Examining the Teachers’ Sense of Efficacy Scale at the Item Level With Rasch Measurement Model

Mei-Lin Chang¹ and George Engelhard Jr.²

Abstract
The purpose of this study is to examine the psychometric quality of the Teachers’ Sense of Efficacy Scale (TSES) with data collected from 554 teachers in a U.S. Midwestern state. The many-facet Rasch model was used to examine several potential contextual influences (years of teaching experience, school context, and levels of emotional exhaustion) on item functioning within the TSES. Results suggest that although TSES items are rather easy for teachers to endorse, sufficient variance in the item endorsement hierarchy of the scale exists to support the validity of score interpretations. The items are invariant across years of teaching experience or school locations, but not invariant across levels of emotional exhaustion.

Keywords
teachers’ sense of efficacy, Rasch measurement, item functioning, self-efficacy, emotional exhaustion

Introduction
Derived from social cognitive theory (Bandura, 1986, 1997), self-efficacy is conceptualized as an individual’s beliefs about his or her own capabilities to organize and execute courses of action required for accomplishing a specific task. Teachers’ sense of efficacy can be considered one type of self-efficacy specifically applied to the context of teaching tasks. It is the confidence a teacher holds in his or her capability to organize and execute courses of action required to successfully accomplish a specific teaching task in a particular context (Tschannen-Moran, Woolfolk Hoy, & Hoy, 1998).

Several studies have demonstrated that a strong sense of efficacy is related to enthusiasm, higher motivation, greater effort, persistence, and resilience across the span of a teaching career (Coladarci, 1992; Evans & Tribble, 1986; Glickman & Tamashiro, 1982; Jennett, Harris, & Mesibov, 2003). Despite the positive findings in teacher efficacy studies, several researchers have raised concerns about the psychometric quality of teacher efficacy scales (Henson, 2002;
Klassen, Tze, Betts, & Gordon, 2011; Pajares, 1996; Tschannen-Moran et al., 1998). The potential measurement problems include (a) a lack of conceptual clarity (Klassen et al., 2011; Tschannen-Moran & Woolfolk Hoy, 2001) and (b) an absence of levels of specificity and difficulty (Bandura, 2006; Henson, 2002).

Tschannen-Moran and Woolfolk Hoy (2001) created and validated the Teachers’ Sense of Efficacy Scale (TSES) with factor analysis, and it has been considered as more congruent with self-efficacy theory than other measures have been (Klassen et al., 2011). This scale includes three dimensions: (a) efficacy for instructional strategies (IS), which captures teachers’ sense of efficacy in developing and implementing IS to meet students’ needs; (b) efficacy for classroom management (CM), which captures teachers’ sense of efficacy in maintaining classroom order and helping students follow rules; and (c) efficacy for student engagement (SE), which captures teachers’ sense of efficacy in engaging and motivating students to learn. Recently, Klassen et al. (2009) validated it again with confirmatory factor analysis using cross-national samples with six groups of teachers from five countries: Canada, Cyprus, Korea, Singapore, and the United States.

To our knowledge, current researchers have predominantly relied on factor analysis to evaluate the psychometric quality of TSES. Only Klassen and Chiu (2010) have used item response theory models, with a Canadian teacher sample, to examine the TSES. Klassen and Chiu have argued for the need to continue exploring the scale with teachers from different settings. The present study extends their work by examining the invariance of the scale with a U.S. teacher sample, and by exploring a few other explanatory variables that were not examined in their study (i.e., emotional exhaustion).

Bandura (2006) called for more rigorous procedures related to item analysis in his Guide for Constructing Self-Efficacy Scales and argued for the importance of exploring the difficulty levels of self-efficacy scales. To address his call, the purpose of this study is to perform an item analysis of the TSES (Tschannen-Moran & Woolfolk Hoy, 2001) with the Rasch measurement model to explore the hierarchy of difficulty level of the items and measurement invariance. Item response theory emphasizes invariance using model-based measurement (Baker & Kim, 2004; Embretson, 2010), and the Rasch model allows for item-invariant person estimates and person-invariant item estimates (Engelhard, 2013). It also allows investigating whether the items’ order in terms of the “easiness” or “difficulty” of endorsement aligns with the authors’ conceptualization of the order structure underlying the construct. In addition, we investigated the influences of a few explanatory variables on measurement invariance by using a many-facet (MF) Rasch model (Linacre, 1989). Specifically, the study is guided by the following research questions:

**Research Question 1:** Does the internal structure of the TSES represent gradations of item difficulty?

**Research Question 2:** Do the TSES items exhibit acceptable model-data fit that supports the validity of inferences regarding teachers’ sense of efficacy?

**Research Question 3:** Do the TSES items exhibit measurement invariance when explored with explanatory variables such as teaching experience, school location, and emotional exhaustion?

## Teachers’ Sense of Efficacy and Other Explanatory Variables

### Teachers’ Sense of Efficacy and Years of Experience

Researchers have found mixed results regarding the relationship between teachers’ sense of efficacy and years of teaching experience. Teachers’ sense of efficacy increased as students progressed through teacher preparation programs in some countries, such as Canada, Korea, and the United States (Fives, Hamman, & Olivarez, 2007; Woolfolk Hoy & Burke, 2005). Woolfolk Hoy
and Burke (2005) found that preservice teachers’ sense of efficacy increased after their student teaching; however, a declining trend appeared to occur after they finished their first year of teaching. In Israel and Taiwan, researchers found no changes over the years of preparation (Woolfolk Hoy & Burke, 2005). For in-service teachers, some researchers reported that teacher efficacy decreases with years spent as a teacher (Ghaith & Yaghi, 1997; Woolfolk Hoy & Burke, 2005). However, in a case study, Mulholland and Wallace (2001) found that efficacy increased with experience as the teachers became better able to manage students’ behaviors and science inquiry activities. Klassen and Chiu (2010) found a nonlinear relationship between teachers’ years of experience and teacher efficacy, with the three factors of teacher efficacy increasing with experience for early- and mid-career stage teachers and declining for late-career teachers.

**Teachers’ Sense of Efficacy and School Contexts**

Only a few studies have documented the relationships between teachers’ sense of efficacy and school context (Tschannen-Moran & Woolfolk Hoy, 2007). In Klassen et al.’s (2011) study, teachers who teach elementary grades and kindergarten had higher levels of self-efficacy for CM and SE. Siwatu (2011) found preservice teachers have a lower sense of efficacy to teach in urban schools than in suburban schools. Several scholars have argued the importance of including local context in teacher efficacy research (Klassen et al., 2011; Pajares, 2007). In the present study, we include school locations (urban, suburban, and rural) as one of the facets in explaining variance in teachers’ sense of efficacy.

**Teachers’ Sense of Efficacy and Emotional Exhaustion**

A fairly consistent finding is that higher levels of burnout, or emotional exhaustion, are associated with lower perceptions of teachers’ sense of efficacy (Brissie, Hoover-Dempsey, & Bassler, 1988; Greenglass & Burke, 1988; Friedman & Farber, 1992). Other scholars have argued that teachers who hold a higher sense of efficacy might experience emotional exhaustion (EE) as well because they tend to have higher standards and expectations of themselves and students (Fives et al., 2007). Although researchers have identified a general pattern between teachers’ sense of efficacy and teacher emotional exhaustion, inconsistent findings suggest a need to further examine the interaction between these two variables. Klassen and Chiu (2010) found not all three dimensions of teachers’ sense of efficacy are directly linked to workload stress in the same way. Particularly, teachers who reported higher workloads also reported greater self-efficacy in CM. The present study further investigates the interaction between teacher EE and teachers’ sense of efficacy.

Research has shown inconsistent results when exploring the relationships of teachers’ sense of efficacy with years of experiences and emotional exhaustion. We believe these factors may vary depending on the context of the samples. Therefore, it warrants our efforts to explore how these factors influence the measurement invariance of the TSES items.

**Method**

**Participants**

A total of 2,710 teachers within their first four teaching years were randomly selected through e-mail contact lists the Teacher Quality Project (TQP), a state-level research project team in the Midwest, provided. The TQP investigates the effects of initial preparation and professional development on teacher effectiveness and self-efficacy. Seven hundred seventeen teachers (26.45%) responded to the online survey. Participants
include 554 full-time elementary ($n = 186$), middle-school ($n = 135$), and high-school ($n = 233$) teachers (male = 113, female = 441) with a wide range of years of teaching experience (37.18% with 1-2 years, 19.6% with 3-5 years, 16.7% with 6-10 years, 11.01% with 11-20 years, and 15.16% with more than 20 years) and with a mixture of school areas (39.1% in urban schools, 29.9% in suburban schools, and 25.2% in rural schools). In terms of ethnicity, the present study has an overrepresentation of White teachers (94.5%) compared with the regional and national average. From 2004 to 2006, approximately 92% of teachers in the Midwest and 85% of teachers in the United States were White (U.S. Department of Education, National Center for Education Statistics, 2004).

**Instruments**

**Teachers’ Sense of Efficacy Scale (TSES).** Tschannen-Moran and Woolfolk Hoy (2001) developed the 12-item TSES to measure the following three dimensions of teachers’ sense of efficacy: IS ($\alpha = .80$ in original study, $\alpha = .71$ in the current sample), CM ($\alpha = .86$ in original study, $\alpha = .65$ in the current sample), and SE ($\alpha = .81$ in original study, $\alpha = .74$ in the current sample). In this study, we analyzed these 12 items individually in the Rasch analysis as they all contributed to explain teachers’ sense of efficacy holistically. Coefficient alpha in the present study is .90 for the 12-item scale (see Table 2 for full list of the items). The teachers responded to the items through a nine-category rating scale with five anchor labels: *nothing*, very little, *some influence*, *quite a bit*, and *a great deal*.

**Emotional exhaustion.** The modified Teacher Burnout subscale by Schaufeli and Salanova (2007) measured emotional exhaustion. This subscale has four items measuring the extent to which teachers feel emotionally exhausted from their work. Participants were asked to report the frequencies of their experiences of EE on a scale from 0 to 6 (0 = *never*, 6 = *almost daily*). A sample item measuring EE is “I felt emotionally drained by my work.” The Cronbach’s alpha coefficient is .87 in the current study.

**Analytic Approach**

Rasch measurement theory (Rasch, 1960-1980) forms the foundation for the examination of the measurement properties and functioning of the TSES scale presented in this article. All analyses were performed using the Facets program (Linacre, 2013). We chose to use the Rasch model because of the benefits of such an approach. First, Rasch model allows for the conversion of ordinal raw data into interval measures (Wright & Mok, 2004). Second, if the data fit the Rasch model, then one can obtain item difficulty and person ability estimates that are independent of the particular sample used to obtain them (Engelhard, 2013). Third, it yields a variable map that provides a visual display of the locations of persons and items on the latent variable. It allows an investigation of the ordered structure of the items in the scale that aids our understanding of the underlying construct. As such, we chose to use the Rasch model over other approaches (e.g., confirmatory factor analysis) because it allows us to focus on the functioning of the items and the persons simultaneously, and it also allows us to explore the item difficulty of the TSES, which has not been addressed in previous studies.

The unidimensionality of TSES scale was confirmed with an exploratory factor analysis (Kaiser-Meyer-Olkin measure of sampling adequacy = .874, Bartlett’s Test of Sphericity, $p < .001$), and the results showed a 45.21% of variance explained by a single factor model. The amount of variance explained by the first factor exceeds the 20% cutoff Reckase (1979) suggested for using Rasch models to analyze scales that may be composed of other dimensions. For reference and interpretation purposes, we have labeled the items to reflect the three subareas
included in the original scale as Tschannen-Moran and Woolfolk Hoy (2001) proposed to help with the interpretations of the items conceptually.

Two separate analyses were performed. To address Research Questions 1 and 2, the first analysis included the estimation of only the person and item locations to investigate the functioning of the items without including any potential interaction effects from other variables. To address Research Question 3, MF was used in the second analysis to examine the effects of additional facets: years of experiences, locations of schools, and levels of emotional exhaustion. The MF model can be used in an exploratory way and can be thought of as an additive linear model that is based on a logistic transformation of observed ratings to a logit scale (Engelhard, 2013).

The following MF Rasch model was used for the TSES:

\[
\text{Ln}(P_{nijk} / P_{nijk-1}) = \theta_n - \delta_i - \Delta_j - \alpha_l - \beta_m - \tau_k,
\]

where

\( P_{nijk} \) = the probability of teacher \( n \) responding in category \( k \) on item \( i \);
\( P_{nijk-1} \) = the probability of teacher \( n \) responding in category \( k - 1 \) on item \( i \);
\( \theta_n \) = the sense of self-efficacy for teacher \( n \);
\( \delta_i \) = the location of item \( i \);
\( \Delta_j \) = years of teaching experience \( j \);
\( \alpha_l \) = school location \( l \);
\( \beta_m \) = degree of emotional exhaustion \( m \); and
\( \tau_k \) = the difficulty of responding in category \( k \) relative to \( k - 1 \) on the rating scale.

The first two facets (teachers and items) are considered measurement facets, whereas the other three facets (school location, experience, and emotional exhaustion) are considered explanatory facets that are used to explore the measurement invariance of the TSES items by subgroups.

Software and Summary Statistics

The Facets program (Linacre, 2007, 2013) provides estimates of locations for each individual item and person as well as summary statistics for all the items and persons. Statistics that are typically used to determine how well the data fit the Rasch model include the reliability of separation for persons and items that describes how well the persons and items are able to be differentiated on the underlying construct. In addition, fit statistics are given for each of the items and persons. Both Infit and Outfit mean squares give estimates of the mismatch between data and model; however, Infit gives less weight to remote responses (Wright & Mok, 2004). For these fit statistics, the expected mean is 1.0. Values lower than 1.0 represent less variation than expected by the model, whereas values greater than 1.0 represent more variation than expected by the model. Rules of thumb have typically described Infit and Outfit statistics between 0.8 and 1.2 as acceptable (Engelhard, 2013).

Data Preparation

Data were first entered into SPSS and then exported to the software program FACETS (Linacre, 2007) to run the MF Rasch analyses. The preliminary data analysis showed lower usage in the extreme categories of the lower end on the TSES and EE scales. Wright and Linacre (1992)
argued for the need to examine and analyze the respondent rating behaviors and to redefine the categories if some rating categories are underutilized (also see Linacre, 2004). Due to the lack of use of the extreme lower end response categories of the TSES in this study (less than 0.8% usage of categories 1 and 2 in the sample), the original 9-point rating scale was transformed into a 7-point scale, in which we collapsed ratings from 1, 2, and 3 to indicate the lowest level of efficacy. After collapsing the categories, the mean scores for each item were almost identical, and the skewness and kurtosis were significantly improved (ranged from −0.26 to −1.41 for skewness across the items in original ratings, and from −0.07 to −1.07 in collapsed ratings; ranged from 0.008 to 4.70 for kurtosis in original ratings, and from −0.16 to 1.5 in collapsed ratings). Similarly, for the variable measuring the frequencies of EE teachers felt in a year, we collapsed the lowest ratings 0 (never), 1 (a few times a year), and 2 (a few times a semester) into a single category, which indicated a very low degree of emotional exhaustion. The other four rating categories in measuring EE used original ratings from low (a few times a month), medium (once a week), high (a few times a week), to very high (almost daily).

Results

Analysis 1: Rasch Analysis of Internal Structure and Model-Data Fit of TSES

The first analysis was conducted to address the Research Questions 1 and 2: Does the internal structure of the TSES represent gradations of item difficulty? Do the TSES items exhibit acceptable model-data fit that supports the validity of inferences regarding teachers’ sense of efficacy?

Overall, the Rasch model explained 50.42% of variance of the TSES. The results of the MF Rasch model indicated the TSES has a relatively high reliability (closer to 1) of person separation (Table 1), \(\text{Rel}_{\text{Teachers}} = .87, \chi^2(712) = 4786.1, p < .05\), which indicates that the scale was sensitive to differences between teachers with high versus low self-efficacy. The scale also has a relatively high reliability (closer to 1) of item separation (Table 1), \(\text{Rel}_{\text{Items}} > .99, \chi^2(11) = 3686.3, p < .05\). The high reliability of item separation indicates that the items are distinctive in their locations on the latent variable. In other words, there is good separation in terms of item ordering and hierarchy used to define the TSES.

Gradation of difficulty levels. The variable map in Figure 1 includes the item locations of the 12 items in the TSES and the person locations for each teacher who completed the survey. This variable map presents graphically the separation noted in the previous paragraph. Column 1 is the logistic scale that represents the unit of measurement underlying TSES, whereas columns 2 and 3 represent the locations of the two measurement facets items and teachers. Higher numbers on the logit scale refer to lower levels of endorsement of the items, which means the item or task is considered more difficult for teachers to endorse. For example, the hardest item is SE2, “How much can you do to get students to believe they can do well in schoolwork?” with a location at .97 logits (as also reported in Table 2). The easiest item to endorse is IS4, “How much can you do to implement alternative strategies in your classroom?” with −.78 logits. Overall, the variable map indicates that within each construct category, difficulty levels are well represented within certain ranges on the logit scale.

Model-data fit. Figure 2 shows graphically the model-data fit between these data and the model shown in Equation 1. The correspondence between the expected values and observed values are quite close in this display; thus, the figure supports the inference that the variable maps are good representations of the underlying latent variables that the measure is designed to represent. Also, the average logit measures for each recoded category from 1 to 7 (with observed percentages in
parentheses) are −0.74 (2%), −0.51 (4%), −0.16 (11%), 0.26 (14%), 0.80 (28%), 1.40 (20%), and 2.00 (21%), respectively. The increase in average logit measures by category supports the inference that the rating scale is performing as expected based on the model because the logit measures increase as the rating categories also increase from 1 to 7.

**Item quality.** In addition, as shown in Table 2, most of the items appear to have a good fit using the rules of thumb for expected values of Infit and Outfit values between 0.8 and 1.2 (Engelhard, 2013). “How much can you do to get students to follow classroom rules?” (CM1) is the only item in the measure that appears to exhibit some misfit (Infit = 1.43, Outfit = 1.37). This item measures the extent to which teachers believe in their ability to get students to follow classroom rules, and the teachers seem to exhibit differential rating behavior when they respond to this item. This misfit indicates that even if a teacher rated high on this item, the person would not necessarily endorse easier items in classroom management such as “How much can you do to calm a student who is disruptive or noisy?” (CM3) or “How much can you do to establish a classroom management system with each group of students?” (CM4).

### Analysis 2: MF Rasch Model of TSES Items With Explanatory Variables

MF was conducted to address Research Question 3: Do the TSES items exhibit measurement invariance when explored with explanatory variables such as teaching experience, school location, and emotional exhaustion?

The differences on TSES items between levels of years of experience and school locations are not statistically significant \((p > .05)\) as shown in Table 3 and Figure 3. However, the differences among levels of EE are statistically significant, \(Rel_{EE} = .97, \chi^2(4) = 115.5^*, p < .05\). Mean values vary from a high of 0.24 logits for teachers with low EE to −0.17 logits for teachers with high emotional exhaustion. Generally, teachers with higher EE tend to report lower levels of efficacy.

<table>
<thead>
<tr>
<th>Table 1. Summary Statistics for Teachers, Items, Experience, School Context, and Emotional Exhaustion.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Measurement facets</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Measure</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Outfit</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Infit</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Separation statistic</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

∗\(p < .05\).
Figure 1. Variable map of teachers’ sense of efficacy and item locations.
Note. Items were labeled based on categories that Tschannen-Moran and Woolfolk Hoy (2001) identified in factor analyses of the TSES. SE = student engagement; IS = instructional strategies; CM = classroom management; TSES = Teachers’ Sense of Efficacy Scale.
The differences among levels of EE warrant further examination of its interaction effects with the TSES items. In Figure 4, as shown in Panel A, the functioning of the TSES items for teachers in very low, medium, and high EE groups is consistent across these three groups, and the $t$ values are among the expected range of −2 to 2, but not for the low and very high EE groups. As highlighted in Panel B, in the high EE group ($n = 150$), four items were more than the expected range (CM3, calm a disruptive student; CM4, establish a CM system with each group of students; SE4, assist families in helping children do well; and IS4, implement alternative strategies), and three items were lower than the range (IS1, use a variety of assessment strategies; CM2, control disruptive behavior; and SE3, help students value learning). In contrast, teachers with low EE ($n = 74$) exhibited a somewhat opposite pattern from teachers with very high emotional exhaustion, rating higher on IS1 (use a variety of assessment strategies) and CM2 (control disruptive behavior) and lower on CM3 (calm a disruptive student).

**Discussion**

**Item Difficulty Hierarchy in the TSES**

Results from the present study indicate that the TSES captures a range of item difficulties that reflect a hierarchy within class management, IS, and SE. However, we also found that most teaching tasks measured in the scale are rather easy to endorse. To fully capture a wider range of
Table 3. Summary of the Two-Way Interactions of TSES Items With Years of Experience, School Context, and Emotional Exhaustion.

<table>
<thead>
<tr>
<th>Explanatory and contextual facets</th>
<th>Infit</th>
<th>Outfit</th>
<th>$\chi^2$</th>
<th>df</th>
<th>$p$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MS</td>
<td>MS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Years of experience</td>
<td>1.0</td>
<td>1.0</td>
<td>66.0</td>
<td>60</td>
<td>.28</td>
</tr>
<tr>
<td>SD</td>
<td>0.2</td>
<td>0.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>School context</td>
<td>1.0</td>
<td>1.0</td>
<td>29.3</td>
<td>36</td>
<td>.78</td>
</tr>
<tr>
<td>SD</td>
<td>0.2</td>
<td>0.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emotional exhaustion</td>
<td>1.0</td>
<td>1.0</td>
<td>97.5</td>
<td>60</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>SD</td>
<td>0.2</td>
<td>0.2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. TSES = Teachers’ Sense of Efficacy Scale; MS = mean square.

difficulty levels, efficacy researchers should consider developing items that could capture gradations of challenges at higher levels on the variable (Bandura, 2006). In his guide for constructing scales, Bandura (2006) suggested increasing the difficulty level by raising the level of challenge in the item. For example, the item, “How much can you do to calm a student who is disruptive or noisy?” (CM 3), may be easy to endorse because it did not specify the content. The respondent may respond to the item as if to calm a student in some general situations without any challenges.
When the item includes an impediment, it will raise the difficulty level of the item (Bandura, 2006). For instance, the previous item could be expanded as “How much can you do to calm a student who is constantly disruptive or noisy?” or “How much can you do to calm a defiant student’s disruptive behaviors?”

**Measurement Invariance of TSES**

The findings support that the TSES is invariant in terms of school locations and years of experiences because these two factors were not significant to differentiate teachers’ levels of sense of efficacy. However, the interaction effects shown in Panel B of Figure 4 between teacher efficacy items and the levels of EE (very high vs. low) are significant to the teacher efficacy studies and warrant further discussion. The results in this study indicate a nonlinear relationship between teachers’ sense of efficacy and emotional exhaustion, which adds a new understanding about the relationship between these two variables established in the literature. It appears that teachers’ emotional states might influence their judgments about their abilities in teaching tasks, particularly for those teachers who reported the highest levels of emotional exhaustion.

**Figure 3.** Variable map for explanatory variables.

Note. The rating categories in measuring emotional exhaustion ranged from very low (never, a few times a semester), low (a few times a month), medium (once a week), high (a few times a week), to very high (almost daily).
The mirroring interaction effects also raise questions about why these two groups of teachers interpret some of the TSES items (IS1, CM2, and CM3) differently. The classroom context may account for how teachers envisioned the classroom situations and how they interpreted these items while responding. For example, for IS1 (use a variety of assessment strategies), teachers with high EE might have been situated in a more stressful, high-stakes testing environment,
which made them more vulnerable, and they felt less efficacious about using a variety of assessment strategies.

The opposite directions of the two CM items (low on CM2 and high on CM3) in the high EE group are also particularly interesting. Although both CM2 (control disruptive behavior) and CM3 (calm a student who is disruptive) capture the same construct of classroom management, the opposite responses suggest that these items might not function the same for these highly exhausted teachers. One speculation for such discrepancy is that these items measure different scopes of classroom management. CM3 measures teachers’ efficacy in CM to a limited scope in which one deals only with “calming” “a” student one at a time. When responding to CM3, teachers might be envisioning students whom they can rather easily redirect. CM2 measures teachers’ efficacy in CM in a broader scope in which they need to “control” disruptive “behaviors” and which may refer to multiple students at a time. Previous research has established that teachers’ EE is highly related to CM issues (Chang, 2009). Thus, it is not surprising that highly emotionally exhausted teachers rated low in controlling disruptive behaviors. However, to improve the quality of the item in terms of ensuring that it is invariant, as we suggested earlier, CM3 could be modified by specifying the context as “How much can you do to calm a student who is ‘constantly’ disruptive or noisy?” or “How much can you do to calm a ‘defiant’ student’s disruptive behaviors?” This change will not only increase the difficulty levels of the scale but also make the item more invariant across groups.

Additional research employing techniques such as cognitive interviews may further explore the interaction effects so that we may fully understand how to interpret such results.

**Implications for Future Research**

This study contributes to our understanding of the psychometric quality of a scale for measuring teachers’ self-efficacy at the item level, and the measurement invariance of the items across several explanatory variables. The measurement of this construct with a strong model-based theory, such as the Rasch measurement theory, strengthens the validity of the score interpretations by examining the gradations of challenges of the items and the invariance of the items. To more closely capture the efficacy levels among individuals, future researchers should consider adopting the item response model scoring techniques that Klassen and Chiu (2010) used in their study when using the TSES. Additional research is also needed on the number of categories used in the rating scale for TSES. We collapsed the lower categories and found the skewness and kurtosis statistics were improved significantly. Future research should consider dropping the lowest ratings when adopting the TSES.

**Declaration of Conflicting Interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**

The author(s) received no financial support for the research, authorship, and/or publication of this article.

**References**


