MA Kang, Robert TREVETHAN, LU Shuhong

Measuring Teacher Sense of Efficacy: Insights and Recommendations Concerning Scale Design and Data Analysis from Research with Preservice and Inservice Teachers in China

© The Author(s) 2020, corrected publication February 2020.

Abstract In this study, issues concerning the design of scales for measuring teacher sense of efficacy (TSE) are first identified with particular attention to the Teacher Sense of Efficacy Scale (TSES). Psychometric issues concerning analysis and reporting of TSE data are subsequently identified. Recommendations are offered about all identified issues, and these recommendations are taken into account when obtaining and analyzing TSE data from Chinese mainland preservice and inservice teachers. Exploratory factor analyses yielded a single factor for both samples as well as for four subgroups within the inservice teacher sample. Results also provided insights about scale design as well as the TSES being limited for capturing the breadth of TSE. Suggestions are made for improvements in the assessment of TSE.

Keywords teacher sense of efficacy (TSE), teacher self-efficacy, Teacher Sense of Efficacy Scale (TSES), scale design, response rate, factor analysis, China

The original version of this article was revised due to a retrospective Open Access order.

MA Kang
Department of Education, Southwest University, Chongqing 400715, China

Robert TREVETHAN (✉)
Independent academic researcher and author, Albury, NSW 2640, Australia
E-mail: robertrevethan@gmail.com

LU Shuhong
Faculty of Education, Shandong Normal University, Jinan 250014, China

Introduction

Teacher sense of efficacy (TSE) commonly refers to the beliefs teachers hold about the influence they can exert on learning outcomes and the behavior of students (see Tschannen-Moran & Woolfolk Hoy, 2001), even for students who are unmotivated or difficult (Armor et al., 1976; Guskey & Passaro, 1994). Research concerning this construct has occurred within two broad eras, the first spanning the final quarter of the 20th century, and the second occurring within the first two decades of the 21st century. The initial era might be characterized as a vigorous array of activities in which theorists and researchers attempted to identify the most effective ways to conceive of and measure TSE. Tschannen-Moran, Woolfolk Hoy, and Hoy (1998, p. 202) referred to this era as a “celebrated childhood” followed by an “identity crisis of adolescence” concerning TSE.

When the second era commenced, it was described by Tschannen-Moran et al. (1998) as being on the “verge of maturity” (p. 202) and by Henson (2002) as having prospects for moving to adulthood. A major event early in this second era was publication of the Ohio State Teacher Efficacy Scale (OSTES; Tschannen-Moran & Woolfolk Hoy, 2001), subsequently renamed by its originators as the Teachers’ (or, more recently, *Teacher*) Sense of Efficacy Scale (TSES).¹ This second era has been characterized by predominant use of that scale or close variants of it, including translations. The variants have exhibited a variety of instructions, number and wording of items, number and nature of response options, and groupings of items in analyses (see, e.g., Charalambous, Philippou, & Kyriakides, 2008; Cheon, Reeve, Yu, & Jang, 2014; Cheung, 2008; Duffin, French, & Patrick, 2012; Fackler & Malmberg, 2016; Fives & Buehl, 2010; Kennedy & Hui, 2006; Klassen & Chiu, 2010; Nie, Lau, & Liau, 2012; O’Neill & Stephenson, 2012; Pfitzner-Eden, Thiel, & Horsley, 2014; Poulou, 2007; Ruan et al., 2015; *Teacher*, n.d.; Turner, Jones, Davies, & Ramsay, 2004).

Other scales that could be regarded as alternatives to the TSES have been proposed, for example by Bandura (2006), Chan (2008a, 2008b), Friedman and

¹ Several variations of this scale’s name exist, including *Teacher Self-Efficacy Scale*. We use the singular word *teacher* in *Teacher Sense of Efficacy Scale* as an attributive noun because of its simplicity and lack of ambiguity as well as its having been used by Tschannen-Moran concerning the scale on her website (<http://wmpeople.wm.edu/site/page/mxtsch/researchtools>), where the most recent versions of the scale, including translations, are available.

Kass (2002), and Skaalvik and Skaalvik (2007), but, whatever their merits, these scales have attracted little empirical attention. Effort expended in assessing the attributes of the TSES and considering ways in which it might be improved therefore seems warranted and is one of the major aims of this study.

Within the introduction of this article, we initially provide background information concerning the origin, importance, associated issues, and scope of research concerning TSE. We then describe the TSES, raise issues related to its design and the design of similar scales, and canvass a number of psychometric issues, first, by investigating the factor structure of the TSES and scales that are strongly affiliated with it, and then by describing a range of problems associated with analysis and representation of data concerning TSE. Within this introduction, we also embed summaries, comments, and recommendations about the design of scales and analyses appropriate to measurement of TSE. As the final part of the introduction, we provide background that is specific to the empirical component of this study.

Research about TSE: Origin, Importance, Issues, and Scope

Empirical research concerning TSE commenced in the 1970s when David Armor and his colleagues used two questions to investigate facilitators of students' reading and found that TSE was one of those facilitators (Armor et al., 1976). Since then, TSE has increasingly featured in theory and research as a variable that appears to influence the motivations, behavior, and accomplishments of both teachers and their students in a variety of contexts. A major division within this research has been between the investigation of preservice teachers (PSTs) and inservice teachers (ISTs). Different sets of TSE-related attributes and outcomes have been claimed for these two groups. According to Chesnut and Burley (2015), high levels of TSE tend to increase PSTs' commitment to teaching as an occupation, something that is indicated by their intention to remain in the profession longer after graduation (Bruinsma & Jansen, 2010; Pendergast, Garvis, & Keogh, 2011). This intention could have important consequences for reducing the rate of novice teachers abandoning the profession (Hong, 2010; Pfitzner-Eden, 2016).

For ISTs, the range of proposed attributes and outcomes is broader, with a variety of authors and researchers referring to teachers who have higher levels of

TSE exhibiting a range of positive attributes. These include greater enthusiasm and more effort put into teaching; higher job commitment and satisfaction; higher goals and aspirations; greater motivation for learning and willingness to reflect on their own teaching; a stronger belief that they can influence how well students learn; greater openness to new ideas and innovations; more persistence in the face of difficulties; higher levels of effort in classroom planning and organization; more time spent delivering content; greater involvement in students' learning experiences, greater encouragement of students, responsiveness to the needs of students, use of positive strategies (including peer cooperation) to support students, and understanding of mistakes made by students; more time working with students who encounter learning difficulties instead of referring them for special education; greater use of adaptive student-teacher interaction techniques as well as humanistic and positive approaches to achieve or maintain desirable classroom behavior; greater creation of a high sense of efficacy in students as well as higher motivation and academic achievement; greater inclination to share experiences with colleagues; and less job-related stress, burnout, and attrition from the profession (see Alt, 2018; Berg & Smith, 2016; Chesnut, 2017; Cheung, 2008; Duffin et al., 2012; Fackler & Malmberg, 2016; Fives & Buehl, 2010; Ho & Hau, 2004; Hoy & Woolfolk, 1993; Kennedy & Hui, 2006; Klassen & Chiu, 2010; Nie et al., 2012; O'Neill & Stephenson, 2012; Perera, Calkins, & Part, 2019; Pfitzner-Eden et al., 2014; Skaalvik & Skaalvik, 2007; Woolfolk Hoy & Burke Spero, 2005).

Although the array of findings is impressive, the field has been beset by a number of problems, substantial identification of which first occurred two decades ago by Tschannen-Moran, Woolfolk Hoy, and Hoy (1998). Since then, authors such as Henson (2002), Wheatley (2005), Klassen, Tze, Betts, and Gordon (2011), and Malinen (2016) have provided elaborated accounts of these problems and have identified additional problems. In reviewing TSE research from 1986 to 2009, Klassen et al. (2011) summarized the problems as falling into four groups: insufficient research concerning the sources of teacher efficacy, unresolved issues concerning the measurement of TSE, insufficient evidence concerning associations between TSE and student outcomes, and insufficient investigation into how TSE might be related to actual educational practice.

We embarked on this study with four major aims, all related to the second group of problems identified by Klassen et al. (2011), namely the measurement

of TSE. Specifically, these aims involved identifying problems within the design of scales intended to measure TSE, attempting to overcome those design problems when creating scales for measuring TSE, investigating the outcomes in relation to participant responses on the scales we had created, and applying appropriate psychometric procedures to explore the nature of TSE.

A subsidiary aim in our research was to start describing data obtained from PSTs and ISTs on revised scales within the context of Chinese mainland. We believe these data could provide a valuable foundation for examining efficacy perceptions of Chinese PSTs and ISTs. These perceptions could be of particular interest because of educational features in China such as the range of responsibilities carried by teachers; expectations held by teachers in relation to students and their families; relationships between teachers, students, and parents; opportunities for and encouragement of collegial feedback and support; differential funding for schools; and ways of interacting with authorities (see Cheung, 2008; Darling-Hammond et al., 2017; Gao & Watkins, 2002; Ho, 2004; Ho & Hau, 2004; Organisation for Economic Co-operation and Development, 2011; Paine & Ma, 1993). Most of these features appear to be influenced by the Confucian/collectivist traditions and philosophy that infuse China's culture and education system (see Chan & Rao, 2009; Gu, Chen, & Li, 2015; He & Miao, 2006; Malinen, 2016; Pratt, 1992; Song, Zhu, & Liu, 2013; Watkins, 2000).

Our interest was heightened because of a claim by Malinen (2016) that there is little empirical research concerning TSE in the Chinese mainland within the English-language literature. In order to examine this claim, which is supported by Kleinsasser (2014), we conducted an ERIC database search limited to items in English that had become publicly available from the start of 2010 until mid-2018 about TSE in China. We used keywords and combinations comprising (*teacher* sense of efficacy* or *teacher* efficacy* or *teacher* self-efficacy*) and [*China*], examined reference lists within the items that materialized, and noted additional relevant items as we researched this topic. We located only 19 empirical studies from our search. Most of these are limited in scope in that six (Feng & Wang, 2014; Lu, Hao, Chen, & Potmėšilc, 2018; Malinen, 2013; Malinen et al., 2013; Malinen, Savolainen, & Xu, 2012; Wang, Zan, Liu, Liu, & Sharma, 2012) are primarily concerned with teaching students who have special needs; four (Manzar-Abbas & Lu, 2015; Sang, Valcke, van Braak, & Tondeur, 2010; Shi, 2014, 2016) with teaching related to specific areas such as Chinese, mathematics, and information technology; another four (Dou, Devos, & Valcke, 2017; Tian,

2011; Yin, Lee, Jin, & Zhang, 2013; Yu, Wang, Zhai, Dai, & Yang, 2015) with topics that do not involve classroom activities; four (Hoi, Zhou, Teo, & Nie, 2017; Malinen, Savolainen, & Xu, 2013; Ruan et al., 2015; Zheng, Yin, & Li, 2019) with scale development; and one (Yuan & Zhang, 2017) with commitment to teaching as a profession. Only one study (Ruan et al., 2015), in which the primary focus is scale development, includes general aspects of face-to-face teaching. Of the 19 studies, 14 are based on data from ISTs, one on data from both PSTs and ISTs, and only four solely on data from PSTs. The present study could, therefore, provide the foundations for a relatively substantial contribution to English-language publications about TSE in China.

The Teacher Sense of Efficacy Scale

Tschannen-Moran and Woolfolk Hoy (2001) created the TSES in a series of procedures involving extensive examination of conceptualizations and research concerning TSE; several stages of item generation, selection, and revision; progressively refined factor analyses; and assessment of internal consistency and some aspects of construct validity. The outcome of these procedures, based on a final sample comprising both PSTs and ISTs, was a set of 24 items couched as questions, for example *How much can you do to get through to the most difficult students?* The 24 items were designed to comprise three factors, each with eight items, and a 12-item short form was created by selecting the four highest-loading items from the three long-form factors. Tschannen-Moran and Woolfolk Hoy (2001) claimed that their scale had a “unified and stable factor structure” (p. 801). The three factors have conventionally been regarded as comprising three moderately correlated but independent domains, characterized as engaging students (engagement), instructional strategies (instruction), and managing student behavior (management).

The original versions of both the long and short forms were subsequently altered with different instructions, item sequence, and wording above the response options.² However, both forms of the TSES have nine response options. The most recent version of both the long and short forms has option extremes of

² The original version of the TSES is available at <http://stelar.edc.org/sites/stelar.edc.org/files/TSES-%2B-scoring-zted8m.pdf>. Given text within the Tschannen-Moran and Woolfolk Hoy (2007) publication, the instructions were altered within five years of the publication in which the TSES first appeared.

None at all (1) and *A great deal* (9), with intermediate odd-numbered options labeled *Very little*, *Some degree*, and *Quite a bit*. Even-numbered response options are not labeled.

The TSES is usually scored by calculating the mean of all unweighted responses under consideration, thus producing scores that, theoretically, can range from 1 to 9. Most commonly, three separate scores are calculated, each representing one of the scale's three domains. However, a composite score is often produced, usually in addition to, but sometimes instead of, individual domain scores.³ In practice, the way in which scores on the TSES and its close variants are represented is inconsistent and can depend on the nature of the data obtained,⁴ the calculations that the data are subjected to,⁵ and the beliefs, usually unstated, of researchers about whether or not scales can simultaneously measure separate domains as well as an overarching construct.

Duffin et al. (2012) regarded the TSES to be the most promising scale for assessing TSE because of its alignment with both Bandura's theory of self-efficacy (Bandura, 1997) and recommendations of other theorists about TSE. Furthermore, Duffin et al. (2012) stated that the TSES had become the dominant means of assessing TSE among PSTs.

The popularity of the TSES, whether for use with PSTs or ISTs, appears to be

³ Some researchers (e.g., Pfitzner-Eden et al., 2014; Wang, Hall, & Rahimi, 2015; Wolters & Daugherty, 2007) provide separate scores for each domain, but not a composite score, presumably in accordance with the principle that a composite score is inappropriate if its components are "conceptually meaningful and empirically useful" (Briggs & Cheek, 1986, p. 129). Other researchers (e.g., Fives & Buehl, 2010; Heneman, Kimball, & Milanowski, 2006; Pendergast et al., 2011; Tschannen-Moran & Woolfolk Hoy, 2001; Yousuf Zai, 2016; Yousuf Zai & Munshi, 2016) provide individual domain scores, therefore indicating that they regard the TSES as comprising separate domains, but they also provide a composite score. It is therefore unclear whether they regard TSE to be a multidimensional or unified construct. Sometimes only a composite score is sought when factor analysis has indicated a single factor is present in the data (see, e.g., Fives & Buehl, 2010), but sometimes separate domain scores are sought *despite* factor analysis having indicated the presence of only one factor (see Duffin et al., 2012).

⁴ The nature of the data can be affected by such things as the number of response options provided to participants. For example, Nie et al. (2012) and Poulou (2007) provided only five response options and therefore obtained lower means than are usually obtained on the TSES. Furthermore, the full complement of TSES items is not always used in analyses. For example, Klassen and Chiu (2010) removed one item from the engagement domain, thus almost inevitably obtaining lower scores on that domain relative to the other two domains in some of their analyses.

⁵ Some researchers, for example, Klassen et al. (2009), Klassen and Chiu (2010), and Wang et al. (2015) provide total (summed), not averaged, scores for all, or subsets, of their data.

robust. For example, within the journal *Teaching and Teacher Education*, TSE was a major variable in 14 empirical articles during 2018, and not only was the TSES used in eight of those studies, but no other scale for measuring TSE was used more than once among the remaining six studies. Within the same journal for the three years spanning 2016 to 2018 (inclusive), no fewer than 26 articles were based on the TSES or a close adaptation of it. Careful examination of at least some aspects of the TSES and highly similar scales therefore seems warranted, particularly because Tschannen-Moran and Woolfolk Hoy (2001) acknowledged that their scale needed “further testing and validation” (p. 802) and generously waived copyright restrictions on it when used in scholarly research, tacitly inviting further development.

Comprehensive and detailed examination of measurement issues concerning the TSES and related scales would require more than a single journal article. We therefore limit our examination within this article by considering what we regard to be two major areas of concern related to TSE scales. The first area refers to the design of those scales, particularly why their design might contribute to the perception by a number of researchers that measurement of TSE is problematic. The second area concerns a range of issues that are more psychometric in nature.

Issues Concerning Design of Scales for Measuring TSE

One of the fundamental principles of scale design is that a scale should appear, even on brief initial inspection, to be capable of measuring its target construct effectively. This is referred to as face validity (Boateng, Neilands, Frongillo, Melgar-Quinonez, & Young, 2018; Trevethan, 2009). Without this form of validity, respondents may ignore or choose not to respond to items (Streiner & Norman, 1995) or they may not take the task of completing a scale seriously and may therefore respond in hasty, careless, haphazard, or misleading ways (Murphy & Davidshofer, 1998), sometimes referred to as “satisficing” (Krosnick & Presser, 2010, p. 265).

Although face validity usually refers to whether a scale *appears* to measure what it is intended to measure, an equally important aspect of face validity is whether respondents feel confident, when they encounter a scale, that it is likely to measure something successfully. A scale needs to make sense to, and elicit respect from, prospective respondents. Establishing initial rapport with

respondents is therefore an important component of face validity (Trevethan, 2009).

For PSTs, there is an immediate problem concerning face validity in the latest version of the TSES because they are asked to “respond to each of the questions by considering the combination of your *current* ability, resources, and opportunity to do each of the following in your present position” (italics in original; all text bolded). This request, with three explicit indications that it focuses on a presently experienced within-school context (*current* resources, *current* opportunity, and *present* position), is clearly intended for ISTs, so the “present position” of PSTs, as students, might be irrelevant, particularly if they have not yet had any, or have had only minimal, practicum experience. That discrepancy could arouse suspicions about the validity and appropriateness of the scale. The earlier version of the TSES was not as obviously focused on ISTs in that it asked respondents to “indicate your opinion about each of the statements below,” but, in terms of making sense, respondents were confronted with the problem that the items comprised neither statements nor opinions. They were questions. Respect for the scale could therefore have been immediately challenged on that version simply because of inappropriate word choice.

These problems (the mismatch between instructions and reality for PSTs on the latest version of the TSES, and the mismatch between the instructions and the nature of the items for both PSTs and ISTs on the earlier version) might be neither overlooked nor trivial for many respondents and could immediately incline both PSTs and ISTs to doubt the validity of the TSES. For PSTs, confidence in the scale could be reduced further because some items, including three that refer to “your students,” are clearly intended for ISTs.

Added to the above, a major set of requirements for face validity indicates that items should be succinct, unambiguous, and grammatically correct (Portney & Watkins, 2009). The TSES does not fully satisfy the first two of these requirements. First, the items are not succinct in that 12 of the 24 commence with *How much can you*, seven with *How well can you*, and the remaining five with *To what extent can you*. This repetition deprives the scale of crispness and could easily have been avoided, and, furthermore, might disincline people from continuing to respond to the scale even if they had not noticed, or had decided to overlook, deficiencies in the initial instructions.

Second, a problem of ambiguity arises with the words *most difficult* in the first

item: *How much can you do to get through to the most difficult students?* Similar wording on the TSE scale of Gibson and Dembo (1984) led Henson (2002) to ask: “How difficult are the students? Is difficulty related to behavior, motivation, or instruction, or to all of these? Are learning disabilities relevant to the situation?” (p. 140). Fives and Buehl (2010) expressed similar concerns. For some respondents, a conceptually difficult first item such as this might generate, or provide further grounds for forming, a negative impression of the TSES.

Although the requirement of grammatical correctness is satisfied on the TSES, a more subtle issue arises for astute respondents (which, it might be anticipated, many PSTs and ISTs are). None of the items can be responded to appropriately by respondents who want to choose the first, and probably initially inspected, option of *None at all* on the most recent version of the TSES. More explicitly, following the question *How much can you do to get through to the most difficult students?* a response of *None at all* is patently incongruent. On the original version of the TSES, the first response option of *Nothing* was meaningful as a response to all 24 items, but the fifth option of *Some influence* would have been an incompatible response to most items. With both versions of the scale, therefore, a number of respondents might react with disrespect because of the response options even if they had not noticed problems with the instructions and item succinctness. These problems might again appear trivial, but the consequences of researchers not avoiding them could be significant for respondent compliance and commitment (Blair & Zinkhan, 2006; Wyatt, 2014).

The TSES could also be problematic in terms of the number of response options. Although no definitive conclusions have been established about the number and type of response options within Likert-type scales (see Portney & Watkins, 2009), there is evidence that between five and seven options are desirable (see Lozano, García-Cueto, & Muñiz, 2008; Robinson, 2018), and Carifio and Perla (2007, p. 111) have claimed that seven response points are preferable, as have Krosnick and Presser (2010, p. 272). Some researchers have argued that they will acquire a greater spread of scores, and therefore greater sensitivity, with a larger number of response options (see Bandura, 2006; Durksen, Klassen, & Daniels, 2017). However, Nunnally (1978) pointed out that scales with seven response options reach the limit of reliability. Furthermore, Darbyshire and McDonald (2004) have argued that a major consideration should be what appeals to respondents and that more than seven options might be seen

by respondents as requiring too fine-tuned decision making.⁶ Similarly, Jacoby and Matell (1971) argued that respondents should feel comfortable with a scale, and respondents might perceive more than seven response options as lack of consideration on the scale developers' part.⁷ Prospective participants might therefore be less inclined to respond to a scale, or to accord it less thoughtful responses, than would otherwise have been the case.

Furthermore, Portney and Watkins (2009) have indicated that response options such as *Don't know* and *Unsure* should always be provided for respondents. The absence of these kinds of option on TSE scales, including the TSES, could be particularly problematic for PSTs, many of whom might be unable, or unwilling, to anticipate how effective they would be in a teaching environment and would therefore appreciate having an "escape" option.

The above deficiencies in the TSES and similar scales could incline participants to regard such scales as not having been competently created and therefore as undeserving of respect to the point that those prospective participants refuse to take part in the research or, if they take part, provide hasty or even sabotaging responses. The consequences for data quality, as well as for desirable response and retention rates, can probably be disregarded only at peril. Although a number of variables could influence response and retention rates, and an attendant problem is that those rates can be defined in different ways, it is telling that TSE research is typically unimpressive in terms of those rates. For example, Pendergast et al. (2011) invited PSTs, during lectures, to participate in a survey that included the TSES. The researchers reported an initial response rate of 63 %, but on a second administration of the scale only 43 % of the original participants provided data, resulting in a participation rate for both occasions of only 27 %. Swan, Wolf, and Cano (2011) invited 17 novice ISTs to participate in online surveys that included the TSES across the first 3 years of their teaching experience. Of those 17 ISTs, despite anticipatory and follow-up emails

⁶ Poulou (2007) indicated explicitly that she found PSTs preferred five response options when she was adapting the TSES.

⁷ Perhaps out of consideration for participant preferences, a number of researchers have used fewer than seven response options on their TSE scales. In addition to Poulou (2007), for example, Nie et al. (2012), Ottley et al. (2015), and Stipek (2012) used five options. Hatlevik (2017) used five and six options in two different scales within her research, and other researchers have used six options (see, e.g., Gavora, 2010; Moore-Hayes, 2011). Still other researchers have based their research on scales with only four options when measuring TSE (see, e.g., Overbaugh & Lu, 2008; Vieluf, Kunter, & van de Vijver, 2013), and Shi (2014) based her research on a scale that had only three response options.

encouraging participation, only three provided data on all three occasions. Pfitzner-Eden (2016) administered an adaptation of the TSES to two groups of PSTs on three successive occasions, with administrations at Times 1 and 2 occurring in seminars, and at Time 3 online. Initial response rates could not be definitively determined but appeared to be 84 % and 83 % for the two groups, respectively. The retention rates at Time 2 were 60 % and 72 %, and at Time 3 were 50 % and 45 %. However, retention rates across the whole study period were only 32 % and 36 %. Pfitzner-Eden referred to other research in which retention rates were lower. Under these circumstances, the internal validity of longitudinal research could easily be called into question (see Campbell & Stanley, 1966), and conclusions based on data collected on a single occasion could also be suspect because of uncertainty concerning the characteristics of those who did, and those who did not, take part in the research.

Summary, Comments, and Recommendations

Even a cursory inspection of the TSES reveals it to have several design flaws, some of which also characterize other scales intended to measure TSE. These problems highlight the need for instructions, items, and response options to “make sense” to, and be appropriate for, intended respondents. Furthermore, there should be congruence between instructions, items, and response options, care should be taken with semantics and word choice, perplexing items should be avoided, and overall construction and wording should be sparing. All options that might be needed by respondents should be provided and care should be taken to ensure that the number of response options is most likely to yield data that are psychometrically valid. Lack of attention to these elements of scale design could contribute to poor response and retention rates—features that seem to be endemic in research about TSE—as well as to poor quality of data because participants satisfice rather than provide thoughtful responses on scales. Researchers should be encouraged to alter scales if they believe that alterations would improve scale presentation, increase response and retention rates, and enhance data quality.

Psychometric Issues Associated with the Measurement of TSE

A number of psychometric issues pertain to the measurement of TSE. After considering issues associated with the factor structure of the TSES, we broaden

our discussion to issues of a more general nature that relate to factor identification and representation, and we follow that with consideration of issues concerning inadequate and inappropriate reporting of findings.

Factor Structure of the TSES and Close Variants

In this subsection, factor analytic results from research based on responses from both PSTs and ISTs on the TSES and some close variants are presented. We have attempted to provide an accurate representation of these results, but in order to avoid obscuring the most important points, attention will not be drawn to some details unless there is reason for doing so. Therefore, we will not refer to minor alterations made to the original versions of the TSES. However, at times we will refer to whether analyses were based on the long or short form of the TSES, whether exploratory or confirmatory factor analyses had been conducted, and whether major alterations had been made to item wording, number of items, and wording and number of response options. The findings, which are based on data from a number of different countries, will be presented first for data from PST samples, then for data from IST samples.

Factor Analytic Results from PSTs. Factor analyses of the TSES and close variants conducted with PSTs fall into three groups. Within the first group, three distinct factors have emerged (see Pfitzner-Eden et al., 2014; Poulou, 2007; Yousuf Zai, 2016). In the second group, two factors have been found, with one factor sometimes comprising a combination of the engagement and instruction items and the other factor comprising only management items (see Turner et al., 2004); however, sometimes the two factors comprised a mixture of the three purported TSES domains (see Mergler & Tangen, 2010) or items needed to be deleted to attain a clear solution (see Charalambous, Philippou, & Kyriakides, 2008). In the third group, the data comprised a single factor. This was first found by Tschannen-Moran and Woolfolk Hoy (2001) when they developed the TSES, thus supporting their claim that the TSES has a unified factor structure. Since then, other researchers have also found a single factor among PSTs' responses (see, e.g., Duffin et al., 2012; Fives & Buehl, 2010; O'Neill & Stephenson, 2012).

Factor Analytic Results from ISTs. Factor analytic results from ISTs have also

been varied, but much less clearly demarcated. Some researchers (e.g., George, Richardson, & Watt, 2018; Klassen et al., 2009; Klassen & Chiu, 2010; Tschannen-Moran & Woolfolk Hoy, 2001) found three factors within the TSES. However, Kennedy and Hui (2004) found only two factors (a combination of engagement and instruction items as one factor, and management items as a separate factor) but also obtained a single-factor solution. Tsui and Kennedy (2009) also found two factors, but one item moved to an unexpected factor and another item had low loadings that spanned both factors. In two separate pieces of research, Tschannen-Moran and Woolfolk Hoy (2001, 2007) obtained a single factor following second-order factor analyses, although there were some PSTs in the first of these studies. A single factor was also found by Tsigilis, Grammatikopoulos, and Koustelios (2007) as well as by Chang and Engelhard (2016).

Cheung (2008) obtained more discrepant results. She found that Shanghai ISTs' data comprised the two factors that Kennedy and Hui (2004) had found, but Hong Kong ISTs' data loaded on a single factor. However, when the data from the Shanghai and Hong Kong participants were combined, a single factor was produced, and, in earlier research (Cheung, 2006), found that the female Hong Kong ISTs' data loaded on a single factor whereas the males' data separated into two factors. Ruan et al. (2015), when surveying ISTs in China, South Korea, and Japan, found that a three-factor solution was not a good fit for the long-form TSES data in any of the three countries, but the short form yielded an acceptable fit comprising three factors in all three countries on the condition that one item (Item 11) was deleted. Heneman, Kimball, and Milanowski (2006) found that three factors characterized their data most effectively, but that the fit was "improved slightly by allowing some items to load on factors other than those they were intended to measure, or by allowing some correlated uniqueness of items within subscales" (p. 10).

Other researchers have identified three factors in data from ISTs, but only after alterations, sometimes substantial in nature, were made to the conceptualization or presentation of the TSES. For example, Yousuf Zai and Munshi (2016) found that an instruction item had the highest loading on the engagement factor. Wolters and Daugherty (2007) obtained results with exploratory factor analysis that initially indicated the presence of two, three, or four factors within their data. After deciding to conform with the conventional three-factor solution for the TSES, they found that six of the 24 items "did not load as expected, had weak

primary loadings ($< .50$), or had elevated crossover loadings ($> .40$)” (p. 185). Using confirmatory factor analysis with a different sample, they found that the original three TSES domains could be reproduced only by eliminating six items, thus leaving the engagement, instruction, and management domains with 4, 6, and 8 items, respectively, and with the first of these domains having a slightly different thematic focus from that of the domain on which it is conventionally based. Fives and Buehl (2010) moved items between the three TSES domains after they found that five items cross-loaded at > 0.40 and three additional items loaded on unanticipated domains. As a result, in contrast to the TSES having eight items on each domain, Fives and Buehl’s 24-item scale had nine items on one domain, 10 items on the second, and five on the third. In their research, the short form had a clearer factor structure, with only one item exhibiting a high cross loading (across the engagement and instruction domains). Nie et al. (2012) used five rather than nine response options, reworded the response options, reworded all 24 TSES items so that they were, in the researchers’ view, appropriate to a Singaporean context, removed three of the 24 items judged to be inappropriate for the engagement domain, and deleted a further nine items because of cross loadings and low loadings. The resulting 12-item inventory contained three clear factors, but a second-order analysis reduced the data to a single factor.

Summary, Comments, and Recommendations. The above factor analyses indicate that, for both PSTs and ISTs, the number and composition of domains that comprise TSE—at least as represented by the TSES and its close variants—are inconsistent, even fragile. Of particular interest in terms of the claim by Tschannen-Moran and Woolfolk Hoy (2001) that the TSES “has a unified and stable factor structure” (p. 801), even the data collected by those researchers themselves when the TSES was created, had clearly demonstrated that their claim of a unified factor structure was unjustified (there were three distinct factors in the ISTs’ data) and the factor structure differed between ISTs and PSTs (revealing that the factor structure was not stable). As indicated above, the lack of a unified and stable factor structure in the TSES and similar scales has been clearly demonstrated in other research, so the original claim by Tschannen-Moran and Woolfolk Hoy was not only anomalous and disconfirmed by their own data, but has also been unsupported by subsequent research.

That few researchers acknowledge this is puzzling, but the empirical evidence sends distinct signals that the factor structure of the TSES is by no means stable and that there should not be such a pervasive and little-questioned assumption that the TSES comprises three distinct factors. If distinct factors are not present in the data, conducting analyses as if there were separate domains is likely to lead to type 1 errors, particularly if no Bonferroni-type adjustment is applied to p values in null hypothesis statistical testing.

Although some of the inconsistencies in factor structure could result from genuine differences in the samples—a possibility that we pursue in the empirical component of the present research—some inconsistency could also be associated with psychometric issues concerning the way in which researchers identify and represent factors. This possibility is examined in the next section along with a number of additional psychometric issues related to factor analysis that appear to feature in the research about TSE.

Issues Concerning Factor Identification and Representation

Psychometric issues associated with factor identification and representation can be divided into two groups: first, sample size and determining the number of factors in the data, and, second, the way in which factors are consolidated. Each group of issues is considered separately in the following two subsections, usually with general relevance to factor analysis but also, at times, with specific reference to research involving the TSES and its close variants. Throughout this overall subsection, which we finish with a summary, comments, and recommendations subsection, we have attempted to provide information that is simultaneously authoritative and accessible. We acknowledge that more elaborated and sophisticated accounts are available elsewhere (we provide citations for a number of these), but our main aims are to encourage researchers of TSE to abandon some seemingly entrenched practices and, in their place, adopt better procedures and perspectives for measuring TSE. In order to realize that aim, we have included this information as a preface to the empirical component of this article, affording us the opportunity to demonstrate much of what we are advocating in a context that is relevant to researchers of TSE.

Sample Size and Determining the Number and Nature of Factors in Factor Analyses. Because much research about the TSES and its close variants has

focused on exploratory factor analysis, some of the basic and recommended principles regarding that procedure are worth noting.⁸ Although a mixture of informed practices and subjective judgements is often required on the part of researchers when they use factor analysis, a number of moderately specific guidelines as well as rules of thumb have become generally endorsed over the last three decades (see Gaskin & Happell, 2014; Hair, Black, Babin, & Anderson, 2014; Matsunaga, 2010; Yong & Pearce, 2013). Several of these appear to be unfamiliar to, or disregarded by, researchers who investigate TSE.

One of the most important principles is that factor analysis usually requires large samples to produce valid and stable results (Nunnally & Bernstein, 1994). Recommendations about sample size vary, however. Although Pallant (2016) has proposed that a sample size only five times the number of items might be sufficiently large for some factor analyses, Comrey and Lee (1992) described samples with fewer than 50 people as very poor, samples smaller than 100 as poor, and samples with 200 participants as only fair. Dixon (2005) recommended that the sample size for factor analysis should be at least 10 times the number of items involved in the analysis; Hair et al. (2014) recommended, similarly, that the cases-per-variable ratio should be as high as possible, and Tabachnick and Fidell (2007) recommended that there be at least 300 people in the sample. Other, more complex, methods for determining a satisfactory sample size exist. These methods are based on the size of communalities, number of items in the factors, and the size of loadings (see Gaskin & Happell, 2014; Hogarty, Hines, Kromrey, Ferron, & Mumford, 2005; Mundfrom, Shaw, & Ke, 2005). However, an unnecessarily large sample might need to be recruited before those parameters can be assessed. Despite acknowledging these complexities, inspection of TSE research indicates that participant numbers are often unsatisfactorily low for factor analysis, even when only 12 items are involved. The likelihood of poor validity as well as instability in results is therefore high.

Additionally, appropriate methods should be used to determine the number of factors in a data set. Several methods exist for this process (see Fives & Buehl, 2010; Gaskin & Happell, 2014; Hair et al., 2014), and seasoned researchers will

⁸ Confirmatory factor analysis, in which researchers assess the accuracy of their previously established expectations about the number and composition of factors in a data set, is considered briefly later this article. It appears to be used much less frequently than is exploratory factor analysis in research concerning TSE, but is no less valuable.

often run several analyses of different kinds before making a decision. The two most commonly used methods at the initial stage of factor analysis are based on the Kaiser criterion and the scree test. The Kaiser criterion, which refers to the number of eigenvalues greater than 1, is renowned for suggesting the presence of more factors than really exist in the data (Preacher & MacCallum, 2003; Zwick & Velicer, 1986) and, according to Briggs and Cheek (1986), many factor analysts “spurn this procedure” (p. 119). Despite this, use of the Kaiser criterion is commonplace when researchers factor analyze TSE data. The scree test is usually considered preferable to the Kaiser criterion and has been demonstrated to provide an accurate indication of the number of factors in a data set most of the time (Cattell & Vogelman, 1977; Costello & Osborne, 2005; Tzeng, 1992), and more sophisticated and complex techniques involving parallel analysis or similar procedures are recommended if the scree test is indeterminate (see Crawford et al., 2010; Hayton, Allen, & Scarpello, 2004). Nevertheless, the scree test appears to be used infrequently in TSE research and parallel analysis is extremely uncommon, and, as a result, there is a likelihood that many researchers attempt to extract an inappropriate number of factors, usually too many, from their data. Parallel analysis is not available in major statistics packages, but the procedure is not so daunting that researchers should avoid it.⁹

When two or more factors appear to be present in the data, an additional requirement is that the factors be rotated in a way that draws out their distinguishing features most accurately, informatively, and meaningfully. Orthogonal, usually varimax, rotations are frequently used, presumably because these rotations increase the likelihood of separate factors being extracted (Fabrigar, Wegener, MacCallum, & Strahan, 1999; Portney & Watkins, 2009). However, orthogonal rotations increase the possibility of separate factors *appearing* to be present in the data, rather than *actually* being present. Furthermore, orthogonal rotations are based on an assumption that the domains being analyzed are uncorrelated, an assumption that is often unfounded (Fabrigar et al., 1999; Gaskin & Happell, 2014; Preacher & McCallum, 2003)—and it is obviously untenable with regard to the TSES given the moderate to high correlations that are typically obtained among its subscales. For example, on the original version of the TSES, Tschannen-Moran and Woolfolk Hoy (2001)

⁹ Helpful information is provided by Hayton et al. (2004), Matsunaga (2010), and Çokluk and Koçak (2016).

reported correlations between subscales ranging from 0.58 to 0.70. These correlations cannot be regarded as indicating a lack of association. Nevertheless, varimax rotations are common in TSE research. Oblique rotations, which more effectively allow for domains being correlated, appear to be unattractive to researchers despite their greater validity. Because of the inappropriate use of orthogonal rotations when analyzing the TSES, the production of distinct factors could well be appealing, particularly for researchers who are attracted to categorization, but the individuality of those factors might be specious.

Consolidation of Factors. In conjunction with, and following, the initial identification of factors within a set of data, a number of issues arise concerning the way in which those factors are consolidated. The main issue pertains to the size of item loadings within factors. Nunnally and Bernstein (1994) have argued that, although it is desirable to have some items with high loadings (at least 0.70), it is also desirable to have other items with loadings as low as 0.40 on the condition that loadings for the latter items are considerably lower on other factors. High-loading items most clearly define a domain, whereas moderate- to low-loading items enhance domain width and complexity, and therefore probably increase content validity, within the construct(s) being assessed (Hair et al., 2014; Panayides, 2013).

Despite these principles, TSE measurement is frequently deprived of domain breadth and depth because of what appears to be a preoccupation with creating or using scales that consist solely of items with high factor loadings. Tschannen-Moran and Woolfolk Hoy (2001) exhibited this preoccupation by selecting items with the highest loadings from the TSES long form when they created the short form, and other researchers have followed suit when they seem intent on providing reassurance that there are only high-loading items on the separate factor(s) in their research.

Summary, Comments, and Recommendations. Research involving the TSES and related scales could be so plagued by the application of inappropriate psychometric practices that inconsistency in findings concerning the factor structure of TSE is inevitable, and, as a result, a satisfactory conceptualization, or even reconceptualization of TSE is difficult to establish. Particularly when inappropriate practices in factor analyses are used in combination with each

other—for example, use of both the Kaiser criterion and orthogonal rotations, as is often exhibited by researchers in TSE—the possibility of misleading results concerning the number and composition of factors in a dataset is high (Fabrigar et al., 1999). Researchers should ensure that their sample sizes are sufficiently large if they intend to conduct factor analyses; abandon use of the Kaiser criterion for determining the number of factors in their data and replace it with the scree test or, if the scree test is indeterminate, parallel analysis; use oblique rather than orthogonal rotations; and consider retaining items if they have moderate or even low loadings (down, perhaps, to 0.40) as long as there are no cross-factor loadings.

Inadequate and Inappropriate Reporting

Apart from the issues related specifically to sample size and determination of factors within factor analyses, a number of additional psychometric issues are worthy of consideration because they appear to attract insufficient or inappropriate attention from researchers of TSE and therefore permit undesirable practices to thrive or constrain the perceptions and evaluations that consumers can form about data quality, the validity of analytical procedures, the value of research findings, and possibilities for improving research practices. Of concern within this section are four areas about which TSE researchers typically provide inadequate information. These areas relate first to information about response and retention rates as well as to circumstances of data acquisition, second to reporting of data attributes and data cleaning, third to reporting of several aspects concerning factor analyses, and fourth to inappropriate reporting of information about individual items and coefficient alpha.

Inadequate Reporting of Response and Retention Rates and Circumstances of Data Acquisition. A major set of problems concerns transparency associated with response and retention rates. The text of many publications suggests that all people who were asked to participate in a particular study agreed to do so. However, when response rates are provided, a 100 % response rate in research about TSE appears extremely unlikely. This is revealed by response rates not being impressive even when people are asked to participate under circumstances in which compliance might be high. For example, Klassen and Chiu (2010) reported obtaining a response rate approximating only 75 % when delegates at a

Canadian teacher conference were approached directly with a request to participate in research involving TSE, and Pfitzner-Eden et al. (2014) reported a similar response rate from two cohorts of German PSTs who were asked to complete surveys in seminars.

Difficulties in defining response and retention rates might provide some excuse for not reporting these rates. If a class had an enrollment of 100 students, but if only 75 students attended a lecture in which surveys were distributed and only 50 of the 75 students returned their surveys in a lecture one week later, and 15 of the 50 surveys had large amounts of missing data, there are several ways in which response and retention rates could be determined depending on which numbers were regarded as valid numerators and denominators. Most TSE researchers do not report any form of response or retention rates, and, if they do, the basis for calculating those rates is almost never indicated.

Furthermore, the circumstances under which people are asked to participate could influence research outcomes but might be insufficiently examined. Duffin et al. (2012) studied TSE in two ostensibly similar groups of PSTs. One group was provided with an information letter about the research and was asked to sign consent forms; the other group participated as a compulsory course requirement. An independent-samples *t*-test can be calculated from the published means, standard deviations (*SDs*), and numbers of participants in each group. This reveals noticeably higher TSE in the first group relative to the second, with means of 6.69 and 5.87, respectively, $t(450) = 7.00$, $p < 0.001$, $d = 0.66$. Within the research by Pfitzner-Eden et al. (2014), a group of New Zealand PSTs were invited to complete an online survey. In contrast to the response rate of approximately 75 % from German students in the same study, the response rate from the New Zealand students was only 33 %. However, the latter group had significantly higher TSE relative to the German PSTs who had completed the surveys in seminars (the difference is again calculable from information within the original publication). Although the results might have been influenced by nationality, age, and educational experience, the circumstances under which the data were collected might have played a role. In neither piece of research was any comment made, examination conducted, or speculation entertained about the group differences in TSE scores.

Inadequate Reporting of Data Attributes and Data Cleaning. Researchers seldom provide sufficient information about their data. In order to ensure that data are

likely to differentiate between participants, there should be some indication that standard deviations are not undesirably narrow and, possibly, that the data are not leptokurtic. Furthermore, information about skewness, often most effectively obtained from inspection of histograms, should also be provided to indicate the likelihood of outliers, which could subsequently be identified by boxplots.

Despite the importance of judicious data cleaning and appropriate procedures for handling of missing data (see “Data cleaning,” n.d.; Hellerstein, 2008; Osborne & Overbay, 2004; Van den Broeck, Cunningham, Eeckels, & Herbst, 2005), little mention is made within the TSE literature about procedures and outcomes relating to either data cleaning or missing data. In particular, despite calls for transparency concerning not only treatment of outliers but also different outcomes resulting from outlier adjustments (see Aguinis, Gottfredson, & Joo, 2013; Zijlstra, van der Ark, & Sijsma, 2011), mention of dealing with outliers is scant. This is particularly concerning given the strong negative skewness usually evident in composite and subscale scores obtained with the TSES and close variants such as the Scale for Teacher Self-Efficacy (Pfitzner-Eden et al., 2014) because that skewness creates a high likelihood of outliers at the low end of the distribution.¹⁰ This, in turn, raises the disturbing possibility that many results concerning TSE are distorted, perhaps deceptively so, and would be different if outliers had been dealt with appropriately.

Issues relating to missing data are likely to need addressing not only because of what appear to be characteristic poor response and retention rates in research about TSE but also because of nonresponses on specific items or strings of items. Missing data can have different attributes, for example the extent to which they are random or systematic, and they can be handled in a variety of ways, for example by imputing values or by using pairwise or listwise deletion (see Brick & Kalton, 1966), but little mention is made of any of these processes or the extent to which they were applied.

Inadequate Reporting of Data, Procedures, and Results in Relation to Factor Analysis. In association with factor analyses, sufficient reassurance is often not provided about a number of metrics that indicate the adequacy of the data,

¹⁰ The noticeable skewness and concomitant likelihood of outliers that is evident in most composite and subscale scores is almost inevitably a product of skewness on individual items. This can be seen in publications that include means and *SDs* of individual items (see, e.g., George et al., 2018; O'Neill & Stephenson, 2012).

procedures, and outcomes. In exploratory factor analysis, consumers of research should be assured of the absence of multicollinearity in the data (no interitem correlations > 0.90) as well as satisfactory results having been attained—even if specific details are not provided—on the Kaiser-Meyer-Olkin (KMO) index (preferably, > 0.80 but as low as 0.60) and Bartlett's test of sphericity ($p < 0.05$, but preferably $p < 0.001$). Furthermore, the range and mean of communalities should be reported to indicate the amount of variance that individual items share with other variables in the analysis. Although hard and fast rules do not exist, communalities < 0.50 might be regarded as unsatisfactory (Hair et al., 2014). Furthermore, in light of the results emerging from exploratory factor analysis, information should be provided about the number of eigenvalues > 1 , the size of the largest of those eigenvalues, and where the largest drop occurred among them so that readers are aware of the distinctiveness and likely number of factors. Information should also be provided about the percent of variance explained by the retained factors—an amount that ideally should exceed 40 % and preferably approach 70 % (Hair et al., 2014). Justification should be provided concerning choice of factor rotation, for example, that the absence of noticeable skewness or kurtosis permitted rotation with maximum likelihood rather than principal axis factoring (Fabrigar et al., 1999).

Although interitem correlations should have been inspected in order to identify undesirably high correlations in the data, there is seldom any indication about those correlations having been considered prior to factor analysis—a point made two decades ago by Fabrigar et al. (2009)—or about the correlations among items that are retained subsequent to factor analysis. This is perhaps surprising given that interitem correlations can easily reveal lack of association between items on the one hand, and item redundancy on the other. Clark and Watson (1995) have recommended that interitem correlations on scales should lie between 0.15 and 0.50, and Briggs and Cheek (1986) have recommended that the mean of those correlations should ideally lie between 0.20 and 0.40. Briggs and Cheek asserted that scales with mean interitem correlations greater than 0.50 “tend to be overly redundant” (p. 115). Despite the ease with which interitem correlations can be obtained and interpreted, they do not seem to feature in research associated with the TSES. Indeed, we could find no evidence of them having been sought or examined in the literature concerning TSE.

Confirmatory factor analysis requires provision of different information.

Although some methodologists argue that a wide range of metrics should be provided to indicate the outcomes from this type of factor analysis (see, e.g., Schreiber, Nora, Stage, Barlow, & King, 2006), as a minimum, researchers should state what they regard to be acceptable criteria and they should also provide their findings concerning the normed chi-square (e.g., < 3 preferable, 3 to 5 permissible), Tucker–Lewis index (TLI; e.g., ≥ 0.95 preferable, 0.90 to 0.95 permissible), comparative fit index (CFI; e.g., ≥ 0.95 preferable, 0.90 to 0.95 permissible), root mean square error of approximation (RMSEA; e.g., < 0.06 preferable, 0.06 to 0.10 permissible), and standardized root mean square residual (SRMR; e.g., < 0.05 preferable, 0.05 to 0.08 permissible). A range of other metrics exist, but those metrics appear to be increasingly unreported.

Inappropriate Reporting of Information about Individual Items and Coefficient Alpha. In this subsection, we deal with two topics. The first of these is inappropriate reporting of information and analyses about individual items. Carifio and Perla (2007) have referred to a focus on individual items as akin to a “laundry list and fuzzy jumble” (p. 115). Although inspection at the individual-item level can yield useful insights, some researchers provide extensive information about, and conduct analyses based on, individual items. This not only runs counter to the purpose of multi-item scales as a means of capturing the complexity and richness of constructs, but, even more than when two or more supposed domains are unjustifiably analyzed separately, it raises the probability of generating type 1 errors.

A more pervasive problem in research about TSE relates to the previously mentioned desire to create scales that have high factor loadings. This desire is often accompanied by an equally strong desire to obtain and report high coefficient alphas¹¹ as evidence of scale reliability. An apparent veneration of alpha throughout research concerning TSE occurs despite alpha having been known for some time to be both more complex and less useful than seems to be generally appreciated (Boyle, 1991; McNeish, 2018; Panayides, 2013) and despite it not being an indicator of unidimensionality (Clark & Watson, 1995;

¹¹ In the context of scale reliability, coefficient alpha is usually associated with the name of Lee Cronbach, and therefore as Cronbach’s alpha. Cho and Kim (2015) present reasons for not associating Cronbach with coefficient alpha and refer to it simply as alpha. We do the same, as do other researchers (Crutzen & Peters, 2017).

Hattie, 1985; Sijtsma, 2009). Cho and Kim (2015) have even argued that alpha carries a number of meanings, none of which is clear or justified.

Nevertheless, alphas are commonly cited within the literature about TSE, almost as a mantra,¹² presumably in order to claim unidimensionality among items, so it is important that some of that statistic's features are understood. First, alpha values can be highly dependent on the number of items in a particular analysis (Cho & Kim, 2015; Cortina, 1993; Streiner, 2003). Because of that, a small number of items (say, four to eight) can easily have a low alpha value despite being quite highly interrelated, and, conversely, a larger number of items (say, 24) can easily produce a high alpha value despite many of those items bearing little relationship with each other.

The associated implications should not be ignored. A high alpha value obtained with a small number of items indicates that a high level of consistency exists among those items such that they are likely to be tapping the same domain and are therefore unnecessarily repetitive (Cho & Kim, 2015; Taber, 2018)—at least in relation to the way that respondents perceive those items, regardless of the items' overt appearance. High alpha values with a small number of items are therefore not achievements. They are probably better regarded as confessions of methodological deficiency in that much the same measurement outcomes could be achieved more economically by removing redundant items. High indications of consistency might also indicate that operationalization of a construct has been too constrained (see Crutzen & Peters, 2017).

In addition, researchers should bear in mind that a high alpha value obtained with a large number of items is often merely a mathematical inevitability (Schmitt, 1996; Trevethan, 2009) and therefore, although not manifestly problematic, high alphas can be devoid of meaning (Cortina, 1993; Taber, 2018). McDowell and Newell (1996) have recommended that "as with food and other good things, moderation in internal consistency is best" (p. 40).

There are no definitive criteria concerning alpha values even in contexts where they might be valid and informative. However, according to Jackson (2003), if researchers aim for a moderate amount of association among a set of items, an

¹² Sijtsma (2009) has argued that "the only reason to report alpha is that top journals tend to accept articles that use statistical methods that have been around for a long time such as alpha" (p. 118). Cho and Kim (2015) expressed a similar viewpoint and accompanied it with a request that journal editors recommend authors use superior alternatives to, or in conjunction with, alpha.

alpha of approximately 0.65 is probably satisfactory for four or five items; Cortina (1993) indicated that an alpha of 0.84 would be permissible for 12 items; and, according to Trevethan (2009), alphas of approximately 0.75 and 0.80 would be appropriate for 10 and 15 items, respectively. This means that an alpha in the region of 0.80 would be appropriate for each eight-item factor in the long-form of the TSES, and an alpha in the region of 0.65 would be appropriate for each four-item factor in the short form of the TSES.

There appears to be little awareness of these principles in research associated with the TSES and similar scales. As a result, researchers commonly cite high alpha values when TSES data are based on the 24 long-form items without realizing that doing so is essentially meaningless,¹³ and they cite high alpha values for separate subscales in the TSES short form without realizing that those values could indicate a high degree of redundancy among the items at the cost of conceptual breadth.

Tschannen-Moran and Woolfolk Hoy (2001) might have generated some impetus for citing high alpha values by providing those metrics in their seminal research when they reported alphas ranging from 0.87 to 0.91 on the three eight-item subscales of the long form,¹⁴ and from 0.81 to 0.86 on the three four-item subscales of the short form.¹⁵ Their purpose in citing these values is unclear, but given the numbers of items on the subscales, these alphas suggest that each of the domains is limited in scope regardless of whether those domains are being accessed by the long- or short-form items. Tschannen-Moran and

¹³ This meaninglessness is revealed by the alpha values obtained when TSES subscales are combined (presumably resulting in a combination of three sets of disparate items) being consistently higher than the alphas obtained when each of those subscales is analyzed independently (with each set presumably containing its own set of similar items). For example, Yousuf Zai (2016) obtained an extremely high alpha of 0.93 on the 24 TSES long-form items, but considerably lower alphas of 0.84, 0.83, and 0.83 on the three eight-item subscales. Similarly, but on the TSES short form, Chang and Engelhard (2016) obtained a high alpha of 0.90 on the 12 items, but considerably lower alphas of 0.74, 0.71, and 0.65 on the three four-item subscales.

¹⁴ Sometimes even higher alphas are associated with the eight-item TSES subscales. For example, Herman, Hickmon-Rosa, and Reinke (2018) obtained alphas at or above 0.95 on three administrations of the eight-item management subscale, displaying the participants' perception of extreme construct narrowness among that subscale's items.

¹⁵ The unsatisfactorily high alpha values obtained by Tschannen-Moran and Woolfolk Hoy on the three separate four-item subscales within the TSES short form are not unusual. For example, Fives and Buehl (2010) reported alpha values ranging from 0.81 to 0.89 on those subscales, and von Suchodoletz, Jamil, Larsen, and Hamre (2018) reported alphas of 0.78 to 0.85.

Woolfolk Hoy (2001) also reported an alpha of 0.94 on the composite 24 long-form items, which is almost inevitably a function of the large number of items in the analysis and therefore meaningless, and they reported an alpha of 0.90 on the composite 12 short-form items, which indicates an unsatisfactory degree of redundancy among those items despite them presumably comprising three disparate domains. These findings are not unique. For example, alphas on the composite 12 short-form items were 0.92 for PST and 0.86 for IST samples in the USA (Fives & Buehl, 2010) and 0.95 for combined IST samples from Hong Kong and Shanghai (Cheung, 2008).

Summary, Comments, and Recommendations. Too frequently, insufficient information is provided about response and retention rates and how they were determined. Consumers of research should be made aware of relevant definitions, for example whether a response rate is based on the number of people from a specific pool who might have participated in a study, those who were actually invited to participate, or those who commenced responding to a survey. Rates of retention might also be of interest, and should be defined, possibly in relation to data that were analyzable subsequent to data cleaning or at progressive stages of data cleaning. Because response and retention rates could be dependent on the circumstances under which data are acquired, and those rates might also influence research results, providing information about these rates is not only important for interpreting results, but could provide useful information for researchers seeking to improve design and administration of surveys that might, in turn, enhance these rates. Whether the circumstances under which data are acquired are important or not, and whether circumstances can be adopted that improve the prospect of valid data, can be investigated more effectively if researchers are transparent and inquisitive about response and retention rates.

Researchers should also provide transparent information about the nature of their data and the extent and nature of data cleaning, and they should consider reporting outcomes of analyses if those outcomes differed before and after alterations to the original data. Because outliers on individual items can influence correlations (Zijlstra, van der Ark, & Sijtsma, 2011); alphas (Liu, Wu, & Zumbo, 2010; Zijlstra et al., 2011), particularly if the outliers are asymmetrically distributed (Liu & Zumbo, 2007); factor analysis (Liu, Zumbo, & Wu, 2012); and, indeed, a range of statistics related to correlation, including regression and structural equation modeling (Goodwin & Leech, 2006), information should be

provided about the extent of outliers in the data and that any outlying data points had been dealt with appropriately. Furthermore, the extent, nature, and handling of missing data should be reported and consideration should be given, if appropriate, to how missing data might have influenced the outcomes. Missing data might not be problematic if minimal, trivial, and random, but, if missing data are nonproblematic, that should be indicated.

With regard to exploratory factor analysis, researchers should employ procedures that are appropriate and reflect best practice, and they should also provide sufficient information for others to be able to assess the quality of data and results. In addition to seemingly entrenched but undesirable practices concerning exploratory factor analysis, researchers should cease harboring and perpetuating misconceptions associated with the nature and reporting of coefficient alpha. Furthermore, because of the unstable number and composition of factors on the TSES (and possibly on closely allied scales), researchers who use those scales should employ confirmatory factor analyses as a prelude to their more substantive analyses in order to ascertain the most appropriate way in which to conduct those analyses. Because confirmatory factor analysis is used so infrequently in TSE studies, three scenarios are possible: Researchers are willing to make assumptions about their data that might be unjustified, they do not appreciate the importance of testing those assumptions, or they have tested those assumptions but are unwilling to reveal the results. Each of these scenarios should be avoided.

Empirical Directions for This Research

As a result of the above considerations, in the empirical component of this study we made what we believe to be improvements to the design of the TSES and produced translations for administration to PSTs and ISTs in China. In doing so, a major aim was to create scales that were more attractive to respondents in order to increase response and retention rates and, ultimately, to obtain data with enhanced validity and representativeness from samples within a cultural context that is underreported in English-language publications. We intended to examine the nature of responses on the adapted scales, particularly in order to assess the effect of the changes we had made on them. We also factor analyzed the data and reported the results in ways that we believe reflect best psychometric practice. In addition, we intended to identify how this study could provide insights about the

most appropriate way to analyze TSE between and within PSTs and ISTs in China, and also to obtain general insights about the measurement of TSE.

Method

Participants

All participants, comprising both PSTs and ISTs, were from the Chinese mainland. The PSTs were students currently enrolled in teacher education programs, and ISTs were assumed to be practicing schoolteachers given the way in which invitations to take part in the research were distributed.

Surveys

Survey Contents

Separate surveys were created for the PSTs and ISTs, commencing with demographic questions appropriate to each group. Among these questions, PSTs were asked to indicate their age and sex, the qualification they were enrolled for, and their year of study. Inservice teachers were asked to indicate their sex, number of years as a teacher, whether they were teaching at a standard or advanced school,¹⁶ the level of their school (preschool, primary, etc.), where their school was located, and whether they were subject teachers (*renke laoshi* 任课老师) or head teachers (*ban zhuren* 班主任).¹⁷

On both surveys, the demographic questions were followed by a scale intended to assess respondents' sense of efficacy in a teaching environment. The full

¹⁶ The main difference between standard and advanced schools is that the latter are "given additional resources and assigned better teachers" (Organisation for Economic Co-operation and Development, 2011, p. 95). Although the distinction is no longer approved of or maintained officially, the underlying differences persist and are widely recognized. Most teachers would know whether they were teaching at standard or advanced schools.

¹⁷ In the Chinese school system, there are two main categories of teacher, neither of which is readily translated into English in a way that is both accurate and succinct. One category, subject teachers (*renke laoshi* or *renke jiaoshi*), deliver subject content and might be referred to in English as *subject teachers*. The other category, head teachers (*ban zhuren* or *banzhuren laoshi*), teach, coordinate other teachers, and assume broad coordinating and pastoral roles with regard to students, often extending for more than a year. These teachers might be referred to in English as *head teachers*.

24-item TSES was used as a foundation for these scales, with the original scale being altered to produce slightly different versions for the PSTs and ISTs. In accordance with Hatlevik (2017), the PSTs were asked to respond to the items with respect to how effective they *anticipated* being as teachers (indicating a future orientation), whereas ISTs were asked to respond with respect to how effective they *currently* perceived themselves to be (indicating a present orientation). Both versions had seven response options designed to focus on respondents' sense of efficacy, with options at 1, 3, 5, and 7 labeled *Minimally effective*, *Only moderately effective*, *Quite effective*, and *Extremely effective*, respectively.¹⁸ Options 2, 4, and 6 were not labeled. Unlike the TSES, both versions also had a final option of *Not applicable* in order to allow for, and detect, cultural nonequivalence of items, and the PSTs' scale had a penultimate (eighth) option of *I really have no idea at all* to allow for the possibility that some PSTs might not yet have had any teaching experience, might therefore find it difficult to anticipate how effective they would be in a teaching environment, and could, as a result, be disinclined to respond to many or all of the items or respond to items carelessly (see Trevethan, 2009). An option corresponding to *None at all* on the TSES was not offered for several reasons: Inspection of data obtained by the first author (Ma, 2017) indicated that PSTs almost never chose an option as extreme as that on the Scale for Teacher Self-Efficacy (Pfitzner-Eden et al., 2014); Vieluf et al. (2013) had found a similar phenomenon when investigating the TSE of ISTs across 23 countries; and Chang and Engelhard (2016) combined responses on the three lowest options of the TSES because of the infrequent use by ISTs of the lowest two response options.

On both scales, items for assessing TSE were identical and were presented in the same sequence as in the long form of the TSES,¹⁹ but the introductory wording for each item (e.g., *How much can you do to ...*) was removed as being unnecessary given the initial instructions and in order to make the scale more attractive to prospective respondents as a result of streamlining.²⁰ Item wording was identical on both TSE scales.

Both surveys were prepared for digital administration. All respondents were assured of anonymity, they were not asked to indicate their specific educational

¹⁸ Because the focus of TSE is efficacy, we used the word *effective* in response options. This is in contrast to most other research in which response options are couched in terms of extent (e.g., "A lot"), confidence, or agreement.

¹⁹ Items on the long and short forms of the TSES are sequenced differently.

²⁰ Similar modifications were made to the TSES by George et al. (2018).

institution or school, no incentives were offered for participation in the research, and participation was regarded as an indication of consent. English versions of the TSE scales for the PSTs and ISTs are provided in Appendices A and B, respectively. Chinese versions of instructions, items, and response categories for the PST survey, and instructions for the IST survey, are provided in Appendix C.

Translation of the Teacher Sense of Efficacy Scale

Although a translation of the TSES into Chinese is publicly available,²¹ it was not used because of the changes we had made to the scale's instructions, items, and response options. Instead, based on the procedures recommended by Wild et al. (2005) and Peña (2007), we produced translations of the TSE scales using translation and back translation.

Translation into Chinese. The original English versions of the scales were independently translated into simplified Chinese by two postgraduate Chinese students studying at an Australian university, both of whom spoke Chinese as their mother tongue and English as a second language. Because the 24 items were worded identically on the PSTs' and ISTs' scales, each translator provided only one translation of those items. However, each translator provided separate translations for the PST-specific and IST-specific instructions and response options. All translations were sent to a teacher of English in China who discussed them face-to-face with a Chinese American who had more than 10 years of teaching experience in the USA. As a result of that interaction, they indicated a preference for one rather than the other, or a combination, of the initial translations, but they also made a number of observations, including possible resolutions, regarding nuances and contextual appropriateness of some words (e.g., *difficult* in Item 5 and *individual* in Item 17). All of their suggestions were then considered by the first author (MK) and a second teacher with proficiency in English from China, both of whom worked together to evaluate the linguistic equivalence and cultural appropriateness of the translations. In this process, two prospective ambiguities (associated with the words *failing* and *very capable* in Items 14 and 20, respectively) were resolved, and the Chinese versions were then finalized for back translation.

²¹ This translation is provided at: <http://wmpeople.wm.edu/site/page/mxtsch/researchtools>

Back Translation to English. The interim final Chinese translations were sent to a native Chinese-speaking teacher of English at a foreign-language university in China. He back translated the scales to English. The second author (RT) compared the original English versions with the Chinese back translations, and one small discrepancy was resolved.

Procedure

In order to obtain a sample of PSTs, the first author contacted colleagues at teacher-education institutions in China seeking assistance to access prospective participants. Subsequently, teacher educators at a national university, a provincial university, and a city college offered to invite their students, via social media, to take part in the research. These educators were sent a link for placing on the students' social media sites with the invitation to access and complete the survey from those sites. Enrolment information from the educators indicated that a total of 654 students might have been contacted through the student sites.

In order to obtain the sample of ISTs, the first author sent the IST survey to 12 acquaintances who were currently schoolteachers in China with a request that they invite colleagues to distribute the link to the survey among other teachers via social media. Eleven of these acquaintances responded with an offer to help. This resulted in snowball sampling of ISTs.

Because there was a rapid decrease in submissions to both survey sites 10 days after the invitations were first issued, by which time a total of 803 surveys had been received, access to both surveys was discontinued at that point.

Analyses

Data were analyzed using SPSS Version 22[®] (IBM Corp., Armonk, NY, USA). Descriptive statistics comprised frequencies, percentages, means, *SDs*, correlations, and alphas. We employed exploratory rather than confirmatory factor analyses for three reasons: First, the variety of findings regarding the factor structure of the TSES above provided us with no a priori expectations; second, we had altered the TSES with regard to instructions, item wording, and number and type of response options; and third, the two samples were from a culture with perspectives concerning education that differ from the perspectives

prevalent in cultures—usually Western—in which the TSES has most frequently been administered. We used principal axis factoring rather than maximum likelihood as the method of extraction to conform with common practice in TSE research and, if there were two or more factors in the data, we anticipated using oblique rather than orthogonal rotations in order to avoid artificially forcing the data into separate factors. All responses of 8 and 9 were converted to missing data prior to calculation of means, *SDs*, correlations, and alphas as well as in the factor analyses reported in this article, and cases were excluded pairwise in the factor analyses.

Results

Participants

Analyzable data were received from 395 PSTs. Their ages ranged from 18 to 27 years (mean = 20.1, *SD* = 1.47). The majority were female ($n = 366$, 92.7 %). The two main degrees enrolled in were a 4-year bachelor's degree ($n = 342$, 86.6 %) and a 2-year master's degree ($n = 51$, 12.9 %). Most bachelor's students ($n = 286$, 83.6 % of that group) were in the second year of their courses, and most master's students ($n = 43$, 84.3 % of that group) were in their first year.

Analyzable data were received from 279 ISTs. Demographic information about these participants is provided in Table 1. The majority were female, and most had been teaching for less than 10 years and were teaching at standard schools. Most were also teaching at primary and junior high schools, with very few at preschools and senior high schools. They were spread reasonably equally across city, town, and village schools, but there was a noticeably smaller percentage at county schools. Approximately half (48 %) were subject teachers, with the remainder head teachers.

Response Characteristics

Preservice Teachers

A total of 451 PSTs initially responded to the invitation to participate in the research, representing a 69 % response rate if all prospective participants had

seen the invitations. However, of the 451 respondents, 56 (12 %) were removed from the data prior to the substantive analyses. Among these 56 respondents, 27 (6 % of the original respondents) were removed because they did not provide any responses beyond the demographic questions, 15 (3 % of the original respondents) because they responded to fewer than 12 of the TSE items (most of these participants ceased responding at the end of the first screen of the survey with its seven items), and 14 (3 % of the original respondents) because they responded to all TSE items identically and therefore were assumed to have regarded completing the survey as a trivial rather than serious task. Removal of the 56 participants resulted in the 395 PSTs who provided analyzable data, and therefore a retention rate among the initial respondents of 88 %.

Table 1 Demographic Characteristics of ISTs (*N* = 279)

Variable		Number	%
Sex	Female	208	74.6
	Male	71	25.4
Years of teaching	Less than 5	125	44.8
	5–10	62	22.2
	11–25	65	23.3
	25+	27	9.7
School type	Advanced	91	32.6
	Standard	188	67.4
School level	Senior high	11	3.9
	Junior high	107	38.4
	Primary	149	53.4
	Preschool	12	4.3
School location	City	91	32.6
	County	38	13.6
	Town	80	28.7
	Village	70	25.1
Category of teacher	Subject teacher	134	48.0
	Head teacher	145	52.0

Among the 395 PSTs whose data were retained for analysis, Option 8, *I really have no idea at all*, was used on all items. Nine PSTs chose that option for more than three items, with the greatest usage accorded that response (on 11 items) by

one of those nine participants. For most of the items, eight (2 %) or fewer PSTs chose this option, with the greatest number ($n = 20$, 5 %) selecting it on the first item, *Getting through to the most difficult students*, followed by the second-greatest number ($n = 13$, 3 %) on the fourth item, *Motivating students who show low interest in school work*. Even fewer PSTs chose the *Not applicable* option (Option 9), with only one of them choosing that option more than three times and the greatest number ($n = 6$, 1.5 %) choosing that option on Item 17, *Adjusting your lessons to the proper level for individual students*. Until Item 7, all PST participants responded to all items. Until Item 16, on the second screen of the survey, a maximum of five items had missing data on them (some of which resulted from two participants having apparently overlooked all nine items on the second screen of the survey), but 14 participants did not respond to any of the final eight items, all on the third screen of the survey. Disregarding those 14 participants, the final item on the scale had the greatest number of missing responses ($n = 8$). If the 16 participants who failed to respond to a complete screen on the survey were disregarded, only eight participants' responses contained missing data, and for those participants the missing data occurred on only one or two of the 24 items. Overall, therefore, there was extremely little, and sporadic, missing data apart from nonresponses to complete screens of the survey.

Inspection of item frequencies revealed a strong tendency for respondents to choose options above which wording was provided. For example, on the first item, 74.2 % chose the first and third options. In contrast, only 6.8 % chose the second and fourth options (labeled simply 2 and 4, respectively). That pattern, although not always as noticeable, persisted across all 24 items.

The options of 1 to 7 that referred to levels of perceived TSE were used for all 24 items, across which the means ranged from 2.94 to 3.88 and *SDs* from 1.30 to 1.52. Means and *SDs* for these respondents' data on all 24 items are included in Table 2. Congruent with item means being close to the midpoint of the option range and *SDs* being broad (refer to Table 2), inspection of histograms indicated no problems with either skewness or leptokurtosis. However, boxplots revealed outliers on Items 1, 2, and 4. Item 1 was the most problematic, with 12 high outliers and 7 low outliers. We decided not to adjust those outliers because the degree of asymmetry was tolerable. However, we flagged that item in case it became prominent in subsequent analyses. Items 2 and 4 each had 11 outliers and

for both items those outliers exhibited a sufficient degree of symmetry (seven high; four low) that we considered adjusting them to be unwarranted.

Table 2 Means and SDs of PSTs and ISTs on the 24 TSE Items^a

Code ^b	No. ^c	Scale item	PSTs		ISTs	
			Mean	SD	Mean	SD
E	1	Getting through to the most difficult students	2.94	1.32	2.99	1.47
E	2	Helping your students to think critically	3.35	1.34	3.64	1.55
M	3	Controlling disruptive behavior in the classroom	3.43	1.45	4.36	1.60
E	4	Motivating students who show low interest in school work	3.24	1.44	3.87	1.65
M	5	Making your expectations clear about student behavior	3.54	1.45	4.11	1.55
E	6	Getting students to believe they can do well in school work	3.63	1.49	4.37	1.64
I	7	Responding to difficult questions from your students	3.64	1.53	4.78	1.63
M	8	Establishing routines to keep activities running smoothly	3.70	1.45	4.45	1.53
E	9	Helping your students to value learning	3.68	1.44	4.28	1.45
I	10	Gauging student comprehension of what you have taught	3.71	1.47	4.47	1.40
I	11	Crafting good questions for your students	3.80	1.52	4.63	1.45
E	12	Fostering student creativity	3.60	1.41	4.27	1.67
M	13	Getting students to follow classroom rules	3.88	1.46	4.78	1.60
E	14	Improving the understanding of a student who is failing	3.65	1.47	4.35	1.52
M	15	Calming a student who is disruptive or noisy	3.44	1.50	4.70	1.65
M	16	Establishing a classroom management system with each group of students	3.39	1.48	4.14	1.68
I	17	Adjusting your lessons to the proper level for individual students	3.02	1.30	3.77	1.55
I	18	Using a variety of assessment strategies	3.40	1.41	3.98	1.57

(To be continued)

(Continued)

Code ^b	No. ^c	Scale item	PSTs		ISTs	
			Mean	SD	Mean	SD
M	19	Keeping a few problem students from ruining an entire lesson	3.38	1.46	4.40	1.57
I	20	Providing an alternative explanation or example when students are confused	3.56	1.40	4.58	1.50
M	21	Responding to defiant students	3.46	1.47	4.30	1.57
E	22	Assisting families in helping their children do well in school	3.75	1.47	4.36	1.59
I	23	Implementing alternative strategies in your classroom	3.82	1.46	4.50	1.55
I	24	Providing appropriate challenges for very capable students	3.85	1.48	4.58	1.63

Note. ^a These items correspond to the 24 items on the long form of the Teacher Sense of Efficacy Scale.

^b Factor domains from the Teacher Sense of Efficacy Scale: E = Engagement of students; I = Instructional strategies; M = Management of student behavior.

^c These numbers refer to the items on the original TSES.

Inservice Teachers

A total of 352 ISTs initially responded to the invitation to participate in the research. However, 73 (21 %) were removed from the data before the final analyses. Of these, five (1 % of the original respondents) were removed because they were not teaching at schools, 23 (7 % of the original respondents) because they did not provide any responses beyond the demographic questions, 29 (8 % of the original respondents) because they responded to only the seven items on the first screen of the TSE items, and 16 (5 % of the original respondents) because they responded to all TSE items uniformly and, as for the PSTs, were regarded as having put insufficient thought into the survey or not taking the task seriously. Removal of the 73 participants resulted in the 279 ISTs with analyzable data, and therefore a retention rate among the initial respondents of 79 %.

Among the 279 ISTs whose data were retained for analysis, few chose the *Not applicable* option (Option 8) and only four participants used that option more

than three times, with the greatest usage accorded that response (on 8 items) by two of those four participants. Greatest use of that option occurred on two items, *Helping your students to think critically* and *Adjusting your lessons to the proper level for individual students* ($n = 10$, 3.6 % in each case). For these participants (ISTs), there were no missing data because, if they had responded to the first seven TSE items, they went on to provide responses for all 24 items.

As had been the case for the PSTs, there was a strong tendency among ISTs to choose options for which explicit wording had been provided. For example, on the first item, 68.1 % chose the first and third options. In contrast, only 11.8 % chose the second and fourth options. As was the case for the PSTs, that pattern continued across all 24 items. Furthermore, as with the PSTs, all response options from 1 to 7 that referred to levels of perceived TSE were used for all 24 items. Means ranged from 2.99 to 4.78, and *SDs* from 1.39 to 1.68 (see Table 2). Entries in Table 2 reveal that, relative to PSTs, the ISTs had higher TSE on all items and wider *SDs* on all but two items. As might have again been expected given that the item means were close to the midpoint of the option range and *SDs* were broad, inspection of histograms indicated no problems with either skewness or leptokurtosis. However, boxplots revealed 11 high outliers on the first item. We decided not to adjust those outliers because they had occurred on only one item and were therefore unlikely to exert a noticeable influence on the results.

Factor Analyses

Separate exploratory factor analyses were initially conducted for the PSTs' and ISTs' data on all 24 items of their respective TSE scales. The highest interitem correlations were 0.64 and 0.74, respectively, thus indicating absence of multicollinearity in the data. For both data sets, the KMO index (0.93 and 0.95, respectively) and Bartlett's test of sphericity (both $p < 0.001$) were satisfactory. Communalities in the PSTs' data ranged from 0.21 to 0.56 (mean = 0.41) and in the ISTs' data ranged from 0.25 to 0.70 (mean = 0.52). There was a sharp decrease between the first and subsequent three eigenvalues, from 10.27 to 1.65, 1.33, and 1.11 in the PSTs' data, and from 13.06 to 1.35, 1.23, and 0.92 for the ISTs' data. The pattern of eigenvalues and accompanying scree plots (refer to Figures 1 and 2) so strongly indicated the presence of a single factor in both sets of data that we did not employ parallel analysis. The first factor accounted for 42.78 % of the variance in the PSTs' data and 54.41 % in the ISTs' data.

Figure 1 Scree Plot for PSTs' Sense of Efficacy Scores

Figure 2 Scree Plot for ISTs' Sense of Efficacy Scores

Single-factor solutions were then sought from each data set to determine whether any items failed to load satisfactorily. The results for both solutions are provided in Table 3 where the items have been ordered according to decreasing size of loadings in the ISTs' data. For neither PSTs nor ISTs were any loadings below 0.45, confirming all 24 items as belonging on the sole factor. The loadings ranged from 0.46 to 0.75 for the PSTs and from 0.50 to 0.84 for the ISTs, suggesting that domain coverage was not undesirably restricted.

Table 3 Factor Loadings for PSTs and ISTs on the 24 TSE Items^a

Code ^b	No ^c	Scale item	Loadings	
			PSTs	ISTs
I	23	Implementing alternative strategies in your classroom	0.69	0.84
I	10	Gauging student comprehension of what you have taught	0.65	0.80
M	13	Getting students to follow classroom rules	0.69	0.78
I	20	Providing an alternative explanation or example when students are confused	0.65	0.77
I	24	Providing appropriate challenges for very capable students	0.64	0.77
E	12	Fostering student creativity	0.60	0.77
E	9	Helping your students to value learning	0.75	0.76
I	11	Crafting good questions for your students	0.64	0.76
I	18	Using a variety of assessment strategies	0.61	0.76
M	8	Establishing routines to keep activities running smoothly	0.71	0.75
E	22	Assisting families in helping their children do well in school	0.68	0.75
M	16	Establishing a classroom management system with each group of students	0.69	0.74
E	14	Improving the understanding of a student who is failing	0.65	0.74
E	6	Getting students to believe they can do well in school work	0.63	0.74
M	19	Keeping a few problem students from ruining an entire lesson	0.65	0.72
M	15	Calming a student who is disruptive or noisy	0.67	0.72
I	17	Adjusting your lessons to the proper level for individual students	0.65	0.71
M	21	Responding to defiant students	0.64	0.71
M	5	Making your expectations clear about student behavior	0.60	0.69
I	7	Responding to difficult questions from your students	0.57	0.68
M	3	Controlling disruptive behavior in the classroom	0.58	0.64
E	4	Motivating students who show low interest in schoolwork	0.54	0.60
E	2	Helping your students to think critically	0.55	0.58
E	1	Getting through to the most difficult students	0.46	0.50

Note. ^a These items correspond to the 24 items on the TSES long form and are sequenced according to loadings in the IST analysis.

^b E = engagement of students; I = instructional strategies; M = management of student behavior.

^c These numbers refer to the item numbers on the long form of the TSES.

When comparing the factor loadings between the PSTs' and ISTs' data, these loadings were not always similar, or parallel, but most items that loaded between 0.60 and 0.70 in the PST data loaded between 0.70 and 0.80 in the IST data. Management items tended to attract high loadings for the PSTs, instruction items tended to attract high loadings for the ISTs, and three engagement items attracted the lowest loadings for both participant groups. For both groups, items from each of the three TSES domains were otherwise distributed throughout the single factor.²²

Following the practice of Fives and Buehl (2010) and Ruan et al. (2015), the factor structure of items from the TSES short form was also explored. For both PSTs and ISTs, the KMO index (0.90 and 0.94, respectively) and Bartlett's test of sphericity (both $p < 0.001$) were again satisfactory. Communalities in the PSTs' data ranged from 0.28 to 0.53 (mean = 0.42) and in the ISTs' data ranged from 0.34 to 0.70 (mean = 0.55). There was a sharp decrease between the first and subsequent three eigenvalues, from 5.59 to 1.07, 0.85, and 0.76 in the PSTs' data and from 7.00 to 0.89, 0.80, and 0.56 in the ISTs' data. The pattern of eigenvalues and the scree plots again indicated a single factor in each data set so strongly that we did not conduct parallel analysis. The single factor accounted for 46.54 % of the variance in the PSTs' data and 58.30 % in the ISTs' data. Loadings from both single-factor solutions are provided in Table 4 where the items have again been ordered according to decreasing size of loadings in the ISTs' data. These loadings ranged from 0.53 to 0.73 for the PSTs and from 0.59 to 0.84 for the ISTs, indicating again that domain breadth was not unsatisfactorily constrained. Given that all loadings exceeded 0.50, the 12 items could be regarded as belonging within the single factor for each group of participants.

Overall, the loadings were similar to those obtained with the 24 long-form items, particularly in that engagement items tended to have lower loadings than did instruction and management items. However, for both groups of participants, items from all three TSES domains were distributed throughout the single factor. Alphas were 0.90 and 0.93 for the PSTs' and ISTs' data, respectively.

The size of any subsamples within the two participant groups placed constraints on follow-up factor analyses to determine whether the initial factor

²² The alphas were 0.94 and 0.96 for the PSTs' and ISTs' data, respectively. We report these alphas merely for the sake of convention and do so in a footnote rather than in the body of the text as a way of indicating that the number of items on the scales (24 in each case) almost definitely renders these metrics meaningless.

solutions were robust. However, separate analyses for primary and high school teachers, and subject and head teachers, were conducted using the short-form items because it was then possible to satisfy the condition that there should be at least 10 participants for each item in an analysis.

Table 4 Factor Loadings for PSTs and ISTs on the 12 TSE Items^a

Code ^b	No. ^c	Scale item	Loadings	
			PSTs	ISTs
M	3	Controlling disruptive behavior in the classroom	0.70	0.84
M	15	Calming a student who is disruptive or noisy	0.68	0.79
M	13	Getting students to follow classroom rules	0.61	0.78
E	22	Assisting families in helping their children do well in school	0.63	0.76
I	23	Implementing alternative strategies in your classroom	0.69	0.76
I	20	Providing an alternative explanation or example when students are confused	0.60	0.76
I	11	Crafting good questions for your students	0.73	0.75
I	18	Using a variety of assessment strategies	0.71	0.74
E	9	Helping your students to value learning	0.60	0.73
M	16	Establishing a classroom management system with each group of students	0.68	0.71
E	4	Motivating students who show low interest in schoolwork	0.57	0.63
E	6	Getting students to believe they can do well in school work	0.53	0.59

Note. ^a These items correspond to the 12 items on the TSES short form and are sequenced according to loadings in the IST analysis.

^b E = engagement of students; I = instructional strategies; M = management of student behavior.

^c These numbers refer to the item numbers on the long form of the TSES.

These IST subgroup analyses, summarized in Table 5, consistently indicated the presence of only one factor within each of the four subsamples, with loadings ranging from 0.52 to 0.85. There were tendencies for higher loadings on instruction items and lower loadings on management items among the primary school teachers. There were also tendencies for higher loadings on instruction items and lower loadings on engagement items among subject teachers, and higher loadings on both instruction and engagement items among head teachers.

However, the loadings indicated that items from the three TSES domains were, as for the previous analyses, interspersed among each other.

Table 5 Factor Loadings for Primary and High School Teachers, and for Subject and Head Teachers, on the 12 TSE Items^a

Code ^b	Scale item	Loadings			
		Prim	High	ST	HT ^c
I	Implementing alternative strategies in your classroom	0.82	0.87	0.84	0.85
I	Crafting good questions for your students	0.78	0.79	0.76	0.80
M	Getting students to follow classroom rules	0.77	0.82	0.79	0.78
I	Providing an alternative explanation or example when students are confused	0.78	0.76	0.77	0.76
E	Assisting families in helping their children do well in school	0.73	0.81	0.78	0.74
E	Getting students to believe they can do well in school work	0.73	0.73	0.73	0.72
E	Helping your students to value learning	0.74	0.77	0.76	0.72
M	Calming a student who is disruptive or noisy	0.70	0.70	0.69	0.72
I	Using a variety of assessment strategies	0.77	0.74	0.79	0.71
M	Establishing a classroom management system with each group of students	0.71	0.80	0.81	0.64
M	Controlling disruptive behavior in the classroom	0.60	0.66	0.60	0.64
E	Motivating students who show low interest in schoolwork	0.61	0.54	0.65	0.52

Note. ^a These items correspond to the 12 items on the TSES short form and are sequenced according to loadings in the head teacher analysis.

^b E = engagement of students; I = instructional strategies; M = management of student behavior.

^c Prim = primary school teachers; High = high school teachers; ST = subject teachers; HT = head teachers.

In order to examine why the alpha values for the 12 short-form items were so high, Spearman’s rank-order correlation coefficients between those items were obtained. These correlations revealed that, for the PSTs, only eight of the 66 correlations (i.e., 12 %) exceeded 0.50, and the mean interitem correlation was 0.42. Four of the items had a correlation greater than 0.55 with another item, but only one correlation exceeded 0.60. For the ISTs, 38 of the 66 correlations (i.e.,

58 %) exceeded 0.50, and the mean interitem correlation was 0.52; five of the items had correlations greater than 0.55 with at least three other items, and the three highest correlations, all of which exceeded 0.65, occurred among instruction items.

Discussion

This study provides a number of insights concerning the measurement and nature of TSE, not only in a Chinese context, but also within broader international contexts. Our discussion focuses first on the design of TSE scales, with particular emphasis on the responses of the PSTs and ISTs on their respective TSE scales. We intend to reveal successes, failures, and insights that might be informative for other researchers if they are designing TSE scales, whether for administration in China or elsewhere. We then turn our attention to issues of a more psychometric nature. We include summary subsections that contain comments and recommendations, and finish with a brief conclusion.

Issues and Insights Associated with Design of Scales

Response Rates

We regard a response rate to be the proportion of people who, having been invited to participate in research, commence participating even if some of the data they subsequently provided prove to be unusable and were therefore discarded. For both PSTs and ISTs in this study, therefore, the response rate would have been influenced by some prospective respondents not having seen the invitations in the 10-day period during which the surveys were made available as well as other prospective participants being unwilling to take part despite seeing the invitation. Additional prospective respondents might have found the demographic questions on the first screen of the survey too invasive and might therefore have not even commenced responding, in which case a higher response rate might have been achieved had those questions been placed at the end of the survey.

Despite the above challenges to response rates, one of the most noticeable features of this research is the estimated response rate of 69 % for the PSTs assuming that all enrolled PSTs had seen the invitation to participate. This is high

compared with published response rates relating to PSTs on other web-based surveys. As mentioned earlier in this article, Pfitzner-Eden et al. (2014) reported a response rate of only 33 %, and the response rate in research by O'Neill and Stephenson (2011) was much lower at only 14 % despite the researchers offering respondents the prospect of a gift voucher worth \$150 in a prize draw.

Accounting for the higher PST response rate in this research can only be speculative, but one reason might be that PSTs from Western cultures experience greater survey fatigue because they are more often asked to complete surveys, predominantly as a means of rating their lecturers and the adequacy of the teaching they receive, but also because of their ready availability for researchers—and they may therefore ignore requests for participation in survey research. There might also, however, be a sociocultural origin to the relatively high response rate among the PSTs in our research. That origin relates to the Chinese phenomenon of relationship (*guanxi* 关系) by which cooperation and assistance are gained, and offered, on the basis of respect and social connections. This phenomenon has been identified as raising the likelihood that people will respond positively to requests to take part in research (Li, 2015; Zhou & Nunes, 2013). In our research, the PSTs were invited to participate via postings placed by their lecturers on the students' social websites, so the response rate might have been enhanced by admiration and respect of the students for their lecturers.

Although the response rate of the ISTs could not be determined because snowball sampling had been used with that sample and no indication could be obtained about exposure to invitations, their response rate might also have been relatively high as a result of relationship because the participants could have been inclined to respond on the basis of acquaintance or even friendship with the people from whom they received the invitation to participate.

Retention Rates

We regard a retention rate to be the proportion of participants who, having commenced responding to a survey, subsequently responded to a sufficient number of items in a way that appears to be valid. With this definition in mind, there could be several reasons for the high retention rate (88 %) among the PSTs in the present research. Among these reasons could be the instructions on their scale having been deliberately tailored for PSTs rather than, as in the TSES,

framed for ISTs, although, in retrospect, we could have worded the instructions to allow for the possibility that some PSTs might not enter the profession. The high PST retention rate might also be attributable to other scale-specific features, including our having provided response options of *Not applicable* and *I really have no idea*—options that PSTs might appreciate being offered but are seldom, if ever, provided on other TSE scales, and certainly not on the TSES and its immediate derivatives. Although those two options were used infrequently by PSTs, their mere availability might have been reassuring and could have engendered a favorable initial disposition toward the scale that resulted in respondents maintaining their participation.²³ In addition, our having streamlined the format of the scale as well as wording items to be semantically compatible with the response options might have avoided some of the TSES's original features from having served as deterrents for both PSTs and ISTs, and therefore have contributed to the ISTs retention rate not falling lower than its 79 %.

From the opposite perspective, there could also be several reasons for nonretention of respondents' data. An interesting feature of the data was the noticeable percentages (12 % of PSTs and 21 % of ISTs) who provided insufficient data or data that appeared to be inappropriate for analysis. One subgroup within both samples comprised respondents whose records contained demographic information but no subsequent responses. Although the percentages who did this were small (6 % of PSTs and 7 % of ISTs), identification of anything that might have dissuaded them from continuing to respond is worth considering. One possibility is the already-identified ambiguity in the first TSE item, *Getting through to the most difficult students* (see Fives & Buehl, 2010; Henson, 2002), which might have generated, or enhanced, an early negative attitude toward the inventory for some respondents.

Although only 3 % of the PSTs answered fewer than 12 of the 24 TSE items, more than twice that percentage (8 %) of ISTs also answered fewer than 12 items. We thought this invited some scrutiny. One possibility is that our not having provided the ISTs with an option such as *I really have no idea* inclined some of them to log out of the survey prematurely. If so, our not providing that option

²³ Fortunately, these options did not encourage an undue amount of missing data in relation to the assessment of participants' TSE.

was a failure on our part.²⁴ Another possibility is that four of the seven items on the first screen of the inventory referred to the domain of the TSES related to engagement of students, but if the ISTs did not regard that domain as sufficiently relevant to the aspects of teaching that they considered important, they might have abandoned the inventory. This conjecture is supported by Ho and Hau (2004) having found that engaging students was not as high a priority for Chinese teachers as were other aspects of teaching. A similar problem might have occurred for the PSTs if they felt that the engagement items were not important for teachers. With foresight, we might have avoided the preponderance of engagement items at the start of the inventory, a feature that was inherited from the TSES and that we failed to anticipate being problematic.

An unrelated explanation for some participants abandoning the inventory after responding only to items on the first screen is that the second screen contained a further nine items with no indication about how many more items there would be on any subsequent screens. If we had indicated in the instructions that the inventory was restricted to a total of 24 items, and perhaps also that the survey would probably take only approximately 5 minutes to complete, some participants might have been willing to persist with responding. This could be regarded as another failure on our part—a failure that could also relate to the 14 PSTs who abandoned the survey when the third screen of TSE items was presented.

It is more difficult to identify the reason(s) behind some respondents (3 % of the original PSTs and 5 % of the original ISTs) responding identically to all items. It is possible that some of those respondents genuinely had a similar impression of their efficacy about all items—in which case our having removed them from the data set might be regarded as inappropriate unless an argument could be made that those responses would have hampered the valid emergence of separate factors.²⁵ Other respondents, having committed themselves to participate, might have persisted but with an intention to complete the task as quickly and effortlessly as possible, perhaps in conjunction with disrespect for the scale (see Krosnick & Presser, 2010).

²⁴ We had assumed that ISTs would have a sense of their effectiveness concerning all 24 TSE items. That assumption might have been incorrect, particularly with regard to the engagement items because those items refer to subjective perceptions and experiences of students that teachers might have reasonably believed they had insufficient access to.

²⁵ For a discussion of issues concerning removal of participants with identical responses on all items, see Huang, Curran, Keeney, Poposki, and DeShon (2012).

Response Options

A noticeable feature of the responses was a strong tendency for PSTs, and a slightly less strong but nevertheless evident tendency for ISTs, to select options above which specific wording had been provided. This obvious appreciation of guidance concerning the meaning of response options (see Krosnick & Presser, 2010, p. 271) runs counter to the structure exhibited in some scales, for example the Scale for Teacher Self-Efficacy (Pfitzner-Eden et al., 2014), in which verbal anchors are provided at opposite ends of a continuum but with no indication about the meanings that might apply to numbered positions within that continuum. We are currently investigating whether these option differences carry implications for the nature of TSE data obtained from participants.

Because the first seven response options—the options that pertained to gradations of TSE—were used by both PSTs and ISTs on all 24 items, the outermost of those options were evidently not so extreme as to be meaningless for either group of respondents. Furthermore, although the ISTs had wider *SDs* than did the PSTs on all but two items, the range of *SDs* was between 1.39 and 1.68 for the two groups. This consistency is of particular interest with regard to the notion that scales with more response options are desirable because they create greater variability. In research based on 9-point response options, O'Neill and Stephenson (2012) obtained *SDs* that were similar to ours, ranging from 1.32 to 1.62, but Cheung (2008) obtained considerably narrower *SDs*, ranging from 1.11 to 1.48 and from 0.71 to 1.10 in Hong Kong and Shanghai samples, respectively, and Tschannen-Moran and Woolfolk Hoy (2007) also obtained narrower *SDs*, ranging from 0.78 to 1.18.

Having nine response options does not, therefore, automatically confer an advantage over seven response options regarding data variability. The larger *SDs* in our data compared with *SDs* in the research by Tschannen-Moran and Woolfolk Hoy (2007) also disconfirms the expectation proposed by Chen (2008) that Chinese participants use a more restricted range of responses than do US participants—a disconfirmation that is strengthened because the Chinese participants in our study were offered a narrower range of response options than is on the TSES.

Despite our belief that the TSES could be improved by offering the response option of *I really have no idea* for PSTs, use of that option was infrequent,

yielding the interesting insight that most PSTs were willing to venture beliefs about their future efficacy as teachers with regard to all of the items.

Furthermore, availability of the *Not applicable* option might have been reassuring for the ISTs despite few of them using it. However, use of that option provided valuable insights. Its minimal usage indicated that items based on the TSES were regarded by most respondents as being apposite in a Chinese school teaching environment and that it would probably be unnecessary to reword the original TSES items to any greater extent than we had done in order to capture the aspects of teaching that the TSES targets.

Responses on the TSE Items

Compared with the PSTs, the ISTs in this study had higher means on all 24 items—a phenomenon that we intend to explore in subsequent research, particularly in light of the proposal by some researchers that PSTs have higher TSE relative to ISTs and experience “reality shock,” and therefore lowered TSE, on entering the profession (Weinstein, 1988). Another noticeable feature in the data was PSTs’ regarding Item 1, *Getting through to the most difficult students*, to be the item on which they most frequently said they had no idea at all, and both PSTs and ISTs regarding the same item as being associated with their lowest level of perceived efficacy. Other researchers (e.g., McCampbell, 2014; O’Neill & Stephenson, 2012) have also found this item to be the one on which respondents believed themselves to be least efficacious. The uniqueness of this item, which is regularly cited as a hallmark of TSE and frequently appears in TSE scales, including the TSES, suggests that it might be advantageously reworded or even abandoned. Reasons for not retaining this item in its present form could be its having attracted the greatest number of outliers in this research, its manifest ambiguity being disconcerting for respondents, and its focus on a possibly irksome aspect of teaching—an unfortunate focus because placing this item at the start of the inventory could disincline respondents from continuing beyond the first few items. It is perhaps fortunate that this item is not on the short form of the TSES. However, that could not be taken advantage of in the present research because the short form was not administered independently and, therefore, this problematic item was the first that participants encountered.

Summary, Comments, and Recommendations

Although the amount of information we provided about response and retention rates within the results section of this article might be more elaborated than is often needed, we believe we have provided useful exemplars about provision of information, and, more particularly in this case, because our having done so could lead to improvements in scale design. Specifically, response and retention rates might be enhanced if participants are provided with initial information about the number of items on a survey and how long the survey would take to complete; if instructions are appropriate for the targeted respondents; if a cumbersome format and ambiguity are avoided; if features of language such as word usage and grammar are polished; if items of low salience to participants, or that are challenging, are not placed at the start of a scale; if response options correspond semantically with the wording of items; and if participants are offered all response options they might want to use. Furthermore, items that possess undesirable psychometric attributes and persistently appear to be problematic for respondents should not continue to be used unaltered, or should be abandoned. As a side note, this research provides evidence that a larger number of response options does not necessarily produce greater variability within data and that wording above more response options than only the extremes might be advantageous.

Psychometric Insights, Issues, and Implications

In this section, we focus on psychometric insights, issues, and implications that arose for us when we conducted this study. These fall into three main categories: general psychometric issues, insights, and implications; insights about single-factor results; and insights about construct conception and coverage.

General Psychometric Issues, Insights, and Implications

A variety of general psychometric issues, insights, and implications arise from this research, some of which are not commonly revealed or discussed in the TSE literature. In this study, response and retention rates, missing data, skewness and kurtosis, and outliers all raised no, or few, concerns about the nature of our data,

but that kind of information, or reassurance, is infrequently provided in other research about TSE. Information about response and retention rates, when provided, is usually undefined, and of particular concern is the TSES almost always yielding scores that are well above the midpoint of its option range, as evidenced by the mean of those scores frequently being at least one *SD* above that midpoint. As a result, data are not only negatively skewed but also likely to produce outliers at the lower end of distributions that could distort substantive analyses. Despite the high likelihood of outliers, we could find no mention of them in the TSE literature.

In this study, a number of metrics associated with factor analysis were satisfactory. The KMO index and Bartlett's test of sphericity were highly satisfactory in all analyses, and all item loadings were greater than 0.45. These desirable attributes are not uncommon in research involving the TSES. However, communalities for the PSTs' data were unsatisfactorily low on both long and short TSE forms, and communalities for the ISTs' data were only a little better on both forms.²⁶ Furthermore, the percentage of variance accounted for by the single factor on the PSTs' data was unsatisfactorily low on both long and short forms (approximately 43 % and 47 %, respectively) and only a little better on both forms for the IST data (approximately 54 % and 58 %, respectively). These percentages of variance are lower than in some other research with the TSES, but amounts of variance in TSE research seldom approach the desirable 70 % (see, e.g., Cheung, 2006; Fives & Buehl, 2010; Htang, 2018; Tschannen-Moran & Woolfolk Hoy, 2001). In our experience, mention of communalities does not feature in accounts of research involving the TSES and similar scales, and low percentages of variance are simply reported without being commented on. Under these circumstances, it is easy for deficiencies in scales to be unexposed and for researchers as well as research consumers to have misplaced confidence in those scales.

Although the spread of loadings for all factor solutions in this study indicated that domain breadth is not undesirably limited within either the long- or short-form items from the TSES, deficiency in domain breadth is suggested by the alphas for the PSTs and ISTs on the 12 short-form items being 0.90 and 0.93, respectively. These high alphas from a relatively small number of items indicate

²⁶ The low communalities could be offset on all but one of the items (*Getting through to the most difficult students*) by them having loadings > 0.50 for both PSTs and ISTs.

that the 12 short-form items contain an undesirable degree of conceptual redundancy in both samples because, for 12 items, alphas in the region of 0.75 to 0.85 would have been not only adequate, but more desirable, as indicators of nonrepetitive tapping of the targeted domain. Most interitem correlations were acceptable in the PSTs' data, thus countering the prospect of redundancy there. However, the high interitem correlations in the ISTs' data, where 58 % of those correlations exceeded 0.50 and their mean was 0.52, suggest redundancy among some items—and even visual inspection of the TSES items reveals some of those items to have manifest repetitiveness.

Insights Arising from Single-Factor Results

A major inescapable outcome of this research is that, for both PSTs and ISTs among our Chinese participants, a single factor embraced the items on both the long and short forms of the TSES-based items. Therefore, the common belief that the TSES and its close variants always comprise three factors is strongly disconfirmed, as, indeed, it had been in many studies reviewed in the introduction of this article. Although it would be unreasonable to require identical cross-sample loadings before measurement invariance could be claimed, we obtained such similar results with data from the PSTs and ISTs, from primary and high school teachers, and from subject and head teachers, that the existence and composition of a single factor within our participants appears to be robust. This robustness is more impressive because of item-loading similarity for subject and head teachers between whom a difference might have been expected given the dissimilar responsibilities they carry. Overall, therefore, our results strengthen the likelihood of comparisons being valid if these scales were used to investigate Chinese PSTs and ISTs.

The PSTs' items loading on a single factor might not be regarded as surprising given other researchers having obtained a similar result with PSTs (see, e.g., Duffin et al., 2012; Fives & Buehl, 2010; O'Neill & Stephenson, 2012; Tschannen-Moran & Woolfolk Hoy, 2001). However, the ISTs' items also loading on a single factor in this study suggests that Chinese teachers regard, perhaps subconsciously, the assessed aspects of teaching to be integrated in terms of their efficacy as teachers, thus representing a merged combination of engaging students, providing instruction, and managing student behavior. This lack of

compartmentalization among Chinese ISTs has been indicated in other research. For example, Ho and Hau (2004) found that, for teachers in Hong Kong, instructing students, managing their behavior, and providing them with guidance appeared to be “identical constructs” (p. 320), whereas, by comparison, the Australian teachers in their research tended to regard those three domains as separate from each other.

Heine and Buchtel (2009) discussed a similar tendency toward non-compartmentalization in collectivist as opposed to individualist cultures when considering how personality traits and self-conceptions are constructed. In doing so, they proposed that “one cannot fully understand the nature of people without considering the cultural context within which they exist” (p. 370).²⁷ In a similar vein, Chen (2008) highlighted the need to exercise caution when comparing people from different cultures on the basis of data acquired from inventories. She argued strongly for the importance of researchers ensuring scale invariance. Her warning is worth heeding given the results reported in this article that demonstrate a higher level of integration among the TSES items than is often reported or assumed in non-Chinese research.

Insights about Construct Conception and Coverage

The single-factor results in this study do more than provide insights about the nature of TSE among Chinese PSTs and ISTs and the validity of TSE comparisons across differing cultural contexts. They also raise issues about the TSES in particular and the measurement of TSE more broadly.

Our results indicate that the common, although not universal, finding of a single TSE factor for PSTs might benefit from reinterpretation. Some researchers have proposed that single-factor results from PSTs indicate a deficiency on the part of PSTs. For example, Fives and Buehl (2010) suggested that PSTs are unable to distinguish between the different aspects of teaching and, in their naivety, do not perceive teaching to be a highly complex task. Tschannen-Moran and Woolfolk Hoy (2001) suggested, similarly, that it would only be with actual teaching experience that PSTs could appreciate the differentiated components of teaching.

²⁷ See Klassen (2004) for comprehensive and insightful information about people’s sense of personal efficacy in collectivist and individualist cultures.

These might not be the only ways to interpret single-factor results. The presence of a single factor to represent an activity as complex as teaching might relate to more than PSTs' conceptions of teaching, and, in this research, to ISTs' conceptions of teaching. Rather, the single factor can raise concerns about the diversity of items on the TSES and similar scales. The high alphas that are frequently cited in relation to the three separate eight-item TSES subscales indicate an undesirable degree of redundancy within each of those subscales, and the high alphas for the combined 12 items on the TSES short form in this and other research indicates the same phenomenon.

A basis for this redundancy could be the almost-total focus on classroom activities that have direct relevance to students, which, for PSTs and ISTs in a Chinese teaching environment, appear to be integrated. For them, therefore, researchers might consider not only preferencing the short-form over long-form items from the TSES when examining TSE in China, but they might also consider removing some of those 12 short-form items as being superfluous for capturing Chinese PSTs' and ISTs' sense of efficacy for teaching, and replacing them with items representing a wider range of perspectives or facets concerning classroom activities. For example, some of the overtly similar items that focus on controlling unwanted student behavior could be replaced with items about encouraging courteous, cooperative, or facilitative behavior among students. Replacement items could also relate to broader aspects of teaching such as perceived efficacy with regard to preparation and planning of teaching activities and materials as well as to proficiency in the subject matter (see Heneman et al., 2006).

The sources from which teachers gauge their sense of efficacy within the context of their profession are likely to extend beyond the classroom, however. Friedman and Kass (2002) added a domain that refers to organizational aspects of teaching. Bandura (2006) suggested 28 items spanning six overtly different domains within which TSE could be measured. Three of those domains (decision making in the school, enlisting community involvement, and creating a positive school climate) are not represented in the TSES. Nor is guiding and counselling students (see Chan, 2008a). Skaalvik and Skaalvik (2007) identified the domains of adapting education to individual students' needs, cooperating with colleagues and parents, and coping with change and challenges as being absent from the TSES. The last of these domains might be a particularly undesirable oversight

given the now well-established importance of technology in teaching (see Darling-Hammond, 2006; Dede, Yilmaz, & Ilhan, 2017; Gong & Lai, 2018; Tsai & Chai, 2012). Sang et al. (2010) even referred specifically to TSE with computers. Malinen (2013) identified efficacy domains of inclusive instruction, collaboration, and managing behavior; Yin et al. (2013) identified efficacy domains in relation to professional growth, participation in decision making, and perceived impact on colleagues; and Mischo (2015) identified a range of activities that he categorized as either child-related competencies or environment-related competencies. Although some of these domains overlap each other and with the domains represented by the TSES, the recommendation made more than 10 years ago by Poulou (2007, p. 214) that “future research should encompass additional dimensions of teaching efficacy in order to reflect the multidimensionality and complexity of teaching” continues to resonate.

Summary, Comments, and Recommendations

Research concerning TSE could be improved if researchers were more willing to clearly define and report aspects of their data as well as the results obtained from those data. Greater attention should also be paid to details within results so that improvements in the conception and measurement of TSE can be made. Using that information as a foundation, development of scales that capture the breadth of teachers’ professional lives more effectively should be encouraged, even though considerable effort needs to be exerted to establish scales with satisfactory reliability and validity. Without that development, and to return to the metaphors used in the opening paragraphs of this article, TSE research might be deprived of a rich adulthood and an even richer old age. That the TSES is still used so widely suggests a momentum characterized by researchers using scales off the shelf without sufficiently questioning, examining, or improving, those scales’ effectiveness, reliability, content validity, and usefulness.

Conclusions

This research indicates that placing confidence in scales merely because of their availability, wide use by other researchers, and assumed trustworthy provenance might be misguided and that response and retention rates as well as data quality

would improve if there was greater adherence to basic principles of survey design associated with formulation of instructions, response options, and item wording and placement. Furthermore, researchers should be encouraged to examine and report their results more comprehensively and transparently in order to assess the quality of their data and results, in order to allow and encourage others to do the same, and in order to reveal possibilities for improvements in scales. This might be advantageous for the utility of research in which perennial issues such as content of teacher education programs, enthusiasm of teachers for teaching, teacher job satisfaction, teacher burnout, and retention of teachers in the profession are investigated.

Although the representativeness of the samples in this research cannot be guaranteed and there is acknowledged item redundancy as well as constricted construct coverage, we believe that the data we obtained provide a sufficiently solid foundation for exploring TSE, within a framework consonant with the TSES, among Chinese PSTs and ISTs because of the sizeable number and heterogeneity of participants, removal of dubious responses, high retention rates, and small amount of missing data. We have conducted some exploration of that nature in a subsequent publication (Ma & Trevethan, 2020) in which we acknowledge that items from the TSES are confined to a sense of efficacy for teaching (SET)—the latter being deliberately focused on classroom activities rather than on broader aspects of teachers' professional lives. In conducting that research as well as in conducting this present study, we believe that, as Malinen (2016, p. 122) has argued, "a closer collaboration with international researchers and bringing research more accessible to non-Chinese readers would benefit both international and mainland Chinese research."

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Acknowledgements Sylvia McAra and Michael Moxham suggested improvements to earlier drafts of this manuscript. Colleagues in Australia and China, including Qing Xu, Yichi Zhang,

and Tengfei Gong, provided assistance with translating and back translating the surveys, and other colleagues helped with accessing participants.

References

- Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods*, 16(2), 270–301. <https://doi.org/10.1177/1094428112470848>
- Alt, D. (2018). Science teachers' conceptions of teaching and learning, ICT efficacy, ICT professional development and ICT practices enacted in their classrooms. *Teaching and Teacher Education*, 73, 141–150. <https://doi.org/10.1016/j.tate.2018.03.020>
- Armor, D., Conroy-Oseguera, P., Cox, M., King, N., McDonnell, L., Pascal, A., ... Zellman, G. (1976). *Analysis of the school preferred reading programs in selected Los Angeles minority schools* (Report No. R-2007-LAUDS). Retrieved from <https://www.rand.org/content/dam/rand/pubs/reports/2005/R2007.pdf>
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York, NY: Freeman.
- Bandura, A. (2006). Guide for creating self-efficacy scales. In F. Pajares & T. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (Vol. 5, pp. 307–337). Greenwich, CT: Information Age Publishing.
- Berg, D. A. G., & Smith, L. F. (2016). Preservice teacher self-efficacy beliefs: An opportunity to generate “good research” in the Asia-Pacific region. In S. Garvis & D. Pendergast (Eds.), *Asia-Pacific perspectives on teacher self-efficacy* (pp. 1–17). Rotterdam, The Netherlands: Sense.
- Blair, E., & Zinkhan, G. M. (2006). Non-response and generalizability in academic research. *Journal of the Academy of Marketing Science*, 34(1), 4–7. <https://doi.org/10.1177/0092070305283778>
- Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quinonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health*, 6, 149. <https://doi.org/10.3389/fpubh.2018.00149>
- Boyle, G. J. (1991). Does item homogeneity indicate internal consistency or item redundancy in psychometric scales? *Personality and Individual Differences*, 12(3), 291–294. [https://doi.org/10.1016/0191-8869\(91\)90115-R](https://doi.org/10.1016/0191-8869(91)90115-R)
- Brick, J. M., & Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, 5(3), 215–238. <https://doi.org/10.1177/096228029600500302>
- Briggs, S. R., & Cheek, J. M. (1986). The role of factor analysis in the development and evaluation of personality scales. *Journal of Personality*, 54(1), 106–148. <https://doi.org/10.1111/j.1467-6494.1986.tb00391.x>
- Bruinsma, M., & Jansen, E. P. W. A. (2010). Is the motivation to become a teacher related to pre-service teachers' intentions to remain in the profession? *European Journal of Teacher*

- Education*, 33(2), 185–200. <https://doi.org/10.1080/02619760903512927>
- Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Boston, MA: Houghton Mifflin.
- Carifio, J., & Perla, R. J. (2007). Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes. *Journal of Social Sciences*, 3(3), 106–116. <https://doi.org/10.3844/jssp.2007.106.116>
- Cattell, R. B., & Vogelmann, S. (1977). A comprehensive trial of the scree and KG criteria for determining the number of factors. *Multivariate Behavioral Research*, 14(3), 289–325. https://doi.org/10.1207/s15327906mbr1203_2
- Chan, C. K. K., & Rao, N. (2009). The paradoxes revisited: The Chinese learner in changing educational contexts. In C. K. K. Chan & N. Rao (Eds.), *Revisiting the Chinese learner: Changing context, changing education* (pp. 315–349). Hong Kong, China: Springer.
- Chan, D. W. (2008a). Dimensions of teacher self-efficacy among Chinese secondary school teachers in Hong Kong. *Educational Psychology*, 28(2), 181–194. <https://doi.org/10.1080/01443410701491833>
- Chan, D. W. (2008b). General, collective, and domain-specific teacher self-efficacy among Chinese prospective and in-service teachers in Hong Kong. *Teaching and Teacher Education*, 24(4), 1057–1069. <https://doi.org/10.1016/j.tate.2007.11.010>
- Chang, M-L, & Engelhard, G. (2016). Examining the Teachers' Sense of Efficacy Scale at the item level with Rasch measurement model. *Journal of Psychoeducational Assessment*, 34(2), 177–191. <https://doi.org/10.1177/0734282915593835>
- Charalambous, C. Y., Philippou, G. N., & Kyriakides, L. (2008). Tracing the development of preservice teachers' efficacy beliefs in teaching mathematics during fieldwork. *Educational Studies in Mathematics*, 67(2), 125–142. <https://doi.org/10.1007/s10649-007-9084-2>
- Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology*, 95(5), 1005–1018. <https://doi.org/10.1037/a0013193>
- Cheon, S. H., Reeve, J., Yu, T. H., & Jang, H. R. (2014). The teacher benefits from giving autonomy support during physical education instruction. *Journal of Sport & Exercise Psychology*, 36(4), 331–346. <https://doi.org/10.1123/jsep.2013-0231>
- Chesnut, S. R. (2017). On the measurement of preservice teacher commitment: Examining the relationship between four operational definitions and self-efficacy beliefs. *Teaching and Teacher Education*, 68, 170–180. <https://doi.org/10.1016/j.tate.2017.09.003>
- Chesnut, S. R., & Burley, H. (2015). Self-efficacy as a predictor of commitment to the teaching profession: A meta-analysis. *Educational Research Review*, 15, 1–16. <https://doi.org/10.1016/j.edurev.2015.02.001>
- Cheung, H. Y. (2006). The measurement of teacher efficacy: Hong Kong primary in-service teachers. *Journal of Education for Teaching*, 32(4), 435–451. <https://doi.org/10.1080/02607470600982134>
- Cheung, H. Y. (2008). Teacher efficacy: A comparative study of Hong Kong and Shanghai

- primary in-service teachers. *Australian Educational Researcher*, 35(1), 103–123. <https://doi.org/10.1007/BF03216877>
- Cho, E., & Kim, S. (2015). Cronbach's coefficient alpha: Well known but poorly understood. *Organizational Research Methods*, 18(2), 207–230. <https://doi.org/10.1177/1094428114555994>
- Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, 7(3), 309–319. <https://doi.org/10.1037/1040-3590.7.3.309>
- Çokluk, Ö., & Koçak, D. (2016). Using Horn's parallel analysis method in exploratory factor analysis for determining the number of factors. *Educational Sciences: Theory & Practice*, 16(2), 537–551. <https://doi.org/10.12738/estp.2016.2.0328>
- Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78(1), 98–104. <https://doi.org/10.1037/0021-9010.78.1.98>
- Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, 10(7), 1–9.
- Crawford, A. V., Green, S. B., Levy, R., Lo, W.-J., Scott, L., Svetina, D., & Thompson, M. S. (2010). Evaluation of parallel analysis method for determining the number of factors. *Educational and Psychological Measurement*, 70(6), 885–901. <https://doi.org/10.1177/0013164410379332>
- Crutzen, R., & Peters, G.-J. Y. (2017). Scale quality: Alpha is an inadequate estimate and factor-analytic evidence is needed first of all. *Health Psychology Review*, 11(3), 242–247. <https://doi.org/10.1080/17437199.2015.1124240>
- Darbyshire, P., & McDonald, H. (2004). Choosing response scale labels and length: Guidance for researchers and clients. *Australasian Journal of Market Research*, 12(2), 17–26. <https://doi.org/10.1017/S1551929500051968>
- Darling-Hammond, L. (2006). Constructing 21st-century teacher education. *Journal of Teacher Education*, 57(3), 300–314. <https://doi.org/10.1177/0022487105285962>
- Darling-Hammond, L., Burns, D., Campbell, C., Goodwin, A. L., Hammerness, K., Low, E. L., ... Zeichner, K. (2017). *Empowered educators: How high-performing systems shape teaching quality around the world*. San Francisco, CA: Jossey-Bass.
- Data cleaning. (n.d.). *Practical analysis of nutritional data (PANDA)*. Tulane University. Retrieved from <http://www.tulane.edu/~panda2/Analysis2/datclean/dataclean.htm>
- Dede, H., Yilmaz, Z. A., & Ilhan, N. (2017). Investigation of the self-efficacy beliefs of pre-service science teachers in terms of following and using the innovations in the field of education. *Journal of Education and Training Studies*, 5(2), 21–30. <https://doi.org/10.11114/jets.v5i2.2025>
- Dixon, J. K. (2005). Exploratory factor analysis. In B. H. Munro (Ed.), *Statistical methods for health care research* (5th ed., pp. 321–349). Philadelphia, PA: Lippincott Williams &

Wilkins.

- Dou, D., Devos, G., & Valcke, M. (2017). The relationships between school autonomy gap, principal leadership, teachers' job satisfaction and organizational commitment. *Educational Management Administration & Leadership*, 45(6), 959–977. <https://doi.org/10.1177/1741143216653975>
- Duffin, L. C., French, B. F., & Patrick, H. (2012). The Teachers' Sense of Efficacy Scale: Confirming the factor structure with beginning pre-service teachers. *Teaching and Teacher Education*, 28(6), 827–834. <https://doi.org/10.1016/j.tate.2012.03.004>
- Durksen, T. L., Klassen, R. M., & Daniels, L. M. (2017). Motivation and collaboration: The keys to a developmental framework for teachers' professional learning. *Teaching and Teacher Education*, 67, 53–66. <https://doi.org/10.1016/j.tate.2017.05.011>
- Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4(3), 272–299. <https://doi.org/10.1037/1082-989X.4.3.272>
- Fackler, S., & Malmberg, L.-E. (2016). Teachers' self-efficacy in 14 OECD countries: Teacher, student group, school and leadership effects. *Teaching and Teacher Education*, 56, 185–195. <https://doi.org/10.1016/j.tate.2016.03.002>
- Feng, Y. J., & Wang, M. (2014). A survey study of Chinese pre-service teachers' self-efficacy and attitude to inclusive education. In CIEd, University of Minho (Eds.), *Proceedings of Braga 2014: Embracing inclusive approaches for children and youth with special needs* (pp. 695–699). Braga, Portugal: CIEd, University of Minho.
- Fives, H., & Buehl, M. M. (2010). Examining the factor structure of the Teachers' Sense of Efficacy Scale. *Journal of Experimental Education*, 78(1), 118–134. <https://doi.org/10.1080/00220970903224461>
- Friedman, I. A., & Kass, E. (2002). Teacher self-efficacy: A classroom-organization conceptualization. *Teaching and Teacher Education*, 18(6), 675–686. [https://doi.org/10.1016/S0742-051X\(02\)00027-6](https://doi.org/10.1016/S0742-051X(02)00027-6)
- Gao, L. B., & Watkins, D. A. (2002). Conceptions of teaching held by school science teachers in P. R. China: Identification and cross-cultural comparisons. *International Journal of Science Education*, 24(1), 61–79. <https://doi.org/10.1080/09500690110066926>
- Gaskin, C. J., & Happell, B. (2014). On exploratory factor analysis: A review of recent evidence, an assessment of current practice, and recommendations for future use. *International Journal of Nursing Studies*, 51(3), 511–521. <https://doi.org/10.1016/j.ijnurstu.2013.10.005>
- Gavora, P. (2010). Slovak pre-service teacher self-efficacy: Theoretical and research considerations. *The New Educational Review*, 21(2), 17–30.
- George, S. V., Richardson, P. W., & Watt, H. M. G. (2018). Early career teachers' self-efficacy: A longitudinal study from Australia. *Australian Journal of Education*, 62(2), 217–233. <https://doi.org/10.1177/0004944118779601>
- Gibson, S., & Dembo, M. H. (1984). Teacher efficacy: A construct validation. *Journal of Educational Psychology*, 76(4), 569–582. <https://doi.org/10.1037/0022-0663.76.4.569>

- Gong, Y., & Lai, C. (2018). Technology integration into the language classroom: Developmental trajectory of beginning teachers. *Frontiers of Education in China*, 13(1), 1–27. <https://doi.org/10.1007/s11516-018-0001-5>
- Goodwin, L. D., & Leech, N. L. (2006). Understanding correlation: Factors that affect the size of r . *Journal of Experimental Education*, 74(3), 249–266. <https://doi.org/10.3200/JEXE.74.3.249-266>
- Gu, L. M., Chen, J., & Li, J. C. (2015). The leadership of “banzhuren” in Chinese school: Based on the sample survey in Changzhou city of China. *Journal of Education and Human Development*, 4(4), 102–114. <http://dx.doi.org/10.15640/jehd.v4n4a13>
- Guskey, T. R., & Passaro, P. D. (1994) Teacher efficacy: A study of construct dimensions. *American Educational Research Journal*, 31(3), 627–643. <https://doi.org/10.3102/00028312031003627>
- Hair, J. F., Jr., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). Exploratory factor analysis. In J. F. Hair Jr., W. C. Black, B. J. Babin, & R. E. Anderson (Eds.), *Multivariate data analysis* (pp. 89–149). Harlow, UK: Pearson.
- Hatlevik, I. K. R. (2017). The impact of prospective teachers’ perceived competence on subsequent perceptions as schoolteachers. *Teachers and Teaching: Theory and Practice*, 23(7), 810–828. <https://doi.org/10.1080/13540602.2017.1322056>
- Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, 9(2), 139–164. <https://doi.org/10.1177/014662168500900204>
- Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods*, 7(2), 191–205. <https://doi.org/10.1177/1094428104263675>
- He, N., & Miao, D. M. (2006). Retrospect and prospect of studies of teacher efficacy in China. *US-China Education Review*, 3(3), 31–36.
- Heine, S. J., & Buchtel, E. E. (2009). Personality: The universal and the culturally specific. *Annual Review of Psychology*, 60(1), 369–394. <https://doi.org/10.1146/annurev.psych.60.110707.163655>
- Hellerstein, J. M. (2008). *Quantitative data cleaning for large databases*. Retrieved from <http://db.cs.berkeley.edu/jmh/papers/cleaning-unece.pdf>
- Heneman, H. G. III, Kimball, S., & Milanowski, A. (2006). *The Teacher Sense of Efficacy Scale: Validation evidence and behavioral prediction*. WCER Working paper No2006-7. <https://eric.ed.gov/?id=ED496313>
- Henson, R. K. (2002). From adolescent angst to adulthood: Substantive implications and measurement dilemmas in the development of teacher efficacy research. *Educational Psychologist*, 37(3), 137–150. https://doi.org/10.1207/S15326985EP3703_1
- Herman, K. C., Hickmon-Rosa, J., Reinke, W. M. (2018). Empirically derived profiles of teacher stress, burnout, self-efficacy, and coping and associated student outcomes. *Journal of Positive Behavior Interventions*, 20(2), 90–100. <https://doi.org/10.1177/1098300717732066>

- Ho, I. T. (2004). A comparison of Australian and Chinese teachers' attributions for student problem behaviors. *Educational Psychology, 24*(3), 375–391. <https://doi.org/10.1080/0144341042000211706>
- Ho, I. T., & Hau, K.-T. (2004). Australian and Chinese teacher efficacy: Similarities and differences in personal instruction, discipline, guidance efficacy and beliefs in external determinants. *Teaching and Teacher Education, 20*(3), 313–323. <https://doi.org/10.1016/j.tate.2003.09.009>
- Hogarty, K. Y., Hines, C. V., Kromrey, J. D., Ferron, J. M., & Mumford, K. R. (2005). The quality of factor solutions in exploratory factor analysis: The influence of sample size, communality, and overdetermination. *Educational and Psychological Measurement, 65*(2), 202–226. <https://doi.org/10.1177/0013164404267287>
- Hoi, C. K. W., Zhou, M. M., Teo, T., & Nie, Y. Y. (2017). Measuring efficacy sources: Development and validation of the Sources of Teacher Efficacy Questionnaire (STEQ) for Chinese teachers. *Psychology in the Schools, 54*(7), 756–769. <https://doi.org/10.1002/pits.22025>
- Hong, J. Y. (2010). Pre-service and beginning teachers' professional identity and its relation to dropping out of the profession. *Teaching and Teacher Education, 26*(8), 1530–1543. <https://doi.org/10.1016/j.tate.2010.06.003>
- Hoy, W. K., & Woolfolk, A. E. (1993). Teachers' sense of efficacy and the organizational health of schools. *The Elementary School Journal, 93*(4), 355–372. <https://doi.org/10.1086/461729>
- Htang, L. K. (2018). Measurement of teacher sense of efficacy: A study with Myanmar in-service teachers. *Journal of Education and Practice, 9*(35), 39–48.
- Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology, 27*(1), 99–114. <https://doi.org/10.1007/s10869-011-9231-8>
- Jackson, W. (2003). *Methods: Doing social research* (3rd ed.). Toronto, Canada: Prentice Hall.
- Jacoby, J., & Matell, M. S. (1971). Three-point Likert scales are good enough. *Journal of Marketing Research, 8*(4), 495–500. <https://doi.org/10.1177/002224377100800414>
- Kennedy, K., & Hui, S. K. F. (2004). *Self-efficacy as a key attribute for curriculum leadership—Linking individual and organizational capacity to meet the challenges of curriculum reform*. Paper presented at the annual conference of the Commonwealth Council of Educational Administration and Management, Hong Kong.
- Kennedy, K. J., & Hui, S. K. F. (2006). Developing teacher leaders to facilitate Hong Kong's curriculum reforms: Self-efficacy as a measure of teacher growth. *International Journal of Educational Reform, 15*(1), 137–151. <https://doi.org/10.1177/105678790601500107>
- Klassen, R. M. (2004). Optimism and realism: A review of self-efficacy from a cross-cultural perspective. *International Journal of Psychology, 39*(3), 205–230. <https://doi.org/10.1080/00207590344000330>
- Klassen, R. M., Bong, M., Usher, E. L., Chong, W. H., Huan, V. S., Wong, I. Y. F., & Georgiou, T. (2009). Exploring the validity of a teachers' self-efficacy scale in five countries.

- Contemporary Educational Psychology*, 34(1), 67–76. <https://doi.org/10.1016/j.cedpsych.2008.08.001>
- Klassen, R. M., & Chiu, M. M. (2010). Effects on teachers' self-efficacy and job satisfaction: Teacher gender, years of experience, and job stress. *Journal of Educational Psychology*, 102(3), 741–756. <https://doi.org/10.1037/a0019237>
- Klassen, R. M., Tze, V. M. C., Betts, S. M., & Gordon, K. A. (2011). Teacher efficacy research 1998–2009: Signs of progress or unfulfilled promise? *Educational Psychology Review*, 23(1), 21–43. <https://doi.org/10.1007/s10648-010-9141-8>
- Kleinsasser, R. C. (2014). Teacher efficacy in “Teaching and Teacher Education”. *Teaching and Teacher Education*, 44, 168–179. <https://doi.org/10.1016/j.tate.2014.07.007>
- Krosnick, J., & Presser, S. (2010). Question and questionnaire design. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (2nd ed., pp. 263–313). Bingley, UK: Emerald.
- Li, Y. (2015). Accessing fellow academics as research participants: Constraints, collegiality, and “academic citizenship.” *Publications*, 3(2), 131–149. <https://doi.org/10.3390/publications30201312>
- Liu, Y., Zumbo, B. D. (2007). The impact of outliers on Cronbach's coefficient alpha estimate of reliability. *Educational and Psychological Measurement*, 67(4), 620–634. <https://doi.org/10.1177/0013164406296976>
- Liu, Y., Zumbo, B. D., & Wu, A. D. (2012). A demonstration of the impact of outliers on the decisions about the number of factors in exploratory factor analysis. *Educational and Psychological Measurement*, 72(2), 181–199. <https://doi.org/10.1177/0013164411410878>
- Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 4(2), 73–79. <https://doi.org/10.1027/1614-2241.4.2.73>
- Lu, M. H., Hao, L., Chen, X. M., & Potmėšilc, M. (2018). Teacher efficacy, work engagement, and social support among Chinese special education school teachers. *Frontiers in Psychology*, 9, 648. <https://doi.org/10.3389/fpsyg.2018.00648>
- Ma, K. (2017). *Classroom ready? Pre-service teachers' self-efficacy for their first professional experience placement* (Unpublished master's thesis). Macquarie University, Sydney, Australia.
- Ma, K., & Trevethan, R. (2020). Efficacy perceptions of preservice and inservice teachers in China: Insights concerning culture and measurement. *Frontiers of Education in China*. In Press.
- Malinen, O.-P. (2013). *Inclusive education from teachers' perspective: Examining pre- and in-service teachers' self-efficacy and attitudes in mainland China* (Unpublished doctoral dissertation). University of Eastern Finland, Joensuu, Finland.
- Malinen, O.-P. (2016). Self-efficacy research in mainland China. In S. Garvis & D. Pendergast (Eds.), *Asia-Pacific perspectives on teacher self-efficacy* (pp. 115–126). Rotterdam, The Netherlands: Sense.
- Malinen, O.-P., Savolainen, H., & Xu, J. C. (2012). Beijing in-service teachers' self-efficacy

- and attitudes towards inclusive education. *Teaching and Teacher Education*, 28(4), 526–534. <https://doi.org/10.1016/j.tate.2011.12.004>
- Malinen, O.-P., Savolainen, H., Engelbrecht, P., Xu, J. C., Nel, M., Nel, N., & Tlale, D. (2013). Exploring teacher self-efficacy for inclusive practices in three diverse countries. *Teaching and Teacher Education*, 33, 34–44. <https://doi.org/10.1016/j.tate.2013.02.004>
- Malinen, O.-P., Savolainen, H., & Xu, J. C. (2013). Dimensions of teacher self-efficacy for inclusive practices among mainland Chinese pre-service teachers. *Journal of International Special Needs Education*, 16(2), 82–93. <https://doi.org/10.9782/2159-4341-16.2.82>
- Manzar-Abbas, S. S., & Lu, L. J. (2015). Self-efficacy beliefs of Chinese primary school teachers. *Pakistan Journal of Psychological Research*, 30(2), 289–303.
- Matsunaga, M. (2010). How to factor-analyze your data right: Do's don'ts, and how-to's. *International Journal of Psychological Research*, 3(1), 97–110. <https://doi.org/10.21500/20112084.854>
- McCampbell, S. M. (2014). *Pre-service teachers' self-efficacy for teaching mathematics* (Unpublished doctoral dissertation). The University of New Mexico, Albuquerque, NM.
- McDowell, I., & Newell, C. (1996). *Measuring health: A guide to rating scales and questionnaires* (2nd ed.). New York, NY: Oxford University Press.
- McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, 23(3), 412–433. <https://doi.org/10.1037/met0000144>
- Mergler, A. G., & Tangen, D. (2010). Using microteaching to enhance teacher efficacy in pre-service teachers. *Teaching Education*, 21(2), 199–210. <https://doi.org/10.1080/10476210902998466>
- Mischo, C. (2015). Early childhood teachers' perceived competence during transition from teacher education to work: Results from a longitudinal study. *Professional Development in Education*, 41(1), 75–95. <https://doi.org/10.1080/19415257.2014.886282>
- Moore-Hayes, C. (2011). Technology integration preparedness and its influence on teacher-efficacy. *Canadian Journal of Learning and Technology*, 37(3), 1–15. <https://doi.org/10.21432/T2B597>
- Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing*, 5(2), 159–168. https://doi.org/10.1207/s15327574ijt0502_4
- Murphy, K. R., & Davidshofer, C. O. (1998). *Psychological testing: Principles and applications* (4th ed.). Upper Saddle River, NJ: Prentice Hall.
- Nie, Y. Y., Lau, S., & Liao, A. K. (2012). The teacher efficacy scale: A reliability and validity study. *The Asia-Pacific Education Researcher*, 21(2), 414–421. <http://hdl.handle.net/10497/14287>
- Nunnally, J. C. (1978). *Psychometric theory*. New York, NY: McGraw Hill.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw Hill.
- O'Neill, S., & Stephenson, J. (2012). Exploring Australian pre-service teachers sense of efficacy, its sources, and some possible influences. *Teaching and Teacher Education*, 28(4),

- 535–545. <https://doi.org/10.1016/j.tate.2012.01.008>
- Organisation for Economic Co-Operation and Development. (2011). *Strong performers and successful reformers in education: Lessons from PISA for the United States*. Paris, France: OECD.
- Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should ALWAYS check for them). *Practical Assessment, Research & Evaluation*, 9(6), 1–12. <https://pareonline.net/getvn.asp?v=9&n=6>
- Ottley, J. R., Piasta, S. B., Mauck, S. A., O'Connell, A., Weber-Mayrer, M., & Justice, L. M. (2015). The nature and extent of change in early childhood educators' language and literacy knowledge and beliefs. *Teaching and Teacher Education*, 52, 47–55. <https://doi.org/10.1016/j.tate.2015.08.005>
- Overbaugh, R., & Lu, R. (2008). The impact of a NCLB-EETT funded professional development program on teacher self-efficacy and resultant implementation. *Journal of Research on Technology in Education*, 41(1), 43–61. <https://doi.org/10.1080/15391523.2008.10782522>
- Paine, L., & Ma, L. P. (1993). Teachers working together: A dialogue on organizational and cultural perspectives of Chinese teachers. *International Journal of Educational Research*, 19(8), 675–697. [https://doi.org/10.1016/0883-0355\(93\)90009-9](https://doi.org/10.1016/0883-0355(93)90009-9)
- Pallant, J. F. (2016). *SPSS survival manual: A step by step guide to data analysis using SPSS* (6th ed.). Sydney, Australia: Allen & Unwin.
- Panayides, P. (2013). Coefficient alpha: Interpret with caution. *Europe's Journal of Psychology*, 9(4), 687–696. <https://doi.org/10.5964/ejop.v9i4.653>
- Peña, E. D. (2007). Lost in translation: Methodological considerations in cross-cultural research. *Child Development*, 78(4), 1255–1264. <https://doi.org/10.1111/j.1467-8624.2007.01064.x>
- Pendergast, D., Garvis, S., & Keogh, J. (2011). Pre-service student-teacher self-efficacy beliefs: An insight into the making of teachers. *Australian Journal of Teacher Education*, 36(12), 45–58. <https://doi.org/10.14221/ajte.2011v36n12.6>
- Perera, H. N., Calkins, C., & Part, R. (2019). Teacher self-efficacy profiles: Determination, outcomes, and generalizability across teaching level. *Contemporary Educational Psychology*, 58, 186–203. <https://doi.org/10.1016/j.cedpsych.2019.02.006>
- Pfitzner-Eden, F. (2016). I feel less confident so I quit? Do true changes in teacher self-efficacy predict changes in preservice teachers' intention to quit their teaching degree? *Teaching and Teacher Education*, 55, 240–254. <https://doi.org/10.1016/j.tate.2016.01.018>
- Pfitzner-Eden, F., Thiel, F., & Horsley, J. (2014). An adapted measure of teacher self-efficacy for preservice teachers: Exploring its validity across two countries. *Zeitschrift für Pädagogische Psychologie*, 28(3), 83–92. <https://doi.org/10.1024/1010-0652/a000125>
- Portney, L. G., & Watkins, M. P. (2009). *Foundations of clinical research: Applications to practice* (2nd ed.). Upper Saddle River, NJ: Prentice-Hall.
- Poulou, M. (2007). Personal teaching efficacy and its sources: Student teachers' perceptions. *Educational Psychology*, 27(2), 191–218. <https://doi.org/10.1080/01443410601066693>

- Pratt, D. D. (1992). Chinese conceptions of learning and teaching: A Westerner's attempt at understanding. *International Journal of Lifelong Education*, 11(4), 301–319. <https://doi.org/10.1080/0260137920110404>
- Preacher, K. J., & MacCallum, R. C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics*, 2(1), 13–43. https://doi.org/10.1207/S15328031US0201_02
- Robinson, M. A. (2018). Using multi-item psychometric scales for research and practice in human resource management. *Human Resource Management*, 57(3), 739–750. <https://doi.org/10.1002/hrm.21852>
- Ruan, J. N., Nie, Y. Y., Hong, J., Monobe, G., Zheng, G. M., Kambara, H., & You, S. (2015). Cross-cultural validation of teachers' sense of efficacy scale in three Asian countries: Test of measurement invariance. *Journal of Psychoeducational Assessment*, 33(8), 769–779. <https://doi.org/10.1177/0734282915574021>
- Sang, G., Valcke, M., van Braak, J., & Tondeur, J. (2010). Student teachers' thinking processes and ICT integration: Predictors of prospective teaching behaviors with educational technology. *Computers & Education*, 54(1), 103–112. <https://doi.org/10.1016/j.compedu.2009.07.010>
- Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, 8(4), 350–353. <https://doi.org/10.1037//1040-3590.8.4.350>
- Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of Educational Research*, 99(6), 323–338. <https://doi.org/10.3200/JOER.99.6.323-338>
- Shi, Q. M. (2014). Relationship between teacher efficacy and self-reported instructional practices: An examination of five Asian countries/regions using TIMSS 2011 data. *Frontiers of Education in China*, 9(4), 577–602. <https://doi.org/10.1007/BF03397041>
- Shi, Q. M. (2016). *Interrelationships among teacher self-efficacy, collective teacher efficacy, and teachers' pedagogical content knowledge* (Unpublished doctoral dissertation). University of Nevada, Las Vegas, NV.
- Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, 74(1), 107–120. <https://doi.org/10.1007/s11336-008-9101-0>
- Skaalvik, E. M., & Skaalvik, S. (2007). Dimensions of teacher self-efficacy and relations with strain factors, perceived collective teacher efficacy, and teacher burnout. *Journal of Educational Psychology*, 99(3), 611–625. <https://doi.org/10.1037/0022-0663.99.3.611>
- Song, H., Zhu, X. D., & Liu, L. B. (2013). The honourable road and its impact on teacher practice: An analysis of China's national honour system in cultivating professional development. *Asia-Pacific Journal of Teacher Education*, 41(3), 253–270. <https://doi.org/10.1080/1359866X.2013.809055>
- Stipek, D. (2012). Context matters: Effects of student characteristics and perceived administrative and parental support on teachers' self-efficacy. *The Elementary School Journal*, 112(4), 590–606. <https://doi.org/10.1086/664489>
- Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and

- internal consistency. *Journal of Personality Assessment*, 80(1), 99–103. https://doi.org/10.1207/S15327752JPA8001_18
- Streiner, D. L., & Norman, G. R. (1995). *Health measurement scales: A practical guide to their development and use* (2nd ed.). New York, NY: Oxford University Press.
- Swan, B. G., Wolf, K. J., & Cano, J. (2011). Changes in teacher self-efficacy from the student teaching experience through the third year of teaching. *Journal of Agricultural Education*, 52(2), 128–139. <https://doi.org/10.5032/jae.2011.02128>
- Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Pearson Allyn & Bacon.
- Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, 48(6), 1273–1296. <https://doi.org/10.1007/s11165-016-9602-2>
- Teacher sense of self efficacy survey*. (n.d.). Retrieved from https://www.london.gov.uk/sites/default/files/teacher_efficacy_scale.pdf
- Tian, M. (2011). *Distributed leadership and teachers' self-efficacy: The case studies of three Chinese schools in Shanghai* (Unpublished master's thesis). University of Jyväskylä, Jyväskylä, Finland.
- Trevethan, R. (2009). Self-assessment of foot health: Requirements, issues, practicalities, and challenges. *Journal of the American Podiatric Medical Association*, 99(5), 460–471. <https://doi.org/10.7547/0990460>
- Tsai, C-C., & Chai, C. S. (2012). The “third”-order barrier for technology integration-instruction: Implications for teacher education. *Australasian Journal of Educational Technology*, 28(6), 1057–1060. <https://doi.org/10.14742/ajet.810>
- Tschannen-Moran, M., & Woolfolk Hoy, A. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education*, 17(7), 783–805. [https://doi.org/10.1016/S0742-051X\(01\)00036-1](https://doi.org/10.1016/S0742-051X(01)00036-1)
- Tschannen-Moran, M., & Woolfolk Hoy, A. (2007). The differential antecedents of self-efficacy beliefs of novice and experienced teachers. *Teaching and Teacher Education*, 23(6), 944–956. <https://doi.org/10.1016/j.tate.2006.05.003>
- Tschannen-Moran, M., Woolfolk Hoy, A., & Hoy, W. K. (1998). Teacher efficacy: Its meaning and measure. *Review of Educational Research*, 68(2), 202–248. <https://doi.org/10.3102/00346543068002202>
- Tsigilis, N., Grammatikopoulos, V., & Koustelios, A. (2007). Applicability of the Teachers' Sense of Efficacy Scale to educators teaching innovative programs. *The International Journal of Educational Management*, 21(7), 634–642. <https://doi.org/10.1108/09513540710822229>
- Tsui, K-T, & Kennedy, K. J. (2009). Evaluating the Chinese version of the Teacher Sense of Efficacy Scale (C-TSE): Translation adequacy and factor structure. *The Asia-Pacific Education Researcher*, 18(2), 245–260. <https://doi.org/10.3860/taper.v18i2.1326>
- Turner, K., Jones, E., Davies, M., & Ramsay, S. (2004). Student teacher perceptions of preparedness for teaching. *Educating: Weaving Research into Practice*, 3, 184–193.

- Tzeng, O. C. S. (1992). On reliability and number of principal components: Joinder with Cliff and Kaiser. *Perceptual and Motor Skills*, 75(3), 929–930. <https://doi.org/10.2466/pms.1992.75.3.929>
- Van den Broeck, J., Cunningham, S. A., Eeckels, R., & Herbst, K. (2005). Data cleaning: Detecting, diagnosing, and editing data abnormalities. *PLoS Medicine*, 2(10), e267. <https://doi.org/10.1371/journal.pmed.0020267>
- Vieluf, S., Kunter, M., & van de Vijver, F. J. R. (2013). Teacher self-efficacy in cross-national perspective. *Teaching and Teacher Education*, 35, 92–103. <https://doi.org/10.1016/j.tate.2013.05.006>
- von Suchodoletz, A., Jamil, F.M., Larsen, R. A. A., & Hamre, B. K. (2018). Personal and contextual factors associated with growth in preschool teachers' self-efficacy beliefs during a longitudinal professional development study. *Teaching and Teacher Education*, 75, 278–289. <https://doi.org/10.1016/j.tate.2018.07.009>
- Wang, H., Hall, N. C., & Rahimi, S. (2015). Self-efficacy and causal attributions in teachers: Effects on burnout, job satisfaction, illness, and quitting intentions. *Teaching and Teacher Education*, 47, 120–130. <https://doi.org/10.1016/j.tate.2014.12.005>
- Wang, M., Zan, F., Liu, J. Q., Liu, C. L., & Sharma, U. (2012). A survey study of Chinese in-service teachers' self-efficacy about inclusive education. *Journal of International Special Needs Education*, 15(2), 107–119. <https://doi.org/10.9782/2159-4341-15.2.107>
- Watkins, D. (2000). Learning and teaching: A cross-cultural perspective. *School Leadership & Management*, 20(2), 161–173. <https://doi.org/10.1080/13632430050011407>
- Weinstein, C. S. (1988). Preservice teachers' expectations about the first year of teaching. *Teaching and Teacher Education*, 4(1), 31–40. [https://doi.org/10.1016/0742-051X\(88\)90022-4](https://doi.org/10.1016/0742-051X(88)90022-4)
- Wheatley, K. F. (2005). The case for reconceptualizing teacher efficacy research. *Teaching and Teacher Education*, 21(7), 747–766. <https://doi.org/10.1016/j.tate.2005.05.009>
- Wild, D., Grove, A., Martin, M., Eremenco, S., McElroy, S., Verjee-Lorenz, A., & Erikson, P. (2005). Principles of good practice for the translation and cultural adaptation process for patient-reported outcomes (PRO) measures: Report of the ISPOR Task Force for translation and cultural adaptation. *Value in Health*, 8(2), 94–104. <https://doi.org/10.1111/j.1524-4733.2005.04054.x>
- Wolters, C. A., & Daugherty, S. G. (2007). Goal structures and teachers' sense of efficacy: Their relation and association to teaching experience and academic level. *Journal of Educational Psychology*, 99(1), 181–193. <https://doi.org/10.1037/0022-0663.99.1.181>
- Woolfolk Hoy, A., & Burke Spero, R. (2005). Changes in teacher efficacy during the early years of teaching: A comparison of four measures. *Teaching and Teacher Education*, 21(4), 343–356. <https://doi.org/10.1016/j.tate.2005.01.007>
- Wyatt, M. (2014). Towards a re-conceptualization of teachers' self-efficacy beliefs: Tackling enduring problems with the quantitative research and moving on. *International Journal of Research & Method in Education*, 37(2), 166–189. <https://doi.org/10.1080/1743727X.2012.742050>

- Yin, H.-B., Lee, J. C.-K., Jin, Y.-L., & Zhang, Z.-H. (2013). The effect of trust on teacher empowerment: The mediation of teacher efficacy. *Educational Studies*, 39(1), 13–28. <https://doi.org/10.1080/03055698.2012.666339>
- Yong, A. G., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology*, 9(2), 79–94. <https://doi.org/10.20982/tqmp.09.2.p079>
- Yousuf Zai, S. A. (2016). *Investigating the factor structure of the Teachers' Sense of Efficacy Scale with Pakistani in-service and pre-service teachers* (Unpublished doctoral dissertation). University of Arkansas, Fayetteville, AR.
- Yousuf Zai, S. A., & Munshi, P. (2016). Construct validity to examine the latent traits of teacher self-efficacy instrument. *Advances in Social Sciences Research Journal*, 3(4), 74–83. <https://doi.org/10.14738/assrj.34.1944>
- Yu, X. B., Wang, P. Y., Zhai, X. S., Dai, H., & Yang, Q. (2015). The effect of work stress on job burnout among teachers: The mediating role of self-efficacy. *Social Indicators Research*, 122(3), 701–708. <https://doi.org/10.1007/s11205-014-0716-5>
- Yuan, R., & Zhang, L. J. (2017). Exploring student teachers' motivation change in initial teacher education: A Chinese perspective. *Teaching and Teacher Education*, 61, 142–152. <https://doi.org/10.1016/j.tate.2016.10.010>
- Zheng, X., Yin, H. B., & Li, Z. L. (2019). Exploring the relationships among instructional leadership, professional learning communities and teacher self-efficacy in China. *Educational Management Administration & Leadership*, 47(6), 843–859. <https://doi.org/10.1177/1741143218764176>
- Zhou, L., & Nunes, M. B. (2013). Doing qualitative research in Chinese contexts: Lessons learned from conducting interviews in a Chinese healthcare environment. *Library Hi Tech*, 31(3), 419–434. <https://doi.org/10.1108/LHT-11-2012-0104>
- Zijlstra, W. P., van der Ark, L. A., & Sijtsma, K. (2011). Outliers in questionnaire data: Can they be detected and should they be removed? *Journal of Educational and Behavioral Statistics*, 36(2), 186–212. <https://www.jstor.org/stable/29789477>
- Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, 99(3), 432–442. <https://doi.org/10.1037/0033-2909.99.3.432>

Appendix A: Teacher Sense of Efficacy Scale Used with Preservice Teachers

Please indicate how effective you think you will be as a teacher in each of the following activities.
Select a number for each item.

	Minimally effective	Only moderately effective	Quite effective	Extremely effective	I have absolutely <i>no</i> idea	Not applicable			
1 Getting through to the most difficult students	1	2	3	4	5	6	7	8	9
2 Helping your students to think critically	1	2	3	4	5	6	7	8	9
3 Controlling disruptive behavior in the classroom	1	2	3	4	5	6	7	8	9
4 Motivating students who show low interest in schoolwork	1	2	3	4	5	6	7	8	9
5 Making your expectations clear about student behavior	1	2	3	4	5	6	7	8	9
6 Getting students to believe they can do well in school work	1	2	3	4	5	6	7	8	9
7 Responding to difficult questions from your students	1	2	3	4	5	6	7	8	9
8 Establishing routines to keep activities running smoothly	1	2	3	4	5	6	7	8	9
9 Helping your students value learning	1	2	3	4	5	6	7	8	9
10 Gauging student comprehension of what you have taught	1	2	3	4	5	6	7	8	9
11 Crafting good questions for your students	1	2	3	4	5	6	7	8	9
12 Fostering student creativity	1	2	3	4	5	6	7	8	9

(To be continued)

(Continued)

	Minimally effective	Only moderately effective	Quite effective	Extremely effective	I have absolutely <i>no</i> idea	Not applicable			
13 Getting students to follow classroom rules	1	2	3	4	5	6	7	8	9
14 Improving the understanding of a student who is failing	1	2	3	4	5	6	7	8	9
15 Calming a student who is disruptive or noisy	1	2	3	4	5	6	7	8	9
16 Establishing a classroom management system with each group of students	1	2	3	4	5	6	7	8	9
17 Adjusting your lessons to the proper level for individual students	1	2	3	4	5	6	7	8	9
18 Using a variety of assessment strategies	1	2	3	4	5	6	7	8	9
19 Keeping a few problem students from ruining an entire lesson	1	2	3	4	5	6	7	8	9
20 Providing an alternative explanation or example when students are confused	1	2	3	4	5	6	7	8	9
21 Responding to defiant students	1	2	3	4	5	6	7	8	9
22 Assisting families in helping their children do well in school	1	2	3	4	5	6	7	8	9
23 Implementing alternative strategies in your classroom	1	2	3	4	5	6	7	8	9
24 Providing appropriate challenges for very capable students	1	2	3	4	5	6	7	8	9

Appendix B: Teacher Sense of Efficacy Scale Used with Inservice Teachers

Please indicate how effective you think are as a teacher in each of the following activities.
Select a number for each item.

	Minimally effective	Only moderately effective	Quite effective	Extremely effective	Not applicable				
1	Getting through to the most difficult students	1	2	3	4	5	6	7	8
2	Helping your students to think critically	1	2	3	4	5	6	7	8
3	Controlling disruptive behavior in the classroom	1	2	3	4	5	6	7	8
4	Motivating students who show low interest in schoolwork	1	2	3	4	5	6	7	8
5	Making your expectations clear about student behavior	1	2	3	4	5	6	7	8
6	Getting students to believe they can do well in school work	1	2	3	4	5	6	7	8
7	Responding to difficult questions from your students	1	2	3	4	5	6	7	8
8	Establishing routines to keep activities running smoothly	1	2	3	4	5	6	7	8
9	Helping your students value learning	1	2	3	4	5	6	7	8
10	Gauging student comprehension of what you have taught	1	2	3	4	5	6	7	8
11	Crafting good questions for your students	1	2	3	4	5	6	7	8

(To be continued)

(Continued)

		Minimally effective	Only moderately effective	Quite effective	Extremely effective	Not applicable			
12	Fostering student creativity	1	2	3	4	5	6	7	8
13	Getting students to follow classroom rules	1	2	3	4	5	6	7	8
14	Improving the understanding of a student who is failing	1	2	3	4	5	6	7	8
15	Calming a student who is disruptive or noisy	1	2	3	4	5	6	7	8
16	Establishing a classroom management system with each group of students	1	2	3	4	5	6	7	8
17	Adjusting your lessons to the proper level for individual students	1	2	3	4	5	6	7	8
18	Using a variety of assessment strategies	1	2	3	4	5	6	7	8
19	Keeping a few problem students from ruining an entire lesson	1	2	3	4	5	6	7	8
20	Providing an alternative explanation or example when students are confused	1	2	3	4	5	6	7	8
21	Responding to defiant students	1	2	3	4	5	6	7	8
22	Assisting families in helping their children do well in school	1	2	3	4	5	6	7	8
23	Implementing alternative strategies in your classroom	1	2	3	4	5	6	7	8
24	Providing appropriate challenges for very capable students	1	2	3	4	5	6	7	8

Appendix C: Chinese Versions of Teacher Sense of Efficacy Scales

Preservice teachers' survey

当您成为一名教师时，您认为自己在何种程度上可以有效完成下列教学活动？请选择出一个最能代表您回复的数字。

	略微有效	比较有效	十分有效	极为有效	我完全不知道	不适用			
1 成功应对最难教的学生	1	2	3	4	5	6	7	8	9
2 引导学生进行批判性思考	1	2	3	4	5	6	7	8	9
3 控制扰乱课堂的行为	1	2	3	4	5	6	7	8	9
4 激发学习兴趣不浓的学生的学习积极性	1	2	3	4	5	6	7	8	9
5 阐明对学生课堂行为的期待	1	2	3	4	5	6	7	8	9
6 使学生树立做好功课的信心	1	2	3	4	5	6	7	8	9
7 回答学生提出的难题	1	2	3	4	5	6	7	8	9
8 制定规则以保证教学活动顺利开展	1	2	3	4	5	6	7	8	9
9 引导学生重视学习	1	2	3	4	5	6	7	8	9
10 衡量学生对所授内容的理解	1	2	3	4	5	6	7	8	9
11 为学生精心设计好问题	1	2	3	4	5	6	7	8	9
12 培养学生的创造力	1	2	3	4	5	6	7	8	9
13 引导学生遵守课堂纪律	1	2	3	4	5	6	7	8	9
14 加深“差生”对学习内容的理解	1	2	3	4	5	6	7	8	9

(To be continued)

(Continued)

	略微有效	比较有效	十分有效	极为有效	我完全不知道	不适用			
15 制止捣乱嘈杂的学生	1	2	3	4	5	6	7	8	9
16 建立一套适应不同学生的课堂管理体系	1	2	3	4	5	6	7	8	9
17 使课程适应每个学生的水平	1	2	3	4	5	6	7	8	9
18 使用多种评估策略	1	2	3	4	5	6	7	8	9
19 阻止个别问题学生破坏整个课堂	1	2	3	4	5	6	7	8	9
20 当学生感到困惑时提供另一种解释或案例	1	2	3	4	5	6	7	8	9
21 回应违规的学生	1	2	3	4	5	6	7	8	9
22 协助家长帮助子女搞好学业	1	2	3	4	5	6	7	8	9
23 采取多样化的课堂教学策略	1	2	3	4	5	6	7	8	9
24 为学有余力的学生提供适当的挑战	1	2	3	4	5	6	7	8	9

Instructions for inservice teachers' survey

请标出您认为自己能够多么有效地完成下列活动。请选择出一个最能代表您回复的数字。