



## An evaluation of the Behavior and Instructional Management Scale's psychometric properties using Portuguese teachers



Daniel A. Sass<sup>a,\*</sup>, João Lopes<sup>b</sup>, Célia Oliveira<sup>c</sup>, Nancy K. Martin<sup>a</sup>

<sup>a</sup> University of Texas at San Antonio, USA

<sup>b</sup> University of Minho, Portugal

<sup>c</sup> Lusófona University of Porto, Portugal

### HIGHLIGHTS

- The BIMS provides adequate factorial validity with 1520 Portuguese teachers.
- The BIMS predicts student engagement, instructional strategies, and classroom management.
- Demonstrates the benefits of Bayesian versus Frequentist estimation.

### ARTICLE INFO

#### Article history:

Received 17 March 2015

Received in revised form

12 January 2016

Accepted 22 January 2016

Available online xxx

#### Keywords:

Behavior management

Instructional management

Classroom management

Bayesian factor analysis

Validity

Reliability

### ABSTRACT

Classroom management research is an important topic as teachers cannot effectively educate students in unstructured environments. With that said, few psychometrically sound measures are available to measure behavior and instructional management. Using 1520 Portuguese teachers, we evaluated the Behavior and Instructional Management Scale's (BIMS) psychometric properties using Bayesian estimation and found that the original 12-item scale provided reasonable evidence of factorial validity and internal consistency reliability; however, a slightly revised model may be more promising. The BIMS subscales also had strong concurrent validity evidence based on the associations with perceived student engagement, perceived instructional strategies, and perceived classroom management.

© 2016 Elsevier Ltd. All rights reserved.

### 1. Introduction

Within the literature, classroom management is commonly considered a general term that includes teacher efforts to direct individual student conduct, student interactions, and learning (Burden, 2000; Evertson & Harris, 1999; Evertson & Weistein, 2006; Good & Brophy, 2000; Iverson, 2003). Therefore, teachers' behavior management and instructional styles are a direct outgrowth of these guidance efforts. In fact, they purportedly manifest in the teacher's inclinations and efforts to accomplish appropriate learning objectives, while at the same time managing

student behaviors. As teachers form opinions about how classrooms should be managed, they attempt to operationalize their beliefs by institutionalizing a code for classroom interactions and behaviors (Carter, 1985; Epstein, 1973). Thus, the way teachers manage instruction and behavior creates an overall classroom management approach (Martin & Sass, 2010).

Emmer, Evertson, and Anderson's (1980) classic study documented that one of the primary differences between effective and ineffective classroom managers was the manner in which they formulated and implemented classroom rules. Still, classroom rules are of little assistance if students are not motivated to follow them. As Evertson and Weistein (2006) explain, " ... how a teacher achieves order is as important as whether a teacher achieves order" (p. 4). Establishing an effective reward structure and encouraging student input can be useful tools in the prevention of misbehavior and the maintenance of order in the classroom environment.

The manner in which teachers manage instruction is equally

\* Corresponding author. University of Texas at San Antonio, Department of Management Science and Statistics, One UTSA Circle, BB 4.06.02, San Antonio, TX 78249, USA

E-mail address: [daniel.sass@utsa.edu](mailto:daniel.sass@utsa.edu) (D.A. Sass).

relevant, as it contributes to the general classroom atmosphere and classroom management style (Burden, 1995; Kounin, 1970; McNeely & Mertz, 1990; Reeve & Jang, 2006; Weinstein & Mignano, 1993). Nowhere is this better documented than in Kounin's (1970) classic study of orderly and disorderly classrooms. Concepts such as smoothness and momentum of instruction were consistently found to be characteristics of well-planned lessons that prevented off-task behaviors. Further, McNeely and Mertz's (1990) study revealed that student teachers began their field experience by focusing on quality lesson planning. By the end of their experience, however, they had begun to see students as the "enemy" and shifted the focus of lesson planning from activities designed to encourage learning to those likely to discourage disruption.

While one might expect a teacher's approach to instruction to be congruent with behavior management (Woolfolk Hoy & Weinstein, 2006), relatively little is known about the connection (or lack of it) between these two variables. For example, one might expect to find less controlling instructional approaches (e.g., more open to student feedback related to instruction) coupled with less controlling behavior management expectations (e.g., use of cell phones). However, this certainly may not be the case depending on the teacher's personality, approach to managing student behaviors, and teaching pedagogy. For instance, some teachers may welcome student collaboration and interactive learning in the classroom (i.e., less instructional control), but exhibit a high level of control over student misbehavior (e.g., being off task while working in groups). To better understand this relationship from a quantitative perspective and provide teachers, administrators, and researchers with a tool to assess the classroom atmosphere, Author and Author (2010) developed the Behavior and Instructional Management Scale (BIMS).

While no published research has utilized the BIMS to understand the classroom environment to date, initial research by Author and Author (2010) found that the Behavior Management (BM) and Instructional Management (IM) scales are only moderately correlated ( $r = 0.22$ ). This finding suggests that how teachers manage classroom instruction and behavior are relatively unrelated (i.e., teachers cannot be considered controlling or non-controlling, but instead their level of control depends on the classroom activity) and suggests these constructs should be evaluated separately. With that said, the BIMS is a relatively new tool/instrument that requires additional psychometric research (e.g., factor analyses, reliability analyses) and understanding of how these scales relate to and predict other classroom variables. While the latter is certainly important, the current study is focused more on the former given the progression in research (i.e., a scale should have a good internal structure before being used for predictive purposes).

### 1.1. Theoretical framework

Teacher actions are purportedly preceded by their belief systems regarding child development. These perspectives translate into teacher expectations for both learning and behavior. Whether they are consciously aware of it or not, these beliefs are arguably quite powerful as they frame teachers' interactions with students. As a result, Glickman and Tamashiro (1980) conceptualized a continuum of control that provides the theoretical underpinnings of the BIMS. Non-interventionists (least controlling) anchor one end of the continuum, with interventionists (most controlling) anchoring the other. Midway between these extremes, interactionists focus on teacher-student relationships and draw strategies from both non-interventionist and interventionist perspectives (see Martin & Sass, 2010).

Most theoretical models of classroom management and

instruction reflect the authors' beliefs about the range of teacher control in classrooms (Lopes & Santos, 2013). Gordon (2003) and Rogers and Freiberg (1994) models, for instance, advocate much less teacher control than the Canter and Canter's model (2001), which stresses the value of assertive discipline. Glasser's (1998) model, on the other hand, represents an intermediate position. Some authors (e.g., Chall, 2000; Ravitch, 2000) suggest that these models fall somewhere in a continuum between a behaviorist and a constructivist perspective of classroom management.

Non-interventionist (student-centered) and the interventionist (teacher-centered) classrooms vary a great deal regarding the role of the teacher, the role of the students, classroom rules and procedures, rewards and punishments, teacher's expectations, students' autonomy, instruction, time spent on management, etc. For instance, in a non-interventionist classroom the expected role of the teacher is to act as a facilitator. It is also expected that rules and procedures are negotiated between the teacher and the students. Therefore, the instruction is shared with the students, there is frequent use of group work, and the time spent on management is valued (Clarke-Stewart, Lee, Allhusen, Kim, & McDowell, 2006; Terhart, 2003).

The interventionist classroom, on the other hand, is very different. Here, the teacher leads the class, with classroom rules and procedures set by the teacher (without negotiation). Moreover, the teacher is responsible for instruction with a great deal of recitation and individual seatwork and time spent on management is perceived as a waste. Instead, academic performance and classroom discipline are most valued (Olafson & Schraw, 2006).

These two different perspectives about classroom management seem to more generally stem from different views of mankind's nature (Ertesvåg, 2011). Specifically, interventionist teachers seem to hold more pessimistic beliefs about mankind than non-interventionist teachers. Unlike non-interventionist teachers, interventionist teachers do not tend to value students' autonomy and do not expect students to self-regulate. The structure of student-teacher relationships and the expected behaviors of both parties are therefore usually quite different for teachers holding these different beliefs about classroom management. Still, two things must be stressed: (1) most teachers fall somewhere in a continuum between these extreme positions and only a few seem to show extreme beliefs; and (2) it is not unusual to find teachers holding incongruent beliefs about classroom management and order. The TALIS study (OECD, 2009), for instance, found that there are countries (e.g. Italy, Bulgaria, Malaysia, Spain, Portugal) where most teachers hold both constructivist (non-interventionist) beliefs about classroom order and more controlling, teacher-focused beliefs about curriculum contents. This is an important finding since the congruence of beliefs and behaviors are considered vital conditions for classroom order (Oppenedaker & Van Damme, 2006).

### 1.2. Nature and quality of existing measures

Classroom management research has been hindered by its intricacy, as well as the nature and quality of instruments available to measure it. Although both the Pupil Control Ideology (PCI, Willower, Eidell, & Hoy, 1973) and Beliefs on Discipline (BDI, Wolfgang & Glickman, 1980) have enjoyed fairly widespread use, neither scale addresses the broader aspect of instructional management. The Attitudes and Beliefs on Classroom Control Inventory (ABCC, Martin, Yin, & Baldwin, 1998) was drawn from the BDI and addressed both discipline and instructional management. However, both the ABCC and its revised version (ABCC-R, Martin, Yin, & Mayall, 2007) were not without psychometric concerns.

Initial psychometric research on the BIMS has been promising (Martin & Sass, 2010; Sass, 2011) and suggests it could be a useful

tool for research and practical purposes. Nevertheless, further validation research is essential to strengthen the BIMS' psychometric foundation and determine if the psychometric properties remain intact after being translated and administered in different cultures/samples.

Since its original publication in 2010, the BIMS has been used in at least four published studies (Martin et al., 2012; Sass, 2011; Jalali, Panahzade, & Firouzmand, 2014; Unal & Unal, 2012). Unal and Unal (2012) translated the scale into Turkish to investigate differences between novice elementary level teachers and their more experienced counterparts. Two independent professors reviewed the Turkish version of the BIMS to confirm “semantic equivalence across languages, conceptual equivalence across cultures, and normative equivalence” to the original BIMS (p. 46). Although validity evidence was not examined, they did report adequate internal consistency reliability for the IM ( $\alpha = 0.75$ ) and BM ( $\alpha = 0.72$ ) scales.

The BIMS has also been used to investigate the connection between Iranian teachers' management orientations and their attitudes toward computer use in the classroom (Jalali et al., 2014). While they did not evaluate the BIMS' factorial validity, they did report a reasonably acceptable Cronbach's alpha ( $\alpha = 0.67$ ). Unfortunately, the authors evaluated the BIMS' reliability as a single construct (i.e., rather than the IM and BM subscales separately) and used all 24-items (rather than the recommended 12-item version). Author and Author (2010) indicated the full 24-item BIMS was never intended for use, as the 24-items version was only used to determine the best set of 12 items to represent behavior and instructional management. Hence, it is unknown whether the 6-item IM and BM subscales were psychometrically sound with Jalali et al.'s (2014) Iranian sample.

Using the sample from the Author and Author (2010) study, Sass (2011) tested whether the BIMS was invariant across grade level and Schmitt and Sass (2011) investigated whether a structural model using the IM subscale was invariant across gender and grade level. Collectively, results revealed that while the IM scale was invariant across grade levels, the BM scale was only partially invariant (Sass, 2011). Related to concurrent validity, Schmitt and Sass (2011) found that the IM scale was a strong predictor of teacher's student behavior stressors and feelings of personal accomplishment. Schmitt and Sass (2011) also found that the prediction of IM on student behavior stressors was moderated by gender, with this prediction being stronger for males than females.

Given the BIMS' international use and the need for additional psychometric research, the purpose of this paper is fourfold. First, we explored the original 24-item BIMS and the proposed 12-item BIMS via exploratory factor analysis (EFA) with a sample of Portuguese teachers (i.e., we replicated the Martin & Sass, 2010, study). Second, confirmatory factor analyses (CFA) were conducted on the proposed 12-item BIMS and an alternative 12-item factor structure based on the EFA results. During this phase, we also assessed the internal consistency reliability of the IM and BM scales. Third, we used other available data (i.e., perceived student engagement, instructional strategies, and classroom management) to explore concurrent validity. Fourth, we demonstrated the benefits of Bayesian CFA (BCFA) compared to more traditional/Frequentist CFA approaches. Collectively, these analyses should provide a better understanding of the BIMS' psychometric properties and generalizability.

## 2. Methods

### 2.1. Participants

Our sample of 1520 Portuguese volunteer teachers came from

approximately 800 public ( $n = 1276$ ) and private ( $n = 244$ ) schools in the eight educational administrative regions of the country: north coast (46.5%), north interior (5.1%), center (9.6%), Lisbon and Tejo Valley (23.3%), Alentejo (1.3%), Algarve (6.0%), and Atlantic Islands (8.2%). While the sample was not randomly selected, our sampled percentages appear to mirror the actual regional teacher population (north coast, 43%; north interior, 4%; center, 10%; Lisbon and Tejo Valley, 25%; Alentejo, 2%; Algarve, 6%; Atlantic Islands, 10%).

Most teachers were women (76.4%) working in public schools (82.5%), with a relatively equal distribution across the elementary (21%, Primary or grades 1 through 4), 2nd cycle (27%, grades 5 & 6), and 3rd cycle (53%, grades 7, 8, & 9) levels. Note that the 2nd and 3rd cycle Portuguese schools roughly correspond to junior high schools in the United States. Like the United States, Portugal primary school teachers hold a single teacher format, whereas the 2nd and 3rd cycle levels hold a multi-disciplinary format taught by several teachers. Therefore, 2nd and 3rd cycle teachers teach specific disciplines, such as Portuguese Language, Sciences, Physical Education, History, or Mathematics. In terms of class size, approximately 4% of teachers had an average class size below 15 students, 12% between 15 and 20 students, 40% between 21 and 25 students, 42% between 26 and 30 students, and 2% had more than 30 students. The teachers with more than 30 students were from private schools, as Portuguese law prohibits more than 30 students in public school classrooms.

From a teacher education standpoint, approximately 10% held a Bachelor's degree, 82% had a Licensure degree, 7% earned a Master's degree, and 1% obtained a Ph.D. Those who held a Bachelor's degree were most likely older and received their training at a Polytechnic Institute, not at a university. Those who held a Licensure, Master's, or Ph.D. degree received their training in a university program. Approximately 85% of participants taught in regular classrooms, with 15% having a mixed classroom (e.g., vocational classrooms, special education classrooms). Approximately half of the teachers (52%) had more than 20 years of experience, with the remainder having the following breakdown in teaching experience: less than four years (2%), 4–9 years (7%), and 10–20 years (38%). It should be stressed that there is very low turnover in the teaching profession in Portugal. Therefore, teachers at any school level, including the preschool level, tend to have considerable professional teaching experience.<sup>1</sup> Across their teaching experience, about 60% received no specific training in classroom management or disciplinary practices, 12% received 1–5 h of training, 8% received 6–15 h of training, 10% received 16–25 h of training, and another 10% received more than 25 h of training.

### 2.2. Procedures

From the aforementioned eight educational administrative regions, teacher data were collected online via SurveyMonkey<sup>®</sup>. To encourage participation, an email with the link and instructions was sent to 700 public and 120 private school principals asking them to forward the invitation to their teachers. Using the institutional review board (IRB) approved protocol, data were collected within a three month time frame and two reminder emails were sent, each one at the end of the first and second month of data

<sup>1</sup> According to the European Commission (2010), about 13% of the Portuguese primary teachers are less than 30 years old (which eventually means 1–8 years of experience), 29% are between 30 and 39 years old (8–18 years of experience), 29% are between 40 and 49 years old (19–28 years of experience), and 28% are more than 50 years old (more than 28 years of experience). For secondary education estimates are 11%, 36%, 32%, 21%. These estimates do not seem to be far from the years of experience of our own sample.

collection. Unfortunately, there was no method to determine the number of teachers who received the email and, therefore, the response rate is unknown. This research study received university IRB approval and participation was voluntary.

After collecting the sample of 1520, it was randomly divided into two samples of 760 (see Table 1) to assess internal validity/cross-validation (Osborne & Fitzpatrick, 2012). In other words, the first sample was used for the EFAs and the second sample for CFAs. While these subsamples were used for cross-validation via factor analyses, concurrent validity analyses used the full sample to maximize information.

### 2.3. Measures

#### 2.3.1. Behavior and Instructional Management Scale (BIMS, Martin & Sass, 2010)

Validated using a sample of United States teachers, the BIMS is composed of two subscales: BM and IM. While 24-items (see Appendix) were used for scale development, only 12-items (six per subscale) were validated and intended for research purposes. BM is similar to, but is distinguished from discipline in that it includes planned efforts to prevent misbehavior as well as the teacher's response to it. Specifically, BM was operationally defined as establishing rules, forming a reward structure, and presenting

opportunities for student input. Conversely, IM measures aspects related to monitoring seatwork, structuring daily routines, and teacher's use of teacher-centered versus interactive instruction, and participatory approaches to instruction. Each item was rated on a 6-point scale from 'strongly agree' to 'strongly disagree.' The continuum of control proposed by Glickman and Tamashiro (1980) and Wolfgang (1995) underlies both BM and IM. Thus, higher scores indicate a more controlling (interventionist) approach to instruction and behavior.

#### 2.3.2. Ohio state teacher efficacy scale: student engagement (OSTES, Tschannen-Moran & Woolfolk Hoy, 2001)

Tschannen-Moran and Woolfolk Hoy (2001) created the teacher efficacy scale to include three dimensions (perceived Instructional Strategies, perceived Classroom Management, and perceived Student Engagement) and was included to evaluate the BIMS' concurrent validity. Perceived Instructional Strategies is operationally defined as the degree to which teachers alter their teaching style, extent to which they can increase student comprehension, and foster good questions for students. A few example questions include "To what extent can you use a variety of assessment strategies?" and "To what extent can you provide an alternative explanation or example when students are confused?" Perceived Classroom Management is operationally defined as a teacher's perceived ability to control the classroom and disruptive students. Example questions include "How much can you do to get children to follow classroom rules?" and "How well can you keep a few problem students from ruining an entire lesson?" Perceived Student Engagement is operationally defined as a teacher's perceived ability to get students engaged in school and help students perform well in the classroom. Example items include "How much can you do to get students to believe they can do well in schoolwork?" and "How much can you do to improve the understanding of a student who is failing?"

These data were collected to assess concurrent validity evidence for the BIMS and replicate the findings of previous research (Schmitt & Sass, 2011). To reduce the overall survey length, we used the short version of the OSTES, which has four items per scale and uses a 9-point scale from 1 (nothing) to 9 (a great deal). Note, due to researcher error, one item (item 10) was accidentally omitted. Based on reliability theory (i.e., all items are highly correlated and measure the same factor), it is purported that the omission of a single item should have minimal impact on the mean score or conclusions. However, this is an empirical question that was unable to be addressed with our data. It is important to note the correlation between these scales was rather high (correlations ranged from 0.66 to 0.70).

In addition to the research conducted by Tschannen-Moran and Woolfolk Hoy (2001), Klassen et al. (2009) provided evidence of construct validity and measurement invariance across grade levels within culturally similar regions of North America (Canada & U.S.) and East Asia (Korea & Singapore). Despite their research, our EFA and CFA analyses<sup>2</sup> indicated that a one-factor solution might be preferable, which may be expected given the high interfactor correlations. Nevertheless, we used the proposed scales as intended and obtained acceptable internal consistency reliability coefficients for the perceived Instructional Strategies ( $\alpha = 0.79$ ), perceived Classroom Management ( $\alpha = 0.87$ ), and perceived Student Engagement ( $\alpha = 0.85$ ) subscales.

**Table 1**  
Descriptive characteristics for participants in each sample of 760.

	EFA data	CFA data
Gender		
Female	594	567
Male	166	193
Percent of time spent on classroom behavior management		
Less than 10%	158	179
10%–20%	237	223
20%–30%	182	192
30%–40%	85	82
40%–50%	52	37
50%–60%	22	28
60%–70%	9	6
70%–80%	11	7
80%–90%	4	6
On average, how often do you report classroom disruptive behaviors to others?		
Rarely	158	154
Only a few times	179	188
Sometimes	233	208
Frequently	135	144
Did you ever get any specific training to deal with classroom misbehavior?		
No	275	289
1–5 h	59	51
6–15 h	37	34
16–25 h	45	52
Greater than 25 h	46	55
Estimated number of students in your classroom.		
Less than 10	6	4
10 to 15	11	18
16 to 20	68	63
21 to 25	208	229
26 to 40	224	231
31 to 35	6	6
Greater than 35	2	0
As far as you can see, in the last five years classroom misbehavior ...		
Significantly decreased	3	2
Decreased	6	14
Has not changed	53	57
Increased	200	201
Significantly increased	203	215

Note. Numbers within each sample not summing to 760 are a result of missing data. Notice, the last three items have significantly more missing data, which may have resulted from being the last questions on the survey or simply difficult questions to answer.

<sup>2</sup> Please contact the corresponding author for factor analysis and dimensionality analyses.



### 3. Statistical analyses

#### 3.1. Frequentist vs. Bayesian

It is important to remember that Frequentist and Bayesian analyses are not different statistical models per se, but instead simply different estimation procedures that make different data and model assumptions. Unlike Frequentist approaches, Bayesian statistics do not rely on large sample normal theory, can better accommodate non-normal parameter estimate distributions, allow researchers to incorporate prior knowledge (i.e., prior distributions), and do not require strict model assumptions. More specifically, Frequentist CFA approaches often apply unnecessarily strict model assumptions (e.g., all cross-loadings, residual covariances, and other non-theorized coefficients are fixed at zero), thus resulting in unrealistic model assumptions, less than desired model fit, and at times considerable model modifications are required. To remedy these limitations, Bayesian estimation combines prior distributions for each parameter and the data likelihood to generate a posterior distribution for each parameter of interest. Perhaps most importantly, Bayesian analyses allow researchers to incorporate prior information or background information (e.g., previous factor loadings and interfactor correlations) into the model rather than continuously testing the traditional null hypotheses. Moreover, it allows researchers to test models that would otherwise be under-identified using Frequentist estimation methods (e.g., estimating all the cross-loadings and residual covariances). Given that a full exposition is beyond the scope of this paper, interested readers are encouraged to read the work of Asparouhov, Muthén, and Morin (2015), Muthén and Asparouhov (2012), van de Schoot et al. (2014), and Zychur and Oswald (2013).

#### 3.2. Model estimation

Exploratory and confirmatory factor analyses were conducted in Mplus 7.3 (Muthén & Muthén, 1998–2015) on a polychoric correlation matrix. While EFA and CFA analyses were conducted using Frequentist (i.e., Weighted Least-Squares with Mean and Variance adjusted, WLSMV) estimation methods, CFA were also tested using Bayesian estimation for comparison and demonstration purposes (see above for more details). CFA with WLSMV estimation was employed to provide comparable results (e.g., model fit statistics) with previous BIMS research (Martin & Sass, 2010; Sass, 2011) and because it has been shown to perform well with ordered categorical data (Curran, West, & Finch, 1996). Conversely, Bayesian CFA (BCFA) was used to increase model flexibility (i.e., allow cross-loadings and residual covariances), evaluate the stability/performance of Frequentist CFAs, and provide a better understanding of the BIMS' factor structure. Unlike the factor analyses, parallel analyses used maximum likelihood robust (MLR) estimation with a Pearson correlation given that other procedures for ordered categorical data are not provided in Mplus due to poor statistical performance.<sup>3</sup>

##### 3.2.1. Frequentist model estimation

EFA was used to replicate the 24- and 12-item BIMS factor structure found by Author and Author (2010) using WLSMV with an oblique Geomin rotation. For the reasons stated above, Bayesian analyses were the primary focus of the CFA analyses. However, we also estimated the CFA models using WLSMV for comparison and

demonstration purposes. For CFA with WLSMV estimation, all cross-loadings were fixed at zero, the residual variances were assumed to be uncorrelated (i.e., fixed at zero), and the factor variances were fixed at one for model identification.

##### 3.2.2. Bayesian model estimation

Like the CFA with WLSMV estimation, the factor variances were set to one for the purposes of latent variable scaling with BCFA. Setting the factor variances at one also made the prior distributions more intuitive and produced comparable unstandardized and standardized results. The posterior distributions were estimated using four MCMC chains with 30,000 iterations and the Gibbs sampler using the random walks algorithm (Muthén & Asparouhov, 2012). The first half of the chains was discarded as a burn-in phase, with the second half used to estimate the posterior distributions.

Normal priors were used for the cross-loadings [ $N(0,0.02)$ ], thus the 95% CI was  $\pm 0.28$ ] and residual covariances [ $N(0,0.01)$ ], thus the 95% CI was  $\pm 0.20$ ], with program defaults used for the rest of the parameter estimates. Slightly larger prior variances were specified for the cross-loadings to allow greater flexibility in model estimation and to permit the data to play a larger role in estimating the cross-loadings. Note, cross-loadings, residual covariances, and other model prior distributions were rather non-informative given the lack of previous BIMS measurement research with Portuguese teachers. Nevertheless, to assess prior distribution sensitivity, alternative priors were evaluated and did not significantly impact the results.

BCFA models were tested in the following order to understand model misfit. Model 1: restrictive BCFA model (i.e., cross-loadings and residual covariances fixed at zero, which mirrored the WLSMV model estimation), Model 2: BCFA model with the above priors placed on the cross-loadings, but the residual covariances remained fixed at zero, and Model 3: BCFA model with priors placed on the cross-loadings and residual covariances. As can be seen here, the BCFA models continued to become less restrictive and more complex (i.e., more parameters were estimated, rather than being fixed at zero).

#### 3.3. Model fit

##### 3.3.1. WLSMV estimation

Model fit was evaluated using the robust  $\chi^2$ , Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), and Root Mean Square Residual (RMSEA). It is well known that  $\chi^2$  statistics are frequently statistically significant for complex models with large sample sizes, thus more emphasis was placed on the approximate model fit statistics (i.e., CFI, TLI, & RMSEA). Generally speaking, CFI and TLI statistics greater than 0.90 are considered as an “adequate,” whereas values greater than 0.95 are deemed as a “good” (Hu & Bentler, 1999). RMSEA values less than 0.10 and 0.06 are considered “mediocre” and “good,” respectively. For a more detailed description of these statistics please see Hu and Bentler (1999).

##### 3.3.2. Bayesian estimation

Unlike traditional Frequentist model fit statistics ( $\chi^2$ , CFI, TLI, & RMSEA), Bayesian estimation assesses the predictive accuracy of the model and is referred to as *posterior predictive checking* (Gelman, Carlin, Stern, & Rubin, 2004). The premise of this procedure is to evaluate the degree of discrepancy between data generated by the model and the actual data itself, with any difference between these results suggests possible model misspecification. Specifically, Bayesian model fit is assessed using the posterior predictive p-value (PPP) and the associated 95% credibility interval (CI) around the model fit statistic. The fit statistic,  $f$ , is calculated using the likelihood-ratio  $\chi^2$  of a  $H_0$  model compared to

<sup>3</sup> Please see <http://www.statmodel.com/discussion/messages/8/11966.html?1393952682> for more details.

the unrestricted  $H_1$  model. To calculate the 95% CI, Mplus takes the difference in the  $f$  statistic for the data generated by the model and the actual data over each of the 10th iterations (Asparouhov & Muthén, 2010b). Therefore, the PPP measures the proportion of  $\chi^2$  values obtained in the simulated data that exceed that of the actual data. A well-fitting model is expected to be non-significant and ideally around 0.50. However, Asparouhov and Muthén (2010a) indicated values greater than 0.10, 0.05, or 0.01 are acceptable for most applications.

### 3.4. Model modification

Another difference between Frequentist and Bayesian estimation is the use (or lack thereof) of modification indices to determine model misfit. For Frequentists using CFA, modification indices are used to determine those paths (e.g., factor-loadings) that differ significantly from the fixed value (often set at zero) and would improve the model fit if estimated freely. This approach is often considered problematic in that multiple model changes can produce the same change in model fit and this process can result in elevated Type I error rates (MacCallum, Roznowski, & Necowitz, 1992). Rather than using these modification indices, Bayesian estimation allows all parameters to be freed and estimated simultaneously, thus eliminating the need for modification indices. Consequently, one could argue that Bayesian models allow a simpler procedure for testing the significance of cross-loadings, residual covariances, and other coefficients traditionally fixed because all these parameters are estimated simultaneously. Estimates that deviate from theory are then easily seen in the model results and tested for statistical significance.

### 3.5. Sample sizes

From the full Portuguese sample of 1520, a random sample of 760 teachers without replacement was selected for exploratory and confirmatory analyses. MacCallum, Widaman, Zhang, and Hong (1999) found that the sample size requirements for adequate factor analysis results depend on the data characteristics (i.e., communalities), thus general rule of thumb sample size recommendations should be used with caution. Based on the factor analytic results found here and in previous publications using the BIMS, sample sizes of 760 appear more than sufficient to provide stable results. Our sample sizes are also acceptable based on Comrey and Lee (1992) guidelines: 100 = poor, 200 = fair, 300 = good, 500 = very good, and 1000 or more = excellent. It is worth noting that the Portuguese data did not have any missing values.

## 4. Results

### 4.1. Exploratory factor analysis

The 24-item EFA results indicated a five factor solution based on the eigenvalues greater than one rule (eigenvalues were 8.33, 2.60, 1.46, 1.23, & 1.03), arguably two-factors based on the scree plot, and three factors based on the parallel analysis. These results suggest that the 24-item scale does not offer good evidence for a two factor solution, which replicates Author and Author (2010) findings and provides further evidence that the 24-item scale should not be used for research purposes. Moreover, the 24-item factor structure with two factors did not produce a clean and theoretically meaningful factor structure (see Table 2, labeled 24-item WLSMV EFA).

Using the proposed 12-item from Author and Author (2010),

the scree plot and parallel analysis provided evidence of a two-factor solution, with the eigenvalues greater than one rule suggesting a very weak third factor (eigenvalues were 4.69, 1.93, & 1.05). Given that past research suggests the parallel analysis procedure performs best (for a review see Schmitt, 2011), we elected to retain the two factor solution (see Table 2, labeled 12-item WLSMV EFA) that also provided an adequate to good model fit,  $\chi^2(43) = 280.67, p < .001, CFI = 0.95, TLI = 0.92, RMSEA = 0.05$ . These model fit results offer substantiation for a two-factor solution with adequate factorial validity. The only concern was the slightly lower TLI statistic, along with the few slightly elevated cross-loadings (i.e.,  $\lambda \approx |0.15|$ ). Nevertheless, the item level convergent and discriminant validity appear positive (i.e., large factor loadings and relatively small cross-loadings, respectively) and the interfactor correlation provided support for construct validity.

### 4.2. Confirmatory factor analyses

#### 4.2.1. Bayesian estimation

To replicate the Author and Author (2010) factor structure and assess the 12-item BIMS' psychometric properties, CFA were conducted using Bayesian (to allow a more flexible modeling approach) and WLSMV (a more traditional CFA model) estimation. To better understand the factor structure, the three aforementioned BCFA models were tested. Analyses revealed a poor model fit for BCFA Model 1 (95% CI for  $\Delta\chi^2$  is 116.89–209.31, PPP < .001) and BCFA Model 2 (95% CI for  $\Delta\chi^2$  is 27.48–113.34, PPP < .001), which implies the cross-loadings need to be estimated and perhaps the residual covariances. After correlating the residual variances, BCFA Model 3 produced a good model fit, 95% CI for  $\Delta\chi^2$  is –35.74 to 39.22, PPP = .461. For presentation purposes, the standardized factor loadings are displayed in Table 2 (labeled 12-item BCFA), with the significant cross-loadings and residual covariances discussed below.

Although rather small, only BM2 and IM9 had statistically significant cross-loadings after adjusting for Type I error using a Bonferroni adjustment ( $\alpha = .05/24 = .002$ ). Although six residual covariances were statistically significant at 0.05, none were statistically significant after the Bonferroni adjustment ( $\alpha = .05/66 \approx .001$ ) and all the correlations were relatively small (i.e.,  $> |0.15|$ ). Therefore, despite the need to estimate the cross-loadings and residual covariances to for a non-significant  $\chi^2$  (i.e., a PPP < .05), these estimates tended to be rather small in magnitude.

#### 4.2.2. WLSMV estimation

It is worth noting that a traditional CFA (i.e., cross-loadings and residual covariances fixed at zero) produced a mediocre model fit,  $\chi^2(53) = 563.03, p < .001, CFI = 0.91, TLI = 0.89, RMSEA = 0.11$ . The modification indices indicated that BM2 cross-loaded on the IM factor and IM9 loading on the BM factor (see Table 2, labeled 12-item modified WLSMV CFA). Using the DIFFTEST procedure within Mplus to compare nested models, the results revealed a statistically significant improvement in model fit as a consequence of estimating these cross-loadings,  $\Delta\chi^2(2) = 141.52, p < .001$ , and an overall good model fit for the revised model,  $\chi^2(51) = 296.55, p < .001, CFI = 0.96, TLI = 0.94, RMSEA = 0.08$ .

#### 4.2.3. Revised CFA

Due to the less than ideal model fit with the unmodified WLSMV CFA model (i.e., what most researchers would use in practice) and concerns related to common method variance (CMV, see the Common Method Variance section below), we elected to replace BM2 with BM7 and IM9 with IM7 based on the 24-item EFA model. This model change was justified given that BM2 and IM9 had larger

**Table 2**  
Standardized factor loadings and interfactor correlations (*r*) for the EFA, BCFA, and WLSMV CFA models.

24-item WLSMV EFA			12-item EFA WLSMV EFA			12-item BCFA		12-item Modified WLSMV CFA			Revised 12-item WLSMV CFA			
BM	IM		BM	IM		BM	IM	BM	IM		BM	IM		
BM1*	<b>0.47</b>	0.12	BM1	<b>0.71</b>	−0.01	BM1	<b>0.53</b>	0.06	BM1	<b>0.60</b>	0.00	BM1	<b>0.72</b>	0.00
BM2*	<b>0.79</b>	−0.20	BM2	<b>0.60</b>	0.09	BM2	<b>0.72</b>	−0.17	BM2	<b>0.84</b>	−0.36	BM7	<b>0.72</b>	0.00
BM3	0.05	<b>0.32</b>	BM8	<b>0.54</b>	−0.17	BM8	<b>0.70</b>	0.09	BM8	<b>0.76</b>	0.00	BM8	<b>0.50</b>	0.00
BM4	<b>0.37</b>	0.06	BM9	<b>0.62</b>	−0.01	BM9	<b>0.84</b>	0.02	BM9	<b>0.83</b>	0.00	BM9	<b>0.61</b>	0.00
BM5	0.01	<b>0.58</b>	BM11	<b>0.51</b>	0.22	BM11	<b>0.88</b>	−0.01	BM11	<b>0.85</b>	0.00	BM11	<b>0.70</b>	0.00
BM6	−0.39	0.24	BM12	<b>0.73</b>	0.02	BM12	<b>0.70</b>	0.03	BM12	<b>0.69</b>	0.00	BM12	<b>0.78</b>	0.00
BM7	<b>0.72</b>	0.10	IM2	0.18	<b>0.40</b>	IM2	0.01	<b>0.72</b>	IM2	0.00	<b>0.72</b>	IM2	0.00	<b>0.58</b>
BM8*	<b>0.67</b>	0.15	IM3	−0.17	<b>0.75</b>	IM3	0.03	<b>0.71</b>	IM3	0.00	<b>0.73</b>	IM3	0.00	<b>0.81</b>
BM9*	<b>0.83</b>	−0.01	IM5	0.16	<b>0.63</b>	IM5	−0.12	<b>0.60</b>	IM5	0.00	<b>0.50</b>	IM5	0.00	<b>0.77</b>
BM10	<b>0.65</b>	0.19	IM6	0.02	<b>0.84</b>	IM6	0.03	<b>0.57</b>	IM6	0.00	<b>0.62</b>	IM6	0.00	<b>0.82</b>
BM11*	<b>0.87</b>	−0.04	IM9	0.00	<b>0.84</b>	IM9	0.19	<b>0.52</b>	IM9	<b>0.31</b>	<b>0.42</b>	IM7	0.00	<b>0.85</b>
BM12*	<b>0.71</b>	−0.10	IM12	−0.07	<b>0.73</b>	IM12	−0.04	<b>0.81</b>	IM12	0.00	<b>0.77</b>	IM12	0.00	<b>0.68</b>
IM1	<b>0.61</b>	0.09												
IM2*	0.01	<b>0.69</b>												
IM3*	0.07	<b>0.61</b>												
IM4	<b>0.50</b>	0.11												
IM5*	−0.15	<b>0.55</b>												
IM6*	−0.04	<b>0.67</b>												
IM7	−0.01	<b>0.64</b>												
IM8	<b>0.64</b>	0.23												
IM9*	0.24	<b>0.53</b>												
IM10	<b>0.59</b>	0.08												
IM11	<b>0.46</b>	−0.02												
IM12*	0.01	<b>0.71</b>												
R	<b>0.48</b>			<b>0.46</b>			<b>0.48</b>			<b>0.52</b>			<b>0.52</b>	

Note. Items from the 24-item EFA model marked with an \* correspond to those items from the original 12-item BIMS version.

cross-loadings (thus hurting discriminant validity), whereas BM7 and IM7 had large loadings on the primary/theorized factor (i.e., good convergent validity) and small cross-loadings (good discriminant validity) based on the EFA analyses (see Table 2, labeled *24-item WLSMV EFA*).

This revised CFA model (see Table 2, labeled *Revised 12-item WLSMV CFA*) without any estimated cross-loadings or residual covariances produced a good model fit,  $\chi^2(53) = 278.59$ ,  $p < .001$ , CFI = 0.96, TLI = 0.96, RMSEA = 0.08. Therefore, this factor structure (or set of items) appears better suited with Portuguese teachers. In fact, the three revised BCFA models also produced better model fit statistics: BCFA Model 1 (95% CI for  $\Delta\chi^2$  is 33.20–117.45, with PPP < .001), BCFA Model 2 (95% CI for  $\Delta\chi^2$  is 7.71–89.13, with PPP = .100), and BCFA Model 3 (95% CI for  $\Delta\chi^2$  is −38.41–35.60, with PPP = .527). Collectively, these results imply that while small in magnitude, two cross-loadings (IM5 = −0.13 & IM6 = 0.11) still need to be estimated, whereas the residual covariances could remain fixed at zero. The correlation between IM and BM was 0.49 for this BCFA model, thus suggesting the correlation between factors changed very little due to substituting these two items. Due to the relatively clean factor structure and similarity with the WLSMV results, these results are not presented here, but instead available from the corresponding author.

#### 4.2.4. Common method variance

While not discussed in significant detail to conserve space, we also tested both 12-item CFA models with and without adjusting for CMV using the *unmeasured latent method factor* method (Podsakoff, MacKenzie, Lee, & Podsakoff, 2012). After modeling the single unmeasured latent method factor, our analyses revealed that CMV was a concern for the IM factor using the original 12-item scale, whereas CMV was not an issue for the revised 12-item version. This provides additional justification for using the revised 12-item model with Portuguese teachers. A full report of the CMV results is available from the corresponding author.

#### 4.2.5. Internal consistency reliability analyses

To complement the exploratory and confirmatory analyses, we estimated the internal consistency coefficients for each scale and sample separately. For each IM and BM scale, internal consistency reliability was estimated using the traditional Cronbach's alpha ( $\alpha$ ) and the omega ( $\omega$ ) method that is often more statistically appropriate given that the assumption of tau equivalence (i.e., equal factor loadings) and uncorrelated residual variances are not met (Yang & Green, 2011).

Using Bayesian estimation to compute  $\omega$ , the original BM subscale revealed adequate internal consistency using the EFA ( $\omega = 0.81$ , 95% CI [0.79 to 0.83];  $\alpha = 0.80$ ) and CFA ( $\omega = 0.83$ , 95% CI [0.81 to 0.85];  $\alpha = 0.83$ ) sample data. Results for the original IM subscale also showed adequate internal consistency reliability using the EFA ( $\omega = 0.75$ , 95% CI [0.72 to 0.77];  $\alpha = 0.74$ ) and CFA ( $\omega = 0.78$ , 95% CI [0.76 to 0.80];  $\alpha = 0.78$ ) sample data. For the revised CFA model using Sample 2 data, adequate internal consistency reliability estimates were obtained for the BM ( $\omega = 0.85$ , 95% CI [0.83 to 0.87];  $\alpha = .83$ ) and IM ( $\omega = 0.80$ , 95% CI [0.78 to 0.82];  $\alpha = 0.79$ ) subscales and were slightly better than the original set of items. Overall, results provided evidence of acceptable internal consistency based on Nunnally's (1978) standards.

#### 4.3. Concurrent validity

##### 4.3.1. Regression analyses

For concurrent validity purposes, we first assessed whether the IM and BM scales,<sup>4</sup> along with gender (0 = female & 1 = male) and all the three interaction terms, were significant predictors of the three OSTES scales (i.e., perceived student engagement, perceived instructional strategies, and perceived classroom management).

<sup>4</sup> Only results from the original 12-item BIMS scale proposed by Martin and Sass (2010) were provided given that the results did not differ practically from the revised/alternative 12-item BIMS. However, results from revised/alternative 12-item BIMS are available from the corresponding author.

These statistical analyses were conducted using SAS Proc GLM (SAS Institute, 2013).

The first linear model using perceived student engagement ( $M = 6.15$ ,  $SD = 1.38$ ) as the criterion variable and gender, BM, IM, and all three interaction terms as predictor variables resulted in a statistically significant model prediction,  $F(6, 1438) = 68.64$ ,  $p < .001$ ,  $R^2 = 0.22$ . Because only IM and BM were significant predictors of perceived student engagement, a more parsimonious model with only these two variables was tested. Results indicated that combined IM and BM explained 21.9% of the variance in perceived student engagement,  $F(2, 1442) = 202.16$ ,  $p < .001$ ,  $R^2 = 0.22$ , which is a large effect size (i.e.,  $R^2$ ) based on Cohen's (1988) tentative effect size standards (small  $R^2 \approx 0.01$ , medium  $R^2 \approx 0.06$ , large  $R^2 \approx 0.14$ ). Here, both IM ( $p < .001$ , partial  $\eta^2 = 0.17$ , semipartial  $\eta^2 = 0.16$ ) and BM ( $p = .003$ , partial  $\eta^2 = 0.07$ , semipartial  $\eta^2 = 0.06$ ) were statistically significant predictors, although IM was the strongest predictor of perceived student engagement based on the partial and semipartial  $\eta^2$ .

The second linear model using perceived instructional strategies ( $M = 6.69$ ,  $SD = 1.17$ ) as the criterion variable, with gender, BM, IM, and all three interaction terms as predictors, was also statistically significant,  $F(6, 1438) = 102.00$ ,  $p < .001$ ,  $R^2 = 0.30$ , and produced a large effect size. However, after removing the non-significant terms in the model, only BM ( $p < .001$ , partial  $\eta^2 = 0.15$ , semipartial  $\eta^2 = 0.12$ ), IM ( $p < .001$ , partial  $\eta^2 = 0.20$ , semipartial  $\eta^2 = 0.17$ ), and the interaction between IM and BM ( $p < .001$ , partial  $\eta^2 = 0.01$ , semipartial  $\eta^2 = 0.01$ ) were significant predictors of perceived instructional strategies,  $F(3, 1441) = 202.15$ ,  $p < .001$ ,  $R^2 = 0.30$ . From these analyses, IM appeared to be a slightly better predictor than BM based on the partial and semipartial  $\eta^2$ . The interaction between IM and BM displayed a much smaller effect and required additional analyses to evaluate the interaction effect.

To better understand the IM by BM interaction effect on the perceived instructional strategies variable, the IM and BM variables were categorized into three groups: Low level (more than one standard deviation below the mean), medium level (between one standard deviation above and below the mean), and high level (more than one standard deviation above the mean). Bonferroni adjusted post hoc analyses indicated there was no difference on perceived instructional strategies between the medium and high levels of BM at the low level of IM nor was there a difference between low and medium levels of BM at the medium level of IM (see Fig. 1). Moreover, perceived instructional strategy scores were equal across the three BM levels at the high level of IM. As indicated in Fig. 1 caption, all other perceived instructional strategies marginal means were significantly different from each other after a Bonferroni adjustment at the 0.05 level, thus suggesting that higher levels of IM and BM produced higher perceived instructional strategies scores. However as revealed by the significant interaction term, the degree of change in perceived instructional strategies was not always consistent across different levels of BM and IM. In fact, those with low levels of BM and IM displayed much lower perceived instructional strategy scores than the other groups.

The third linear model using perceived classroom management ( $M = 6.68$ ,  $SD = 1.28$ ) as the criterion variable and gender, BM, IM, and all three interaction terms as predictors was also statistically significant,  $F(6, 1438) = 78.80$ ,  $p < .001$ ,  $R^2 = 0.25$ , and displayed a large effect size. After removing the non-significant terms in the model, only BM ( $p < .001$ , partial  $\eta^2 = 0.16$ , semipartial  $\eta^2 = 0.14$ ), IM ( $p < .001$ , partial  $\eta^2 = 0.11$ , semipartial  $\eta^2 = 0.10$ ), and the interaction between BM and IM ( $p < .001$ , partial  $\eta^2 = 0.01$ , semipartial  $\eta^2 = 0.01$ ) were significant predictors of perceived classroom management,  $F(3, 1441) = 153.84$ ,

$p < .001$ ,  $R^2 = 0.24$ . As seen here, the effect size did not change significantly between the full and reduced model ( $R^2 = 0.25$  vs.  $R^2 = 0.24$ ). Based on these analyses, BM appeared to be a slightly better predictor of perceived classroom management than IM when considering the partial and semipartial  $\eta^2$ . The interaction between IM and BM was also statistically significant and, therefore, additional post hoc analyses were conducted to evaluate the cause of the interaction effect.

Using the same BM and IM classifications as the perceived instructional strategies analyses, the marginal mean for perceived classroom management at each BM and IM level are provided in Fig. 2. Once again, post hoc analyses after the Bonferroni adjustment revealed no differences between any of the BM levels at the high IM level on the perceived classroom management variable nor did any differences exist between the medium and high BM level for those individuals classified as low IM (see Fig. 2). Unlike the perceived instructional strategies results that indicated a significant increase between medium and high levels of IM for those individuals classified as high BM, this difference was not statistically significant for the perceived classroom management variable. All other marginal means differed significantly from one another. As seen in Fig. 2, the significant interaction was largely influenced by the low BM level producing larger changes (i.e., a larger slope) than the other two BM levels. Stated differently, the relationship between IM and perceived classroom management was stronger at the low BM level than at the medium and high BM level.<sup>5</sup>

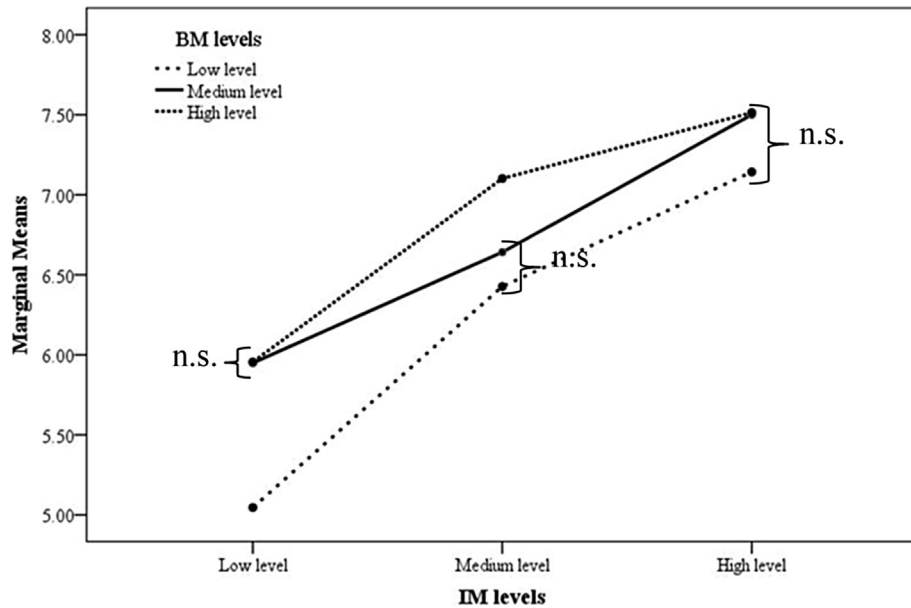
## 5. Discussion

### 5.1. BIMS' psychometric properties

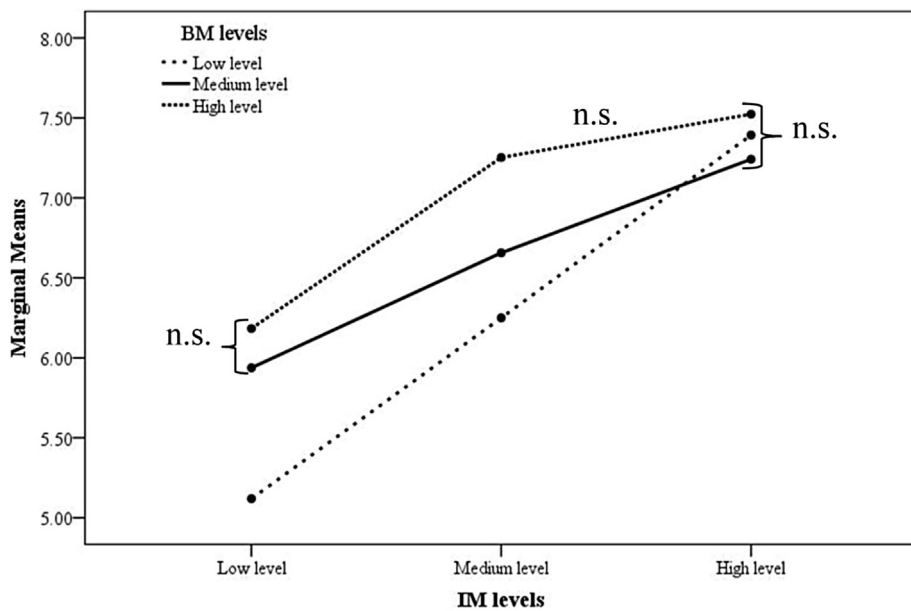
EFA results confirm previous research that the 24-item BIMS should not be used for research purposes, while the original 12-item BIMS appears to have adequate psychometric properties based on exploratory and confirmatory factor analysis results. However, the question remains which 12-item version should be used, as the slightly revised 12-item factor structures showed more promise (i.e., slightly better fit, fewer cross-loadings, & less impacted by CMV). In fact, the results here suggest that BM7 (i.e., I am strict when it comes to student compliance in my classroom) could be used in place of BM2 (i.e., I strongly limit student chatter in the classroom) and IM7 (i.e., I nearly always use inquiry-based learning in the classroom) could replace IM9 (i.e., I nearly always adjust instruction in response to individual student needs). Given that item content does not differ greatly, and one could argue for the inclusion of both sets of items, future research could administer a 14-item BIMS (rather than the 24-item version) and determine which 12-item version provides the strongest factorial validity and internal consistency reliability with their sample. It remains unclear whether the slightly different factor structure from Author and Author (2010) emerged due to cultural differences (Portuguese vs. United States sample) or random variation between samples. While not the original intent of Author and Author (2010) or evaluated here, researchers could use the 14-item BIMS version if adequate psychometric properties emerge. The general thesis of these findings is the BIMS, regardless of which 12-item version is

<sup>5</sup> To reduce the large number of statistical results provided in the text, tables, and figures, the figures only indicated whether the marginal means differed significantly from each other. Other statistical results (e.g., t-statistics, exact p-values, effect sizes) from the post hoc analyses that relate to the BM by IM interactions on the perceived instructional strategies and perceived classroom management are available from the corresponding author.





**Fig. 1.** This diagram provides the estimated marginal means for each IM (i.e., Instructional Management) and BM (i.e., Behavior Management) level to explore the interaction between IM and BM on the perceived instructional strategies variable. Based on the post hoc analyses from the general linear model, means that were not significantly different from each other after a Bonferroni adjustment (at the 0.05 level) were marked with an n.s. (i.e., non-significant), whereas all other mean comparisons being statistically significantly at the 0.05 level.



**Fig. 2.** This diagram provides the estimated marginal means for each IM (i.e., Instructional Management) and BM (i.e., Behavior Management) level to explore the interaction between IM and BM on the perceived classroom management variable. Based on the post hoc analyses from the general linear model, means that were not significantly different from each other after a Bonferroni adjustment (at the 0.05 level) were marked with an n.s. (i.e., non-significant), whereas all other mean comparisons being statistically significantly at the 0.05 level.

used, performed relatively well and appears suitable for use in research and practice.

An examination of the interfactor correlations revealed fairly consistent relationships between the IM and BM factors (ranging

from 0.46 to 0.52) regardless of the estimation method and factor structure estimated. These correlations were higher than previous research ( $r = 0.22$ , Martin & Sass, 2010) and arguably better approximate the relationship between these variables, at least in

theory. Of course, it is unknown if these differences are due to sampling error or culture.

When predicting perceived student engagement, perceived instructional strategies, and perceived classroom management, both BM and IM were consistent predictors of these variables. In addition, the interaction between BM and IM was a statistically significant predictor of perceived instructional strategies and perceived classroom management, which occurred because the relationship between IM and these criterion variables (i.e., perceived instructional strategies and classroom management) was stronger (i.e., a larger slope) at the low BM level. With that said, the contribution of these interaction terms were rather small from a practical standpoint (i.e., the partial  $\eta^2 = 0.01$  and semi-partial  $\eta^2 = 0.01$  for both criterion variables), especially when compared to the direct effects of IM and BM on these criterion variables. In fact, the IM and BM scales were relatively good predictors of these three self-efficacy scales, with most partial  $\eta^2$  being larger than 0.15. This is important because teacher's self-efficacy is associated with the classroom climate, teacher's effectiveness, students' achievement, and other classroom variables (Djigic & Stojiljković, 2011; Klassen, Durksen, & Tze, 2014; Oppedenaker & Van Damme, 2006), thus the BIMS may be a useful tool in predicting these variables as well.

### 5.2. Benefits of BCFA

While the manuscript's central focus was to evaluate the BIMS' psychometric properties, this study also sought to demonstrate the flexibility of BCFA. As seen here, when traditional CFA approaches produce model misfit, researchers may be perplexed by what model changes are most appropriate. This is especially true when multiple cross-loadings and residual covariances would produce, if estimated, similar changes in the overall model fit based on the modification indices. Given that using a stepwise approach via the modification indices can result in elevated Type I errors and often the wrong model, the BCFA approach offers an attractive alternative. As seen in Table 2, the 12-item BCFA provided a better representation of the factor structure (i.e., small cross-loadings, with BM2 and IM9 being slightly elevated), whereas the 12-item traditional CFA (labeled 12-item modified WLSMV CFA) produced alarmingly high cross-loadings for BM2 and IM9 and falsely assumed that all the residual covariances were zero.

As illustrated by Sass (2011), overly restrictive CFA models (i.e., most of the cross-loadings and residual covariances are fixed at zero) can produce unrealistic factor structures and elevated inter-factor correlations. BCFA focuses not only on estimating the model, but generating model changes to produce a better model fit and better representation of the data (Asparouhov et al., 2015). Consequently, the BCFA approach tends to identify cross-loadings, residual covariances, and other parameter estimates that might otherwise be missed with traditional CFA models (see Asparouhov et al., 2015; Muthén & Asparouhov, 2012; van de Schoot et al., 2014; Zypur & Oswald, 2013).

### 5.3. Limitations and future research

Several limitations and areas of future research related to the BIMS psychometric properties are needed. First, while laborious, it is critical to assess the relationship between perceived and actual instructional and behavior management. Specifically, one's perception of their instructional and behavior management as measured by the BIMS may not be highly correlated with their

actual instructional and behavior management (e.g., as measured by instructional and behavior management experts via observational methods).

Second, research is needed to explore what 12- or 14-item BIMS version displays the best factor structure and identify how well this factor structure is uncovered when the 10 (or 12) unused items are omitted. Along a similar vein, while this study provided some results related to CMV, more research in this area is needed. Again, the IM factor using the original 12-item scale did have some concerns related to CMV, whereas the revised 12-item BIMS showed promise.

Third, research is needed to better understand these scales' generalizability and the degree of measurement invariance across various cultures (e.g., Portuguese vs. United States) and teacher characteristics (e.g., novice vs. experienced teachers). For example, it would be interesting to determine whether BM and IM differ by culture and how that translates to the student success and classroom dynamics. Moreover, it would be interesting to explore how teacher's BM and IM change over the course of their teaching careers.

Fourth, it is critical to identify those variables (whether educational or psychological) that predict behavior and instructional management styles, while at the same time determining those variables that behavior and instructional management predict (e.g., student performance, teacher attrition and stress, classroom atmosphere, instructional effectiveness, teacher respect, classroom dynamics, teachers student interactions). Now that the BIMS appears to have adequate factorial validity and internal consistency reliability based on the Author and Author (2010) study and now the current study, the next steps would be to collect additional predictive, concurrent, and convergent validity evidence with the BIMS.

### 5.4. Conclusions

Given that aspects related to behavior and instructional management include student control, instructional style, setting rules, and the regulation of student misbehaviors, the BIMS may provide useful information that increases the understanding of the classroom atmosphere. Consequently, developing a management style profile via the BIMS could be insightful to teachers and administrators in developing an effective management plan (e.g., maybe a teacher is too controlling, thus resulting in student rebellion). Therefore, the BIMS is intended to gain a better understanding of what teachers do in the classroom, with the hope that the IM and BM scales predict other key outcome variables (see *Limitations and Future Research* section for more details).

In summary, these results provided support for factorial/construct validity and internal consistency reliability of the BIMS via Bayesian estimation, but did bring into question whether the original or revised 12-item BIMS is best. While the revised 12-item BIMS did have better psychometric properties (i.e., better model fit, smaller cross-loadings, and less concerns related to CMV), it is possible that the results are sample specific. On a positive note, concurrent validity or internal consistency reliability results did not differ between the two versions. This might be expected given that most (i.e., 5 out of 6) items remained constant across versions. Collectively, these results provide an indication that the BIMS is ready for the next stages of validation (e.g., predictive validity). Finally, this study demonstrated the benefits of BCFA and its ability to provide information that otherwise might not be discovered with traditional CFA methods.

**Appendix. Provides the English and Portuguese version of the BIMS, with the original 12-item version bolded. In practice, the IM and BM items should be alternating.**

English version	Portuguese version
<b>BM1</b> I nearly always intervene when students talk at inappropriate times during class	Na aula, quase sempre intervenho quando os alunos falam de forma inapropriada.
<b>BM2</b> I strongly limit student chatter in the classroom	Durante as aulas, limito fortemente as conversas entre os alunos.
BM3 I reward students for good behavior in the classroom	Costumo recompensar os alunos pelo seu bom comportamento.
BM4 If a student talks to a neighbor, I will move the student away from other students	Se um aluno tem por hábito falar com o colega do lado, mudo-o de lugar.
BM5 I use input from students to create classroom rules	Utilizo os contributos dos alunos na elaboração das regras de sala de aula.
BM6 I allow students to get out of their seat without permission	Permito que os alunos saiam do lugar sem autorização.
BM7 I am strict when it comes to student compliance in my classroom	Sou rigoroso/a no que diz respeito à observância das regras por parte dos alunos
<b>BM8</b> I firmly redirect students back to the topic when they get off task	Quando os alunos estão fora da tarefa, sou firme a redirecioná-los para o assunto.
<b>BM9</b> I insist that students in my classroom follow the rules at all times	Insisto com os alunos para que, na sala de aula, cumpram as regras em todos os momentos.
BM10 I closely monitor off task behavior during class	Durante a aula, monitorizo de perto os comportamentos fora da tarefa.
<b>BM11</b> I strictly enforce classroom rules to control student behavior	Faço cumprir rigorosamente as regras para controlar o comportamento dos alunos.
<b>BM12</b> If a student's behavior is defiant, I will demand that they comply with my classroom rules	Se um aluno é desafiador, exijo que ele cumpra as regras da sala de aula.
IM1 I use whole class instruction to ensure a structured classroom	Costumo dar instruções para toda a turma como forma de garantir a organização da aula.
<b>IM2</b> I nearly always use collaborative learning to explore questions in the classroom	Quase sempre utilizo métodos de aprendizagem cooperativa para explorar questões e assuntos diversos.
<b>IM3</b> I engage students in active discussion about issues related to real world applications	Envolvo os alunos na discussão de assuntos com aplicação no mundo real.
IM4 I establish a teaching daily routine in my classroom and stick to it	Estabeleço uma rotina diária de ensino e cumpro-a.
<b>IM5</b> I nearly always use group work in my classroom	Quase sempre recorro ao trabalho de grupo.
<b>IM6</b> I use student input when creating student projects	Utilizo os contributos dos alunos no desenvolvimento dos seus projectos
IM7 I nearly always use inquiry-based learning in the classroom	Na sala de aula, quase sempre utilizo a aprendizagem por descoberta
IM8 I direct the students' transition from one learning activity to another	Controlo as transições dos alunos entre as várias actividades da aula.
<b>IM9</b> I nearly always adjust instruction in response to individual student needs	Quase sempre ajusto a instrução às necessidades individuais dos alunos.
IM10 I nearly always use direct instruction when I teach	Nas minhas aulas, quase sempre utilizo a instrução directa
IM11 I do not deviate from my pre-planned learning activities	Não me afasto do meu plano de actividades da aula.
<b>IM12</b> I nearly always use a teaching approach that encourages interaction among students	Quase sempre utilizo um método de ensino que promove a interacção entre os alunos.

## References

- Asparouhov, T., & Muthén, B. (2010a). *Bayesian analysis of latent variable models using Mplus* (Technical report). Los Angeles, CA: Muthén & Muthén.
- Asparouhov, T., & Muthén, B. (2010b). *Bayesian analysis using Mplus: Technical implementation* (Technical appendix). Los Angeles, CA: Muthén & Muthén.
- Asparouhov, T., Muthén, B., & Morin, A. J. S. (2015). Bayesian structural equation modeling with cross-loadings and residual covariances: Comments on Stromeier et al. *Journal of Management*, 41, 1561–1577.
- Burden, P. R. (1995). *Classroom management and discipline*. White Plains, NY: Longman.
- Burden, P. R. (2000). *Powerful classroom management strategies: Motivating students to learn*. Thousand Oaks, CA: Corwin Press.
- Canter, M., & Canter, L. (2001). *Assertive discipline*. Los Angeles: Canter & Canter Associates.
- Carter, K. (1985, March–April). Teacher comprehension of classroom processes: An emerging direction in classroom management research. In *Paper presented at the annual meeting of the American educational research association, Chicago, IL*.
- Chall, J. S. (2000). *The academic achievement challenge: What really works in the classroom?* New York: Guilford Press.
- Clarke-Stewart, K. A., Lee, Y., Allhusen, V. D., Kim, M. S., & McDowell, D. J. (2006). Observed differences between early childhood programs in the U.S. and Korea: Reflections of “developmentally appropriate practices” in two cultural contexts. *Journal of Applied Developmental Psychology*, 27(5), 427–443.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, 1, 16–29.
- Djigić, G., & Stojiljković, S. (2011). Classroom management styles, classroom climate and school achievement. In *Paper presented at the procedia – social and behavioral sciences*.
- Emmer, E. T., Evertson, C. M., & Anderson, L. M. (1980). Effective classroom management at the beginning of the school year. *The Elementary School Journal*, 80(5), 219–231.
- Epstein, S. (1973). The self-concept revisited: Or a theory of a theory. *American Psychologist*, 28, 404–416.
- Ertesvåg, S. K. (2011). Measuring authoritative teaching. *Teaching and Teacher Education*, 27(1), 51–61.
- European Commission. (2010). In E. Commission (Ed.), *Developing coherent and system-wide induction programmes for beginning teachers: A handbook for policymakers*. retrieved from [http://ec.europa.eu/education/library/publications/handbook0410\\_en.pdf](http://ec.europa.eu/education/library/publications/handbook0410_en.pdf).
- Evertson, C., & Harris, S. (1999). Support for managing learning-centered classrooms: The classroom management and organization program. In H. J. Freiberg's (Ed.), *Beyond behaviorism: Changing the classroom management paradigm* (pp. 59–74). Boston, MA: Allyn & Bacon.
- Evertson, C. M., & Weinstein, C. S. (2006). Classroom management as a field of inquiry. In C. M. Evertson, & C. S. Weinstein's (Eds.), *Handbook of classroom management: Research, practice, and contemporary issues* (pp. 3–15). Mahwah, NJ: Lawrence Erlbaum Associates.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall.
- Glasser, W. (1998). *The quality school teacher*. New York: Harper Perennial.
- Glickman, C. D., & Tamashiro, R. T. (1980). Clarifying teachers' beliefs about discipline. *Educational Leadership*, 37, 459–464.
- Good, T. L., & Brophy, J. E. (2000). *Looking in classrooms* (8th ed.). New York, NY: Addison-Wesley Educational Publishers.
- Gordon, T. (2003). *Teacher effectiveness training*. New York: Three Rivers Press.
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1–55.
- Iverson, A. M. (2003). *Building competence in classroom management and discipline*. Upper Saddle River, NJ: Merrill/Prentice-Hall.
- Jalali, S., Panahzade, V., & Firouzmand, A. (2014). Attitude towards computers and classroom management of language school teachers. *International Journal of Instruction*, 7(2), 35–50.
- Klassen, R. M., Bong, M., Usher, E., Chong, W. H., Huan, V. S., Wong, I. Y. F., et al. (2009). Exploring the validity of a teachers' self-efficacy scale in five countries. *Contemporary Educational Psychology*, 34, 67–76.
- Klassen, R. M., Durksen, T. L., & Tze, V. M. C. (2014). Teachers' self-efficacy beliefs: Ready to move from theory to practice? In P. W. Richardson, S. A. Karabenick, & H. M. G. Watt (Eds.), *Teacher motivation: Theory and practice* (pp. 100–115). New York: Routledge.
- Kounin, J. S. (1970). *Discipline and group management in classrooms*. NY, NY: Holt, Rinehart & Winston.

- Lopes, J., & Santos, M. (2013). Teachers' beliefs, teachers' goals and teachers' classroom management: A study with primary teachers. *Journal of Psychodactics*, 18(1), 5–24.
- MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modification in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, 111, 490–504.
- MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, 4, 84–99.
- Martin, N. K., & Sass, D. A. (2010). Construct validation of the behavior and instructional management scale. *Teaching and Teacher Education*, 26, 1124–1135.
- Martin, N. K., Sass, D. A., & Schmitt, T. A. (2012). Tests of measurement and structural invariance of a theoretical model predicting a teacher's intent-to-leave. *Teaching and Teacher Education: An International Journal of Research and Studies*, 28, 546–559.
- Martin, N. K., Yin, Z., & Baldwin, B. (1998). Construct validation of the attitudes and beliefs on classroom control inventory. *Journal of Classroom Interaction*, 33(2), 6–15.
- Martin, N. K., Yin, Z., & Mayall, H. (2007). The attitudes and beliefs on classroom control inventory revised and revisited: A continuation of construct validation. *Journal of Classroom Interaction*, 42(2), 11–20.
- McNeely, S. R., & Mertz, N. T. (1990, April). Cognitive constructs of preservice teachers: research on how student teachers think about teaching. In *Paper presented at the annual meeting of the American educational research association, Boston*.
- Muthén, B., & Asparouhov, T. (2012). Bayesian SEM: A more flexible representation of substantive theory. *Psychological Methods*, 17, 313–335.
- Muthén, L. K., & Muthén, B. O. (1998–2013). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.
- OECD. (2009). *Creating effective teaching and learning environments: First results from the OECD teaching and learning international survey (TALIS)*. Paris: OECD.
- Olafson, L. J., & Schraw, G. (2006). Teachers' beliefs and practices within and across domains. *International Journal of Educational Research*, 45, 71–74.
- Oppedenaker, M. C., & Van Damme, J. V. (2006). Teacher characteristics and teaching styles as effectiveness enhancing factors of classroom practice. *Teaching and Teacher Education*, 22, 1–21.
- Osborne, J. W., & Fitzpatrick, D. C. (2012). Replication analysis in exploratory factor analysis. What it is and why it makes your analysis better. *Practical Assessment, Research & Evaluation*, 17, 1–8.
- Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology*, 65, 539–569.
- Ravitch, D. (2000). *Left back: A century of battles over school reform*. New York: Touchstone.
- Reeve, J., & Jang, H. (2006). What teachers say and do to support students' autonomy during a learning activity. *Journal of Educational Psychology*, 98(1), 209–218.
- Rogers, C., & Freiberg, H. J. (1994). *Freedom to learn* (3rd ed.). Columbus, Ohio: Merrill.
- SAS Institute Inc. (2013). *SAS® 9.4 guide to software updates*. Cary, NC: SAS Institute Inc.
- Sass, D. A. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. *Journal of Psychoeducational Assessment*, 29(4), 347–363.
- Schmitt, T. A. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessment*, 29, 304–321.
- Schmitt, T. A., & Sass, D. A. (2011). Rotation criteria and hypothesis testing for exploratory factor analysis: Implications for factor pattern loadings and inter-factor correlations. *Educational and Psychological Measurement*, 71, 95–113.
- Terhart, E. (2003). Constructivism and teaching: A new paradigm in general didactics? *Journal of Curriculum Studies*, 35(1), 25–44.
- Tschannen-Moran, M., & Hoy, W. A. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education*, 17, 783–805.
- Unal, Z., & Unal, A. (2012). The impact of years of teaching experience on the classroom management approaches of elementary school teachers. *International Journal of Instruction*, 5(2), 41–60.
- van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & van Aken, M. A. G. (2014). A gentle introduction to Bayesian analysis: Applications to research in child development. *Child Development*, 85, 842–860.
- Weinstein, C. S., & Mignano, A. J., Jr. (1993). *Elementary classroom management: Lessons from research and practice*. New York, NY: McGraw-Hill.
- Willower, D. J., Eidell, T. L., & Hoy, W. K. (1973). *The school and pupil control ideology*. Penn State Studies No. 24 (2nd ed.). University Park, Pennsylvania: The Pennsylvania State University.
- Wolfgang, C. H. (1995). *Solving discipline problems: Strategies for classroom teachers* (3rd ed.). Boston: Allyn and Bacon.
- Wolfgang, C. H., & Glickman, C. D. (1980). *Solving discipline problems: Strategies for classroom teachers*. Boston: Allyn and Bacon.
- Woolfolk Hoy, A. E., & Weinstein, C. S. (2006). Student and teacher perspectives in classroom management. In C. M. Evertson, & C. S. Weinstein's (Eds.), *Handbook of classroom management: Research, practice, and contemporary issues* (pp. 181–219). Mahwah, NJ: Lawrence Erlbaum Associates.
- Yang, Y., & Green, S. B. (2011). Coefficient alpha: A reliability coefficient for the 21st century? *Journal of Psychoeducational Assessment*, 29, 377–392.
- Zyphur, M. J., & Oswald, F. L. (2013). Bayesian probability and statistics in management research: A new horizon. *Journal of Management*, 39, 5–13.