redie Revista Electrónica de Investigación Educativa
ISSN: 1607-4041
http://redie.uabc.mx

Vol. 21, 2019/e14

Evidencias de validez interna de un instrumento para evaluar la colegialidad docente

Evidence of Internal Validity of an Instrument to Evaluate Teacher Collegiality

Margarita Bakieva (*) margarita.bakieva@uv.es
Jesús Miguel Jornet Meliá (*) jornet@uv.es
José González Such (*) jose.gonzalez@uv.es

(*) Universitat de València
(Recibido: 11 de agosto de 2017; Aceptado para su publicación: 19 de septiembre de 2017)

Cómo citar: Bakieva, M., Jornet, J. M. y González, J. (2019). Evidencias de validez interna de un instrumento para evaluar la colegialidad docente. *Revista Electrónica de Investigación Educativa, 21*, e14, 1-12. doi:10.24320/redie.2019.21.e14.2127

Resumen

Este estudio es parte de una investigación más amplia que busca diseñar un modelo de evaluación de los sistemas educativos basado en la mejora de la cohesión social. Se trata de un estudio de validación métrica de un instrumento de evaluación de la colegialidad docente. El instrumento se aplicó a partir de las conclusiones del estudio de validación por comités de expertos. La muestra estuvo formada por 147 profesores voluntarios de primaria y secundaria de España. La metodología utilizada se basó en la recogida de evidencias métricas de validez interna del instrumento; se estudió la dimensionalidad mediante el análisis factorial exploratorio y la capacidad de discriminación mediante el análisis de conglomerados de K medias; y para determinar la calidad de los ítems se utilizó el Modelo Rasch. Los resultados descartaron algunos ítems en la versión final del instrumento y aportaron evidencias de validez.

Palabras clave: Evaluación del profesor, instrumento de medida, validez, sistema educativo.

Abstract

This study is a part of a broader research that aims to design a model of evaluation of educational systems based on the improvement of social cohesion. It is a study of metric validation of an instrument for evaluation teacher collegiality. The instrument was designed from the findings of the previos validation study in witch the construct was validated by expert judgment. The sample has been formed by 147 volunteer teachers of elementary and secondary Spanish schools. The method is based on the collection of internal metric evidences of validity of the instrument. Dimensionality was studied by exploratory factor analysis and discrimination capacity by means of the analysis of conglomerates of K means. Also the Rasch Model was used to determine the quality of the items. The results revealed some items that were discarded in the final version of the instrument and provided evidences of validity.

Keywords: Teacher evaluation, measurement instrument, validity, educational system.

I. Introducción

La necesidad de una evaluación de sistemas educativos con un modelo diferente al que presenciamos actualmente es evidente. Además de que son escasos los estudios empíricos en el campo de la educación a partir de la recogida de datos de las agencias internacionales (Pereira, Perales y Bakieva, 2016), el modelo social y educativo va hacia otra dirección (Jornet, 2014; Sancho-Álvarez, Jornet y González-Such, 2016), enfocada en elementos que aunque no se basen en indicadores materiales del proceso educativo sí influyen en él de forma significativa.

Uno de estos elementos es la Colegialidad Docente (CD), que juega un rol sustancial en cuanto al formato de organización de trabajo del profesorado, y su evaluación debe ser abordada necesariamente en el marco de diseño de un modelo de evaluación orientado al salón de clases, las escuelas y los sistemas, basado en la mejora de la cohesión social (Bakieva, Jornet y Leyva, 2014). El concepto de CD se define en este estudio como la calidad de la unión entre los docentes del mismo centro educativo, que trabajan de forma coordinada y colaborativa con un fin común de mejora de la cohesión social basada en un fuerte compromiso con los valores y normas compartidas, creando un clima positivo de cohesión y confianza en el grupo a través de la toma de decisiones consensuadas sobre la tarea común.

A su vez, la definición del concepto de "cohesión social" realizada por el Consejo de Europa (2005) sirve tanto para el diseño de acciones de mejora de este aspecto como para el diseño de un nuevo modelo de evaluación –basando la finalidad de ésta en la mejora de inclusividad educativa y equidad en el acceso a recursos y oportunidades.

El modelo de Evaluación de Sistemas Educativos para la mejora de la cohesión social pretende diseñar un modelo que plantee la evaluación de centros y de sistemas educativos desde el análisis de su aportación para el desarrollo de la cohesión social. Los constructos o dimensiones implicados son definidos y validados anteriormente en diversos estudios relacionados (Jornet et al., 2017).

1.1 Necesidad de evaluación de la CD

Numerosos estudios describen elementos de los ambientes colaborativos y colegiales a través de la observación y estudios de casos; entre los más relevantes se cuentan los de Bolívar (1999), Hargreaves (1991), Jarzabkowski (2003), Lavié (2004) y Little (1990). Un paso más en la investigación ha dado Shah (2011), quien realizó el proceso de diseño y validación del instrumento en las escuelas malayas. Sin embargo, en el contexto iberoamericano hasta ahora no se ha identificado ningún estudio empírico sobre una evaluación de la CD mediante un instrumento validado, por ello la realización de este estudio. El hecho de que el instrumento de evaluación de la CD sea uno de los instrumentos que conforman un modelo más global de evaluación de sistemas educativos hace que este trabajo tenga más sentido, pues sólo implicando todo el sistema se podrá conseguir que los procesos cambien.

Cuando hablamos de la colegialidad como elemento base para el trabajo colaborativo lo hacemos considerando que los conceptos de colaboración y colegialidad tienen una relación inseparable, la colegialidad implica compartir objetivos y finalidades educativos en un trabajo conjunto de profesores del centro en el que se vive la cultura colaborativa (Campbell y Southworth, como se citó en Bolívar, 2000).

Por último, cabe resaltar que la presencia de una colegialidad verdadera (lo opuesto de la "colegialidad artificial" de Hargreaves, 1999) aporta una emocionalidad sana (Jarzabkowski, 2003), ambientes de trabajo más satisfactorios y productivos (Santos, 2000) y es un elemento esencial de un centro educativo saludable (Gappa, Austin y Trice, 2007; Walvoord et al., 2000).

Es cierto que la evaluación de colegialidad conlleva cierto tipo de peligros, como castigo de los comportamientos que se consideren poco amables o cívicos, obligando en alguna medida a actuar de forma forzada. Además del peligro que implica el control institucional sobre las relaciones humanas, se suma cierta confusión a la hora de interpretar el término CD. Sin embargo, una colegialidad racionalmente especificada (Weber, 2002, p. 220) puede ser la solución para la dominación tradicional jerarquizada mediante la solidaridad y unanimidad en la toma de decisiones, contrarrestando los efectos negativos de la dirección monocrática (Weber, 2002, p. 223). Por este motivo queremos proponer una definición más detallada y diseñar un instrumento de evaluación que pueda utilizarse en la práctica.

1.2 Un apunte sobre la validez

La validación de un instrumento debe reunir una serie de evidencias a partir de un proceso –en este caso de análisis métrico– de acuerdo con uno o varios modelos teóricos. Según Elosua (2003), las fuentes de evidencia de validez se pueden estructurar como se ilustra en la tabla I.

Tabla I. Fuentes de evidencia

Evidencia	Tipo	Método
Interna	Contenido	Definición del dominio. Representación y relevancia Situación de test (formato, administración, puntuación)
	Proceso de respuesta	Protocolos Entrevistas Modelos componenciales
	Estructura interna	
	Dimensionalidad	Modelos de estructura latente Modelo Factor Común Modelo de Respuesta Ítem Paramétrico No paramétrico
	Funcionamiento diferencial del ítem	Invarianza observada Delta, chi-cuadrado, Mantel-Haenszel, Regresión logística, Log-Lineal, SIBTEST Invarianza Latente Modelo Respuesta al ítem Modelo Factor Común
Externa	Relaciones	
	Convergente/ discriminante *Test/ criterio* *Generalización*	Matriz multirrasgo/multimétodo Factorial Confirmatorio Modelo lineal generalizado Meta-análisis
	Consecuencias	

Fuente: Elosua (2003).

De acuerdo con esta clasificación, el presente estudio se centra en la recogida de evidencias de validez interna del instrumento y los objetivos generales del mismo se sintetizan en tres: 1) validar métricamente el instrumento de evaluación de la CD en el grupo de profesores de Educación Primaria y Secundaria de España, 2) revisar el comportamiento de la escala en un grupo piloto y estimar los niveles de los parámetros más importantes, y 3) reducir el número de ítems a partir del estudio métrico, manteniendo o mejorando la calidad técnica de los ítems y del instrumento.

II. Método

Se estudia el funcionamiento de los ítems y del instrumento global en cuanto a su comportamiento métrico, de acuerdo con dos modelos de medición: la Teoría Clásica del Test (TCT) y la Teoría de Respuesta al Ítem (TRI, Modelo Rasch). Los análisis realizados para tal fin se sintetizan en los siguientes puntos:

a) Validación a partir de la TCT:

i. Escala completa, detección de ítems defectuosos (homogeneidad).
ii. Estudio factorial exploratorio.
iii. Capacidad de discriminación de la escala completa.

b) Validación a partir de la TRI:

i. Escala completa como definición de la variable latente.

Asimismo, se busca obtener otras evidencias de validación (análisis de estructura interna):

c) Estudio Clúster de K medias por dimensiones de la CD.

d) Estudios diferenciales (utilizando el clúster de pertenencia).

i. Por tipología de centros (tomando como variables diferenciadoras: titularidad/ tamaño de centro/ tamaño de la localidad).
ii. Por género.
iii. Por la situación profesional de los profesores (años de experiencia profesional y en el último centro, cargo desempeñado en el centro, nivel de formación).

2.1 Instrumento

El instrumento piloto fue diseñado por Bakieva (2016) a partir de una definición operativa de CD (mediante indicadores) validada por diferentes comités de expertos. Se situó el concepto y sus dimensiones en el marco del modelo de evaluación subyacente, integrándolo junto a otras dimensiones del marco general de la investigación.

El cuestionario aplicado en este estudio se divide en dos partes: variables descriptivas y el instrumento de evaluación de CD, compuesto por 82 ítems distribuidos en 6 dimensiones: A) Valores éticos y profesionales compartidos (13 ítems); B) Cohesión y confianza en el grupo, actitudes de alianza, compañerismo (18 ítems); C) Compromiso con la tarea docente, actitud de mejora profesional continua (13 ítems); D) Toma de decisiones colegiada sobre la tarea docente (7 ítems); E) Relaciones docentes colaborativas: autonomía y colectividad en la tarea docente (11 ítems), y F) Clima dinámico y positivo del centro, ambiente profesional creativo (20 ítems).

Los enunciados fueron valorados con la escala de 4 puntos de Likert que recogía la frecuencia de los comportamientos señalados (de 1 "Nunca" hasta 4 "Siempre"). El instrumento completo de evaluación de CD se puede consultar en Bakieva (2016).

2.2 Recogida y análisis de datos

La recogida de datos se realizó mediante un cuestionario electrónico, diseñado a través de la plataforma freeware Limesurvey. Se enviaron las invitaciones para participar en el estudio a diferentes centros escolares de primaria y secundaria de España. Se pidió a los directores de los centros la distribución de los cuestionarios entre los profesores adscritos a sus centros. Fueron contestados 438 cuestionarios (algunos de forma parcial) y para realizar el estudio se seleccionaron 147 cuestionarios completos.

En cuanto al análisis, se utilizaron las puntuaciones de escala total (CD), así como las puntuaciones de las subdimensiones que componen la escala, con el fin de explorar el comportamiento de éstas y determinar los puntajes correspondientes a diferentes niveles de actitud colegial. Asimismo, se analizó el comportamiento de los ítems, comprendidos como parte de la escala global y dentro de cada una de las subescalas. Este procedimiento se justifica con la estructura del constructo operativo que comprende la actitud colegial como un conjunto de otras variables latentes, que a su vez se componen por el conjunto más detallado de comportamientos mostrados a través de los ítems. El análisis de fiabilidad se realizó recodificando anteriormente aquellos ítems que fueron formulados en el sentido inverso a la actitud colegial (4=1, 3=2; 2=3, 1=4).

El análisis factorial se realizó con el carácter exploratorio de la dimensionalidad del instrumento, utilizando la solución sin rotar de la matriz de las puntuaciones de ítems del instrumento global. El estudio de conglomerados de k medias se realizó con las puntuaciones de subescalas y se pretendió hallar una solución que se basaba en la escalabilidad de los perfiles encontrados y una distribución equilibrada de sujetos por cada clúster (no menor de 10% del total).

En el caso del estudio métrico, de acuerdo con el Modelo Rasch, se indicó que los ítems no comparten estructura de categorías uniformes entre ellos (Modelo de Crédito Parcial para ítems politómicos).

El análisis de datos se realizó con los programas SPSS 22 y Winsteps –que implementa el algoritmo de máxima verosimilitud para obtener los parámetros de los ítems.

2.3 Grupo de estudio

El grupo se integró con 147 docentes, 83 de ellos mujeres (56%). La edad del grupo osciló entre 25 y 68 años; el nivel educativo en su mayoría es equivalente a grado universitario (96.6%), 3.3% de ellos (5 docentes) con nivel posgrado universitario (máster o doctorado).

Los centros escolares a los que están adscritos los docentes del grupo son públicos en su mayoría (74%), el resto son concertados (23.1%) y privados (2%), ubicados en 10 provincias diferentes de España.[1] El contexto en el que se sitúan los centros de adscripción se caracteriza por ser:

- 12.2% (18 personas), rural (centro situado en una población con menos de 3,000 habitantes).
- 15.6% (23 personas), población de 3,000 a 15,000 habitantes.
- 46.9% (69 personas), población de 15,000 a 100,000 habitantes.
- 17.7% (26 personas), población de 100 mil a 1 millón de habitantes.
- 7.5% (11 personas), población de 1 a 5 millones de habitantes.

Los centros en los que están adscritos los profesores participantes, además de ofrecer el tramo de Educación Primaria, ofrecen también formación en otros tramos educativos, así, 93.2% de los profesores trabaja en un centro que ofrece tramos de Educación Infantil y Primaria: de estos, 23.8% lo hace en los centros de Infantil, Primaria y ESO (Educación Secundaria Obligatoria), y 8.8% en los centros que ofrecen Infantil, Primaria, ESO y Bachillerato.

Los encuestados declararon estar ocupados en diferentes tareas; así, 78.9% forma parte del profesorado de Primaria, 11.56% es profesor de Infantil, 7.48% de ESO, 2.72% de Bachillerato y 1.36% es profesor de Universidad. Asimismo, un 4.76% del grupo trabaja como pedagogo, psicopedagogo u orientador de un centro.

Además de las funciones inherentes a sus puestos de trabajo, muchos de los encuestados dijeron cumplir funciones de organización y coordinación en su centro, como: parte del equipo directivo (49.66%), miembro del Consejo escolar (29.25%), miembro del Claustro docente (57.14%), tutorías (38.1%), tareas de coordinación en diferentes actividades (12.93%).

En cuanto a la experiencia docente total y en el último centro que laboraron, la tabla II muestra la distribución de grupos de docentes en función de su experiencia.

Tabla II. Experiencia docente

	Categorías					
	0-5	6-10	11-20	21-30	31-40	≥41
% años en el último centro	22.45	17.69	34.01	16.33	7.48	2.04
% años en su profesión	6.80	11.56	28.57	31.29	19.73	2.04

Según la tabla II, la mayor parte de los profesores tiene entre 11 y 35 años de experiencia; sin embargo, pocos profesionales de educación superan 25 años continuos en el último centro trabajado.

III. Resultados

A lo largo de los análisis, las puntuaciones de las subescalas (A, B, C, D, E, F) componentes de la escala global (CD) se comparan a través de sus puntuaciones estandarizadas (puntuaciones tipificadas Z), para poder cotejar los niveles y distribución de éstas. En la tabla III se observan los estadísticos básicos de distribución de cada subescala y de la escala total CD.

[1] Alicante, Almería, Castellón, Cuenca, Guadalajara, Murcia, Palencia, Sevilla, Tarragona y Valencia.

Tabla III. Estadísticos de puntuaciones de totales de subescalas y total de la escala

Escala/ Sub-escala	Estadísticos descriptivos					Prueba de normalidad K-S*		Estadísticos Z (tipificados)		Alfa Cronbach	Ítems defec-tuosos
	N	Mín.	Máx.	M	D.T.	Estad.	Sig.	Mín.	Máx.		
A	13	31	52	42.52	5.005	1.415	.037	-2.30	1.89	.834	A5, A7
B	18	30	72	58.54	7.797	1.073	.200	-3.66	1.73	.910	B7, B10
C	13	34	52	46.38	4.183	1.538	.018	-2.96	1.34	.839	C4
D	7	13	28	23.59	3.828	1.508	.021	-2.76	1.15	.886	
E	11	21	44	35.31	5.550	.975	.298	-2.58	1.57	.909	
F	20	35	80	63.05	10.36	1.174	.127	-2.71	1.64	.936	F14
CD	82	171	328	269.38	33.78	.932	.350	-2.91	1.74	.978	

Nota: *Prueba Kolmogorov-Smirnov. No. válido (según lista) 147.

Además de los estadísticos descriptivos, en la tabla III se observa también que la distribución de las subescalas B, E, F y de la puntuación total de la escala CD se aproximan a la distribución normal, lo que permite utilizar contrastes paramétricos más adelante. Sin embargo, en el caso de las subdimensiones A, C y D el supuesto no se cumple, lo que obliga a utilizar los contrastes no paramétricos con las puntuaciones de estas subescalas.

Desde la tabla III se confirma que los niveles de consistencia interna de las subescalas y de la CD total, medida por el coeficiente alfa de Cronbach, son altos. Como dato adicional se señalan en la última columna los elementos que han mostrado un comportamiento no homogéneo con la escala de pertenencia (ítems defectuosos), aportando evidencias de calidad. Por último, el alfa compuesto de la CD (calculado a partir de las sumas de cada subescala) es de 0.931.

Las puntuaciones totales de subescalas permiten determinar la capacidad de la escala para discriminar los grupos con diferente nivel de actitud colegial, medida a través del instrumento. Para tal fin, se debe dividir el grupo total en un número de grupos idóneo, conocer los límites de los diferentes grupos de puntuación y en torno a qué puntuación media se concentran (centroides), además de cuántos son los grupos (conglomerados) que se pueden diferenciar de forma clara a partir de las puntuaciones del instrumento. El procedimiento más adecuado para determinar los grupos es el conglomerado de K medias, dadas las características de las variables (puntajes de subescalas del instrumento). En la fase piloto este estudio de conglomerados puede aportar nuevos indicios de validez mediante las evidencias de una adecuada discriminación de diferentes categorías de CD. Además, a partir del estudio de conglomerados se pueden caracterizar los grupos de referencia mediante las variables de contexto que se asocian al estudio de CD. Estas características grupales pueden ayudar a una correcta interpretación de los resultados de aplicación del instrumento, ya que pueden explicar las características del contexto que son particulares de colectivos con diferentes niveles de colegialidad.

Para determinar el número adecuado de conglomerados (o clústeres) se realizaron iteraciones para 2, 3, 4, 5 y 6 conglomerados. Según los criterios de idoneidad (ver metodología) se marcó el número óptimo de conglomerados en 3. Se observó que en todas las dimensiones se dieron diferencias estadísticamente significativas entre los perfiles de los tres grupos (ver tabla IV).

La tabla IV ofrece las puntuaciones directas centrales para cada subescala, correspondientes a los tres conglomerados finales. Debido a la diferenciación escalar de las puntuaciones podemos denominar los conglomerados por su nivel de CD (bajo, medio y alto).

Tabla IV. Centros de los conglomerados finales

Puntuación directa subescala	Conglomerado		
	CD baja	CD intermedia	CD alta
A	35.81	42.15	46.54
B	47.48	57.75	65.37
C	4061	46.12	49.79
D	18.13	23.54	26.60
E	27.65	34.32	40.51
F	47.84	61.88	72.53

De la misma manera, los tres conglomerados se distancian entre sí de manera adecuada (ver tabla V).

Tabla V. Distancias entre los centros de los conglomerados finales

Conglomerado	CD baja	CD media	CD alta	Casos
CD baja		21.142	36.960	31
CD media	21.142		15.868	59
CD alta	36.960	15.868		57

Para confirmar una solución óptima mediante la representación a través de 3 conglomerados se ha realizado el contraste paramétrico Anova de un factor para los 3 conglomerados en las puntuaciones directas de las 6 subescalas, así como en la puntuación total CD. Se añadió la puntuación total de la escala de CD para comprobar la efectividad de agrupación en 3 conglomerados. Anteriormente se confirmó la distribución normal para los 3 conglomerados en estas puntuaciones, salvo la puntuación de la dimensión D. Los contrastes de grupos han mostrado la existencia de diferencias significativas entre los grupos, marcados por pertenencia a conglomerado de correspondencia, tanto para escala total como para las subescalas (tabla VI).

Tabla VI. Contrastes entre grupos

	A	B	C	D	E	F	CD
Estadístico de contraste*	89.172	195.166	89.716	94.390	228.334	271.204	422.785
Grados de libertad	2	2	2	2	2	2	2
Sig. asintótica	.000	.000	.000	.000	.000	.000	.000

Nota: Variable de agrupación: Pertenencia al clúster final (opción de 3 grupos). *En el caso de la puntuación de la subescala D se aplicó el contraste no paramétrico, basado en el cálculo de χ^2, en el caso del resto de las subescalas y la escala total (CD) se utilizó el contraste paramétrico de Anova de un Factor.

A continuación se realizó la caracterización de cada conglomerado mediante los datos personales aportados por los docentes en la parte I del cuestionario de CD. En general, las tendencias de distribución no muestran diferencias observables en la distribución, salvo una leve diferencia en el caso de las variables como Titularidad del Centro de adscripción, Contexto en el que se sitúa el Centro y el Número de años en el último Centro.

Para comprobar las diferencias en la distribución de las variables señaladas se aplicó el procedimiento de tablas cruzadas con el análisis no paramétrico de χ^2 para determinar en qué medida las diferencias que pueden aparecer serán estadísticamente significativas. La única diferencia estadísticamente significativa se demuestra en el caso de la variable Número de años en el último Centro. Para una mejor apreciación, en la figura 1 se puede observar la distribución de las categorías de la variable.

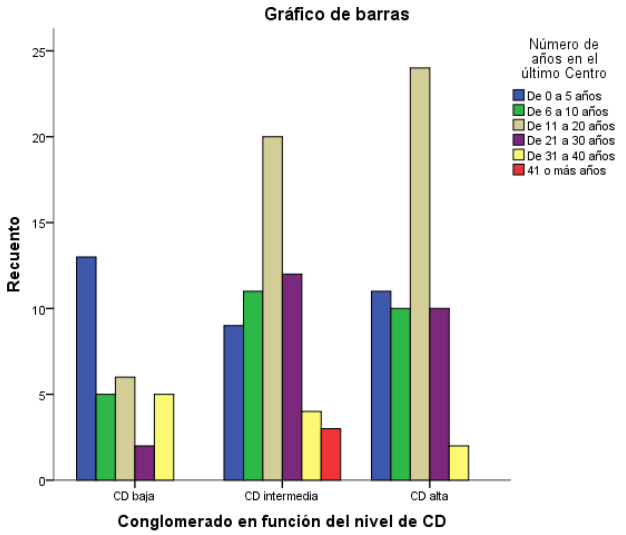

Figura 1. Experiencia en el último centro en función de conglomerado de pertenencia

En la figura 1 se pueden apreciar diferencias en la distribución de las categorías 0-5 años, 6-10 años y 11-20 años, y 31-40 años y 41 y más años; el número de profesores con mayor experiencia en un Centro disminuye en cuanto aumenta el nivel de colegialidad.

El siguiente punto de análisis buscó explorar la dimensionalidad de la escala. Para tal fin, se aplicó el procedimiento de análisis factorial exploratorio. Según McDonald (1999), la metodología de análisis factorial permite una mejora sustancial sobre procedimientos tradicionales para estimación de la fiabilidad, además de un tratamiento unificado de la Teoría de los Tests, tanto de la clásica como de la TRI (Santisteban y Alvarado, 2001, p. 68). El concepto central es que los ítems de la escala sean de contenido homogéneo. En caso contrario la puntuación total carecería de sentido. Así, es necesario que todos los ítems del dominio de actitud o de conducta se ajusten a un modelo unifactorial. En el caso del instrumento que nos ocupa, este supuesto se cumple: aunque en el constructo, y en la escala, se identifican diferentes dimensiones, éstas aportan al mismo de forma acumulativa, siendo elementos de un mismo constructo, el de CD, que se concibe como unidimensional. Para tal fin, se realizó un análisis factorial confirmatorio de componentes principales para los ítems que componen la escala. Consideramos, de acuerdo con Carretero-Dios y Pérez (2007) que esta técnica debe ser sometida a los intereses conceptuales y estar sujeta a las premisas teóricas sobre la dimensionalidad subyacente a los ítems que se utilizan.

La medida de adecuación muestral Kaiser-Meyer-Olkin y la prueba de esfericidad de Barlett y la matriz de correlaciones de Pearson confirman la pertinencia del uso del modelo factorial para explicar los datos.

En cuanto a la información sobre las comunalidades asignadas inicialmente a las variables (inicial) y las comunalidades reproducidas por la solución factorial (extracción), éstas señalan datos sobre la proporción de la varianza que puede ser explicada por el modelo factorial obtenido. En este caso las variables peor explicadas por el modelo son A9, A13, B14, C4, C7, E7, E9, ya que el modelo es capaz de explicar menos de 65% de su variabilidad original.

El primer factor, a su vez, explica hasta un 38.27% de la varianza total. Se extraen 18 factores de la matriz analizada, aunque a partir del segundo pueden considerarse residuales, porque llegan a explicar menos del 3% de la varianza total y carecen del sentido de análisis respecto al contenido. La situación se refleja en la figura 2, que ofrece el gráfico de sedimentación, vemos que el primer factor ofrece el mayor autovalor de todos los posibles, y los restantes son residuales, incapaces de explicar una cantidad relevante de la varianza total. Es decir, a partir del segundo autovalor no se pueden extraer más factores y éstos deben ser desechados.

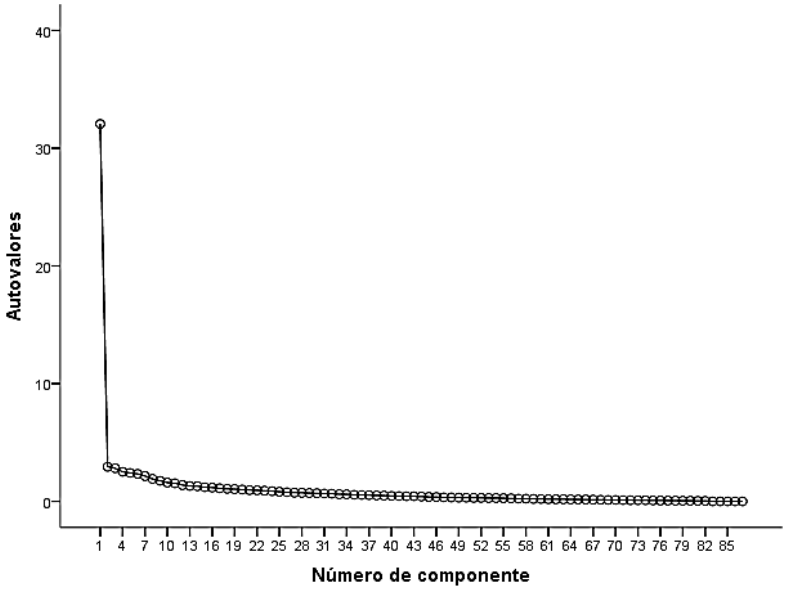

Figura 2. Sedimentación

La unidimensionalidad de la escala puede ser confirmada a través del estudio de análisis factorial. Ello es importante para poder establecer la interpretabilidad de las puntuaciones total y por dimensiones. Podemos señalar a partir de estos resultados que, de confirmarse en el estudio con el instrumento definitivo, el puntaje total sería el mejor indicador de CD, mientras que los de dimensiones servirían para determinar el perfil de la misma. Es decir, si se tuviera que clasificar niveles de CD a partir de las puntuaciones, la referencia sería la puntuación total, mientras que las de dimensiones servirían a nivel descriptivo del modo en que deviene en cada caso dicho puntaje (basado en un perfil u otro).

Complementando el estudio basado en la Teoría Clásica del Test, se realizó el estudio de comportamiento de ítems como elementos de la variable latente CD de acuerdo con el Modelo Rasch. La selección del modelo exige el cumplimiento de tres suposiciones básicas que se cumplieron en este estudio: unidimensionalidad, independencia local, principio de invariancia (Baker, 2001).

En la interpretación de los ajustes de ítem al modelo, el valor de la media cuadrática (MNSQ) en el ajuste interno y externo, según Wright y Linacre (1994) y Boone, Staver y Yale (2014, p.167), debe situarse en el rango razonable entre 0.5 y 1.5 puntos. A partir de >2.0 el ítem distorsiona o degrada el sistema de medición, y si es <0.5 el ítem puede ser menos productivo para la medición, aunque no degradante; además, puede producir engañosamente un buen nivel de fiabilidad y separaciones.

En el caso analizado, las variables que presentan una falta de ajuste interno y externo son: A01, A08, B08, B11, F15 y A06 (considerado este último como ítem crítico, aunque su ajuste interno está en el límite de considerarse aceptable, pero poco funcional). Otros ítems presentan sólo desajuste externo: C05 y D04. La correlación punto-media de los ítems señalados es otro de los indicadores de ajuste y en casi todos los casos presenta valores menores a 0.3. También se presentan otros ítems que con sus puntuaciones marcan el límite con el rango productivo del modelo Rasch, los ítems A05, A07, A09, B16, C02, C11 y F14.

IV. Discusión y conclusiones

De acuerdo con American Educational Research Association, American Psychological Association y National Council on Measurement in Education (AERA, APA y NCME, 2014, p. 9) el concepto unitario de validez se refiere al grado en el que la evidencia y la teoría apoyan la interpretación de los resultados de la prueba para los usos propuestos. En palabras de Messik (1989), la validez es "un juicio global sobre el grado en que la evidencia lógica y empírica apoyan la concepción y conveniencia de las inferencias y acciones realizadas en base a las puntuaciones del instrumento" (p. 19). El proceso de validación involucra la acumulación de evidencia relevante que provee una sólida base científica para la interpretación de los resultados (APA, AERA y NCME, 2014, p. 11). De acuerdo con lo anterior, no se valida el instrumento de evaluación en sí mismo, sino las inferencias e interpretaciones realizadas a partir de las puntuaciones que proporciona. Los instrumentos de evaluación son válidos para algún propósito o proceso en particular por lo que es preciso indicar sus alcances y limitaciones.

Para tener una referencia clara sobre el tipo de evidencias de validez, nos basamos en el estándar de APA, AERA y NCME (2014): contenido, procesos de respuesta, estructura interna, relaciones con otras variables y consecuencias de la prueba. A la vez, de acuerdo con la clasificación que realiza Elosua (2003) reunimos las evidencias de validez interna a partir de las fuentes señalados en la tabla I. En consonancia con lo expresado, el estudio realizado reúne diferentes evidencias; una de ellas se refiere a la estructura interna, –estudiada a partir del análisis factorial que demuestra que el instrumento cuenta con suficientes evidencias de unidimensionalidad y que en general los ítems son homogéneos, con alguna excepción corregible. El hecho de que el instrumento presenta la evidencia de la unidimensionalidad implica el uso e interpretabilidad final de las puntuaciones, y la total resulta la más adecuada para la clasificación individual, siendo en ese caso las de dimensiones orientativas acerca del perfil que fundamenta la colegialidad observada en el puntaje global. Tiene, por tanto, valor diagnóstico, siendo preciso todavía (como prospectiva de trabajo futuro) realizar estudios confirmatorios de análisis factorial para poder revalidar la dimensionalidad del instrumento.

El instrumento ha mostrado, además, la capacidad de discriminación clara entre diferentes niveles de actitud medida, aportando de esta manera las evidencias de un adecuado funcionamiento acorde con los objetivos perseguidos. El estudio de conglomerados de K medias ha revelado un número de 3 clústeres como solución óptima, aunque esto debe ser reafirmado en la siguiente fase del estudio, con un grupo de sujetos más amplio, señalando ésta como otra de las líneas de trabajo futuro.

Además, la descripción de conglomerados permite formular las posibles hipótesis sobre la colegialidad y los contextos relacionados con ella. Por ejemplo, si los ambientes de CD alta se caracterizan por ser propios de instituciones públicas rurales o de ciudades pequeñas y a los docentes de esos centros les resulta más sencillo trabajar colaborativamente y este tipo de escuelas son más propicias a que haya diálogo, se dé la participación en la igualdad de condiciones, de forma que los profesores se sientan más seguros y más respetados, de acuerdo con el planteamiento de Jarzabkowski (2003). No obstante, es necesario un estudio más profundo sobre las relaciones entre diferentes variables, otra línea de trabajo futura.

Se estudiaron las diferencias entre los grupos (en función del conglomerado de pertenencia), aportando de esta manera evidencias sobre la capacidad de discriminación del instrumento diseñado; se estudió también el comportamiento de los ítems como elementos de la variable latente CD en función del Modelo Rasch, demostrando que, en general, el instrumento reúne suficientes evidencias de calidad métrica, pero revelando asimismo algunos ítems que no ajustan correctamente al Modelo Rasch.

Los ítems que han mostrado el comportamiento anómalo en los diferentes análisis son A5, A7, B7, B10, C3, C12, F5, F13, F14 (según la TCT) y A1, A8, B8, B11, F15 (según la TRI). La decisión de eliminación de los ítems anómalos se toma sólo si repetidos análisis revelan su falta de ajuste al modelo utilizado, siempre teniendo en cuenta que la eliminación del elemento o elementos no comprometa la definición del constructo. Debemos remarcar que el diseño de un instrumento de este tipo se debe establecer en un juicio equitativo, basado en los indicios de diferentes tipos, tanto de carácter conceptual como métrico, para construir un instrumento de evaluación con el mayor número de evidencias de validación y con un determinado sentido educativo.

Entre las limitaciones de este trabajo se pueden señalar la baja participación del profesorado en el estudio, lo que limitó el grupo de estudio a una muestra por disponibilidad. La justificación de esta circunstancia se debe a la saturación profesional de los docentes, una característica ampliamente documentada (San Fabián, 2006). Como solución a la excesiva carga de los docentes y a la ansiedad que ésta provoca, se debe encaminar la cultura organizativa escolar hacia la mejora de las condiciones relacionadas con la colegialidad docente, propiciando la colaboración y confianza entre los colegas, y el primer paso debe situarse en una evaluación y diagnóstico de la situación, seguida ésta de acciones de mejora. El instrumento presentado en este estudio ayuda a aproximarse al análisis de la situación, marcando los elementos más relevantes a través de la definición adecuada del constructo de la colegialidad docente en los términos operativos y permitiendo así actuar en su mejora a nivel de los centros, creando ambientes sanos y productivos de trabajo, en línea con los trabajos de Santos (2000), Gappa, Austin y Trice (2007) y Walvoord et al. (2000).

Referencias

American Educational Research Association, American Psychological Association y National Council on Measurement in Education. (2014). *Standards for educational and psychological tests and manuals.* Washington, DC: APA.

Baker, F. (2001). *The basics of Item Response Theory*. EUA: ERIC Clearinghouse on Assesment and Evaluation.

Bakieva, M. (2016). *Diseño y validación de un instrumento para evaluar la colegialidad docente* (Tesis doctoral). Universidad de Valencia. Recuperado de http://roderic.uv.es/handle/10550/56226

Bakieva, M., Jornet, J. y Leyva, Y. (2014). Colegialidad docente: un estudio comparativo (España/México) de validación de constructo para el diseño de un instrumento de evaluación. *Revista Iberoamericana de Evaluación Educativa, 7*(2e), 131-145. Recuperado de http://www.rinace.net/riee/numeros/vol7num2e/art10.html

Bolívar, A. (1999). *Cómo mejorar los centros educativos*. Madrid: Síntesis.

Bolívar, A. (2000). *Los centros educativos como organizaciones que aprenden. Promesas y realidades*. Madrid: Muralla.

Boone, W., Staver, J. y Yale, M. (2014). *Rasch analysis in the Human Sciences*. Nueva York: Springer.

Campbell, T. y Southworth, G. (abril, 1990). *Rethinking collegiality: teachers'views*. Documento presentado en la Annual Meeting of the American Educational Research Association (AERA), Boston.

Carretero-Dios, H. y Pérez, C. (2007). Normas para el desarrollo y revisión de estudios instrumentales: consideraciones sobre la selección de tests en la investigación psicológica. *International Journal of Clinical and Health Psychology, 7*(3), 863-882.

Consejo de Europa (2005). *Concerted development of social cohesion indicators: methodological guide*. Bélgica: Autor.

Elosua, P. (2003). Sobre la validez de los test. *Psicothema, 15*(2), 315-321. Recuperado de http://www.psicothema.com/psicothema.asp?id=1063

Gappa, J., Austin, A. y Trice, A. (2007). *Rethinking faculty work: higher education's strategic imperative*. San Francisco, CA: Jossey-Bass.

Hargreaves, A. (1991). Cultures of teaching. En A. Hargreaves y M. Fullan (Eds.), *Understanding teacher development*, (pp. 216-240). Nueva York: Teachers College Press.

Hargreaves, A. (1999). *Profesorado, cultura y postmodernidad: cambian los tiempos, cambia el profesorado*. Madrid: Morata.

Jarzabkowski, L. (2003). Teacher collegiality in a remote Australian school. *Journal of Research in Rural Education, 18*(3), 139-144. Recuperado de http://jrre.psu.edu/?page_id=1773

Jornet, J. (septiembre, 2014). *Evaluación de sistemas educativos de acuerdo con el modelo de cohesión social*. Conferencia invitada en el V Coloquio Internacional de la Red Iberoamericana de Investigación sobre Evaluación de la Docencia (RIIED), Ensenada, México.

Jornet, J., González-Such, J., Perales, M., Sánchez-Delgado, P., Bisquert, M., Bakieva, M., Sáncho-Álvarez, C., Belda, A., Llorens, A., Bodoque, A. y Ortega, S. (2017, julio). Aproximaciones cualitativas para la definición y validación de constructos de instrumentos estandarizados de medida. Enfoque teórico y metodología. En P. Costa, M. C. Sánchez-Gómez y M. V. Martín (Orgs.), *La Práctica de la investigación cualitativa: ejemplificación de estudios* (pp. 159-196). Portugal: Ludomedia. Recuperado de https://ciaiq.org/wp-content/uploads/2017/09/ebook_Practica_Investigacion_Cualitativa_Espanol.pdf

Lavié, J. M. (2004). Micro-contextos para la colaboración docente: el caso de los equipos de ciclo. *Revista de Educación,* (335), 345-370. Recuperado de http://www.revistaeducacion.mec.es/re335_22.htm

Little, J. W. (1990). Teachers as colleagues. En A. Lieberman (Ed.), *Schools as collaborative cultures: creating the future now* (pp.165-193). Londres: Falmer Press.

McDonald, R. P. (1999). *Test theory: a unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Messik, S. (1989). Validity. En R. L. Linn (Ed.), *Educational Measurement* (pp. 13-103). Nueva York: McMillan.

Pereira, D., Perales, M. J. y Bakieva, M. (2016). Análisis de tendencias en las investigaciones realizadas a partir de los datos del Proyecto PISA. *Revista Electrónica de Investigación y Evaluación Educativa, 22*(1), 1-18. doi:10.7203/relieve.22.1.8248

San Fabián, J. (2006). La coordinación docente: condiciones organizativas y compromiso profesional. *Participación Educativa,* 3, 6-11.

Sancho-Álvarez, C., Jornet, J. y González-Such, J. (2016). El constructo Valor Social Subjetivo de la Educación: validación cruzada entre profesorado de escuela y universidad. *Revista de Investigación Educativa, 34*(2), 329-350. doi:10.6018/rie.34.2.226131

Santisteban, C. y Alvarado, J. (2001). *Modelos psicométricos*. Madrid: UNED.

Santos, M. A. (2000). *La escuela que aprende*. Madrid: Morata.

Shah, M. (2011). Dimensionality of teacher collegiality and the development of teacher collegiality scale. *International Journal of Education, 3*(2), 1-20. doi:10.5296/ije.v3i2.958

Walvoord, B., Carey, A., Smith, H., Soled, S., Way, P. y Zorn, D. (2000). *Academic departments: how they work, how they change*. San Francisco, CA: Jossey-Bass. Recuperado de https://eric.ed.gov/?id=ED447746

Weber, M. (2002). *Economía y sociedad* (2a. ed.). Colombia: Fondo de Cultura Económica.

Wright, B. D. y Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.