Psicothema

Psicothema 2021, Vol. 33, No. 3, 509-517
doi: 10.7334/psicothema2020.262

ISSN 0214 - 9915 CODEN PSOTEG

www.psicothema.com

Adaptation of the Teacher Efficacy Scale to Measure Effective Teachers' Educational Practices Through Students' Ratings: A Multilevel Approach

María-José Lera[1], José-M. León-Pérez[1], and Paula Ruiz-Zorrilla[2]
[1] Universidad de Sevilla, and [2] Universidad Complutense de Madrid

Abstract

Background: There is an increasing evidence of the role that teachers' educational practices have for students' school achievement and their well-being. However, there is a lack of valid measures in Spanish to address effective educational practices based on students' perceptions. In response, this study aims to provide a valid, reliable scale for measuring educational practices in school settings: the Students' ratings of Teachers' Educational Practices Scale (STEPS). **Methods:** We analyzed the scale's internal consistency and reliability, factor solution and invariance, and criterion validity, by using a multilevel approach in a sample of 2,242 students nested in 104 classrooms from 22 Spanish schools. **Results:** Indicated that the scale exhibited good reliability according to the omega coefficient (within =.86 and between level =.98). The multilevel confirmatory factor analysis (MCFA) revealed a hierarchical factor solution: classroom management, instructional strategies, and students' engagement as first-order factors, and a general second-order factor labeled as effective educational practices. The scale demonstrated configural invariance by teaching level, sex, and region. Effective educational practices were associated with student self-esteem at the individual level. **Conclusions:** This study offers a reliable, valid instrument, STEPS, for measuring effective educational practices.

Keywords: Educational practices; teaching evaluation; multilevel confirmatory factor analysis; self-esteem.

Resumen

Escala de Autoeficacia del Profesorado según las Percepciones del Alumnado: Un Enfoque Multinivel. **Antecedentes:** existen evidencias del papel que las prácticas educativas de los docentes tienen en el rendimiento escolar y el bienestar de los estudiantes. Sin embargo, faltan medidas válidas en español que permitan estudiar las prácticas educativas efectivas a partir de las percepciones de los estudiantes. Por ello, este estudio tiene como objetivo proporcionar una escala válida y fiable para medir las prácticas educativas eficaces en entornos escolares (STEPS). **Método:** analizamos, en una muestra de 2.242 estudiantes anidados en 104 aulas de 22 escuelas españolas, la consistencia y fiabilidad interna de la escala, la solución e invariancia de factores y la validez de criterio mediante el uso de un enfoque multinivel. **Resultados:** los resultados indicaron que la escala exhibió una buena fiabilidad de acuerdo con el coeficiente omega (intra = .86, e inter = .98); el análisis factorial confirmatorio multinivel (MCFA) reveló una estructura jerárquica: gestión del aula, estrategias de instrucción y participación de los estudiantes, como factores de primer orden; y un factor general de segundo orden etiquetado como prácticas educativas efectivas. Además, las prácticas educativas efectivas se asociaron con mejor autoestima de los estudiantes. **Conclusiones:** este estudio ofrece un instrumento fiable y válido, STEPS, para medir prácticas educativas efectivas.

Palabras clave: prácticas educativas; evaluación docente; análisis factorial confirmatorio multinivel; autoestima.

During the last three decades researchers have focused on teachers' behaviours as predictors of student achievement, trying to establish which specific educational practices are effective. Therefore, nowadays effective educational practices refer to such practices that teachers implement in the classroom that have consistently been connected to students' outcomes in many previous studies (for a review, see Muijs et al., 2014). In this regard, John Hattie (2009) clearly synthesized over 800 different meta-analyses and provided an overview of educational practices affecting students' outcomes. His results confirmed the effectiveness of some educational practices, such as providing feedback, managing classroom behaviour, teacher clarity, teacher–student relationships, cooperative learning, direct instruction, mastery learning, classroom management, peer tutoring, worked examples, and concept mapping. Indeed, he concluded that teachers' educational practices are the most relevant factor in determining students' outcomes, over other factors grouped into the categories: 'student', 'home', 'school', 'curricula', and 'teaching approaches' (Hattie, 2009).

More recently, a meta-analysis of 167 studies revealed seven key dimensions associated with three domains that predict students' school achievements (Kyriakides et al., 2013): classroom management (class learning environment, time management and assessment); instruction (structuring, modelling, application); and self-regulation (i.e., teachers' attempt to encourage self-

Received: July 20, 2020 • Accepted: March 1, 2021
Corresponding author: María-José Lera
Department of Developmental and Educational Psychology
Universidad de Sevilla
41018 Sevilla (Spain)
e-mail: lera@us.es

regulation and help students understand the reasons for which they should be engaged in certain learning tasks). These dimensions are in line with the findings of the Tripod Project, which included five different observational instruments, plus the assessment provided by students and teacher rates (Kane et al., 2014). Their results indicated the existence of two clear domains: classroom management, and instruction; together with a third unprecise domain that comprised emotional climate and students' engagement. Also, the observational instrument TEACH (Molina et al., 2018) has been designed to help low- and middle-income countries improve teaching quality in three domains: classroom culture, instruction, and socioemotional skills.

Based on the existing evidence, we define effective educational practices as the actions that teachers take to promote a supportive teaching-learning environment that facilitates both students' achievements and psychological development (e.g., increased self-esteem: Watkins, 2000), which can be divided into three domains: (1) keeping an adequate classroom management, which allows creating an environment that facilitates both socioemotional and academic student's progress (Rolland, 2012; Vandenbroucke et al., 2018); (2) providing high quality instruction that encourage students' critical thinking and analysis (Kraft et al., 2018; Kyriakides et al., 2013; Stockard et al., 2018); and (3) establishing supportive teacher-students relationships that encourage students' to value learning and promote their engagement (Korpershoek et al., 2016).

The assessment of such educational practices and its domains has traditionally been carried out through three sources of information: teachers' self-assessments, classroom observation, and students' perceptions. Each method has its own pros and cons in terms of reliability, cost-efficiency, and quality of the feedback to provide ongoing teachers' coaching and training. Therefore, recent calls in the literature advocate for incorporating different sources of information to improve the predictive power to measure effective educational practices (see findings from the MET project: Kane et al., 2014).

However, although there are several reliable instruments based on previous evidence to both obtain teachers' self-assessments (i.e., self-reported scales in which teachers rate their own educational practices), and classroom observation by experts such as inspectors and researchers (e.g., ICALT: van de Grift, 2007; CLASS: Pianta et al., 2008; TEACH: Molina et al., 2018); there is a lack of valid measures based on students' ratings. Moreover, student ratings have not been always considered as valid, either because of possible bias or the lack of a theoretical model that guided measuring educational practices from students' perceptions (van der Lans et al., 2019; van der Scheer et al., 2019). Indeed, previous studies have shown both a low correlation between teachers and students' ratings when similar questionnaires have been administered to them, and a substantial variation from student-to-student rates (Klassen & Tze, 2014; Wagner et al., 2016), which makes difficult integrating different sources of information to gauge and improve educational practices.

In response, this study aimed to examine the psychometric properties of the teacher efficacy scale (TES: Tschannen-Moran & Hoy, 2001) when students' rates are used both at the individual (student level) and the group level (classroom-teacher level) to assess effective educational practices. This new version of the TES scale is labelled as the Scale of Teachers' Effective Practices rated by Students (STEPS), which tries to fill a gap in the effective educational practices literature concerning the lack of available measures that incorporate students' perceptions and their multilevel nature (see Woitschach et al., 2019).

From a theoretical perspective, we selected the TES scale because it follows a similar three-dimensional model of effective teaching practices but from the teachers' perspective or teachers' self-efficacy. In doing so, we connect research on teachers' self-efficacy -or "self-referent judgments of capability to organize and execute actions required to successfully perform teaching tasks and positively impact student learning" (Perera et al., 2019, p. 187)- and research on effective educational practices (i.e., assessing educational practices linked with students' achievements), which may open new venues for further research.

From a methodological perspective, as the data from research conducted in educational contexts is usually hierarchically structured (the responses are from students nested within a variety of levels, such as classrooms, teaching levels or schools), multilevel factorial analysis techniques may help to shed some light on the factor structure at the various levels of the data (see Fernández-Alonso & Muñiz, 2019). In our case, students are likely to live a similar experience within their classroom (i.e., at least part of the variability of the measurement depends on the fact that the respondents pertain to groups or classrooms exposed to the same teacher). Consequently, we analyse the psychometric properties of the STEPS by considering: (a) its internal consistency and reliability (Cronbach's alpha and omega coefficient); (b) its multilevel factor solution through confirmatory factor analysis (MCFA); (c) its invariance configuration across contextual factors such as region, sex, and teaching level; and (d) its criterion validity as a predictor of students' self-esteem, which is correlated with higher quality learning process (for a meta-analysis, see Watkins, 2000).

In sum, this study contributes to the existing literature by adapting a widely used valid and reliable measure that is rooted in a theoretical framework, which allows cross-cultural comparisons and reduces students' bias when rating teachers' effective teaching practices. Moreover, when offering the psychometric properties of the scale, we overcome previous limitations in the literature and apply a multilevel approach that considers that teachers' teaching practices affect a certain group of students that are nested in classrooms, and therefore students' ratings can be aggregated.

Method

Participants

Our sample was composed of 2,242 students (48.4% girls) with a mean age of 12.76 years (SD = 1.95, range between 9 and 20 years old) nested into 104 classrooms from 22 schools in Basque Country and Andalusia (Spain). Regarding the teaching level, 28.1% were enrolled in upper levels of primary school (8-12 years), 68.9% in secondary school (12-16 years), and 3% in high school (16-18 years).

Instruments

Sociodemographic factors. Participants' age (years old), sex (girls vs. boys), teaching level (primary, secondary, and high school), and region (Basque Country vs. Andalusia) were controlled or included to check for measurement invariance.

Effective educational practices. There is a teacher responsible for each classroom, whose educational practices were rated by

students. Accordingly, we adapted the 24 items of the Ohio State Teacher Efficacy Scale (TES: Tschannen-Moran & Hoy, 2001) in order to be answered by students. Students rated in a Likert scale ranging from 1 ("nothing") to 5 ("absolutely") the degree in which their teachers use different educational practices according to three domains: classroom management (e.g., "Does your teacher get you and your classmates to follow classroom rules?" instead of the original "How much can you do to get children to follow classroom rules"), instructional strategies (e.g., "Does your teacher craft good questions in class?" instead of the original item "To what extent can you craft good questions for your students?"), and students' engagement (e.g., "Does your teacher help you and your classmates value learning?" instead of the original "How much can you do to help your students value learning"). See Table 1.

Students' self-esteem. This variable was measured with the Rosenberg Self-Esteem Scale (RSES: Rosenberg, 1965) in its Spanish version (Martín-Albo et al., 2007). The scale consists of 10 items following a Likert response scale ranging from 1 (strongly disagree) to 5 (strongly agree). After recoding negative items, the scale provides a total score ranging from 10 to 40 points, where higher scores indicate a higher general self-esteem.

Procedure

The study followed the American Psychological Association (APA) Ethical Principles and Code of Conduct. This study is part of a larger project aimed at improving educational practices. We approached schools located in two regions of Spain that represent two socio-economic extremes within this country: Gipuzkoa in the North that represents a medium-high socioeconomic level; and Algeciras in the South, an area representing low-medium socioeconomic level. In both areas a school counselor facilitated approaching the schools (i.e., convenience sampling). In Gipuzkoa 40 classrooms belonging to 13 schools agreed to participate in the project (as intervention group). Then, we invited other schools (as comparison group) that were randomly selected for being in the same area than those who agreed to participate. Therefore, 20 classrooms from 6 schools were added to our initial sample as they decided to participate in our project. All the schools were funded by public funds, although half of them were managed by parents' cooperatives (30 classrooms in the cooperative schools), while the other half were run by the regional authorities (30 classrooms in the state schools). Finally, in Algeciras, a total of 53 classrooms from 4 state schools voluntarily decided to take part in the project.

After signing an agreement with the School Council, headmasters informed their teachers verbally about the study. Then, information sheets explaining the purpose of the project were given through teachers to the students and their parents, who gave their written consent to participate in the study. Some research assistants under the supervision of the first author collected the data in each classroom during the school schedule as part of routine schoolwork. Participation was voluntary and confidentiality was guaranteed.

Data analysis

First, descriptive statistics, sample adequacy and multivariate normality tests were conducted to ensure that our data met the needs for the analyses. Second, we performed several multilevel

Table 1
Scale items adapted from Tschannen-Moran and Woolfolk-Hoy (2001)

Item no.	Item content
	Please indicate your opinion about each of the statements below… [Por favor expresa tu opinión sobre las siguientes frases. Indica si **tu profesor**…]
1	*How much can you do to get through to the most difficult students?* [Te explica de manera que lo entiendes]
2	*How much can you do to help your students think critically?* [Te ayuda a pensar de manera crítica]
3	*How much can you do to control disruptive behavior in the classroom?* [Te ayuda a controlar el mal comportamiento (si lo tienes)]
4	*How much can you do to motivate students who show low interest in schoolwork?* [Te motiva cuando tienes poco interés por las tareas]
5	*To what extent can you make your expectations clear about student behavior?* [Te dice cómo te debes comportar]
6	*How much can you do to get students believe they can do well in schoolwork?* [Te ayuda a creer que puedes hacer bien las tareas]
7	*How well can you respond to difficult questions from your students?* [Te responde adecuadamente a las preguntas difíciles que haces]
8	*How well can you establish routines to keep activities running smoothly?* [Te ayuda a tener rutinas para hacer mejor tu trabajo]
9	*How much can you do to help your students value learning?* [Te ayuda a valorar el aprendizaje]
10	*To what extent can you gauge student comprehension of what you have taught?* [Sabe medir lo que has aprendido]
11	*To what extent can you craft good questions for your students?* [Te realiza buenas preguntas]
12	*How much can you do to foster student creativity?* [Fomenta tu creatividad]
13	*How much can you do to get children to follow classroom rules?* [Te ayuda a que sigas las normas del aula]
14	*How much can you do to improve the understanding of a student who is failing?* [Te ayuda a mejorar tu comprensión cuando suspendes]
15	*How much can you do to calm a student who is disruptive or noisy?* [Te calma cuando tienes un mal comportamiento]
16	*How well can you establish a classroom management system with each group of students?* [Trabajáis por grupos]
17	*How much can you do to adjust your lessons to the proper level for individual students?* [Adapta las lecciones a un nivel adecuado para ti]
18	*To what extent can you use a variety of assessment strategies?* [Te evalúa de diferentes maneras]
19	*How well can you keep a few problem students from ruining an entire lesson?* [Puede controlar a los estudiantes más problemáticos]
20	*To what extent can you provide an alternative explanation or example when students are confused?* [Te proporciona una explicación o un ejemplo alternativo cuando no entiendes algo]
21	*How well can you respond to defiant students?* [Responde adecuadamente a los/as alumnos/as difíciles]
22	*How much can you assist families in helping their children do well in school?* [Puede ayudar a tu familia para que tengas buenos resultados en el colegio]
23	*How well can you implement alternative strategies in your classroom?* [Te deja elegir entre distintas maneras de hacer las tareas]
24	*How well can you provide appropriate challenges for very capable students?* [Te proporciona retos adecuados para motivarte más]

confirmatory factor analyses (MCFA), in order to test the dimensionality within the STEPS scale. For this purpose, several common exploratory factor analyses (EFA) and confirmatory factor analyses (CFA) were made to compare differences on the measure structure when we consider the nested (hierarchical) data or not. According to the recommendations of Hu and Betler (1999), model fit was assessed through: (a) the chi squared coefficient (χ^2/df) whose optimal values are below 3; (b) comparative fit index (CFI) and the Tucker-Lewis fit index (TLI), whose recommended values are above .95 (indicating a good fit); (c) root mean square error of approximation (RMSEA), with values below .01 indicating excellent fit, below .05 indicating good fit and below .08 indicating mediocre fit; and (d) standardized root mean square residual (SRMR), which is better as closer to 0 (perfect fit).Third, and invariance analysis was performed, comparing three grouping demographic variables (teaching level, sex, and region). Fourth, a multilevel linear regression was performed to assess whether the measure was related to one critical output (criterion validity), a self-esteem scale. All analyses were performed with the free-access statistical software R Studio version 1.2.1335 and SPSS version 25.

Results

Descriptive statistics and reliability

Table 2 shows descriptive statistics for our study variables at the individual level. As sociodemographic factors, we included students' sex, age, and educational stages. Reliability was assessed through the multilevel omega reliability index (Green & Yang, 2015; Peters, 2014), which showed acceptable values for the subscales' within-group component (ω^w = .58 - .76), good values for the between-group component (ω^b = .89 - .98), and good values for the overall STEPS scale (ω^w = .86; ω^b = .98). Table 3 shows item's descriptive (mean, standard deviation, range, skewness and kurtosis) and reliability values. Item 24 seem to be the one that performs better in terms of reliability.

Preliminary steps

In order to ensure that the data was suitable for factorial analyses, we checked for normality, sample adequacy, and sphericity. Univariate normality (Komogorov-Sminrov test) was not accomplished for neither of the items, nor the overall scale score. Multivariate normality was checked through Mardia test, showing that the data also violates this assumption ($p < .01$ for both multivariate skewness and kurtosis). Finally, sample adequacy and sphericity were tested with the KMO test (Kaiser-Meyer-Olkin values ranged between .95 and .98) and Bartlett's test of sphericity ($p < .01$), showing that our data fitted sample assumptions (Cerny & Kaiser, 1977). Then, according to the recommendations of Muthén (1994), we performed several ordinary exploratory factor analyses and confirmatory factor analyses, to compare results from data treated as independent with posterior multilevel analysis (which assumes non-independency among data). For this preliminary step, the sample was randomly split in two halves (N = 1,121 each). Regarding missing data, we used maximum likelihood (ML) estimation for one-level models, and the expectation-maximization (EM) algorithm for multilevel models (see Fernández-Alonso et al., 2012).

First, a parallel analysis revealed the presence of seven underlying factors in our data, with eigenvalues of 7.84, 0.45, 0.38, 0.27, 0.25, 0.19 and 0.15, respectively. An ordinary EFA with Unweighted Least Squares (ULS) estimation method was then performed. Table 2 also reports ICC values by item. For values closer to one, ICC indicates that there is a significant amount of variance that can be due to level 2 (group) properties, and thus, that multilevel CFA would be an accurate analytic choice (James, 1982).

Second, five competing models were estimated through ordinary CFA (unidimensional, uncorrelated three-factor, correlated three-factor, hierarchical and bifactor models). Table 4 shows the fit report for those models (five top rows). Overall, the bifactor model was the one which achieved better fit values (χ^2/df = 1.95, CFI = .97, TLI = .96, RMSEA = .03, SRMR = .02). However, when focusing on standard errors, z values and significance of loadings, this bifactor model was rejected due to some abnormal parameter estimation (data available upon request to researchers). Then, we turned to both the 3-correlated factors and the hierarchical 1-3 models (χ^2/df = 2.23, CFI = .95, TLI = .95, RMSEA = .04, SRMR = .03). As both models are similarly parsimonious, we decided to retain the hierarchical 1-3 model as it captures better the concept of educational practices, which is composed by 3 correlated factors (see Path diagram in Figure 1).

Table 2
Mean, standard deviation, skewness, kurtosis, reliability, and bivariate correlations among study variables

Variable	M (SD)	ω^w	ω^w	ICC1	ICC2	Skew	Kurt	1	2	3	4	5	6	7	8
1- Sex	1.52 (0.50)	–	–	–	–	-0.07	-2	–							
2- Age	12.76 (1.95)	–	–	–	–	0.05	-0.29	-.01	–						
3- Educational stage	1.81 (0.72)	–	–	–	–	2.25	9.19	-.00	.71**	–					
4- Region	1.40 (0.50)	–	–	–	–	0.09	-1.99	-.00	-.52**	-.53**	–				
5- STEPS	3.90 (0.81)	.86	.98	.31	.90	-0.81	0.71	-.04	-.24**	-.18**	.19**	–			
6- Classroom Management	3.91 (0.88)	.58	.89	.22	.85	0.54	7.33	-.04	-.23**	-.19**	.21**	.87**	–		
7- Instructional Strategies	3.87 (0.88)	.73	.98	.27	.88	-0.53	1.65	-.03	-.22**	-.14**	.18**	.90**	.67**	–	
8- Student Engagement	3.91 (0.91)	.76	.98	.27	.88	-0.34	3.78	-.04	-.20**	-.14**	.14**	.92**	.70**	.77**	–
9- Self-esteem	3.91 (0.71)	.03	.79	.07	.58	-0.20	3.50	.03	-.13**	-.10**	.07**	.20**	.16**	.20**	.18**

* p<.05; ** p<.01; N = 2031
M = Mean; SD = Standard Deviation; ω^w = Omega Reliability Within Groups; ω^w = Omega Reliability Between Groups; ICC = Intraclass Correlation Coefficient; Skew = Skewness; Kurt = Kurtosis

Table 3
Item descriptive statistics and reliability indicators

Item	Mean (SD)	Skew	Kurt	M	Var	Corr1	Corr2	α	ICC
1	4.25 (1.38)	12.82	406.07	89.76	363.60	.42	.22	.899	.23
2	3.74 (1.28)	-.73	-.46	90.28	361.12	.52	.30	.897	.41
3	3.83 (1.87)	11.08	323.25	90.19	357.05	.39	.18	.901	.76
4	3.78 (1.62)	4.13	73.77	90.23	353.20	.53	.33	.897	-.68
5	4.02 (2.15)	12.88	294.77	90.00	368.21	.19	.06	.909	.68
6	4.19 (1.12)	15.43	510.32	89.83	357.00	.70	.54	.895	.75
7	4.24 (1.44)	13.66	342.09	89.78	361.457	.45	.24	.899	-.38
8	3.87 (1.74)	14.20	450.18	90.15	352.77	.49	.26	.898	.97
9	4.02 (1.66)	16.85	562.651	89.99	355.51	.48	.25	.898	.86
10	4.04 (1.07)	-1.02	0.42	89.97	359.45	.67	.51	.895	.96
11	4.03 (1.46)	10.92	326.68	89.98	358.26	.50	.28	.898	.91
12	3.73 (1.22)	-.72	-.33	90.28	355.12	.68	.51	.895	.98
13	4.21 (1.05)	-1.27	.92	89.81	363.01	.60	.39	.897	.60
14	3.99 (1.46)	3.33	82.67	90.02	352.52	.61	.40	.895	.98
15	3.89 (2.54)	13.16	267.13	90.12	351.51	.32	.15	.908	.95
16	3.60 (1.43)	-.64	-.91	90.41	363.54	.41	.24	.900	.99
17	3.91 (1.21)	3.89	95.22	90.11	358.27	.62	.41	.896	.97
18	3.59 (1.43)	.34	10.31	90.42	360.09	.48	.29	.898	.97
19	3.88 (1.24)	8.87	249.54	90.14	358.17	.60	.40	.896	.97
20	4.32 (1.03)	5.45	153.54	89.70	362.21	.63	.46	.896	.89
21	4.19 (1.07)	-1.33	1.14	89.82	360.55	.65	.49	.896	.90
22	3.74 (1.32)	-.73	-.59	90.27	353.44	.66	.47	.895	.99
23	3.36 (1.41)	-.34	-1.11	90.65	354.07	.60	.46	.896	.98
24	3.60 (1.38)	-.57	-.90	90.41	34.58	.70	.57	.893	.98

N = 2242; α (overall scale) = .90

Skew = Item Skewness; Kurt = Item Kurtosis; M = Scale Mean if Item Deleted; Var = Scale Variance if Item Deleted; Corr1= Corrected Item-total Correlation; Corr2= Squared Multiple Correlation; α = Scale Cronbach Alpha if Item Deleted

Table 4
Fitting values for the competing CFA models

Model	χ^2	df	χ^2/df	CFI	TLI	RMSEA	SRMR
1 factor (a)[1]	567.028	252	2.25	.95	.94	.04	.03
3 correlated factors (a)[1]	556.228	249	2.23	.95	.95	.04	.03
3 uncorrelated factors (a)[1]	2059.213	252	8.17	.65	.62	.10	.25
Hierarchical 1-3 (a)[1]	556.228	249	2.23	.95	.95	.04	.03
Bifactor (a)[1]	433.226	222	1.95	.97	.96	.03	.02
1 factor (b)[2]	1340.273	504	2.65	.93	.92	.03	.03 (w) .09 (b)
3 correlated factors (b)[2]	1291.211	498	2.59	.93	.92	.03	.03 (w) .09 (b)
3 uncorrelated factors (b)[2]	4333.579	504	8.59	.66	.632	.07	.21 (w) .58 (b)
Hierarchical 1-3 (b)[2]	1291.211	498	2.59	.93	.92	.03	.03 (w) .09 (b)
Bifactor (b)[2]	924.466	444	2.08	.96	.95	.03	.02 (w) .07 (b)

[1]N = 1121; [2]N = 2242

(a) = from common single-level CFA; (b) = from multilevel CFA
(w) = within groups; (b) = between groups

Multilevel confirmatory factor analysis

Due to the bias implicit in previous analyses (as they do not consider the hierarchical nature of data), we replicated all models following a multilevel CFA approach. This means that while ordinary CFA just takes into consideration within-group level information on the covariance matrix for parameter estimation, the multilevel approach adds between-group level information. The resulting model follows a two-level estimation method, and thus, parameters are doubly computed for each level. As subscales' ICC(1) estimation (Table 2) showed, this technique was justified for our data. In that sense, it is important that notice that both items

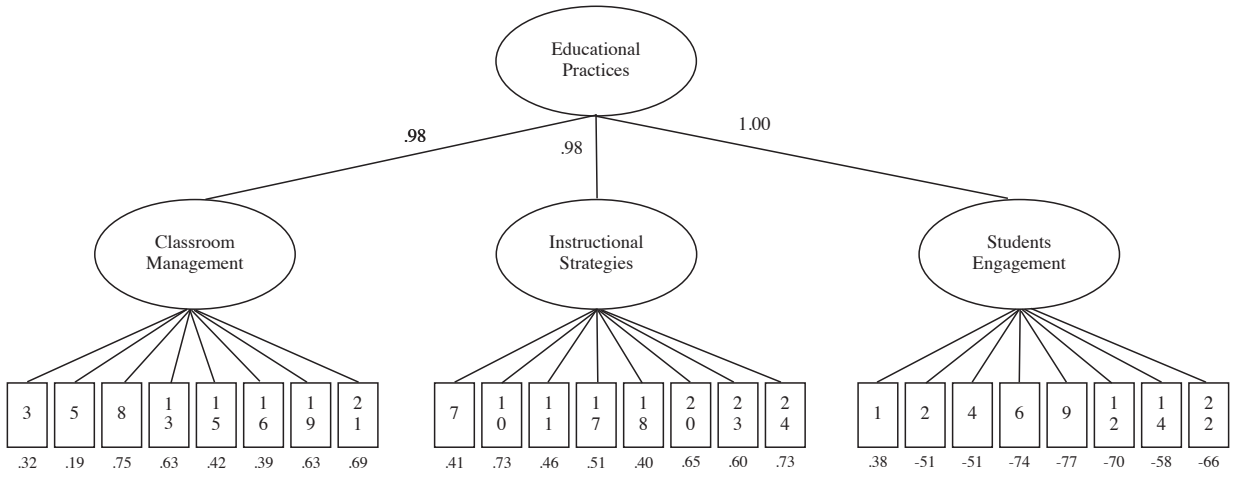

Figure 1. Path diagram for the individual-level hierarchical CFA model

4 and 7 presented negative ICCs. According to Taylor (2010), "Negative ICC estimates are possible and can be interpreted as indicating that the true ICC is low, that is, two members chosen randomly from any class vary almost as much as any two randomly chosen members of the whole population" (p. 8). In our scale, it would indicate that items 4 and 7 capture greater variance at level 1 (individuals) rather than level 2 (group). Nevertheless, in order to preserve the original scale structure, we opted for keeping them.

Table 5 presents the fit report (five last rows) for multilevel CFA models. Again, the bifactor model received more support from data (χ^2/df = 2.08, CFI = .96, TLI = .95, RMSEA = .03, SRMR(w) = .02, SRMR(b) = .07), but, as the model estimation returned abnormal results regarding its standard errors, z values and significance of loadings (results available upon request), we followed the same rational than for ordinary CFA and retained the hierarchical one (χ^2/df = 2.59, CFI = .93, TLI = .92, RMSEA = .03, SRMR(w) = .03, SRMR(b) = .09). All factor loadings were significant on the within-group level structure, but item 5 was not significant at the between-group level. Even so, we retained the item to preserve the original model. Figure 2 shows the path diagram for the multilevel hierarchical solution.

Invariance and criterion validity

Group invariance was tested for educational stages, sex, and region. Several subsequent models were run to test whether the groups shared equivalent factorial structure (configural invariance), factor loadings (metrical invariance), means (scalar invariance) and residuals (strict invariance). Every model was estimated and, if its fit was optimal, it was then compared with the previous one. Whether there were meaningful fit differences, model estimation was stopped, and the former model retained. Following this logic, evidence for configural invariance (same factorial structure across groups) was found for educational stages (CFI = .99, TLI = .99, RMSEA = .00, SRMR = .03), sex (CFI = .99, TLI = .99, RMSEA = .00, SRMR = .03) and region (CFI = .99, TLI = .99, RMSEA = .00, SRMR = .03).

Finally, to test criterion validity we estimated a hierarchical linear regression model with self-esteem as dependent variable and STEPS as predictor (Bliese et al., 2018). As shown in Table 2, ICC(1) for self-esteem was adequate for conducting multilevel regression, as it was .07. In step 1, a null model and a mixed model with random intercepts were compared *(-2Log diff* = 45.55, $p < .01$), which showed that there was meaningful variation between group levels of self-esteem. In step 2, both predictors were introduced (STEPS at individual and aggregated at educational stage). However, STEPS was just significantly predicting self-esteem at the individual level ($t = 8.00$, $p < .01$), which means that group slopes were not significantly differing from those at the individual level. Thus, STEPS scores were significantly predicting individual levels of self-esteem, but not at the group level. This

Table 5
Educational level, sex and region group invariance models' fit and comparison

Variable	Model	χ^2 (df)	*p*	RMSEA	SRMR	CFI	TLI
Educational stage	configural	525.887 (504)	.24	.00	.035	.99	.99
	metrical	688.450 (527)	.00	.02	.04	.99	.99
	ANOVA	**χ^2 diff: 162.56**	**df diff: 23**	***p* < .01**			
Sex	configural	536.381 (504)	.15	.00	.035	.99	.99
	metrical	610.104 (527)	.00	.01	.037	.99	.99
	ANOVA	**χ^2 diff: 73.723**	**df diff: 23**	***p* < .01**			
Region	configural	532.606 (504)	.18	.00	.034	.99	.99
	metrical	777.336 (527)	.00	.02	.04	.99	.99
	ANOVA	**χ^2 diff: 244.73**	**df diff: 23**	***p* < .01**			

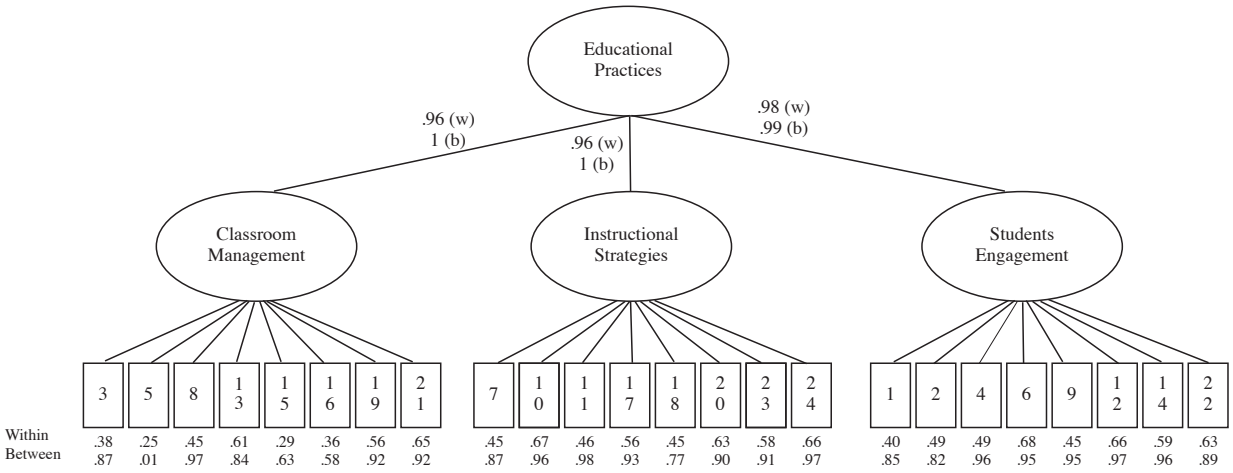

Figure 2. Path diagram for the multilevel CFA hierarchical model

may be due to the nature of our dependent variable (self-esteem) which is intrinsically individual.

Discussion

Giving the importance of evaluating teachers' performance nowadays, there is a common assumption that effective educational practices need to be assessed by integrating information from several sources, including students' ratings. In this regard, this study analysed the psychometric characteristics and multilevel structure of a scale rated by students that measures effective educational practices: the STEPS.

Our results showed that the scale is valid and reliable according to both alpha and omega reliability coefficients. Furthermore, our results supported the theoretical factor structure of the original scale (TES). Hence, a hierarchical factor solution fitted the data best both at the within (students) and between (classrooms) levels, which means that there are three interconnected first-order factors (i.e., classroom management, instructional strategies, and students' engagement) that can be grouped into a second-order core construct: effective educational practices. In addition, this factorial solution remains equal for several sociodemographic and contextual factors, such as sex (girls vs. boys), educational stages (primary, secondary, and high school), and region (Basque Country vs. Andalusia), which underlines the robustness of the multilevel hierarchical factorial solution. In other words, as adopting an individual level approach neglects the nested nature of the phenomena and brings to the wrong assumption of independence of the measures (Bliese et al., 2018), when a multilevel approach is considered, the scale seems a reliable and valid tool for measuring effective educational practices both at the individual and the classroom level.

These results highlight that, although identifying which domains of educational practices contribute more to students' self-esteem might be useful for tailoring interventions aimed at improving teachers' skills and competences (Perera et al., 2019); there is a higher-order general factor comprising classroom management, instructional strategies, and students' engagement that explain more variance that each key domain separately. Therefore, this study contributes to support the integrated approach of effective educational practices (Kyriakides et al., 2013), incorporating at least three generic factors: (1) instruction (i.e., high quality instruction that encourage students' critical thinking and analysis: Kraft et al., 2018; Kyriakides et al., 2013; Stockard et al., 2018); (2) classroom management (i.e., keeping an adequate classroom management, which allows creating an environment that facilitates both socioemotional and academic student's progress: Rolland, 2012; Vandenbroucke et al., 2018); and (3) engagement promotion (i.e., establishing supportive teacher-students relationships that encourage students' to value learning and promote their engagement: Korpershoek et al., 2016). These results also support the generic nature of these factors, as there were not differences between educational stages, students' sex, or regions.

In addition, effective educational practices were associated with higher student's self-esteem (at individual level). In other words, students that perceived their teacher as good in managing the class, instructing in a way that facilitates learning, and caring about them and looking for their motivation, also reported higher levels of self-esteem. This result is in line with previous studies that have emphasize the impact of school experiences on students' self-esteem (Hoge et al., 1990; Watkins, 2000), which is considered a pivotal socio-cognitive resource related to students' academic achievement and well-being (Li et al., 2018). Therefore, although causality need to be addressed in future studies, these results support to some extent that improving teachers' educational practices might be a relevant way to increase students' psychological wellbeing (Ashdown & Bernard, 2012; Fernández-Rodríguez et al., 2019; Suldo et al., 2009).

These results have also interesting implications for practice and policy making. The STEPS can enrich existing training programs by adding an evidence-based evaluation of teachers' educational practices from students' perceptions. Therefore, the STEPS can be used as a research tool, but also as an intervention tool aimed at providing specific feedback to teachers and increase their quality of teaching. In incorporating the STEPS into teachers' training programs, attention should also be paid to the extent that the proposed factors are equally effective across educational stages (as our study suggests).

Limitations and further research

Besides its interesting contributions, our study has some limitations that should be overcame in further research. First,

our sample followed a convenience sampling technique (i.e., some schools were not randomly selected) and therefore is non-representative of the Spanish schools, which limits the generalizability of our findings. Second, we employed a cross-sectional research design. Thus, future studies should incorporate longitudinal designs to evaluate the effects of teaching practices over time on relevant outcomes at different levels (e.g., classroom climate, students' school achievement). In doing so, further research may benefit from other measures beyond self-report scales to both avoid potential common method variance biases (Podsakoff et al., 2003) and offer a more robust view of our findings. Moreover, although self-esteem is correlated with high learning process (for a meta-analysis, see Watkins, 2000) and can be considered a proxy of academic achievement, further studies should directly assess academic achievement or include other variables tapping the motivational or behavioral aspects of students' achievement rather than the psychological one.

Finally, from a methodological perspective, our results overcome previous flaws in the literature by testing the multilevel nature of teachers' educational practices, which allows conducting comparisons between teachers' but also to what extent educational practices vary within the same teacher in different contexts or classrooms. Future research may build on these assumptions and explore two relevant issues: (a) establishing cut-off scores to establish which scores educational practices need to achieve to be labeled as "effective enough" or categorize the quality of teaching; and (b) exploring the external and internal factors that may explain why some teachers reach higher scores than others, or even differences in the scores of the same teacher in one classroom compared to another classroom.

Despite these limitations, this study offers a reliable and valid scale for measuring effective teaching practices in school settings, considering three factors from students' perspective: instructional strategies, classroom management, and help for students' self-regulation and engagement.

Acknowledgements

This study was supported by research transfer contract FIUS 0956/0426. PRZ is a PhD student funded by the FPU program of the Spanish Ministry of Science, Innovation and Universities (FPU18/03536).

References

Ashdown, D. M., & Bernard, M. E. (2012). Can explicit instruction in social and emotional learning skills benefit the social-emotional development, well-being, and academic achievement of young children? *Early Childhood Education Journal*, *39*(6), 397-405. https://doi.org/10.1007/s10643-011-0481-x

Bliese, P. D., Maltarich, M. A., & Hendricks, J. L. (2018). Back to basics with mixed-effects models: Nine take-away points. *Journal of Business and Psychology, 33*(1), 1-23. https://doi.org/10.1007/s10869-017-9491-z

Cerny, C. A., & Kaiser, H. F. (1977). A study of a measure of sampling adequacy for factor-analytic correlation matrices. *Multivariate Behavioral Research, 12*(1), 43-47. https://doi.org/10.1207/s15327906mbr1201_3

Fernández-Alonso, R., & Muñiz, J. (2019). Calidad de los sistemas educativos: modelos de evaluación. *Propósitos y Representaciones, 7*, e347. http://dx.doi.org/10.20511/pyr2019.v7nSPE.347

Fernández-Alonso, R., Suárez-Álvarez, J., & Muñiz, J. (2012). Imputación de datos perdidos en las evaluaciones diagnósticas educativas. *Psicothema*, *24*(1), 167-175.

Fernández-Rodríguez, C., Soto-López, T., & Cuesta, M. (2019). Needs and demands for psychological care in university students. *Psicothema, 31*(4), 414-421. https://doi.org/10.7334/psicothema2019.78

Green, S. B., & Yang, Y. (2015). Evaluation of dimensionality in the assessment of internal consistency reliability: Coefficient alpha and omega coefficients. *Educational Measurement: Issues and Practice*, *34*(4), 14-20. https://doi.org/10.1111/emip.12100

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Hoge, D. R., Smit, E. K., & Hanson, S. L. (1990). School experiences predicting changes in self-esteem of sixth-and seventh-grade students. *Journal of Educational Psychology*, *82*(1), 117.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55.

James, L. R. (1982). Aggregation bias in estimates of perceptual agreement. *Journal of Applied Psychology, 67*(2), 219-229.

Kane, T., Kerr, K., & Pianta, R. (2014). *Designing teacher evaluation systems: New guidance from the measures of effective teaching project*. John Wiley & Sons.

Klassen, R. M., & Tze, V. M. (2014). Teachers' self-efficacy, personality, and teaching effectiveness: A meta-analysis. *Educational Research Review*, *12*, 59-76. https://doi.org/10.1016/j.edurev.2014.06.001

Korpershoek, H., Harms, T., de Boer, H., van Kuijk, M., & Doolaard, S. (2016). A meta-analysis of the effects of classroom management strategies and classroom management programs on students' academic, behavioral, emotional, and motivational outcomes. *Review of Educational Research*, *86*(3), 643-680. https://doi.org/10.3102/0034654315626799

Kraft, M. A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research*, *88*(4), 547-588. https://doi.org/10.3102/0034654318759268

Kyriakides, L., Christoforou, C., & Charalambous, C. Y. (2013). What matters for student learning outcomes: A meta-analysis of studies exploring factors of effective teaching. *Teaching and Teacher Education*, *36*, 143-152. https://doi.org/10.1016/j.tate.2013.07.010

Kyriakides, L., Creemers, B. P., & Panayiotou, A. (2018). Using educational effectiveness research to promote quality of teaching: The contribution of the dynamic model. *ZDM Mathematics Education*, *50*(3), 381-393. https://doi.org/10.1007/s11858-018-0919-3

Li, J., Han, X., Wang, W., Sun, G., & Cheng, Z. (2018). How social support influences university students' academic achievement and emotional exhaustion: The mediating role of self-esteem. *Learning and Individual Differences*, *61*, 120-126. https://doi.org/10.1016/j.lindif.2017.11.016

Martín-Albo, J., Núñez, J. L., Navarro, J. G., & Grijalvo, F. (2007). The Rosenberg Self-Esteem Scale: Translation and validation in university students. *The Spanish Journal of Psychology*, *10*, 458-467. http://dx.doi.org/10.1017/S1138741600006727

Molina, E., Fatima, S. F., Ho, A., Hurtado, C. M., Wilichowksi, T., & Pushparatnam, A. (2018). *Measuring teaching practices at scale: Results from the development and validation of the TEACH classroom observation tool*. Research Working Paper No. 8653. World Bank, Washington, DC.

Muijs, D., Kyriakides, L., Van der Werf, G., Creemers, B., Timperley, H., & Earl, L. (2014). State of the art-teacher effectiveness and professional learning. *School Effectiveness and School Improvement*, *25*(2), 231-256. https://doi.org/10.1080/09243453.2014.885451

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research, 22*(3), 376-398.

Perera, H. N., Calkins, C., & Part, R. (2019). Teacher self-efficacy profiles: Determinants, outcomes, and generalizability across teaching level. *Contemporary Educational Psychology*, *58*, 186-203. https://doi.org/10.1016/j.cedpsych.2019.02.00

Peters, G. J. (2014). The alpha and the omega of scale reliability and validity: Why and how to abandon Cronbach's alpha and the route towards more comprehensive assessment of scale quality. *European Health Psychologist, 16*(2), 56-69. https://doi.org/10.31234/osf.io/h47fv

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System*. Paul H. Brookes Publishing.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879-903.

Rolland, R. G. (2012). Synthesizing the evidence on classroom goal structures in middle and secondary schools: A meta-analysis and narrative review. *Review of Educational Research*, *82*(4), 396-435. https://doi.org/10.3102/0034654312464909

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton University Press.

Stockard, J., Wood, T. W., Coughlin, C., & Rasplica Khoury, C. (2018). The effectiveness of direct instruction curricula: A meta-analysis of a half century of research. *Review of Educational Research*, *88*(4), 479-507. https://doi.org/10.3102/0034654317751919

Suldo, S. M., Friedrich, A. A., White, T., Farmer, J., Minch, D., & Michalowski, J. (2009). Teacher support and adolescents' subjective well-being: A mixed-methods investigation. *School Psychology Review*, *38*(1), 67-85.

Taylor, P. J. (2010). An introduction to intraclass correlation that resolves some common confusions. *Unpublished manuscript, University of Massachusetts, Boston, USA*. Retrieved from http://www.faculty.umb.edu/peter_taylor/09b.pdf

Tschannen-Moran, M., & Hoy, A. W. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education*, *17*(7), 783-805. https://doi.org/10.1016/S0742-051X(01)00036-1

Van de Grift, W. (2007). Quality of teaching in four European countries: A review of the literature and an application of an assessment instrument. *Educational Research, 49*, 127-152. https://doi.org/10.1080/00131880701369651

van der Lans, R. M., van de Grift, W. J., & van Veen, K. (2019). Same, similar, or something completely different? Calibrating student surveys and classroom observations of teaching quality onto a common metric. *Educational Measurement: Issues and Practice*, *38*(3), 55-64. https://doi.org/10.1111/emip.12267

van der Scheer, E. A., Bijlsma, H. J., & Glas, C. A. (2019). Validity and reliability of student perceptions of teaching quality in primary education. *School Effectiveness and School Improvement*, *30*(1), 30-50. https://doi.org/10.1080/09243453.2018.1539015

Vandenbroucke, L., Spilt, J., Verschueren, K., Piccinin, C., & Baeyens, D. (2018). The classroom as a developmental context for cognitive development: A meta-analysis on the importance of teacher-student interactions for children's executive functions. *Review of Educational Research*, *88*(1), 125-164. https://doi.org/10.3102/0034654317743200

Wagner, W., Göllner, R., Werth, S., Voss, T., Schmitz, B., & Trautwein, U. (2016). Student and teacher ratings of instructional quality: Consistency of ratings over time, agreement, and predictive power. *Journal of Educational Psychology*, *108*(5), 705-721. https://doi.org/10.1037/edu0000075

Watkins, D. (2000). Learning and teaching: A cross-cultural perspective. *School Leadership & Management, 20*(2), 161-173. https://doi.org/10.1080/13632430050011407

Woitschach, P., Zumbo, B. D., & Fernández-Alonso, R. (2019). An ecological view of measurement: focus on multilevel model explanation of differential item functioning. *Psicothema, 31*(2), 194-203. https://doi.org/10.7334/psicothema2018.303003