (2ª ed.). Wiley-Interscience.
- Andrews, D. F. (1974). A robust method for multiple linear regression. *Technometrics*, 16(4), 523-531. <https://doi.org/10.1080/00401706.1974.10489233>



- Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., & Tukey, J. W. (1972). *Robust estimates of location: Survey and advances*. Princeton University Press.
- Beaton, A. E., & Tukey, J. W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, 16(2), 147-185. <https://doi.org/10.1080/00401706.1974.10489171>
- Beltrán, L. A. (2017). Didáctica y Currículo. En Corporación Universitaria Minuto de Dios (Eds.). *Didáctica para no didácticos: Reflexiones frente a la didáctica, enseñanzas y experiencias pedagógicas*.
- Bermejo, B., y Ballesteros, C. (2014). *Manual de didáctica general para maestros de educación infantil y de primaria*, Pirámide.
- Bernal, C. A. (2010). Metodología de la Investigación (3^{era} ed.), Pearson Educación.
- Creswell, J. W. (2009). *Research Design, qualitative, quantitative, and mixed Methods Approaches*, SAGE.
- Cook, S., Watson, D., & Webb, R. (2024). Performance evaluation in teaching: Dissecting student evaluations in higher education. *Studies in Educational Evaluation*, 81, 101342. <https://doi.org/10.1016/j.stueduc.2024.101342>
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297-334.
- Danielson, C. (2007). *Enhancing professional practice: A framework for teaching*. Association for Supervision and Curriculum Development.
- Fiallos, L., y Fiallos, M. O. (2024). Diseño y validación de un instrumento de investigación desde los constructos cualitativos hasta los cuantitativos. *Revista Holón*, 2(7), 45-58. <https://doi.org/10.48204/j.holon.n7.a6586>
- Field, A. (2018). *Discovering statistics using IBM SPSS statistics*. SAGE Publications.
- 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.
- Friedman, M. (1940). A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, 11(1), 86-92.
- Frumento, P., & Bottai, M. (2017). Parametric quantile regression: A new approach for robust estimation. *Statistical Methods in Medical Research*, 26(5), 2083-2099. <https://doi.org/10.1007/s10260-021-00557-7>
- Pineda García, G. L., Banegas Retes, D. L., Fiallos Gonzáles, L., Gómez González, A. Y., & Fiallos Gonzáles, M. O. (2025). Exploratory factor analysis of an instrument that evaluates the



relationship between language development theories and methodological processes in third grade of pre-basic education and first grade of basic education. *Salud, Ciencia y Tecnología*, 5, 2455. <https://doi.org/10.56294/saludcyt20252455>

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

Gimeno Sacristán, J. (1998). *El curriculum: Una reflexión sobre la práctica*. Morata.

Hampel, F. R. (1968). *Contributions to the theory of robust estimation* [Tesis doctoral no publicada, University of California, Berkeley].

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1), 73-101. <https://doi.org/10.1214/aoms/1177703732>

Huber, P. J. (1981). *Robust statistics*. Wiley. <https://doi.org/10.1002/0471725250>

IBM Corp. (2026). IBM SPSS Statistics: *M-estimators subcommand*. IBM Documentation. <https://www.ibm.com/docs/it/spss-statistics/31.0.0>

Maronna, R. A., Martin, R. D., Yohai, V. J., & Salibián-Barrera, M. (2019). *Robust statistics: Theory and methods*, Wiley.

Cardeña Ojeda, C. A. (2024). Levene's test for verifying homoscedasticity between groups in quasi-experiments in social sciences. *South Eastern European Journal of Public Health*, 2119–2125. <https://doi.org/10.70135/seejph.vi.2342>

Olivos, T. M. (2016). *Evaluación del aprendizaje y para el aprendizaje: Reinventar la evaluación en el aula*. Universidad Autónoma Metropolitana. https://casadelibrosabiertos.uam.mx/contenido/contenido/Libroelectronico/Evaluacion_del_aprendizaje_.pdf

Organization for Economic Cooperation and Development. (2013). *Teachers for the 21st century: Using evaluation to improve teaching*, OECD Publishing. <https://doi.org/10.1787/9789264193864-en>

Perlman, M. D. (2019). *On the feasibility of parsimonious variable selection for Hotelling's T^2 -test*. arXiv. <https://doi.org/10.48550/arXiv.1910.03669>

Pulido, S., y Rodríguez, J. (2014). *Estadística descriptiva y análisis cualitativo*. Universidad Nacional de Colombia.

Quansah, F., Asamoah, D., Amankwah, B., & Agormedah, E. K. (2024). Validity of student evaluation of teaching in higher education: A systematic review. *Frontiers in Education*, 9, 1329734. <https://doi.org/10.3389/educ.2024.1329734>



- Razali, N. M., & Wah, Y. B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, 2(1), 21-33. <https://www.nrc.gov/docs/ML1714/ML17143A100.pdf>
- Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, 17(3), 354-373. <https://doi.apa.org/doi/10.1037/a0029315>
- Sepúlveda García, J. M., Suárez-Giraldo, A. M., Rodas-Rodríguez, J. M., Ruiz-Ortega, F. J., y Henao-Henao, M. D. (2018). Validación y aplicación de un test modificado de Vandenberg y Kuse de rotación mental para simetría molecular. *Tecné, Episteme y Didaxis: TED*, (43), 155-171. <https://doi.org/10.17227/ted.num43-8656>
- Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research*, 83(4), 598-642. <https://doi.org/10.3102/0034654313496870>
- Tukey, J. W. (1960). A survey of sampling from contaminated distributions. En I. Olkin, *Contributions to probability and statistics: Essays in honor of Harold Hotelling*. (448-485). Stanford University Press.
- 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>
- Wilcox, R. R. (2017). *Introduction to robust estimation and hypothesis testing*, Academic Press.
- Zabaleta, F. (2007). The use and misuse of student evaluations of teaching. *Teaching in Higher Education*, 12(1), 55-76. <https://doi.org/10.1080/1356251060110213>

Síntesis curricular



Melvin Octavio Fiallos González

Profesor de educación técnica industrial, Master en Currículo e Ingeniería, con formación en investigación, sistemas de manufactura, Currículo, experiencia docente a nivel superior.