Unfried *et al.*
*International Journal of STEM Education* (2022) 9:24
https://doi.org/10.1186/s40594-022-00339-x

International Journal of STEM Education

RESEARCH Open Access

# An alternative to STEBI-A: validation of the T-STEM science scale

Alana Unfried[1], Arif Rachmatullah[2], Alonzo Alexander[3] and Eric Wiebe[4*]

## Abstract

**Background:** The Science Teaching Efficacy Belief Instrument A (STEBI-A; Riggs & Enochs, 1990 in Science Education, 74(6), 625-637) has been the dominant measurement tool of in-service science teacher self-efficacy and outcome expectancy for nearly 30 years. However, concerns about certain aspects of the STEBI-A have arisen, including the wording, validity, reliability, and dimensionality. In the present study, we revised the STEBI-A by addressing many concerns research has identified, and developed a new instrument called the T-STEM Science Scale. The T-STEM Science Scale was reviewed by expert panels and piloted first before it was administered to 727 elementary and secondary science teachers. The combination of classical test theory (CTT) and item response theory (IRT) approaches were used to validate the instrument. Multidimensional Rasch analysis and confirmatory factor analysis were run.

**Results:** Based on the results, the negatively worded items were found to be problematic and thus removed from the instrument. We also found that the three-dimensional model fit our data the best, in line with our theoretical conceptualization. Based on the literature review and analysis, although the personal science teaching efficacy beliefs (PTSEB) construct remained intact, the original outcome expectancy construct was renamed science teacher responsibility for learning outcomes beliefs (STRLOB) and was divided into two dimensions, above- and below-average student interest or performance. The T-STEM Science Scale had satisfactory reliability values as well.

**Conclusions:** Through the development and validation of the T-STEM Science Scale, we have addressed some critical concerns emergent from prior research concerning the STEBI-A. Psychometrically, the refinement of the wording, item removal, and the separation into three constructs have resulted in better reliability values compared to STEBI-A. While two distinct theoretical foundations are now used to explain the constructs of the new T-STEM instrument, prior literature and our empirical results note the important interrelationship of these constructs. The preservation of these constructs preserves a bridge, though imperfect, to the large body of legacy research using the STEBI-A.

**Keywords:** Science education, Teaching, Survey, Teacher self-efficacy, Teacher responsibility, Validation

## Introduction

Previous research has shown that effective teachers are the most critical school-related factor to student learning and achievement (Darling-Hammond, 2000; McCaffrey et al., 2003; Muijs et al., 2014). This influence on students is predicated by several teacher-related factors, including external measures such as licensure or participation in a teacher preparedness programme, while others relate to the psychological make-up of the teachers, such as self-efficacy and outcome expectancy (Bandura, 1997; Zee & Koomen, 2016). Though many different psychological measures have been used and reported, teacher self-efficacy has reliably been shown to correlate to student achievement (Darling-Hammond, 2000; Pajares, 1992; Stronge, 2018). Some studies have even identified teacher self-efficacy as the most critical factor in understanding student learning outcomes (Tucker & Stronge, 2005), with research providing evidence of the direct influence of self-efficacy on the strategies that teachers use in the

*Correspondence: wiebe@ncsu.edu
[4] Department of STEM Education, North Carolina State University, Raleigh, NC, USA
Full list of author information is available at the end of the article

classroom (Al Sultan et al., 2018; Albion, 1999). Thus, the measurement of teaching self-efficacy, a teacher's confidence in their ability to teach in a way that affects positive change in their students, and outcome expectancy, or a teacher's belief in their responsibility to produce longer term positive outcomes for students (Lauermann & Karabenick, 2011), is essential to research and evaluation on teachers' impact on student learning and teacher professional growth.

In science education research, the Science Teaching Efficacy Belief Instrument A (STEBI-A; Riggs & Enochs, 1990) has been the dominant measurement tool of in-service science teacher self-efficacy and outcome expectancy for nearly 30 years. However, concerns about certain aspects of the STEBI-A have arisen; some are due to the impact of sociocultural policy shifts over time, while others are fundamental issues of conceptual/theoretical interpretation forming the basis of the survey itself. With regard to item wording, there has been an evolution in how student learning is discussed (i.e. *growth* versus *achievement*; Ho, 2008, Lachlan-Hache & Castro, 2015; Unfried et al., 2014). Empirically, concerns regarding the instrument have manifested in findings regarding the lack of association between the self-efficacy and outcome expectancy subscales (e.g., Lekhu, 2013) and whether the removal of items is needed to increase its reliability (Deehan et al., 2017; Henson et al., 2001). Perhaps most important has been a clear shift conceptually and theoretically in terms of how outcome expectancy is operationalized as a construct in teacher efficacy instruments (Dellinger et al., 2008; Klassen, et al., 2011; Pajares, 1992; Tschannen-Moran & Hoy, 2001; Tschannen-Moran et al., 1998; Zee & Kooman, 2016).

These potential concerns suggest an opportunity for an instrument revision to create a more effective evaluation tool. It also encourages a re-examination of the appropriate dimensionality of the instrument. It is with these concerns in mind that the T-STEM Science instrument was developed (T-STEM Science; Friday Institute for Educational Innovation, 2012). The goal of this instrument was to provide a bridge between the corpus of prior work using the STEBI while providing an instrument reflecting contemporary item wording and a reconceptualization of the outcome expectancy scale. We refer to the instrument as "T-STEM Science" because it is one instrument in the T-STEM family of instruments. There are four versions of the T-STEM instrument, one for each area of STEM (science, technology, engineering, and mathematics). This article focuses specifically on the T-STEM Science version of the instrument. The creation of this new tool for measurement leads to two guiding research questions: (1) Is the T-STEM Science a valid and reliable instrument for capturing science teacher self-efficacy and perceived responsibility for student learning outcomes? (2) How should the subscales of the T-STEM Science instrument be interpreted psychometrically and conceptually?

## Literature review

### Teacher efficacy

For the purposes of this paper's investigation, teacher self-efficacy will be specifically defined as a *science* teacher's belief in their ability to positively impact student's science learning outcomes. Within Bandura's (Bandura, 1997) framework of self-efficacy, personal self-efficacy has been shown to be a strong predictor of a teacher's future actions (Chesnut & Burley, 2015; Tschannen-Moran et al., 1998). When science teacher self-efficacy beliefs are defined as the perception of one's ability to teach in a way that affects positive change in their students, teachers declare it as central to their own feelings of effectiveness (Flores, 2015; Ghaith & Yaghi, 1997; Yoo, 2016). Numerous studies and meta-analyses have found that teacher beliefs are strongly correlated to predictors of teacher effectiveness and typically more important than other common measures, such as measurable content knowledge (e.g., Lui & Bonner, 2016; Pajares, 1992; Zee & Koomen, 2016).

When studying interventions designed to increase science teacher self-efficacy, it is critical to be able to measure self-efficacy in a scalable and robust fashion. Given that self-efficacy is conceptualized as a psychological state, it is not surprising that self-report measures have been among the most common tools used for this measurement. The STEBI, in turn, has been the most common survey instrument used for this purpose for K-12 science teachers (Deehan et al., 2017). As a historical precursor, the RAND Corporation worked on developing survey questions grounded in Bandura's self-efficacy dimensions, resulting in one of the early validated instruments designed to measure teacher self-efficacy, the teaching self-efficacy scale (TES; Gibson & Dembo, 1984). TES included two sub-constructs—general teaching efficacy and personal teaching efficacy. TES studies found discernible links between measures of teacher efficacy and student persistence and showed that even the classroom learning environment is influenced by a teacher's level of self-efficacy (Ghaith & Yaghi, 1997).

In support of Bandura's (Bandura, 1997; Bong, 2006) conceptualization of self-efficacy, research using TES found that teacher efficacy is often situational and content-specific (Pajares, 1992). Therefore, any examination of teacher self-efficacy must gather and interpret data relative to the targeted activities contextualizing the measurement of efficacy, with teaching subject area being

one of the most obvious examples. It is with this contextual specificity in mind that Riggs and Enoch developed the Science Teaching Efficacy Belief Instrument (STEBI; 1990).

### The STEBI instrument

The current STEBI for in-service teachers, or STEBI-A, is a 25-item Likert questionnaire. Respondents answer on a 1- to 5-point scale that ranges from 'strongly disagree' to 'strongly agree'. Remaining consistent with the TES, STEBI maintained two separate, construct-independent subscales: personal science teaching efficacy beliefs (PSTEB) and science teaching outcome expectancy beliefs (STOEB) designed to capture self-efficacy and outcome expectancy from elementary science teachers (Rubeck & Enochs, 1991). PSTEB measures a teachers' beliefs about their own ability to teach science content and develop science skills in students. Items on this scale are both positively and negatively coded; for example, one item states "I find it difficult to explain why experiments work to students." The STOEB subscale measures a teacher's beliefs, more generally, about a teachers' ability to achieve certain results. An example of a STOEB subscale item is "Student achievement is directly related to teacher effectiveness." Reliability of the STEBI-A was established during its creation; the subscales were found to have Cronbach alpha coefficients of 0.90 and 0.76, respectively. Early results also related personal science teaching efficacy to behavioural outcomes like spending more time teaching and dedicating specific time to developing better conceptual understanding (Riggs & Jesunathadas, 1993). PSTEB also correlates to teacher's enjoyment of science-related activities (Watters & Ginns, 1995). In its more than 30-year existence, STEBI-A has consistently risen in use among researchers, based on Google Scholar statistics. While only used (on average) in one published research study per year from 1990 to 1999, in the next decade that average rose to 4.5 studies, and since 2010 the average is more than 14 research studies per year.

Along with its increased use have come concerns about the reliability and validity of the STEBI-A. These concerns include: (1) the lack of association between the self-efficacy and outcome expectancy subscales (Lekhu, 2013); (2) the appropriateness of its use for pre-service teachers (Mulholland et al., 2004); (3) whether the removal of items from the subscales would increase its reliability (Deehan et al., 2017; Henson et al., 2001); (4) the lack of theoretical alignment with the shift towards internally oriented influences on outcome expectancy (Coladarci & Breton, 1997); and (5) the evolution of how student learning is discussed (e.g., *growth* versus *achievement*) (Betebenner, 2009; Ho, 2008; Lachlan-Hache & Castro, 2015; Unfried et al., 2014). A recent EFA from a sample of 1630 Canadian teachers reconfirmed a rather low reliability value for the STOEB factor (alpha = 0.72; Moslemi & Mousavi, 2019).

Perhaps the biggest open question with the STEBI-A and other related teacher efficacy instruments developed during this time period is exactly what the relationship is between the two primary constructs, typically labelled self-efficacy and outcome expectancy (Lekhu, 2013). Returning to the foundations of the STEBI, Gibson and Dembo's (1984) TES subscales (efficacy/GTE and self-efficacy/PTE) are now considered by many contemporary researchers to be completely separate constructs—psychometrically and theoretically (cf., Dellinger et al., 2008; Henson et al., 2001). While researchers contemporary to the development of the STEBI sought to merge Rotter's (1966) theory of locus of control and Bandura's conceptualization of outcome expectancy, this stance is no longer supported (Klassen et al., 2011). Rotter's theorizing around locus of control could be thought of as a type of expectancy, but a more general expectancy with no requisite direct connection between a teacher's actions and student outcomes (Dellinger et al., 2008; Henson et al., 2001). The STEBI-A's STOEB subscale does not connect student outcomes to individual teacher actions, and thus fails Bandura's test for outcome expectancy (Skinner, 1996; Tschannen-Moran & Hoy, 2001). Instead, the STOEB subscale can be interpreted through more current research on models of teacher responsibility (e.g., Lauermann & Karabenick, 2011).

Application of attribution theory (Wang & Hall, 2018; Weiner, 2010) is one approach to understanding how the target student population as characterized in the STOEB items can interact with how a teacher responds to negatively or positively worded items regarding outcomes. Wang and Hall state "Moreover, the present review underscores the double-edged nature of biased attributions in showing teachers to not only report self-protective attributions in failure situations but also self-enhancing attributions following success…." (p. 15). Thus, whether the teacher feels responsible for student outcomes can be influenced by the inferred characteristics of the target student population (e.g., low performing versus high performing students) (Diamond et al., 2004; Gershenson et al., 2016; Rubie-Davies, 2010). As an example, Diamond et al. (2004) demonstrated that teachers sense more responsibility when they think students possess more learning resources than for students who do not possess them; this, in turn, influences teachers' perception of effective teaching.

Similar support for this approach to conceptualizing the construct measured by the STOEB can be found in Lauermann and Karabenick's (2011) literature synthesis

of teacher responsibility. As with self-efficacy, they note researchers have concluded that teachers' sense of responsibility for both positive and negative student outcomes is linked to positive change in student learning and achievement (Guskey, 1984) as well as to a higher likelihood of implementing innovative educational practices after in-service training (Rose & Medway, 1981). Citing Duval and Silvia (2002), they note that teachers' attributions for positive and negative student outcomes are only weakly correlated; on one hand, people often attribute positive outcomes internally and negative outcomes externally to enhance their self-esteem when they succeed and to protect their sense of self-worth in the face of failure. Supporting this, Guskey (1982) found a teacher's efficacy beliefs and how they chose to attribute the cause of student outcomes interacted with whether those student outcomes were positive or negative. This finding, of course, now brings us full circle back to the PSTEB subscale measuring self-efficacy. Lauermann and Karabenick (2011) conclude by stating that further research is needed to clarify the relative importance of teachers' sense of responsibility for above-average versus below-average educational outcomes and, in addition, their relationship to teachers' efficacy beliefs. Other researchers also believe this dimensionality issue has been under-acknowledged in the development of expectancy beliefs-related evaluation tools (Rubie-Davies, 2010).

In summary, prior research has indicated that both teacher-perceived self-efficacy and responsibility are important factors in understanding the impact a teacher has on student outcomes. Thus, a continued interpretation of the PSTEB subscale based on teacher self-efficacy for science instruction and a reconceptualization of the STEBI-A's STOEB subscale centering on teachers' perceived responsibility regarding student science learning outcomes means that both of the STEBI-A's subscale constructs have value in science teacher education research and allows for a productive reconsideration of prior literature utilizing the STEBI instrument. Furthermore, a revisiting of STEBI-A item wording and item inclusion could further improve the instrument's psychometric performance. However, it leaves open the question of whether incremental improvement of items such as this reconceptualization of the STOEB subscale is borne out through a revalidation process. There also are the related, more specific questions of what the relationship is between the below-average and above-average student outcome items in the STOEB subscale and the relationship of the STOEB and the PSTEB subscales using a revised set of items.

#### STEBI piloting and reflection

The authors initially piloted the original STEBI-A instrument (Riggs & Enochs, 1990) for assessing teaching efficacy with just over 400 STEM teachers in North Carolina as part of a 2011 programme evaluation. Four parallel sets of items were created for the different subject areas of STEM (science, technology, engineering, and mathematics) so that teachers working primarily in each of these subject areas could respond to items anchored in their area of instruction. The STEBI items were also altered to allow teachers to respond "I don't know" to any question that they found confusing for piloting purposes.

The pilot administration data for the STEBI-A (and parallel versions) were analysed using subject matter expert (SME) feedback, written teacher feedback, analysis of teacher "I don't know" responses, and exploratory factor analysis (EFA). We found several issues with the STEBI-A after the pilot administration. As a first step, twelve SMEs rated each item on the STEBI-A (and spinoffs), and Lawshe's (1975) Content Validity Ratio was calculated to determine the proportion of experts identifying each item as essential. The majority of SMEs found each personal science teaching efficacy belief (PSTEB) item to be essential. However, for science teaching outcome expectancy beliefs (STOEB), two-thirds of items were identified as being non-essential by SMEs, raising questions about the alignment of these items with current teaching practices. Second, 90 teachers of those surveyed provided written feedback providing suggestions for how to improve the survey(s). Twenty-seven percent of these teachers identified the item wording as confusing (including negatively worded items that were difficult to understand) and six respondents used the phrase "too black and white" to describe their feelings about certain items. Additionally, three survey items in the PSTEB construct had 3% or more of teachers choosing "I don't know" as the response option. These issues indicated to us that there was a discrepancy in the intentions of the survey wording and teacher's interpretations of these items. Finally, the EFA on STEBI items resulted in six items that failed to load on their expected construct at a high enough level (0.4 or higher). Eleven out of 24 survey items exhibited problems across at least one spinoff version.

The findings from the pilot study aligned with many of the prior concerns raised in the literature concerning the reliability and validity of the STEBI-A. It was therefore decided that the current version of the STEBI-A could not be administered with STEM teachers as currently constructed and that a new version was needed that accurately reflected the current climate of teaching and addressed the item construction issues raised by the SMEs and teachers. For these reasons, the researchers created the T-STEM family of instruments

as an informed evolution and adaptation of the original STEBI-A.

### T-STEM science scale

The T-STEM family of instruments, developed by a team of researchers at the Friday Institute for Educational Innovation (2012), was designed to measure teacher efficacy and beliefs for teaching STEM and their use of STEM instructional practices. There are four versions of the T-STEM instrument, one for each area of STEM (science, technology, engineering, and mathematics). As previously mentioned, this article focuses specifically on the T-STEM Science Scale version of the instrument. Since this instrument was based on the STEBI-A, the initial hypothesized assumption is that it consists of two constructs based on the original PSTEB and STOEB item sets.

## Methods

### Revisions to the STEBI-A

Based on the previous discussion of empirical issues with the psychometric properties of STEBI-A, several changes were made to PSTEB and STOEB items for the T-STEM Science instrument.

First, in the original STEBI-A, items from the PSTEB and STOEB constructs were interleaved, seeming to cause confusion among teachers. The PSTEB items ask teachers to reflect on their own personal teaching efficacy, whereas STOEB items ask teachers to reflect on their feelings about teaching in general. In our pilot administration, teachers found it confusing to switch back and forth between these two statement types and thought that they should be reflecting on their personal teaching when reading STOEB items. We therefore altered the survey so that each construct's items are grouped together and given unique instructions; teachers are asked to reflect on their feelings about their own teaching for the PSTEB items, and to reflect on their feelings about teaching in general when answering STOEB items.

Second, most negatively worded items were reworded into positive items to avoid misinterpretation of responses. Aside from issues with respondents reading a negatively worded prompt correctly, some research shows that negatively worded items can lead to improper factor loadings (Krosnick & Presser, 2010).

Third, achievement-focused language was changed to growth-focused language to reflect modern best practices in teaching (Betebenner, 2009; Ho, 2008; Lachlan-Hache & Castro, 2015; Unfried et al., 2014). For example, whereas the original STOEB construct included items focusing on student grades and achievement, revised items instead focus on student learning. It is recognized that teachers may interpret student learning in both the formative and summative sense. Direct student involvement in the goal setting process via formative assessment is a modern educational development that has a positive influence on student outcomes and aligns well with a growth language orientation (Jimerson & Reames, 2015). In addition, minor wording changes were made to better reflect best practices in item wording (cf., Bong, 2006).

Lastly, five items were removed from the original STEBI-A due to confusing wording, problematic factor loadings, or topics that were too specific. Table 1 displays the 20 PSTEB and STOEB construct items from the T-STEM Science Scale, as well as their original wording on the STEBI-A. The STOEB construct items are further organized into two groups based on whether the wording references: (1) above-average student interest or outcomes, or (2) is neutral or below-average student interest or outcomes. Here, a neutral attribution (e.g., STOEB_4, STOEB_6) would be applicable to all students.

### Sample and data collection

The T-STEM Science instrument was administered to K-12 teachers across the state of North Carolina in United States between 2012 and 2015. All data collection was administered under approved human-subjects-research protocols associated with one of the authors of this paper. The administration collected data from 727 teachers. Although some programmes implemented both pre- and post-surveys, data were only analysed from teachers completing the survey for the first time. Moreover, only data from teachers who responded to all the items in the T-STEM Science Scale were analysed in this study. In the data cleaning process, eight teachers were identified not having a complete response and thus were removed from the final data set, resulting in a total of 718 analysable teachers' responses.

Demographically, the data were composed of 77% female, 20% male teachers, and 3% of the teachers did not provide any gender information. Regarding ethnicity, 87% of the teachers who participated in the study were identified as White/Caucasian, 5% Black/African American, 2% Hispanic/Latino and Asian, and 4% identified as Other. These demographics are similar, but not equivalent for the entire state teacher population from this time period (79% female, 82% White, 14% Black; SBE, 2009). The years of experience ranged from 0 to 45 years with an average of 11.67 years (SD = 8.64). Moreover, a plurality of the participants taught students in the grades 6–8 (40%), while the remaining participants taught students in either grades 1–5 (34%) or 9–12 (26%).

**Table 1** T-STEM Science survey items and their original STEBI-A wording

| **(a) Personal science teaching efficacy beliefs (PSTEB) Items** | | | | |
|---|---|---|---|---|
| **T-STEM Science Scale** | | | | **Original STEBI-A** |
| **Dimension/item** | | | **Revised** | |
| Personal Science Teaching Efficacy Beliefs | | PSTEB_1 | I am continually **improving** my science teaching practice | I am continually **finding better ways** to teach science |
| | | PSTEB_2 | I know the steps necessary to teach **science** effectively | I know the steps necessary to teach **science concepts** effectively |
| | | PSTEB_3 | I **am confident that** I can explain to students why science experiments work | I find **it difficult** to explain to students why science experiments work |
| | | PSTEB_4 | I **am confident** that I can teach science **effectively** | I **generally** teach science **ineffectively** |
| | | PSTEB_5 | I wonder if I have the necessary skills to teach science | I wonder if I have the necessary skills to teach science |
| | | PSTEB_6 | I understand science concepts well enough to be effective in teaching **science** | I understand science concepts well enough to be effective in teaching **elementary science** |
| | | PSTEB_7 | Given a choice, I **would invite** a colleague to evaluate my science teaching | Given a choice, I **would not invite** the principal to evaluate my science teaching |
| | | PSTEB_8 | I **am confident** that I can answer students' science questions | I **am typically** able to answer students' science questions |
| | | PSTEB_9 | When a student has difficulty understanding a science concept, I **am confident** that I know how to help the student understand it better | When a student has difficulty understanding a science concept, I am **usually at a loss** as to how to help the student understand it better |
| | | PSTEB_10 | When teaching science, I **am confident enough** to welcome student questions | When teaching science, I **usually** welcome student questions |
| | | PSTEB_11 | I **know what to do to increase student interest** in science | I **don't know what to do to turn students on** to science |
| **(b) Science teaching outcome expectancy beliefs (STOEB) items** | | | | |
| **T-STEM science scale** | | | | **Original STEBI-A** |
| **Dimension/item** | | | **Revised** | |
| Science Teaching Outcome Expectancy Beliefs | Above-average student interest or performance | STOEB_1 | When a student does better than usual in science, it is often because the teacher exerted a little extra effort | When a student does better than usual in science, it is often because the teacher exerted a little extra effort |
| | | STOEB_3 | When a **student's learning in science is greater than expected**, it is **most** often due to their teacher having found a more effective teaching approach | When the **science grades of students improve**, it is often due to their teacher having found a more effective teaching approach |
| | | STOEB_7 | When a low-achieving child progresses **more than expected in science**, it is usually due to extra attention given by the teacher | When a low-achieving child **progresses in science**, it is usually due to extra attention given by the teacher |
| | | STOEB_8 | If parents comment that their child is showing more interest in science at school, it is probably due to the performance of the child's teacher | If parents comment that their child is showing more interest in science at school, it is probably due to the performance of the child's teacher |
| | Neutral or below-average student interest or performance | STOEB_2 | The inadequacy of a student's science background can be overcome by good teaching | The inadequacy of a student's science background can be overcome by good teaching |
| | | STOEB_4 | The teacher is generally responsible for **students' learning** in science | The teacher is generally responsible for the **achievement of students** in science |
| | | STOEB_5 | If **students' learning in science is less than expected**, it is most likely due to ineffective science teaching | If **students are underachieving in science**, it is most likely due to ineffective science teaching |
| | | STOEB_6 | Students' **learning in science** is directly related to their teacher's effectiveness in science teaching | Students' **achievement in science** is directly related to their teacher's effectiveness in science teaching |
| | | STOEB_9 | **Minimal student learning** in science can generally be attributed to their teachers | The **low science achievement** of some students cannot generally be blamed on their teachers |

Wording changes are shown in bold

### Validation procedure

The validation procedure of the T-STEM Science Scale was based on Messick's construct validity Messick (1995) and Standards for Educational and Psychological Testing proposed by American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (AERA et al., 2014). Based on AERA et al. (2014), five sources of validity evidence should be addressed by test developers to validate an instrument, and those sources are evidence based on test content, response processes, internal structure, relations to other variables, evidence for validity and consequences of testing. However, according to Messick (1995), a validation study is an iterative and ongoing process, and thus test developers may start by focusing on gathering one or two specific sources of validity evidence before addressing other sources of evidence. Accordingly, in this study we validated the T-STEM Science Scale by addressing two core sources of validity evidence suggested by AERA et al. (2014), which are evidence based on test content and internal structure.

AERA et al. (2014) define evidence based on test content as "an analysis of the relationship between the content of a test and the construct it is intended to measure" (p. 14). Test content consists of themes and wording of the items, and can be addressed through performing expert judgment. Many studies have used and identified the content and themes of the STEBI-A, resulting in some measure of test content validity for the instrument. However, many of these studies are now dated, and as noted, STEM education goals have changed. We therefore consulted the literature regarding the contemporary issues in teachers' teaching efficacy particularly related to the shortcomings of the STEBI-A. In parallel, we also asked STEM education subject matter experts in our pilot study to provide feedback on the revised instrument by also providing them an explanation of the purpose of the instrument, so that they could properly evaluate the content with the intended purpose. This effort refers to what AERA et al. (2014) call *alignment*. Our changes to the STEBI-A based on the teacher and subject-matter-expert feedback provide validity evidence based on test content for the revised T-STEM Science PSTEB and STOEB constructs.

According to AERA et al. (2014), internal structure validity is based on "the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based." This definition aligns with Messick's structural aspects of construct validity Messick (1995). Therefore, validity related to the number of factors/dimensions, instrument structure, item difficulty level, and item quality were aspects of interest for this study. Moreover, the combination of classical test theory and item response theory-Rasch approaches was used to address the evidence based on internal structure. Due to conflicting norms in these different approaches, we use the terms "factor", "construct" and "dimension" interchangeably in our descriptions.

Although the STEBI-A appears to demonstrate two constructs (PSTEB and STOEB), there are mixed findings regarding the STOEB construct and whether it might comprise two sub-constructs based on whether the item has high or low-achieving students as its target (Duval & Silvia, 2002); (Guskey, 1982; Lauermann & Karabenick, 2011; Wang & Hall, 2018). Therefore, our approach is to analyse the data using confirmatory analyses, assessing several different possible models. We focus on confirmatory methods due to the existing theoretical framework guiding both the STEBI-A and the T-STEM Science Scale, as addressed throughout this paper. The full dataset was utilized for item response theory (IRT)-Rasch, confirmatory factor analysis (CFA), and reliability analyses.

The outcomes from the multidimensional Rasch analysis were used to also evaluate the structural aspect of the T-STEM Science Scale. Multidimensional Rasch analysis allows not only to identify the best model, but also to identify misfitting items within each dimension. Adams and Wu (2010) suggested looking at the lowest Chi-square, final deviance (FD) and Akaike Information Criterion (AIC) to identify the best model. Using this IRT approach, three competing models were tested:

1. one-dimension/factor (baseline model) with all PSTEB and STOEB items on the same dimension,
2. two-dimensions/factors, with PSTEB and STOEB as the two dimensions (see Table 1),
3. three-dimensions/factors with PSTEB as one dimension and STOEB construct items broken into two dimensions for above-average and below-average student interest/outcomes.

In addition to a dimensionality test, Rasch analysis also provides mean-square (MNSQ) values to assess the quality of the item, particularly regarding whether or not the items based on difficulty levels can differentiate the higher and lower achievers (Boone et al., 2014). The items which had MNSQ outside the range of 0.60 to 1.40 were considered misfitting items and removed from the instrument (Wright & Linacre, 1994).

Ordinal confirmatory factor analysis with robust diagonally weighted least squares was also implemented to

compare three competing models (Desjardins & Bulut, 2018; Yang-Wallentin et al., 2010), using the lavaan (Rosseel, 2012) package. Compared to the Rasch analysis, CFA allows for the consideration of a higher-order factor structure. Again, based on our literature review, we explored the STOEB construct as a single factor, as two separate factors and as two factors nested under a higher-order STOEB construct. In addition, based on our IRT findings, the one-dimension/factor model was dropped from consideration. Thus, the three models under consideration with CFA were:

1. two-factor CFA model, parallel to IRT approach,
2. three-factor CFA model, parallel to IRT approach,
3. higher-order CFA model with the PSTEB construct as a single factor, a STOEB construct, and two STOEB sub-constructs for above-average and below-average student interest/outcomes.

We use the cut-off values suggested by Hu and Bentler (1999) and Schreiber et al. (2006) to assess the models. They suggested that a good and acceptable model has CFI > 0.95, TLI > 0.95 and RMSEA < 0.08.

Finally, Cronbach's alpha, along with reliability values (person/plausible value and separation reliability) computed through IRT-Rasch were used to assess the internal consistency of each subscale after any items identified as problematic were removed. The cutoff suggested by DeVellis (2017), which is < 0.70, was used to evaluate the reliability values. Factor analysis methods were conducted in RStudio (RStudio Team, 2018); Item response theory was conducted in ConQuest version 4.14.2 (Adams et al., 2015).

## Results

### Multidimensional Rasch analysis

Multidimensional Rasch analyses were run to evaluate the model and the items of the T-STEM Science Scale. Table 2 shows the results of the multidimensional Rasch analysis. It can be seen from Table 2 that the three-dimensional model had the lowest $X^2$, *FD* and *AIC* compared to the two competing models, indicating the three-dimensional model was identified as the best fitting model. In addition to indicating the three-dimension model as the best model, the IRT analysis identified two misfitting items. These two items were PSTEB_5 (infit and outfit MNSQ 2.54 and 2.80, respectively) and PSTEB_7 (infit and outfit MNSQ 1.50 and 1.47, respectively). We then removed these two items from the model and re-ran the three-dimensional model. The results showed that a three-dimensional model without the two items improved and continued to be better than any of the other models run. No further misfitting items were identified. Based on this analysis, we used this three-dimensional model for further analyses.

Table 3 presents the item measure and quality residing on the three-dimension model based on multidimensional Rasch analysis after the two items were removed. Note that since this instrument is not measuring performance, *item measure* should be interpreted at the degree of agreement with the item. It can be seen from Table 3 that the values of both infit and outfit MNSQ are in the range of acceptable values of 0.60 – 1.40 suggested by Wright and Linacre (1994). This indicates that all the items were well-behaved in terms of their ability to distinguish teachers with differing levels of response to the three constructs. A Wright map (Fig. 1) produced from the multidimensional Rasch analysis shows an acceptable spread of item response and participant scores. High scores on the Wright map indicate more agreement.

### Confirmatory factor analysis

After we removed two problematic items suggested by the multidimensional Rasch analysis (PSTEB_5 and 7), CFA was performed to further examine the structure of the factors residing in the T-STEM instrument. Table 4 presents the comparison of the fit indices for the three models investigated. First, we compared the two-factor model to the three-factor model indicated by the multidimensional Rasch analysis and our conceptualization of the instrument. The results indicated that the three-factor model was better than the two-factor model, with a difference in $\chi^2$ resulting in a p-value close to zero. Next, based on our literature review, we compared the three-factor model to a higher-order model where the two factors of STOEB are part of a higher-order latent STOEB factor. We again found that the three-factor model was better than the higher-order model (p-value close to zero). These tests, along with fit indices, indicate that the T-STEM Science Scale was best fitted to the three-factor model. Figure 2 visualizes the structure of the T-STEM Science Scale three-factor model.

### Reliability values

We used Cronbach's alpha values and plausible-value (PV or person) reliability from the multidimensional Rasch analysis to evaluate this aspect of validity. The T-STEM Science Scale with a three-dimensional model had Cronbach's alpha values of 0.931, 0.778, and 0.767 for PSTEB, STOEB above-average student interest and outcome, and STOEB below-average student interest and outcome, respectively. Based on PV reliability, the T-STEM Science Scale had values 0.881, 0.775, and 0.773 for the three constructs, respectively. Given all the values are above the cut-off of 0.70 (DeVellis, 2017), this indicated a stable

instrument. In addition, Rasch analysis also produces another reliability value called "separation reliability" that evaluates how reproducible the spread of the response levels is. The separation reliability for the instrument was 0.990 indicating a good spread of item responses.

## Discussion

The development and validation of the T-STEM Science Scale in this study was motivated by several concerns around the well-known instrument used to measure in-service science teacher self-efficacy, STEBI-A (Riggs & Enochs, 1990). These concerns include: the evolution of how student learning is codified in items (i.e. growth versus achievement; Unfried et al., 2014), whether the removal of items, particularly negative worded items, from the subscales would increase reliability (Deehan et al., 2017; Henson et al., 2001), and most importantly the lack of resolution concerning the instrument's dimensionality (Lekhu, 2013) and conceptualization of the STOEB construct (Lauermann & Karabenick, 2011). We addressed these concerns by (1) rewording the items to address a more growth orientation of students' learning, (2) showing how rewording and removing poorly worded items improves the reliability and quality of the instrument, and most importantly (3) re-examining the dimensionality and constructs underlying the revised instrument through a new, more contemporary theoretical lens.

STEBI-A was grounded in student achievement-oriented teacher self-efficacy beliefs, leaving it out of step with more growth-oriented conceptualization of student learning. The use of achievement-oriented language may make the teachers' focus of efficacy more on students' final products (e.g., test-scores), rather than their confidence in affecting students' learning process (Schweder et al., 2019). Part of our revision of the STEBI-A included rewording achievement-focused language to growth-focused language to reflect modern best practices in teaching (Betebenner, 2009; Ho, 2008; Lachlan-Hache & Castro, 2015; Unfried et al., 2014). Guided by our pilot study, we both removed and reworded negatively worded items, as suggested by several studies that

**Table 2** Comparison between one, two and three-dimensional models of the T-STEM science scale

| Model | $X^2$ | df | FD | AIC | # Misfitting |
|---|---|---|---|---|---|
| One-dimension | 2143.72 | 19 | 14,427.99 | 14,475.99 | 1 |
| Two-dimension | 963.27 | 18 | 13,197.22 | 13,249.23 | 2 |
| Three-dimension | 938.03 | 17 | 13,129.99 | 13,187.99 | 2 |
| Three-dimension without PSTEB_5 and 7 | 703.66 | 15 | 10,889.00 | 10,943.00 | 0 |

**Table 3** Rasch item fit indices and Cronbach's alpha if item deleted for the three-dimensional model

| Dimension | | Item | Cronbach's Alpha if item deleted | Measure | Infit MNSQ | Outfit MNSQ |
|---|---|---|---|---|---|---|
| Personal science teaching efficacy and beliefs | | PSTEB_1 | .933 | − 1.592 | 1.37 | 1.15 |
| | | PSTEB_2 | .923 | 0.207 | 0.85 | 0.74 |
| | | PSTEB_3 | .922 | 0.410 | 0.80 | 0.65 |
| | | PSTEB_4 | .921 | 0.191 | 0.78 | 0.60 |
| | | PSTEB_6 | .930 | 0.221 | 1.32 | 1.15 |
| | | PSTEB_8 | .927 | 0.556 | 0.94 | 0.88 |
| | | PSTEB_9 | .924 | 0.543 | 0.74 | 0.72 |
| | | PSTEB_10 | .927 | − 0.823 | 0.84 | 0.62 |
| | | PSTEB_11 | .936 | 0.287 | 1.16 | 1.37 |
| Science teaching outcome expectancy beliefs | Above— average student interest or performance | STOEB_1 | .789 | 0.314 | 1.17 | 1.25 |
| | | STOEB_3 | .721 | − 0.124 | 0.94 | 0.95 |
| | | STOEB_7 | .731 | 0.144 | 0.96 | 1.00 |
| | | STOEB_8 | .734 | − 0.335 | 0.96 | 0.98 |
| | Neutral or below-average student interest or performance | STOEB_2 | .736 | − 1.483 | 1.35 | 1.35 |
| | | STOEB_4 | .737 | − 0.168 | 1.29 | 1.27 |
| | | STOEB_5 | .663 | 0.987 | 1.09 | 1.04 |
| | | STOEB_6 | .701 | − 0.234 | 1.12 | 1.09 |
| | | STOEB_9 | .707 | 0.897 | 1.26 | 1.19 |

PSTEB_5 and PSTEB_7 were removed based on the prior analysis

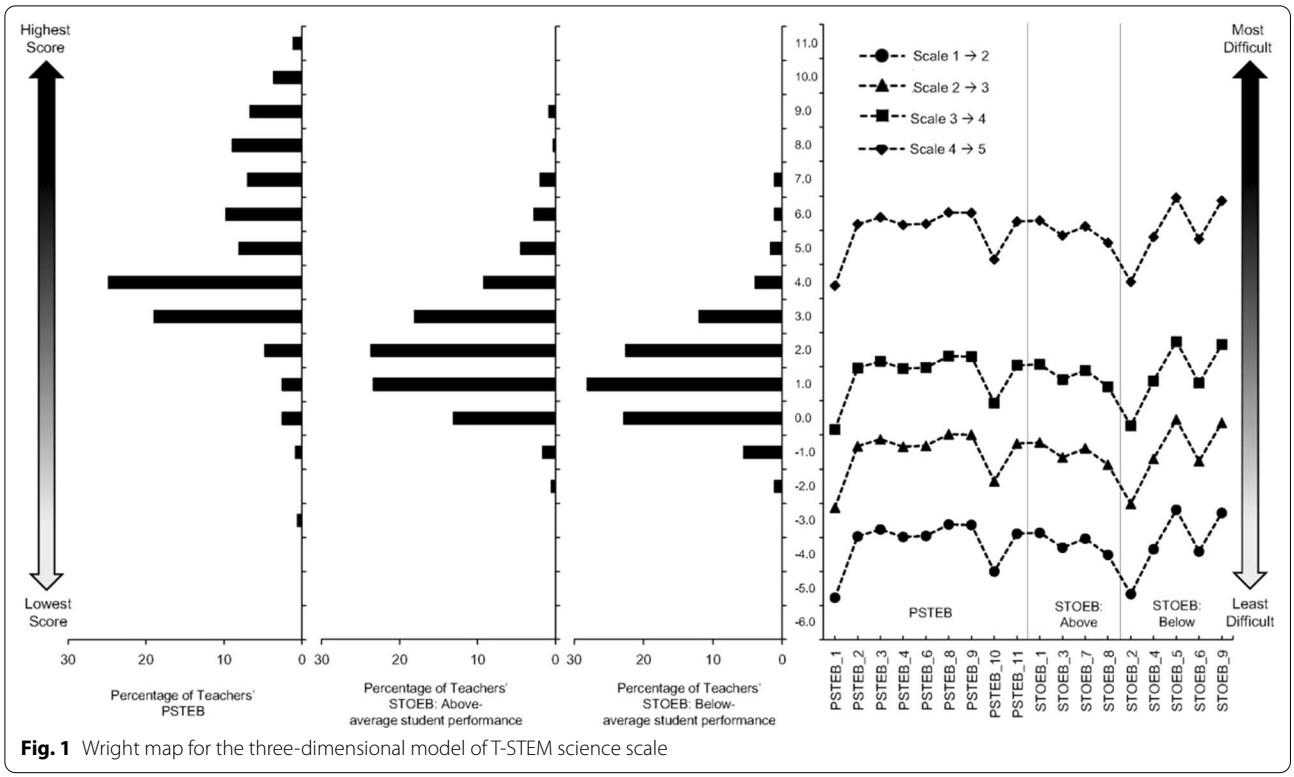


**Fig. 1** Wright map for the three-dimensional model of T-STEM science scale

**Table 4** Confirmatory factor analysis model fit statistics after removing PSTEBS_5 and PSTEBS_7

| Model | CFI | TLI | RMSEA | RMSEA Upper 90% CI | SRMR | DF | $X^2$ Diff from 3-factor | *p*-value |
|---|---|---|---|---|---|---|---|---|
| Two-factor | 0.973 | 0.970 | 0.091 | 0.097 | 0.073 | 134 | 76.74 | < .0001 |
| Higher-order factor | 0.979 | 0.976 | 0.081 | 0.086 | 0.064 | 133 | 19.66 | < .001 |
| Three-factor | 0.981 | 0.978 | 0.078 | 0.084 | 0.058 | 132 | – | – |

show how negatively worded items lead to increased test-fatigue and distort concentration, potentially leading to improper factor loadings (Groves et al., 2009; Krosnick & Presser, 2010). Collectively, we believe these changes both shortened the instrument and helped improve the reliability statistics (i.e. Cronbach's alpha values and plausible-values) of the T-STEM Science subscales over the values of the original STEBI-A PSTEB subscale and on par or better for the STOEB, as reported in the literature (Albion & Spence, 2013; McKinnon et al., 2014; Moslemi & Mousavi, 2019; Riggs & Enoch, 1990).

We expected that after the removal of items following the pilot phase, we would not need to remove any additional items. However, this was not the case. We removed PSTEB_5 due to a high MNSQ value and lowest factor loading. According to Boone et al. (2014), a high MNSQ value means that the item could not differentiate teachers with high and low self-efficacy and thus can distort the interpretation of scores generated from such an item. We then investigated the item and concluded that the word "wonder" in the item does not properly operationalize the concept of self-efficacy, with regard to its relationship to the concept of confidence (Bandura, 1997; Bong, 2006), thus it would make sense to remove the item. We also removed another item, PSTEB_7, having a high MNSQ value. We concluded that PSTEB_7 was contextually problematic because self-efficacy is an internal psychological trait of an individual (Bandura, 1997), and by introducing an external factor, such as "invite a colleague", it made self-efficacy less internally guided (Wang & Hall, 2018) and only indirectly related to one's confidence in science instruction.

Appropriately, the investigation of the STOEB construct provided some of the most interesting findings of the study. While Bandura (1997) continues to be the primary theoretical guide for the PSTEB, Weiner's (2000)

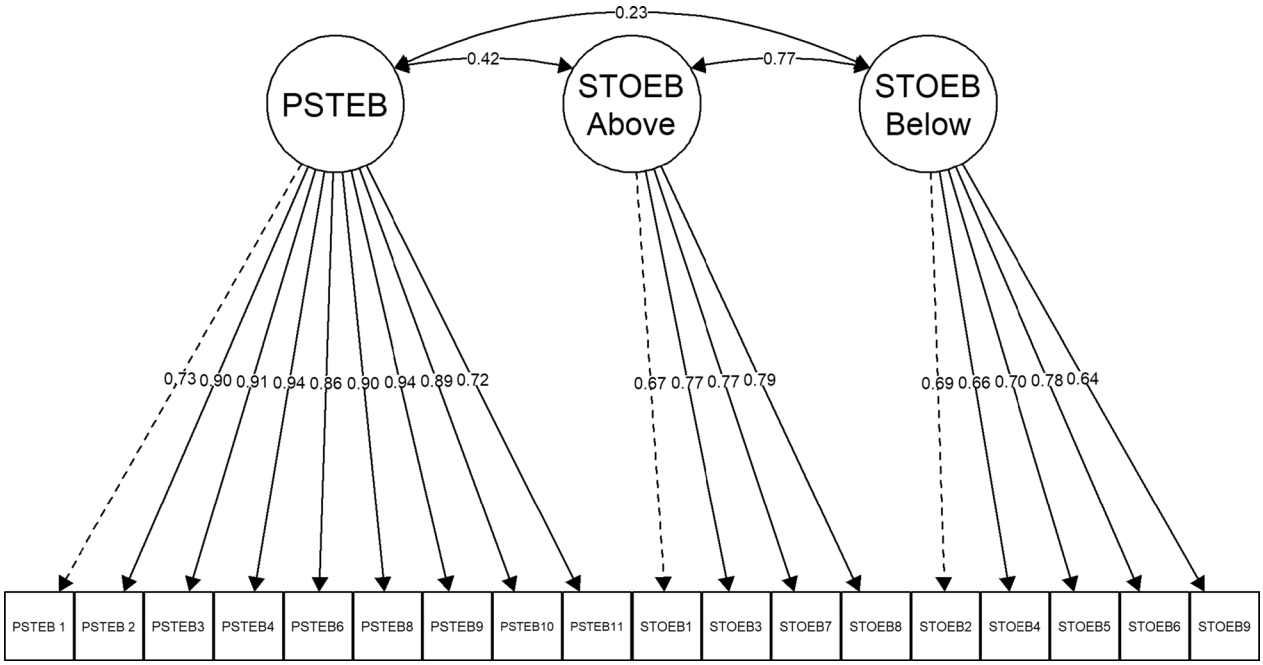


**Fig. 2** CFA path diagram with standardized loadings for 3-factor model

attribution theory and related work on teacher perceptions of responsibility provides a more appropriate theoretical basis for the STOEB. This conclusion is drawn through both our analysis of the literature and findings based on our psychometric analysis. First, current theoretical conceptualization of outcome expectancy clearly indicates that the items in the STOEB subscale(s) are not in alignment with this construct (Skinner, 1996; Tschannen-Moran & Hoy, 2001). Empirically, the CFA and IRT analyses point to two constructs psychometrically distinct but related to the PSTEB. Researchers applying attribution theory (Lauermann & Karabenick, 2011; Wang & Hall, 2018; Weiner, 2010) to teacher's sense of responsibility to the success or failure of their students' learning outcomes have made the case for defining this construct, that we shall now call science teacher responsibility for learning outcomes beliefs (STRLOB). In addition, this same literature base supports conceptualizing this construct as having two separate dimensions—responsibility for above-average performing students and those students performing below-average. This sub-division is parsimonious with Weiner's (2010) concept of attribution bias and confirmed in other cited empirical studies (Diamond et al., 2004; Gershenson et al., 2016; Rubie-Davies, 2010; Wang & Hall, 2018). Thus, we conceptualize the T-STEM Science Scale as having three constructs: science teacher self-efficacy (PSTEB; 9 items) and STRLOB (9 items), which is divided into two separate constructs of teachers' responsibility for above- and below-average interested or performing students (4 and 5 items, respectively), for a total of 18 items. The results from the multidimensional Rasch analysis and CFA showed that the three-factor model is best suited to the instrument. The higher-order model, which groups the above-average and below-average constructs under a broader STRLOB construct, performed only marginally worse than the three-factor model. While the higher order model seems the more elegant interpretation theoretically, empirical evidence has us siding with a flat, 3-dimensional model. Future studies exploring this decision are encouraged. The combination of these analyses demonstrates that, broadly speaking, there is evidence that the T-STEM Science Scale does differentiate between PSTEB and STRLOB constructs, and that the STRLOB items can be broken down into two separate dimensions for items focused on above- and below-average student outcome/interest. These findings support our conceptualization that science teachers are indeed having different expectations for different students, based on attributes such as perceived academic outcomes.

## Conclusion

With these results, we believe that we have addressed some critical concerns emergent from prior research concerning the STEBI-A. Psychometrically, the refinement

of the wording, item removal, and the separation into three constructs have resulted in better reliability values compared to STEBI-A. The resulting T-STEM Science Scale is a more compact and stable instrument than the STEBI-A. While two distinct theoretical foundations are now used to explain the constructs of the new T-STEM instrument, prior literature and our empirical results note the important interrelationship of these constructs (cf., Guskey, 1982). In addition, the preservation of these constructs preserves a bridge, though imperfect, to the large body of legacy research using the STEBI-A.

Messick (1995) argued that instrument validation is an iterative and ongoing process, and we did not address all the validity evidence proposed by AERA et al. (2014) in this study. We plan further validity studies of the T-STEM Science Scale, such as instrument and item bias through differential item functioning, and criterion validity. We acknowledge that the teacher participants in this study were from one U.S. state, which may influence the results of the constructs' separation. According to Mason and Morris (2010), culture plays an integral role in an individual's perceptions of attributes. Hence, different results may emerge from different states or countries, given the impact of culture there. This may also be considered as our direction for future data collection work to confirm whether a similar psychometric structure would appear from a more diverse, international sample.

**Acknowledgements**
We would like to thank Malinda Faber for her input on the original version of the T-STEM instrument.

**Authors' contributions**
AU and AR analyzed and interpreted the validation data. Both were major contributors to writing in the manuscript. AA contributed to the background literature review and theoretical framing of the manuscript. EW was PI of the NSF project supporting the project work, supervised the development of the manuscript and was a major contributor to writing the manuscript. All authors read and approved the final manuscript.

**Funding**
This work is supported by the National Science Foundation through the grant DUE-1038154. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Portions of the work were also supported by the Golden LEAF Foundation.

**Availability of data and materials**
The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

## Declarations

**Competing interests**
The authors declare that they have no competing interests.

**Author details**
[1]Department of Mathematics & Statistics, California State University Monterey Bay, Seaside, CA, USA. [2]Center for Education Research & Innovation (CERI), SRI International, Menlo Park, CA, USA. [3]The Science House, North Carolina State University, Raleigh, NC, USA. [4]Department of STEM Education, North Carolina State University, Raleigh, NC, USA.

Received: 30 September 2021 Accepted: 22 February 2022
Published online: 08 March 2022

**References**

Adams, R., & Wu, M. (2010). *Notes and tutorial ConQuest: Multidimensional model*. Retrieved from https://www.acer.org/conquest/notes-tutorials.
Adams, R. J., Wu, M. L., & Wilson, M. R. (2015). *ACER ConQuest: Generalised item response modelling software Version 4 [Computer software]*. Camberwell: Australian Council for Educational Research.
Al Sultan, A., Henson, H., Jr., & Fadde, P. J. (2018). Pre-service elementary teachers' scientific literacy and self-efficacy in teaching science. *IAFOR Journal of Education, 6*(1), 25–41.
Albion, P. (1999). Self-efficacy beliefs as an indicator of teachers' preparedness for teaching with technology. In: *Proceedings of the 10th international conference of the society for information technology & teacher education (SITE 1999)* (pp. 1602–1608). Association for the Advancement of Computing in Education (AACE)
Albion, P. R., & Spence, K. G. (2013). Primary Connections in a provincial Queensland school system: Relationships to science teaching self-efficacy and practices. *International Journal of Environmental and Science Education, 8*(3), 501–520.
American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
Bandura, A. (1997). *Self-efficacy: The exercise of control*. WH Freeman.
Betebenner, D. (2009). Norm-and criterion-referenced student growth. *Educational Measurement: Issues and Practice, 28*(4), 42–51. https://doi.org/10.1111/j.1745-3992.2009.00161.x
Bong, M. (2006). Asking the right question: How confident are you that you could successfully perform these tasks. In T. Urdan & F. Pajares (Eds.), *Self-efficacy beliefs of adolescents* (pp. 287–306). Information Age Publishing.
Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Springer.
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
Coladarci, T., & Breton, W. A. (1997). Teacher efficacy, supervision, and the special education resource-room teacher. *The Journal of Educational Research, 90*(4), 230–239. https://doi.org/10.1080/00220671.1997.10544577
Darling-Hammond, L. (2000). Teacher quality and student achievement. *Education Policy Analysis Archives, 8*(1), 1–44. https://doi.org/10.14507/epaa.v8n1.2000
Deehan, J., Danaia, L., & McKinnon, D. H. (2017). A longitudinal investigation of the science teaching efficacy beliefs and science experiences of a cohort of preservice elementary teachers. *International Journal of Science Education, 39*(18), 2548–2573. https://doi.org/10.1080/09500693.2017.1393706
Dellinger, A. B., Bobbett, J. J., Olivier, D. F., & Ellett, C. D. (2008). Measuring teachers' self-efficacy beliefs: Development and use of the TEBS-Self. *Teaching and Teacher Education, 24*(3), 751–766.
Desjardins, C. D., & Bulut, O. (2018). *Handbook of educational measurement and psychometrics using R*. CRC Press.
DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). Sage.
Diamond, J. B., Randolph, A., & Spillane, J. P. (2004). Teachers' expectations and sense of responsibility for student learning: The importance of race, class, and organizational habitus. *Anthropology & Education Quarterly, 35*(1), 75–98. https://doi.org/10.1525/aeq.2004.35.1.75
Duval, T. S., & Silvia, P. J. (2002). Self-awareness, probability of improvement, and the self-serving bias. *Journal of Personality and Social Psychology, 82*(1), 49–61. https://doi.org/10.1037/0022-3514.82.1.49.
Flores, I. M. (2015). Developing Preservice Teachers' Self-Efficacy through Field-Based Science Teaching Practice with Elementary Students. *Research in Higher Education Journal, 27*, 1–19.
Friday Institute for Educational Innovation (2012). *Teacher Efficacy and Beliefs toward STEM (T-STEM) Survey*. Raleigh, NC: Author. Retrieved from https://

www.fi.ncsu.edu/pages/about-the-teacher-efficacy-and-attitudes-toward-stem-surveys-t-stem/

Gershenson, S., Holt, S. B., & Papageorge, N. W. (2016). Who believes in me? The effect of student–teacher demographic match on teacher expectations. *Economics of Education Review, 52*, 209–224. https://doi.org/10.1016/j.econedurev.2016.03.002

Ghaith, G., & Yaghi, H. (1997). Relationships among experience, teacher efficacy, and attitudes toward the implementation of instructional innovation. *Teaching and Teacher Education, 13*(4), 451–458. https://doi.org/10.1016/S0742-051X(96)00045-5

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

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (2nd ed.). John Wiley & Sons Inc.

Guskey, T. R. (1982). Differences in teachers' perceptions of personal control of positive versus negative student learning outcomes. *Contemporary Educational Psychology, 7*(1), 70–80. https://doi.org/10.1016/0361-476X(82)90009-1

Henson, R. K., Kogan, L. R., & Vacha-Haase, T. (2001). A reliability generalization study of the teacher efficacy scale and related instruments. *Educational and Psychological Measurement, 61*(3), 404–420.

Ho, A. D. (2008). The problem with "proficiency": Limitations of statistics and policy under No Child Left Behind. *Educational Researcher, 37*(6), 351–360. https://doi.org/10.3102/0013189X08323842

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Jimerson, J. B., & Reames, E. (2015). Student-involved data use: Establishing the evidence base. *Journal of Educational Change, 16*(3), 281–304.

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*, 21–43.

Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of Survey Research* (2nd ed.). Emerald Group Publishing Limited.

Lachlan-Haché, M., & Castro, M. (2015). *Proficiency or growth: An exploration of two approaches for writing student learning targets*. American Institutes for Research. Retrieved from http://www.air.org/sites/default/files/ExplorationofTwo-Approaches-Student-Learning-Targets-April-2015.pdf.

Lauermann, F., & Karabenick, S. A. (2011). Taking teacher responsibility into account(ability): Explicating its multiple components and theoretical status [Article]. *Educational Psychologist, 46*(2), 122–140. https://doi.org/10.1080/00461520.2011.558818

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*, 563–575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Lekhu, M. A. (2013). Relationship between self-efficacy beliefs of science teachers and their confidence in content knowledge. *Journal of Psychology in Africa, 23*(1), 109–112. https://doi.org/10.1080/14330237.2013.10820602

Lui, A. M., & Bonner, S. M. (2016). Preservice and inservice teachers' knowledge, beliefs, and instructional planning in primary school mathematics. *Teaching and Teacher Education, 56*, 1–13. https://doi.org/10.1016/j.tate.2016.01.015

Mason, M. F., & Morris, M. W. (2010). Culture, attribution and automaticity: A social cognitive neuroscience view. *Social Cognitive and Affective Neuroscience, 5*(2–3), 292–306. https://doi.org/10.1093/scan/nsq034

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). Evaluating value-added models for teacher accountability. *PsycEXTRA Dataset*. https://doi.org/10.1037/e658712010-001

McKinnon, M., Moussa-Inaty, J., & Barza, L. (2014). Science teaching self-efficacy of culturally foreign teachers: A baseline study in Abu Dhabi. *International Journal of Educational Research, 66*, 78–89.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749.

Moslemi, N., & Mousavi, A. (2019). A psychometric re-examination of the science teaching efficacy and beliefs instrument (STEBI) in a Canadian context. *Education Sciences*. https://doi.org/10.3390/educsci9010017

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

Mulholland, J., Dorman, J. P., & Odgers, B. M. (2004). Assessment of science teaching efficacy of preservice teachers in an Australian university. *Journal of Science Teacher Education, 15*(4), 313–331. https://doi.org/10.1023/B:JSTE.0000048334.44537.86

Pajares, M. F. (1992). Teachers' beliefs and educational research: Cleaning up a messy construct. *Review of Educational Research, 62*(3), 307–332. https://doi.org/10.3102/00346543062003307

Riggs, I., & Jesunathadas, J. (1993, April). Preparing elementary teachers for effective science teaching in diverse settings. Paper presented at the National Association for Research in Science Teaching, Atlanta, GA.

Riggs, I. M., & Enochs, L. G. (1990). Toward the development of an elementary teacher's science teaching efficacy belief instrument. *Science Education, 74*(6), 625–637. https://doi.org/10.1002/sce.3730740605

Rose, J. S., & Medway, F. J. (1981). Measurement of teachers' beliefs in their control over student outcome. *The Journal of Educational Research, 74*(3), 185–190.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied, 80*(1), 1–28. https://doi.org/10.1037/h0092976

RStudio Team. (2018). *RStudio: Integrated development for R [Software]*. RStudio Inc.

Rubeck, M., & Enochs, L. (1991). *A path analytic model of variables that influence science and chemistry teaching self-efficacy and outcome expectancy in middle school science teachers*. Paper presented at the Annual Meeting of the National Association of Research in Science Teaching, Lake Geneva, WI.

Rubie-Davies, C. M. (2010). Teacher expectations and perceptions of student attributes: Is there a relationship? *British Journal of Educational Psychology, 80*(1), 121–135. https://doi.org/10.1348/000709909X466334

SBE, State Board of Education & NC Department of Public Instruction (2009). *North Carolina Public Schools Statistical Profile: 2008–2009*. Raleigh, NC: Authors. Retrieved from https://digital.ncdcr.gov/digital/collection/p249901coll22/id/92200/rec/2

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & 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

Schweder, S., Raufelder, D., Kulakow, S., & Wulff, T. (2019). How the learning context affects adolescents' goal orientation, effort, and learning strategies. *The Journal of Educational Research, 112*(5), 604–614. https://doi.org/10.1080/00220671.2019.1645085

Skinner, E. A. (1996). A guide to constructs of control. *Journal of Personality and Social Psychology*, *71*(3), 549–570. https://doi.org/10.1037/0022-3514.71.3.549.

Stronge, J. H. (2018). *Qualities of effective teachers* (3rd ed.). Association for Supervision and Curriculum Development.

Tschannen-Moran, M., & Hoy, A. W. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education, 17*(7), 783–805.

Tschannen-Moran, M., Hoy, A. W., & Hoy, W. K. (1998). Teacher efficacy: Its meaning and measure. *Review of Educational Research, 68*(2), 202–248.

Tucker, P. D., & Stronge, J. H. (2005). *Linking teacher evaluation and student learning*. Association for Supervision and Curriculum Development.

Unfried, A., Faber, M., Townsend, L., and Corn, J. (2014). *Validated Student, Teacher, and Principal Survey Instruments for STEM Education Programs*. Presented at the Annual Meeting of the American Evaluation Association, Denver, CO.

Wang, H., & Hall, N. C. (2018). A systematic review of teachers' causal attributions: prevalence, correlates, and consequences. *Frontiers in Psychology, 9*, 1–22. https://doi.org/10.3389/fpsyg.2018.02305

Watters, J. J., & Ginns, I. S. (1995). Origins of, and changes in preservice teachers' science teaching self efficacy. In *Annual Meeting of National Association for Research in Science Teaching*.

Weiner, B. (2000). Intrapersonal and interpersonal theories of motivation from an attributional perspective. *Educational Psychological Review, 12*, 1–14. https://doi.org/10.1023/A:1009017532121

Weiner, B. (2010). The development of an attribution-based theory of motivation: A history of ideas. *Educational Psychologist, 45*(1), 28–36. https://doi.org/10.1080/00461520903433596

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.

Yang-Wallentin, F., Jöreskog, K. G., & Luo, H. (2010). Confirmatory factor analysis of ordinal variables with misspecified models. *Structural Equation Modeling*, *17*(3), 392-423. https://doi.org/10.1080/10705511.2010.489003.

Yoo, J. H. (2016). The effect of professional development on teacher efficacy and teachers' self-analysis of their efficacy change. *Journal of Teacher Education for Sustainability, 18*(1), 84–94. https://doi.org/10.1515/jtes-2016-0007

Zee, M., & Koomen, H. M. (2016). Teacher self-efficacy and its effects on classroom processes, student academic adjustment, and teacher well-being: a synthesis of 40 years of research. *Review of Educational Research, 86*(4), 981–1015. https://doi.org/10.3102/0034654315626801

## Publisher's Note