Inclusion in bilingual education: Assessment of the SPI-BE scale for measuring student perceptions of inclusive practices

José Buz
Ramiro Durán-Martínez
Eva González Ortega (Corresponding Author)
Elena Martín-Pastor
University of Salamanca

Received: 2024-12-18 / Accepted: 2025-11-06
https://doi.org/10.30827/portalin.vi45.32321
Porta Linguarum ISSN paper edition: 1697-7467, ISSN digital edition: 2695-8244

ABSTRACT: Attention to diversity in the bilingual classroom has become a major challenge that is exacerbated by the lack of tools to measure its implementation in primary education. This study aimed to develop and evaluate the Student Perception of Inclusion in Bilingual Education (SPI-BE) scale in order to measure the frequency of different inclusive teaching practices as perceived by students. The sample included 2,626 primary students attending 27 bilingual schools in Spain. A modern IRT-related factorial analysis was used to assess the validity and reliability of the SPI-BE scale by focusing on the quality of the scores resulting from different factor solutions rather than on their goodness of fit. Analyses revealed two well-defined and replicable factor structures that justify the use of two partial scores and one total score. Female students, students in the lower grades, students participating in a British Council programme, and students without special educational needs had greater awareness of the use of inclusive teaching practices than other groups of students. It can be stated that the SPI-BE scale has excellent psychometric properties and can be used by researchers and practitioners as an instrument to measure inclusion in bilingual primary schools.
Keywords: Factor analysis, bilingual education, inclusion, primary students, assessment.

Inclusión en la enseñanza bilingüe: evaluación de la escala SPI-BE para medir las percepciones de los estudiantes sobre las prácticas inclusivas

RESUMEN: La atención a la diversidad en el aula bilingüe se ha convertido en un gran reto que se ve agravado por la carencia de herramientas para medir su aplicación en educación primaria. Este estudio tiene como objetivo desarrollar y evaluar la escala Student Perception of Inclusion in Bilingual Education (SPI-BE) para medir la frecuencia del uso de diferentes prácticas docentes inclusivas según la percepción del alumnado. La muestra incluye 2.626 estudiantes de primaria de 27 colegios bilingües en España. Se utilizó un moderno análisis factorial relacionado con la TRI para evaluar la validez y fiabilidad de la escala SPI-BE que se centra en la calidad de las puntuaciones resultantes de las diferentes soluciones factoriales antes que en su bondad de ajuste. Los análisis revelan dos estructuras factoriales bien definidas y replicables que justifican el uso de dos puntuaciones parciales y una total. Las alumnas, el alumnado de cursos inferiores, los de programas British Council y los que no tienen necesidades educativas especiales mostraron una mayor conciencia del uso de prácticas docentes inclusivas. La escala SPI-BE muestra excelentes propiedades psicométricas y puede ser utilizada por investigadores y profesionales como instrumento para medir la inclusión en centros de enseñanza bilingüe de primaria.

Palabras clave: Análisis factorial, enseñanza bilingüe, inclusión, estudiantes de primaria, evaluación.

1. Introduction

Spain is one of the European countries at the forefront in the implementation of bilingual programmes in compulsory education (Lasagabaster & Ruiz de Zarobe, 2010). For instance, during the 2021-2022 school year, more than one million students used English as a language of instruction in primary education, with 41% of the students attending public schools (Ministerio de Educación, Formación Profesional y Deportes, 2024a). Despite being widespread, compliance with the fourth goal of the 2030 Agenda for Sustainable Development (UNESCO, 2017) regarding quality education requires that bilingual education programmes provide measures that cater to diversity and guarantee inclusion. Teaching practices should ensure that the needs and characteristics of all students are considered, including those with special educational needs (SEN), learning difficulties, immigrants, and socially and culturally disadvantaged students. However, the use of inclusive practices in bilingual education has become a major challenge for teachers (Durán-Martínez et al., 2020; Szczesniak & Luna, 2022), especially when combined with the task of teaching and assessing in a foreign language.

With this in mind, it has become evident there is a need to develop instruments that can assess the extent to which a bilingual programme focuses on inclusion and the types of inclusive practices that are frequently or infrequently used. To date, most research has explored the general perceptions that teachers, parents, and students have about bilingual programmes (e.g. Madrid et al., 2018; Pladevall-Ballester, 2015). However, less attention has been paid to the way students perceive the inclusive practices used by their teachers in the bilingual classroom (Subban et al., 2022). The aforementioned research approach is commonplace even though consideration of the perspectives of learners is essential for developing a successful programme (Coyle, 2013). In addition, from among the self-assessing instruments available, only the Index for Inclusion (Booth & Ainscow, 2011) comprises suitable psychometric properties. However, this index was not specifically designed for the bilingual classroom and has only been validated using data gathered from secondary students in Spain (Fernández-Archilla et al., 2020). According to available information, the only instrument to assess attention to diversity in bilingual education in several European countries, including Spain, has been developed for secondary school students (see Pérez-Cañado et al., 2021), and although it suggests the existence of five dimensions, psychometric data about them is not available. Other instruments, such as the one developed in Spain by Martínez-Hita et al. (2022), aimed at Physical Education teachers in bilingual schools, identified the presence of three dimensions for teaching practices centred on interaction with students, organisational aspects, and perceived self-efficacy.

1.1. The present study

Despite the upsurge in bilingual education both in Europe and Spain, there is still a lack of instruments available for assessing student perceptions of support received in the bilingual classroom, especially in primary education (Bauer-Marschallinger et al., 2023). To address this shortcoming, a multidimensional scale has been developed based on a review of the scientific literature and various measurement instruments. This tool, called the Student Perception of Inclusion in Bilingual Education or SPI-BE scale, was created to measure the degree to which different inclusive practices are implemented by teachers delivering classes in English. In addition, the psychometric properties of the scale have been examined using a novel methodological approach.

Recent developments in psychometrics have highlighted the limitations of relying mainly –and sometimes exclusively– on strict goodness-of-fit criteria to determine the dimensionality of an instrument (e.g., Montoya & Edwards, 2021) or to justify which is the best competing factorial solution (e.g., Marcoulides & Yuan, 2017; Rodríguez et al., 2016; Sellbom & Tellegen, 2019). For example, as argued by authors such as Reise et al. (2015), the best fit of a factor solution can be achieved by including weak factors or factors with only a few items, which would result in models of little substantive interest. Others, such as Garrido et al. (2019), have argued that in addition to acceptable fit and being clearly interpretable, factor solutions must be replicable across studies. The resulting estimated scores must also be determinate and accurate to consistently and effectively rank people along the construct. In the case of multidimensional solutions such as ours, the subscale scores must demonstrate their 'added value' concerning the total score. Otherwise, such scores would be of little use and be confusing in terms of interpretation. Therefore, and in line with the recommendations of the American Educational Research Association (AERA), our interest focuses on the quality of the estimated scores resulting from the factor models examined rather than on the goodness of fit of these models.

Thus, this study has three objectives. The first is to examine the multidimensionality, quality, and appropriateness of the estimated scores resulting from the instrument employed. Specifically, the perceptions of primary school students, in English-Spanish bilingual programmes, were measured regarding the frequency with which their teachers used inclusive teaching practices. The second is to examine whether the scale is essentially unidimensional, which would justify the use of the total score. Lastly, the third objective is to assess criterion validity by examining the differences among perceptions about inclusiveness based on gender, grade, type of bilingual programme, and educational needs.

2. Method

2.1. Sample

Data were gathered from a convenience sample of 2,626 students enrolled at 15 state-run schools and 12 charter schools in Spain following an English-Spanish bilingual curriculum (M_{age}: 10.2 years, *SD*: .92; range: 8-13 years). More than three-quarters of the students (78.8%) belonged to a bilingual programme, known as a bilingual section, created in Castile and Leon, the region in which the study was carried out. In total, 21.2% belonged to the bilingual programme originally established in 1996 through an agreement between the Spanish Ministry of Education (MEC) and the British Council (BC), known as the MEC/BC programme. In both types of bilingual programmes, non-MEC/BC and MEC/BC, the students learn a foreign language as a subject as well as through learning other subjects following the Content and Language Integrated Learning (CLIL) approach (Coyle & Meyer, 2021). These subjects, taught in the same foreign language (English), are usually Natural and Social Sciences, Physical Education, Music, and Arts and Crafts. The data also showed that 49.6% were female students (*n* = 1302) and the sample was distributed as follows: 30.3% were in 4th grade, 34.5% in 5th grade, and 35,2% in 6th grade. Of the total sample, 3.5% (*n* = 93) were students with educational needs (SEN), a percentage that is similar to the data provided by the Spanish government: 3.6% for state-run schools and 3.2% for charter schools (Ministerio de Educación, Formación Profesional y Deportes, 2024b).

2.2. Procedure

All primary schools in the study area with an English-Spanish bilingual programme were contacted via email and telephone to inform the respective management teams about the purpose of the study. After obtaining consent from parents and students, the survey was administered anonymously during school hours, under the supervision of a research team member and/or school staff to address any questions or difficulties. Responding to all items on the scale was mandatory to prevent the occurrence of missing data. Participation was voluntary, non-rewarded, and with the option to withdraw at any time. Also, teachers discreetly marked the surveys taken by the SEN students. Data were gathered in the 2022-2023 school year.

2.3. Scale development

Based on the literature review, a group of six experts with teaching and research experience in bilingual education identified four dimensions linked to inclusive teaching practices:

- *Language* refers to the use of the foreign language (English) for clarifying cognitively complex content (‘When I don’t understand something, they explain it to me again using easier words in English’) or that allows the learner to show what they have learned (‘The teacher helps us to speak in English, giving us useful words and phrases’), among other purposes.
- *Formative Assessment* includes both the feedback provided by the teacher when carrying out any type of activity (‘The teacher explains what we’ve done wrong [...] and how to do it right’) as well as the use of different types of activities that demonstrate the competences acquired (‘The teacher asks us to do different activities to give us a grade, [e.g., a presentation in English, an outline, a test]’).
- *Inclusive Support* includes both scaffolding teaching strategies that stimulate learning and facilitate content acquisition and competence development (‘If I don’t know how to do an activity, the teacher helps me [e.g., clarifying how to do it, giving me examples, giving me more time, translating words, etc.]’), as well as teamwork to foster peer-to-peer support (‘We work in groups or pairs and help each other […]’).

- *Resources* refers to the use of materials, both analogue and digital, that facilitate learning (e.g. 'The teacher uses ICT applications in the classroom [e.g., Kahoot, Plickers, ClassDojo]').

From these four dimensions, an initial set of 38 items was generated with the following general instruction: 'Read these sentences and tell us if they are never (0), sometimes (1), or always (2) true. Remember that they are all about subjects taught in English.: Science, Music, Arts and Crafts, etc.)'. Higher scores represented greater awareness about the frequency of inclusive teaching practices.

2.4. Data analysis

2.4.1. Content validity

To calculate the Coefficient Validity Ratio (CVR), a panel of experts (*n* = 10) was asked to assess whether each item was necessary or important for the theoretical structure. In addition, these experts were required to assess the item content validity (I-CVI) in terms of relevance. The content validity of the whole scale (S-CVI) was computed by using an averaging method (Polit et al., 2007). For CVR and CVI, the minimum acceptable values were > .62 and > .79 (adjusted according to the number of judges), respectively.

2.4.2. Construct validity

An Item Response Theory-related categorial-variable methodology factorial analysis (CVM-FA) was used (Ferrando and Lorenzo-Seva, 2013), where the item scores are treated as ordered-categorical variables, and the FA is fitted to the inter-item polychoric correlation matrix. The number of factors to be retained was assessed by using a parallel analysis (PA). Computations were based on a robust minimum rank factor analysis (R-MRFA) with bias-corrected and accelerated (BCa) bootstraps (500 samples), and the factors were obliquely rotated using Promin. In addition to exploring different correlated factor solutions, since the domains comprising the SPI-BE scale might also be contributing uniquely to the construct, a bifactor solution was fitted. Below we describe the three stages for performing CVM-FA in this study.

1) *Item-level analysis.* The robust measure of sampling adequacy (MSA) was used to examine whether each item was suitable for carrying out a factorial analysis. Statistics such as the item explained common variance (I-ECV) index and the item residual absolute loadings (I-REAL) index indicated how each item reflected the content of a general factor. For I-ECV and I-REAL, values > .85 and < .30, respectively, were expected. The relative difficulty index (RDI) was observed to assess the position of the items along the construct. In this case, values close to 1 (i.e., adequate for measuring students with low scores) were desired. Finally, we conducted a differential item analysis (DIF) to assess whether the items had invariant properties based on gender, grade (4th, 5th, or 6th), the type of bilingual programme (non-MEC/BC vs. MEC/BC), and educational needs (non-SEN vs. SEN). The DIF CONTRAST and Welch´s t-test (with p-value adjustment) was employed. All values above |0.64| logits

and significant t-differences were considered evidence of DIF (Linacre, 2025). As a result of this stage, four underperforming items in terms of CVR, CVI, MSA, I-ECV, I-REAL, and/or RDI were identified. Moreover, item DIF was present in two of them. Students who were enrolled in a non-MEC/BC bilingual programme were less inclined to assign high ratings to the item pertaining to the presence of a support teacher (DIF contrast = -.70; t = -6.62, p < .001), as well as to the item referring to being allowed additional time to complete an exam (DIF contrast = -.86; t = -10.70, p < .001), when compared to students participating in a MEC/BC programme. Consequently, the following analyses were conducted on the final scale, which was comprised of 16 items (see Table A1 in Appendix).

2) *Basic internal assessment***.** Goodness-of-fit was assessed using root mean squared error of approximation (RMSEA), the comparative fit index (CFI), and the population standardized root mean residual (SRMSR) as an indicator of global misfit. The following cut-off criteria were used as indicative of good fit: CFI ≥ .95, RMSEA ≤ .06, and SRMSR values close to Kelley's criterion (this value was .019 in this study). At the item calibration stage, dimensionality was assessed by computing the percentage of explained common variance (ECV) and the Mean of Item Residual Absolute Loadings (MIREAL). ECV values in the range of .70 to .85 indicated the dominance of the first MRFA factor over the primary factors (Rodríguez et al., 2016). For MIREAL, a value < .30 is expected. Replicability was assessed using the generalized H (G-H). For G-H, values > .70 indicated that the factor was well-defined and replicable (Hancock & Mueller, 2001).

   At the scoring stage, we assessed (a) the consistency of differences between students across the construct using the factor determinacy index (FDI), (b) the accuracy of the scores using the marginal reliability estimate (Brown & Croudace, 2015), and (c) the ability of the scores to differentiate levels in the construct using the statistic sensitivity ratio (SR). We also assessed, by plotting the conditional reliability (information) curves, at which levels of the construct the scale provided more accurate measures. For marginal reliability, values around .70 were suitable for group comparisons and values > .90 were suitable for FDI (Rodríguez et al., 2016). Subsequently, we fitted a one-dimensional solution using Samejima's graded response model (GRM) from 1969, since multidimensional data with a strong general factor can be adequately fitted to one-dimensional IRT models.

3) *Added-value assessment.* Finally, we assessed whether the subscale scores provided added value to the total score by calculating the proportional reduction in the mean squared error (PRMSE) of the subscales (PRMSEs) and the total scale (PRMSEt) (Ferrando and Lorenzo-Seva, 2019). Value added is considered to exist when the lower limit of the confidence interval of each factor (i.e., PRMEs) is higher than the estimate obtained for the overall factor (PRMEt).

2.4.3. Criteria validity

For this type of validity, the scale was tested to determine if it could detect differences amongst groups. One-way analysis of variance (ANOVA) was conducted to compare differences in inclusivity based on gender, grade (4th, 5th, or 6th), educational programme (non-MEC/BC vs. MEC/BC), and educational needs (non-SEN vs. SEN). Based on previous findings (Schwab et al. 2018), female students and students in the lower grades were expected to have higher scores than the opposite groups of students. Although there were no initial expectations about differences according to the type of bilingual programme, this variable was included for experimental purposes.

Factor analyses were performed using FACTOR 12.04 (Ferrando & Lorenzo-Seva, 2017), and the rest of the analyses were performed using SPSS v.28 and Winsteps 5 (Linacre, 2025).

3. Results

3.1. Construct validity

3.1.1. Calibration stage

The estimates obtained from the PA suggested that one or two dimensions were retained depending on whether the 95th percentile or the mean was considered. The ECV and MIREAL values (ECV = 74.8%, MIREAL = .24) indicated that the items comprising the SPI-BE scale measured a common overall dimension of inclusive practices.

It was found that the four-factor solution, and later the three-factor solution, had similar factor characteristics and their fit was extremely good or excellent (RMSEA = .021 and .022, respectively). However, the common variance explained by some factors was quite low (values ranged from 10.1% to 34.6%), as was the reliability of some factors (values around .60). Moreover, in line with the G-H statistic, the factor structure was not clear and interpretable. In both solutions, one of the factors consisted of only three items, with some weights being around .20. Moreover, some of the factors were too far from the cut-off point to be considered well-defined and replicable.

Thus, a two-factor solution showed an excellent fit according to all the statistics with a simple structure (Bentler´s simplicity index S = .99) (see Table 1). The common variance explained by each factor was considered suitable. Both factors had satisfactory replicability values, with a clearly interpretable structure, and suitable but moderately low weight distribution, without the presence of complex items (see Table 2).

The first factor was comprised of nine items (mean λ = .46), mainly concerning the use of various materials and activities with which to teach and assess (e.g., ‘The teacher suggests different tasks so that everyone can participate’). This factor was denominated ‘Methodology and Resources’.

The second factor was comprised of seven items (mean λ = .58) related to the use of different teaching aids when teaching and assessing content. This included the use of positive reinforcement and additional explanations (e.g., ‘When I don’t understand something, they explain it to me again using easier words in English’). This factor was denominated ‘Support and Feedback’.

With an estimated inter-factor correlation of .55, it seemed appropriate to fit a two-factor solution with one overall factor and two correlated primary factors. This solution showed an excellent fit and a clearly interpretable structure where the weights were generally well distributed. The weights of the overall factor were higher than the weights of the primary factors, and the overall factor was defined by most of the items of the primary factors. In addition, the correlation between the primary factors, once the overall factor was modelled, remained significant (r = .10), which may be due to the size of the sample. The overall factor strength and replicability were acceptable, but the primary factors did not reach the expected values.

Regarding the bifactor solution, the marginal reliability for the general factor and the primary factor indicated that their ability, in terms of accuracy and consistency, to effectively order individuals along the construct could be considered adequate for research purposes. However, the reliability of the score estimates for the primary factor 'Support and Feedback' was below the acceptable minimum value.

Finally, a good fit was observed when fitting the one-factor solution, although with discrete performance on the residuals. The G-H index showed that it was a well-defined and replicable solution. The factor weights were adequate, although slightly low (λ range = .29 -.57), especially for the factor 'Methodology and Resources'.

Individual scoring stage and score-based measures for accuracy and appropriateness

The general accuracy of the EAP score estimates of the two-factor solution and the one-factor solution even exceeded the cut-off point established for individual assessments (see Table 2).

Table 1. *Factorial Weights of the Solutions*

ITEM WORDING	TWO-FACTORS		TWO-FACTORS BIFACTOR			ONE-FACTOR
	F1	*F2*	GF	*F1*	*F2*	
Factor 1: Methodology and resources						
1. We work in groups or pairs and help each other (…)	**.62**	-.26	.06	***.61***	.15	**.29**
2. The teacher suggests different tasks (…)	**.28**	.15	**.37**	.01	-.04	**.29**
3. The teacher uses different materials to help us understand (…)	**.60**	.02	**.58**	**.30**	-.06	**.53**
4. When we do tasks, the teacher gives materials to help us (…)	**.68**	-.01	**.39**	***.53***	**.21**	**.57**
5. When we work in groups or pairs, we change partners.	**.54**	-.10	**.25**	***.43***	.09	**.37**
6. The teacher uses ICT applications in the classroom (…)	**.44**	-.03	**.41**	**.21**	-.10	**.35**
7. When assessing, the teacher gives some students more time (…)	**.40**	-.01	**.45**	.15	-.16	**.33**
8. When assessing, the teacher asks about the topics (…)	**.26**	.11	**.20**	**.21**	**.20**	**.32**
9. The teacher asks us to do different activities (…)	**.39**	.01	**.36**	**.21**	-.02	**.35**
Factor 2. Support and feedback						
10. The teacher encourages me to participate in class.	-.04	**.55**	**.41**	-.14	**.32**	**.47**
11. If I don't know how to do an activity, the teacher helps me (…)	-.02	**.55**	**.39**	-.11	**.37**	**.48**
12. The teacher helps us to speak in English (…)	.01	**.64**	**.47**	-.09	**.44**	**.59**
13. When I don't understand something, they explain it to me (…)	.06	**.54**	**.37**	-.01	***.46***	**.54**
14. When I don't know how to say something, the teacher (…)	-.15	**.70**	**.31**	-.16	***.57***	**.51**
15. The teacher explains what we've done wrong (…)	.00	**.59**	**.42**	-.09	**.42**	**.54**
16. The teacher congratulates us when we do the tasks correctly.	.16	**.48**	**.50**	.03	**.30**	**.57**

Note: Loading values larger than .20 are printed in boldface. Loadings above |.09| were significant according to BCa confidence intervals.

Items with substantial and similar loadings on the overall factor and primary factors would indicate the suitability of using total score and subscale scores (n = 10 items).

Primary factors greater than the weight of an item in the overall factor are in italics.

Table 2. *Basic Internal Assessment of the Factorial Solutions and Quality and Effectiveness of the Factor Score Estimates*

MODEL	*RMSEA*	*CFI*	*RMSR*	*ECV*	*G-H*	*FDI*	*RELIABILITY*	*SR*
	Value	Value	Value		Value	Value	Value	Value
One-factor	.047	.954	.070	74.8%	.82	.91	.83	2.2
Two-factor	.027	.987	.040					
Factor 1				30.3%	.77	.91	.82	2.1
Factor 2				34.9%	.80	.91	.83	2.2
Two-factor bi-factor	.024	.991	.034					
Factor 1				16.8%	.62	.82	.67	1.4
Factor 2				20.2%	.59	.79	.62	1.3
GF				35.4%	.70	.85	.73	1.6

Note: RMSEA = root mean squared error of approximation; CFI = Comparative Fit Index; RMSR = root mean square of residuals; ECV = explained common variance (MRFA-based); G-H = generalized H index; FDI = factor determinacy index (MRFA-based); Reliability = marginal reliability of estimates; SR = sensitivity ratio.

As can be seen in Figure 1, the two factors showed slightly different patterns within the range of measurement. The estimated scores of the factor ‘Methodology and Resources’ were found to be reliable for measuring the medium and high levels of the construct (between θ = -1.69 and θ = 1.83), while the estimated factor scores for the factor ‘Support and Feedback’ only provided reliable measures for those students at the low and middle levels of the construct (between θ = -2.66 and θ = 0.65). In contrast, the total score provided reliable measures, sufficient for between-group comparisons, in the low and middle range (between θ = -2.62 and θ = 1.95, which represented 87% of the range of measurement), and somewhat more reliable, even suitable for individual assessments, in a narrower range of measurement (between θ = -2.44 and θ = 1.03, which represented 70% of the range of measurement).

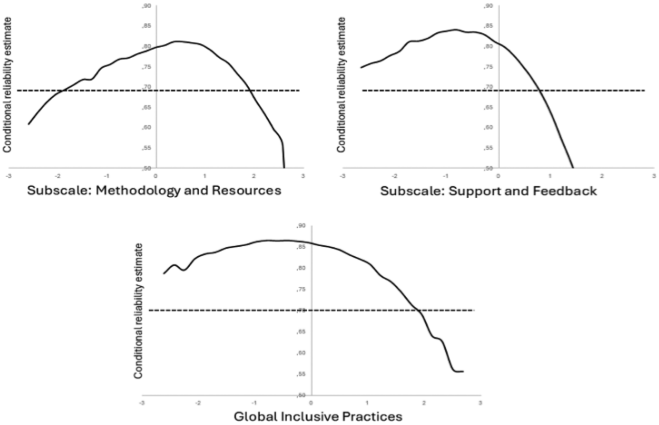

Figure 1. *Conditional Reliabilities for the Total Scale Score and the Subscale Scores from the One-Factor Model and the Two-Correlated Factor Model*

Note: The horizontal dotted line represents the .70 cut-off point for conditional reliability.

3.1.2. Added Value

In the two-factor solution, the statistics for the first factor (PRMSEs = .82, 95% CI [.80; .83]; PRMSEt = .38) and the second factor (PRMSEs = .83, 95% CI [.82; .84]; PRMSEt = .39) showed that the estimated scores of the subscale scores provided 'added value' to a total score, their use was considered suitable.

3.2. Evidence of known-groups validity

Concerning the validity through expected differences in means, Table 3 shows the results of the ANOVA tests for the total score and the subscale scores obtained. Concerning gender, female students were more perceptive of inclusive teaching practices than male students according to the total scale and the two subscales. In terms of the school grade, the total score and the subscales indicated that inclusive teaching practices were perceived more frequently by students in the lower grades than in the higher grades, especially in 6th grade. Regarding educational needs, it was found based on the total score and partial scores that SEN students were not as aware of inclusive practices as the non-SEN students. Regarding the type of bilingual programme, the total score and the support subscale score indicated that there was no difference in student perceptions of inclusive teaching practices between the two types of programmes (non-MEC/BC and MEC/BC). However, the MEC/BC students reported having more access to methodological support than students in non-MEC/BC programmes.

Table 3. *The Ability of SPI-BE Scores to Detect Differences Between Groups*

		TOTAL SCORE		SUBSCALE SCORE: METHODOLOGY		SUBSCALE SCORE: SUPPORT	
VARIABLES	*N*	*M (SD)*	*F*	*M (SD)*	*F*	*M (SD)*	*F*
Gender			13.67***		22.27***		4.19 *
Male	1,324	-0.04 (0.95)		-0.07 (0.77)		-0.03 (0.82)	
Female	1,302	0.09 (0.86)		0.07 (0.87)		0.03 (0.81)	
Grade			9.27 ***		8.08***		7.10 ***
4th [a]	796	0.11 (0.86)		0.09 (0.74)		0.07 (0.77)	
5th [b]	906	0.05 (0.95)		-0.01 (0.88)		0.02 (0.81)	
6th [c]	924	-0.07 (0.91)		-0.07 (0.84)		-0.07 (0.84)	
Student			15.90***		18.21***		12.91***
Non-SEN	2,533	0.04 (0.90)		0.01 (0.82)		0.01 (1.04)	
SEN	93	-0.34 (1.07)		-0.36 (0.94)		-0.29 (0.81)	
Programme			0.84		15.90***		0.22
Non-MEC/BC	2,069	0.20 (0.92)		-0.02 (0.85)		0.01 (0.82)	
MEC/BC	557	0.60 (0.88)		0.10 (0.74)		-0.01 (0.78)	

Note: Post-hoc analysis for: Total score, a, b > c; Methodology, a > b, c; Support, a, b > c.
* < .05. ** < .01. ***< .001.

4. Discussion

The first objective of this work was to create a multidimensional scale to measure the perceptions of primary school students about their teachers' use of inclusive practices to teach subjects in English. This initial objective also included examining the quality and suitability of the scores obtained using the SPI-BE scale.

After examining the goodness of fit of different factorial solutions, it was found that their fit was, by all standards, reasonably good. It would therefore have been highly questionable to choose one solution over others based on this criterion. In fact, a four-factor solution with an excellent fit had to be discarded in favour of the two-factor correlated solution and the two-factor solution, which had a somewhat worse fit but a more interpretable structure. As several authors have argued (e.g., Montoya & Edwards, 2021), this seems to confirm the limited usefulness of fit indices in determining the dimensionality of an instrument. It also highlights the need to pay more attention to the properties of the score estimates derived from any solution. Nevertheless, these estimates must be complemented by other statistics and more substantive arguments (Garrido et al., 2019). At this point, a brief reflection on the observed factor weights is presented.

Before factoring the solutions, commonalities in the range between .30 and .40 had already been observed, which suggested the weights would be low. However, it should be noted that the scale was developed and applied among children who in 25% of the cases were between 8 and 10 years old. Some validation studies have previously shown that factor weights can be low for such young children. For example, Zhao et al. (2024) argue that small changes in age could represent significant differences in the level of development, which makes each item understood and relevant in a different way. This could result in a lower homogeneity of responses, leading to solutions with lower factor weights. Similarly, emotional and cognitive differences may lead to inconsistencies in response patterns, which may also result in factor weights that are lower than those obtained when surveying older children (Mellor & Moore, 2014). In the present study, the level of concreteness/abstraction of each statement may have been interpreted differently according to age. Also, the difficulty in answering the survey using a Likert-type scale may have varied depending on the student's age, or even the management of emotions during the application of the scale may have been more challenging for the youngest children. Likewise, it is also possible that the presence of the examiner and the teacher responsible for the subject in the classroom may have been a contextual variable that positively influenced the consistency of the responses.

In summary, the consistency of the factors with the theory surrounding inclusive teaching practices, the percentages of common variance explained by the two factors found, and their good definition and replicability, as well as the fit, lead us to accept the multidimensionality of the SPI-BE based on two correlated factors. These results confirm that this objective has been achieved.

Regarding the second objective, the good fit and acceptable, well-defined, and replicable factor weights, together with the evidence from ECV and MIREAL, lead us to consider that the SPI-BE scale has an essentially unidimensional structure. Therefore, this objective has been accomplished. It should be noted that multidimensional scales which fulfil the criteria of essential unidimensionality are fully compatible and frequent in the field of education and psychology (Reise et al., 2015).

In terms of the use of the resulting estimated factor scores, it has been found that the two-factor correlated solution and the one-factor solution provide accurate and determined estimated values and allow students to be effectively ordered along the construct. In addition, they can differentiate two levels in the construct and, in the case of the subscale scores, show added value for the total score and thus can be used independently.

Accordingly, it is recommended to use the SPI-BE as follows. The total score may be appropriate when accurate scores are needed to effectively sort students along a range of measurement that is as wide as possible. Subscale scores may be appropriate for assessing and comparing different inclusive teaching practices, although it should be noted that their effective range of measurement is smaller than that of the total score. The researcher should be alert to the presence of inconsistencies in some analyses, especially if group scores are in the upper-middle range of the construct.

Regarding the third objective, this study also examined whether the perception of inclusive teaching practices varied between groups. It was found that female students perceive, during their bilingual lessons, that inclusive teaching practices occur more frequently. This is consistent with the study by Schwab et al. (2018), who showed that female students perceive inclusive environments slightly more positively than males. In terms of the educational level, students in the lower grades perceive more use of inclusive practices. This finding is also in line with that reported by Schwab et al. (2018) who found that students in higher educational levels have a more negative perception about inclusion. Perhaps, increased curricular pressure at higher levels leads teachers to focus more on content than on attention to diversity. Additionally, older students might be more aware of their learning process, and thus, more critical about the teaching practices implemented. With regard to different types of educational needs, non-SEN students consider that inclusive practices are more frequently implemented. This may be the case because the greater needs and/or difficulties of SEN students make them more sensitive to the presence and/or absence of support. Lastly, regarding the type of bilingual programme, students enrolled in a MEC-BC programme perceive inclusive practices as being more frequent. This may be explained in part by the fact that MEC-BC programmes are usually more demanding. The use of English is deemed compulsory, the presence of language assistants is more commonplace, as well as more scaffolding strategies and other learner-centred teaching methods. Thus, the objective related to the criterion validity can be considered fully achieved.

Some limitations of this study should be acknowledged. Although analyses were conducted to test the scale's ability to differentiate groups between which differences were expected, it would have been desirable to have scores using other scales with which to assess the concurrent validity of the SPI-BE scale. Similarly, the obstacle in accessing student profiles has prevented the classification of each student according to their specific learning difficulties. Differences in learning difficulties are understood to generate greater variability regarding educational needs. Thus, obtaining this information would have enabled the design of an instrument that could more accurately assess student diversity in the classroom.

Finally, although considerable attention was devoted to maintain the same level of explicitness across statements and to employ the Likert-type scale format most recommended for children of this age (i.e., 3-point Likert scale based on words reflecting the frequency of behaviours) (Mellor & Moore, 2014), further exploration of alternative response options and their potential impact on the scale's validity may be warranted.

5. Conclusions

This article presents a novel factorial approach for validating SPI-BE scores to assess how primary students perceive inclusive teaching practices in bilingual education programmes. Although bilingual education in Spain has expanded significantly it continues to face challenges concerning inclusion (Pérez-Cañado, 2021; Szczesniak & Muñoz, 2022). The SPI-BE contributes to addressing the need for tools that measure how effectively attention to diversity is attended in bilingual programmes (Subban et al., 2022; Coyle, 2013).

The SPI-BE captured two correlated but interpretable latent factors that represent complementary aspects of inclusive teaching. Beyond demonstrating acceptable model fit, the analyses emphasized the quality of the estimated scores derived from different factorial solutions. In line with current psychometric recommendations (Montoya & Edwards, 2021; Garrido et al., 2019), the study centred on multifaceted analysis, focusing on the quality of scale scores rather than fit indices. This methodological approach addresses the need to prioritize interpretability and precision in latent score estimates when validating educational assessment instruments. Although factorial models with excellent goodness of fit are desirable, they may yield factors of limited substantive meaning (Reise et al., 2015; Marcoulides & Yuan, 2017).

The findings demonstrated that the SPI-BE's total and subscale scores provided valuable insights into students' experiences of inclusion. The reliability and replicability indices suggest that the scale could effectively distinguish between students at different levels of the construct. This confirmed the added value of assessing both general and specific aspects of inclusive teaching.

In practice, the scale allows teachers to identify the strengths and weaknesses of their inclusive teaching practices and allows educational leaders to refine strategies that ensure equitable access to bilingual learning environments. Future research should explore the validity of SPI-BE scores in different multilingual educational contexts, examining how accurately they reflect performance at different stages of education. Finally, the importance of ensuring that measurement instruments yield high-quality, meaningful scores rather than merely acceptable fit statistics should be highlighted. This is essential for promoting evidence-based inclusion in bilingual education.

Funding information

This paper was supported by the Ministry of Innovation and Science of the Spanish Government under Grant PID2020-113956RB-I00.

6. References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (Eds.) (2014). *Standards for Educational and Psychological Testing.* American Educational Research Association.

Durán-Martínez, R., Martín Pastor, M. E., & Martínez Abad, F. (2020). ¿Es inclusiva la enseñanza bilingüe? Análisis de la presencia y apoyos en los alumnos con necesidades específicas de apoyo educativo. *Bordón: Revista de* Pedagogía, *72*(2), 65-82. https://doi.org/10.13042/BORDON.2020.71484

Bauer-Marschallinger, S., Dalton-Puffer, C., Heaney, H., Katzingera, L., & Smith, U. (2023). CLIL for all? An exploratory study of reported pedagogical practices in Austrian secondary schools. *International Journal of Bilingual Education and Bilingualism, 26*(9), 1050-1065. https://doi.org/10.1080/13670050.2021.1996533

Booth, T. & Ainscow, M. (2011). *Index for inclusion. Developing learning and participation in schools* (3rd ed.). CSIE.

Brown, A., & Croudace, T. J. (2015). Scoring and estimating score precision using multidimensional IRT models. In S. P. Reise & D. A. Revicki (Eds.), *Multivariate applications series. Handbook of item response theory modeling: Applications to typical performance assessment* (p. 307-333). Routledge/Taylor & Francis Group.

Coyle, D. (2013). Listening to learners: an investigation into 'successful learning' across CLIL contexts. *International Journal of Bilingual Education and Bilingualism, 16*(3), 244-266. https://doi.org/10.1080/13670050.2013.777384

Coyle, D., & Meyer, O. (2021). *Beyond CLIL: Pluriliteracies teaching for deeper learning.* Cambridge University Press.

Fernández-Archilla, J. A., Álvarez, J. F., Aguilar-Parra, J. M., Trigueros, R., Alonso-López, I. D., and Echeita, G. (2020). Validation of the Index for Inclusion Questionnaire for compulsory secondary education students. *Sustainability, 12*(6), 2169. http://doi.org/10.3390/su12062169

Ferrando, P. J., & Lorenzo-Seva, U. (2013). *Unrestricted item factor analysis and some relations with item response theory*. Technical Report. Department of Psychology, Universitat Rovira i Virgili.

Ferrando, P. J., & Lorenzo-Seva, U. (2017). Program FACTOR at 10: Origins, development and future directions. *Psicothema*, *29*(2), 236–240. https://doi.org/10.7334/psicothema2016.304

Ferrando, P. J., & Lorenzo-Seva, U. (2019). On the added value of multiple factor score estimates in essentially unidimensional models. *Educational and Psychological Measurement*, *79*(2), 249–271. https://doi.org/10.1177/0013164418773851

Garrido, C. C., González, D. N., Seva, U. L., & Piera, P. J. F. (2019). Multidimensional or essentially unidimensional? A multi-faceted factor analytic approach for assessing the dimensionality of tests and items. *Psicothema*, *31*(4), 450–457. https://doi.org/10.7334/psicothema2019.153

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudek, S. H. C. du Toit & D. F. Sorbom (Eds.), *Structural equation modeling: Present and future* (pp. 195-216). Lincolnwood, IL: Scientific Software.

Lasagabaster, D., & Ruiz de Zarobe, Y. (Eds.) (2010). *CLIL in Spain: Implementation, results and teacher training.* Cambridge Scholars Publishing.

Linacre, J. M. (2025). *A user's guide to Winsteps: Rasch-model computer program. Program Manual 5.0.* Winsteps.com

Madrid, D., Parra, E., & Ortega-Martín, J. L. (2018). Evaluación de los programas de AICLE en Andalucía. In J. L. Ortega-Martín, S. Hughes, & D. Madrid (Eds.), *Influencia de la política educativa en la enseñanza bilingüe* (pp. 41–53). Ministerio de Educación, Ciencia y Deporte (MECD).

Marcoulides, K. M., & Yuan, K. H. (2017). New ways to evaluate goodness of fit: a note on using equivalence testing to assess structural equation models. *Structural Equation Modeling*, *24*(1), 148–153. https://doi.org/10.1080/10705511.2016.1225260

Martínez-Hita, F. J., Granero-Gallegos, A., & Gómez-López, M. (2022). Design and validation of a tool to evaluate CLIL in physical education sessions. *Porta Linguarum*, *2022*(37), 193–210. https://doi.org/10.30827/portalin.vi37.17795

Mellor, D., & Moore, K. A. (2014). The use of Likert scales with children. *Journal of Pediatric Psychology*, *39*(3), 369–379. https://doi.org/10.1093/jpepsy/jst079

Ministerio de Educación, Formación Profesional y Deportes (2024a). *Las cifras de la educación en España. Estadísticas e indicadores. Edición 2024.* Secretaría General Técnica

Ministerio de Educación, Formación Profesional y Deportes (2024b). *Enseñanzas no universitarias. Alumnado con necesidad específica de apoyo educativo. Curso 2022-2023.* https://www.educacionfpydeportes.gob.es/servicios-al-ciudadano/estadisticas/no-universitaria/alumnado/apoyo/2022-2023.html

Montoya, A. K., & Edwards, M. C. (2021). The poor fit of model fit for selecting number of factors in exploratory factor analysis for scale evaluation. *Educational and Psychological Measurement*, *81*(3), 413–440. https://doi.org/10.1177/0013164420942899

Pérez-Cañado, M. L. (2021). Inclusion and diversity in bilingual education: A European comparative study. *International Journal of Bilingual Education and Bilingualism, 26*(9), 1129–1145. https://doi.org/10.1080/13670050.2021.2013770

Pladevall-Ballester, E. (2015). Exploring primary school CLIL perceptions in Catalonia: Students', teachers' and parents' opinions and expectations. *International Journal of Bilingual Education and Bilingualism*, *18*(1), 45–59. https://doi.org/10.1080/13670050.2013.874972

Polit, D. F., Beck, C. T., & Owen, S. v. (2007). Focus on research methods: Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing and Health*, *30*(4), 459–467. https://doi.org/10.1002/nur.20199

Reise, S. R., Cook, K. F., & Moore, T. M. (2015). Evaluating the impact of multidimensionality on unidimensional item response theory model parameters. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 13-40). Routledge.

Rodríguez, A., Reise, S. P., & Haviland, M. G. (2016). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment*, *98*(3). https://doi.org/10.1080/00223891.2015.1089249

Roiha, A. (2014). Teachers' views on differentiation in content and language integrated learning (CLIL): Perceptions, practices and challenges. *Language and E ducation , 28*(1), 1-18. https://doi.org/10.1080/09500782.2012.748061

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*(4, Pt. 2), 100.

Schwab, S., Sharma, U., & Loreman, T. (2018). Are we included? Secondary students' perception of inclusion climate in their schools. *Teaching and Teacher Education, 75*, 31-39. https://doi.org/10.1016/j.tate.2018.05.016

Sellbom, M., & Tellegen, A. (2019). Factor analysis in psychological assessment research: Common pitfalls and recommendations. *Psychological Assessment*, *31*(12), 1428–1441. https://

doi.org/10.1037/pas0000623

Sharma, U., Sokal, L., Wang, M., & Loreman, T. (2021). Measuring the use of inclusive practices among pre-service educators: A multi-national study. *Teaching and Teacher Education*, *107*(4), 103506, https://doi.org/10.1016/j.tate.2021.103506

Subban, P., Woodcock, S., Sharma, U., & May, F. (2022). Students' experiences of inclusive education in secondary schools: A systematic review of the literature. *Teaching and Teacher Education, 119*, 103853. http://doi.org/10.1016/j.tate.2022.103853

Szczesniak, A., & Muñoz Luna, R. (2022). Teachers' perceptions of Content and Language Integrated Learning (CLIL) in primary schools in Andalusia. *Porta Linguarum,* 37, 237-257. https://doi.org/10.30827/portalin.vi37.18414

UNESCO (2017). *Education transforms lives*. United Nations Education, Scientific, and Cultural Center. https://unesdoc.unesco.org/ark:/48223/pf0000247234_spa

Zhao, Y., Niu, J., Huang, J., & Meng, Y. (2024). A bifactor representation of the Center for Epidemiological Studies Depression Scale for children: gender and age invariance and implications for adolescents' social and academic adjustment. *Child and Adolescent Psychiatry and Mental Health*, *18*(1), 27. https://doi.org/10.1186/s13034-024-00717-z

6. Appendix

Table A1. *English and Spanish version of the SPI-BE scale*

1. We work in groups or pairs and help each other when we have questions.

   [*Los compañeros trabajamos en grupo o parejas y nos ayudamos cuando tenemos dudas*.]
2. The teacher suggests different tasks so that everyone can participate.

   [*El profesor propone tareas distintas para que todos podamos hacerlas*.]
3. The teacher uses different materials to help us understand him/her (e.g. images, videos, diagrams, etc.)

   [*El profesor utiliza diferentes materiales para ayudarnos a entenderle (ej. imágenes, vídeos, esquemas, objetos, etc.*)]
4. When we do tasks, the teacher gives materials to help us (e.g. pictures, texts, diagrams, etc.)

   [*Cuando hacemos actividades, el profesor nos da materiales para ayudarnos (ej. imágenes, textos, esquemas, etc.*)]
5. When we work in groups or pairs, we change partners.

   [*Cuando trabajamos en grupo o parejas cambiamos de compañeros*.]
6. The teacher uses ICT applications in the classroom (Kahoot, Plickers, ClassDojo, etc.)

   [*El profesor utiliza en el aula aplicaciones informáticas (Kahoot, Plickers, ClassDojo, etc.)*]
7. When assessing, the teacher gives some classmates more time or uses different tests so that everyone can finish them.

   [*En las evaluaciones, el profesor deja más tiempo a algunos compañeros o hace pruebas distintas para que todos podamos acabarlas*.]
8. When assessing, the teacher asks about the topics that have been covered in the classroom.

   [*En las evaluaciones, el profesor nos pregunta sobre los temas que hemos visto en clase*.]
9. The teacher asks us to do different activities to give us a grade (e.g., a presentation, an outline, a test, etc.)

   [*El profesor nos pide hacer varias actividades diferentes para ponernos la nota (ej. una presentación, un esquema, un examen, etc.*)]
10. The teacher encourages me to participate in class.

    [*El profesor me anima a participar en clase*.]
11. If I don't know how to do an activity the teacher helps me (e.g., clarifying how to do it, giving me examples, etc.)

    [*Si no sé hacer una actividad, el profesor me ayuda (ej. aclarándome cómo se hace, poniéndome ejemplos, etc.)*]
12. The teacher helps us to speak English, giving us useful words and phrases.

    [*El profesor nos ayuda a hablar en inglés en clase, dándonos palabras y frases útiles*.]
13. When I don't understand something, the teacher explains it to me again using easier words in English.

    [*Cuando no entiendo algo, el profesor me lo vuelve a explicar en inglés de manera más fácil*.]
14. When I don't know how to say something, the teacher helps me say it in English.

    [*Cuando no sé decir algo, el profesor me ayuda a decirlo en inglés*.]
15. The teacher explains what we have done wrong after doing an activity and how to do it right.

    [*El profesor nos explica lo que hemos hecho mal después de hacer una actividad y cómo hacerlo bien*.]
16. The teacher congratulates us when we do the tasks correctly, when we improve, when we learn new things, etc).

    [*El profesor nos felicita cuando hacemos las cosas bien, cuando mejoramos, cuando aprendemos cosas nuevas, etc.*]

Note: Category responses are: Never, sometimes, always [Nunca, algunas veces, siempre].