J Sci Teacher Educ (2013) 24:1133–1156
DOI 10.1007/s10972-013-9351-2

Construct Validation of the Self-Efficacy Teaching and Knowledge Instrument for Science Teachers-Revised (SETAKIST-R): Lessons Learned

Linda A. Pruski · Sharon L. Blanco · Rosemary A. Riggs · Kandi K. Grimes · Chase W. Fordtran · Gina M. Barbola · John E. Cornell · Michael J. Lichtenstein

Published online: 30 May 2013

Abstract Described herein is the academic lineage and independent validation of the Self-Efficacy Teaching and Knowledge Instrument for Science Teachers-Revised (SETAKIST-R). Data from 334 K-12 science teachers were analyzed using Partial Credit Rasch models. Principal components analysis on the person-item residuals suggest two latent dimensions: Knowledge and Teaching Self-Efficacies. Item-fit statistics were used to select items for each subscale. Person and item separation (reliability) indices were quite low, and we noted disordered response patterns on the person-item maps that revealed problems with item content and/or scaling for both subscales. These issues include the presence of: verbal negatives, ambiguous modifiers, counter-intuitive scaling, and an "undecided/uncertain" option. The SETAKIST-R, in its current form, cannot be recommended as a measure of science teacher self-efficacy.

L. A. Pruski (✉) · S. L. Blanco · R. A. Riggs · K. K. Grimes · C. W. Fordtran · G. M. Barbola · M. J. Lichtenstein
Division of Geriatrics, Gerontology, and Palliative Care, Department of Medicine, Teacher Enrichment Initiatives, University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Drive, MSC 7780, San Antonio, TX 78229-3900, USA
e-mail: Pruski@uthscsa.edu

S. L. Blanco
e-mail: BlancoSL@uthscsa.edu

R. A. Riggs
e-mail: RiggsR@uthscsa.edu

K. K. Grimes
e-mail: GrimesK@uthscsa.edu

C. W. Fordtran
e-mail: Fordtran@uthscsa.edu

G. M. Barbola
e-mail: Barbola@uthscsa.edu

Keywords Self-efficacy teaching and knowledge instrument for science teachers-revised · SETAKIST-R · Teacher self-efficacy beliefs · Teacher efficacy · Teacher professional development · Rasch analysis

Introduction

Effective teachers are a primary fulcrum for student success. As with other professionals, teachers must acquire the confidence and belief (self-efficacy) that they can undertake a set of work-related tasks that encompass their field of endeavor—education (Bandura 1977a). A strong sense of efficacy influences one's choices, effort, perseverance, and resilience (Bandura 1997), and is a well documented aspect of effective teachers (Henson et al. 2001). The tasks of creating engaging learning environments that promote cognitive growth "rests heavily on the talents and self-efficacy of teachers" (Bandura 1997, p. 240).

Along with various measures of teacher knowledge and skills (Guskey and Passaro 1994; Guskey and Yoon 2009; Scher and O'Reilly 2009), one component of teacher professional development (TPD) program evaluation is assessing teacher self-efficacy (Caprara et al. 2006; Gibson and Dembo 1984; Shidler 2009; Tschannen-Moran and Woolfolk Hoy 2001; Tschannen-Moran et al. 1998; Zielinski et al. 2000).

In particular, it is useful to assess the impact of TPD on teacher efficacy as it relates to K-12 *science* teacher ability (Riggs and Enochs 1990, 2000).

Our project sought to examine the usefulness of science teacher efficacy scales in evaluating our science TPD program for kindergarten-through-twelfth grade (K-12) educators. A review of the literature found two inservice science teacher efficacy scales. We explored the academic evolution of these two scales, the Science Teaching Efficacy Belief Instrument (STEBI, Form A; hereinafter referred to as STEBI-A) (Riggs and Enochs 1989, 1990) and the Self-Efficacy Teaching and Knowledge Instrument for Science Teachers (SETAKIST) (Roberts and Henson

M. J. Lichtenstein
e-mail: Lichtenstei@uthscsa.edu

L. A. Pruski · S. L. Blanco · R. A. Riggs · K. K. Grimes · C. W. Fordtran · G. M. Barbola · M. J. Lichtenstein
Barshop Institute for Aging and Longevity Studies, University of Texas Health Science Center at San Antonio, 15355 Lamda Drive, MSC 7755, San Antonio, TX 78245, USA

J. E. Cornell
Department of Epidemiology and Biostatistics, University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Drive, MSC 7933, San Antonio, TX 78229-3900, USA
e-mail: Cornell@uthscsa.edu

M. J. Lichtenstein
Institute for Integration of Medicine and Science, University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Drive, MSC 7759, San Antonio, TX 78229-3900, USA

2000). A portion of the SETAKIST was extracted from the STEBI-A. Both of these scales were used with elementary teachers. The STEBI-A was applied at the middle school level (Desouza et al. 2004).

Because our program encompasses K-12 science teachers, we sought a scale that would serve as an efficacy measure for all precollegiate grade levels. We selected the shorter SETAKIST instrument (16 items in contrast to the STEBI's 25) and made minor modifications of three scale items to better align with our K-12 science teacher experiences. We conducted a validation study of this revision—the SETAKIST-R—and present those results in this paper.

General Self-Efficacy

Key constructs in Bandura's (1978, 1997) social learning theory—self efficacy beliefs and reciprocal determinism—illustrate that humans are motivated by three interrelated forces: external environmental influences, internal personal factors (cognitive, affective, and physiological processes), and our current and past behaviors. We are products of the interplay of these forces. Bandura (1982, 1983, 1986, 1993, 1996, 1997) refined the idea that beliefs in our own abilities have powerful effects on our behavior, motivation, and success or failure. Belief, the "information that a person accepts to be true" (Koballa and Crawley 1985, p. 223), is viewed as different from attitude, which is "a general positive or negative feeling toward something" (Riggs and Enochs 1990, p. 625). Attitudes can be formed on the basis of beliefs; both attitudes and beliefs influence behavior. Riggs and Enochs (1990) explained the association among belief, attitude, and behavior in the following example: "An elementary teacher judges his/her ability to be lacking in science teaching (belief) and consequently develops a dislike for science teaching (attitude). The result is a teacher who avoids teaching science if at all possible (behavior)" (Riggs and Enochs 1990, p. 625–626).

Bandura (1977b) solidified self-efficacy as a theoretical component of behavior change, describing self-efficacy as a probable determinant of our actions when presented with complex situations and how effectively we pursue steps relative to these situations (Bandura 1977b). Stating that behavior is based on beliefs, he further suggests that efficacy beliefs can be measured on different dimensions including level, generality, and strength (Bandura 1977a, p. 42–44). He proposed that self-efficacy assessments include both the "affirmation of capability and the strength of that belief" (Bandura 1977a, p. 382). Two belief components, labeled *self-efficacy* and *outcome expectancy*, are predictors of behavior change relative to dimensions such as occupational tasks, decision making, and learning (Bandura 1977a) and can be situation specific (Bandura 1982, 1997; Guskey 2000; Riggs and Enochs 1990).

Teacher Self-Efficacy

Bandura acknowledges that self-belief may not ensure success, but implies its importance when he says, "self-disbelief assuredly spawns failure" (1977a, p. 77). Classroom teachers are challenged by high-stakes testing, prescribed curricula,

over-crowded classrooms, and decreasing budgets. In this tumult, 'self-disbelief' (low efficacy) can undermine teachers and classroom instruction. In contrast, 'self-belief' (high efficacy) can elevate teachers' confidence and resourcefulness in response to modern classroom challenges. Teacher efficacy is thought to be "one of the key motivation beliefs influencing teachers' professional behaviors and student learning" (Klassen et al. 2011, p. 21).

Teacher efficacy became known as a judgment of a teachers' capabilities to bring about desired outcomes (e.g., student engagement, knowledge, and skill acquisition) (Tschannen-Moran and Woolfolk Hoy 2001). The definition extended from the individual teacher to a "collective" capacity to influence learning (Klassen et al. 2011; see also, Ross and Gray 2006).

Published and unpublished efficacy scales have been used to explore teacher self-efficacy in relationship to various aspects of the teaching enterprise. Self-efficacy responses have been used in studying teacher motivation (Ashton and Webb 1982; Ross 1992), job satisfaction (Caprara et al. 2006), and teachers' views of their roles within an educational system (Desouza et al. 2004). Additional self-efficacy studies have explored teacher contributions to school improvement (Dembo and Gibson 1985), and how much perceived control teachers have over that educational environment (Tschannen-Moran et al. 1998).

Self-efficacy scales have been used in studies reporting how teachers construct student learning activities and the time spent in teaching content (Gibson and Dembo 1984), how they persist in and adapt instruction to meet student learning needs (Guskey 1988; Stein and Wang 1988), and how curriculum is implemented and impacts student achievement (Armor et al. 1976; Ashton and Webb 1982; Caprara et al. 2006; Shidler 2009). Responses to self-efficacy scales were used also to determine the amount of time necessary to conduct cost-effective TPD (Roberts et al. 2000, 2001).

Researchers advocate measuring teacher self-efficacy over time, noting that significant changes in teacher attitudes and beliefs occurred after there was evidence of improved student learning (Guskey 2000; Moreno and Tharp 2006; Shidler 2009). However, a ceiling effect or maximum at which changes in teacher efficacy may be seen is noted (Shidler 2009).

Science Teacher Self-Efficacy Scales: From Rotter to SETAKIST

Rotter's Internal and External Locus of Control and Rand Efficacy

Figure 1 outlines the developmental progression of science teacher self-efficacy scales, including their underlying theoretical constructs. Incorporating Rotter's (1966) locus of control theory, teacher efficacy studies formally began with the Rand report (Armor et al. 1976) (Fig. 1) which noted that efficacious teachers contributed to the success of a reading program used in Los Angeles schools. The portion of the study devoted to efficacy simply provided two reflective prompts; these assumed that student learning and motivation reinforced teachers' actions. The prompts were affixed to a five point Likert scale, "Strongly agree = 1" to "Strongly Disagree = 5." The statements used were:

Rand Efficacy Item 1: "When it comes right down to it, a teacher really can't do much – most of a student's motivation and performance depends on his or her home environment," and
Rand Efficacy Item 2: "If I try really hard, I can get through to even the most difficult or unmotivated students" (Armor et al. 1976, p. 33).

Item responses were combined into a single efficacy score to gauge the extent that a teacher believed he/she had the capacity to effect student learning (Armor et al. 1976). The Rand stated that the more efficacious teachers felt, the higher the student scores rose in reading achievement.

Bandura's Teacher Self-Efficacy Scale (TSES)

Bandura's Teacher Self-Efficacy Scale (http://people.ehe.ohio-state.edu/ahoy/files/2009/02/bandura-instr.pdf) is a 30 item efficacy measure written in interrogative rather than declarative statements, affixed to a nine point scale anchored at five points ("nothing," "very little," "some influence," "quite a bit," and "a great deal") (Fig. 1). Bandura posited seven subscales that probed the degree of influence and efficacy teachers have over certain classroom issues, such as acquiring materials/equipment, instructional skills, and disciplinary concerns.

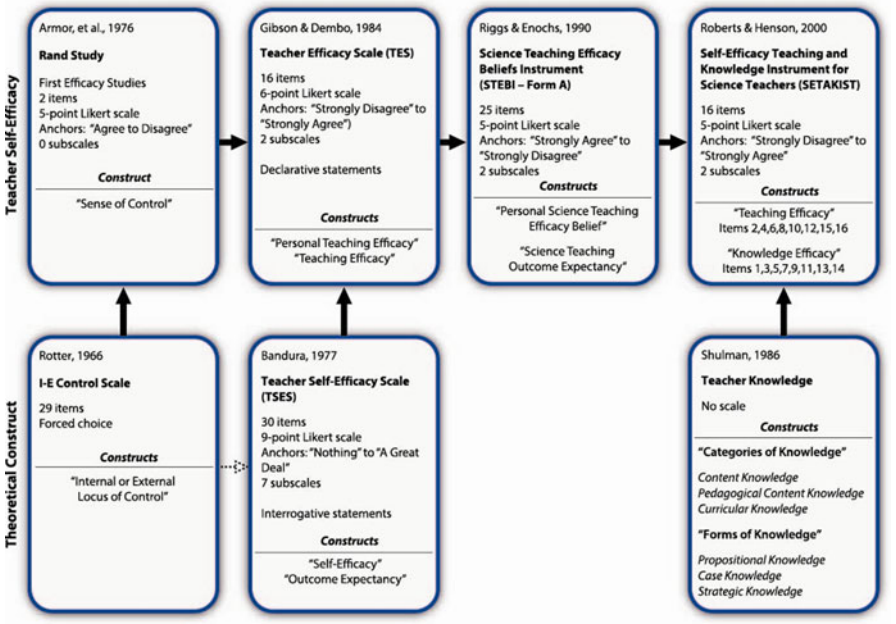

Fig. 1 Academic evolution of science teacher efficacy scales

Gibson's Teacher Efficacy Scale (TES)

The Rand items and the themes in Bandura's scale influenced the creation of the Teacher Efficacy Scale (TES) by Sherri Gibson (Gibson and Dembo 1984)—the first published scale of its kind (Fig. 1). Gibson and Dembo (1984) applied Bandura's self-efficacy theory to teaching outcomes. They state that teachers

> "…who believe student learning can be influenced by effective teaching (outcome expectancy beliefs) and who also have confidence in their own teaching abilities (self-efficacy beliefs) should persist longer, provide a greater academic focus in the classroom, and exhibit different types of feedback than teachers who have lower expectations concerning their ability to influence student learning" (p. 570).

In their study (Gibson and Dembo 1984), Gibson's 30 item TES was reduced to 16 items that factored into two constructs: Personal Teaching Efficacy and Teaching Efficacy. The Teaching Efficacy dimension was later termed General Teaching Efficacy by other users (Henson et al. 2001). The TES provided the seed for growing *subject-specific* scales and influenced the development of the Science Teaching Efficacy Belief Instrument (STEBI-A) (Riggs and Enochs 1989, 1990) (Fig. 1).

Science Teaching Efficacy Belief Instrument, Form A (STEBI-A)

Using a thoroughly detailed process, the 25-item STEBI-A, was developed specifically for inservice elementary teachers of science (Riggs and Enochs 1989, 1990); and concurrently modified for use with preservice elementary science teachers as STEBI, Form B (Enochs and Riggs 1990). STEBI-A retained some of Gibson's items with minor modifications for science content and/or to reflect teachers as a whole, rather than the individual respondent. For example, the modifications of Gibson's TES item 1 to STEBI-A item ' illustrates both points: TES Item 1, "When a student does better than usual, many times it is because I exerted a little extra effort." (Gibson and Dembo 1984, p. 581), as compared with STEBI-A Item 1, "When a student does better than usual in *science*, it is often because *the teacher* exerted a little extra effort." (Riggs and Enochs 1990, p. 634). As well as utilizing Gibson's TES items, Riggs and Enochs (1989, 1990) created additional items for the STEBI-A. All items were linked to a five point Likert scale: "Strongly Agree" (5) to "Strongly Disagree" (1). Negatively worded items were scored in the opposite direction.

When analyzed, the STEBI-A, maintained two belief components suggested by Bandura. These were labeled *Personal Science Teaching Efficacy Beliefs* (PSTE, 13 items) and *Science Teaching Outcome Expectancy* (STOE, 12 items). Riggs and Enochs (1990) suggested ways the scale could inform teacher inservice design and published their final scale. With their scale, they found that teachers with high PSTE scores spent more time developing science concepts for students (Riggs and Jesunathadas 1993); those with low PSTE spent less time (Riggs 1995). The STEBI-A was used to determine the optimal length of TPD programs (Roberts et al. 2000, 2001) and stimulated the development of the SETAKIST (Roberts and Henson 2000) (Fig. 1).

Self-Efficacy Teaching and Knowledge Instrument for Science Teachers (SETAKIST)

In a paper presented at a regional educational research meeting, Roberts and Henson (2000) raised concerns about the STOE subscale of the STEBI-A (Riggs and Enochs 1990) and the Teaching Efficacy subscale of the TES (Gibson and Dembo 1984). Of the STEBI-A, they questioned the reliability of "a two-factor solution that explained no more than 60 % of the overall variance" (Roberts et al. 2000, p. 7) and cautioned others about using the STOE subscale. They chose to create a new scale, the SETAKIST, by removing the STOE construct of the STEBI-A, and replacing it with items intended to reflect Shulman's promotion of pedagogical content knowledge (Roberts and Henson 2000; Shulman 1986, 1987, 1998; Hutchings and Shulman 1999) (Fig. 1), meaning the way in which subject matter knowledge is transformed into the art of teaching that subject (Shulman 1986). A combination of 10 STEBI-A items taken from the PSTE subscale (six copied verbatim, four re-worded), one item copied verbatim from the STOE, and five new items, similarly related to the PSTE, made up the 16 item SETAKIST proposed by Roberts and Henson (2000).

The SETAKIST consisted of two constructs, eight items each: *Teaching Efficacy* and *Knowledge Efficacy* (Roberts and Henson 2000). Developers claimed that the Teaching Efficacy construct in the SETAKIST was similar to the STEBI-A's PSTE subscale. The subscale has two items taken verbatim from the PSTE, three slightly modified from the PSTE, one item from the STEBI-A's STOE, and two new items. Together SETAKIST items 2, 4, 6, 8, 10, 12, 15, and 16 make up Teaching Efficacy.

In place of the STOE construct of the STEBI-A, Roberts and Henson (2000) offer the Knowledge Efficacy subscale. The SETAKIST's Knowledge Efficacy subscale explores the concept that "content knowledge is part and parcel with … teaching ability" (Roberts and Henson 2000, p. 12). The Knowledge Efficacy subscale incorporated four verbatim items from the STEBI-A's PSTE subscale, one modified PSTE item, and three new items.

Roberts and Henson (2000) used the same five point scale as the STEBI-A. However, they do not mention reversely scoring negatively stated items as did Riggs and Enochs (1989, 1990). The factor analysis of their pilot study confirmed two subscales: Teaching Efficacy and Knowledge Efficacy (Roberts and Henson 2000, p. 7).

In this paper, we build on the academic lineage and evolution of science teacher efficacy scale development, with the primary goal of reporting the development and validation of the SETAKIST-R as a possible addition to our TPD evaluation toolkit. Our team revised the original SETAKIST for use with K-12 teachers by rewording three items; herein, the scale is named the SETAKIST-R.

Measuring self-efficacy is only one component of evaluating TPD (Guskey 2000; Moreno and Tharp 2006; Zielinski et al. 2000), with much discourse regarding the appropriate methods for measuring the self-efficacy construct and determining the validity of self-efficacy instruments (Boone et al. 2010; Desouza et al. 2004; Henson et al. 2001, Roberts and Henson 2000). Although classical confirmatory factor (CFA) analysis was used to validate the SETAKIST (Roberts and Henson 2000), we have opted to apply newer item-response theory (IRT) techniques, including Rasch item-response models, to the SETAKIST-R.

Methods

Development of the SETAKIST-R

Items

The SETAKIST-R retained thirteen of the original sixteen SETAKIST items (Roberts and Henson 2000). We changed the wording in three items (6, 9, and 12) to better align the items with the way our K-12 science teachers speak of their instructional practices. The changed items are described in Table 1.

SETAKIST-R Response Format

We used the same 5-item response categories as the original SETAKIST, however, to align the SETAKIST-R with a format already being utilized in our programs, we assigned a score of one to Strongly Agree and five to Strongly Disagree. This weighting is opposite that of Roberts and Henson (2000). For analysis and scoring purposes, we recoded all items so that high scores reflected higher levels of self-efficacy.

Table 1 SETAKIST-R item revisions and rationale

SETAKIST	SETAKIST-R	Rationale for change
Item 6	Item 6	Item 6
Even when I try very hard, I do not teach science as well as I *teach most other subjects*	Even when I try very hard, I do not teach science as well as I *would like*	Most teachers, elementary or secondary, in our data set and in our projects indicate that they are the science teacher, rather than a multi-disciplinary teacher. If teaching more than one subject, it is typically another science, not content from another discipline
Item 9	Item 9	Item 9
I know the *steps* necessary to teach science concepts effectively	I know *how* to teach *important* science concepts effectively	In our area elementary teachers are tasked with 'steps' in a lesson cycle, not 'steps' in teaching science concepts. They are provided with curricula or training that suggests methods (the *how to*) for use in teaching science
Item 12	Item 12	Item 12
I generally teach science ineffectively	I *might be better* teaching something other than science	Even though our teachers are named as the science teachers on their campuses, many hold either a generalist's or other blended certification (e.g., composite) or dual (even multiple) certifications. Hence, by certificate, they could be called upon to teach another subject. We changed the option to reflect that possibility

SETAKIST-R Validation Study Protocol

The SETAKIST-R ("Appendix") was administered at our project's informational booth at the Conference for the Advancement of Science (CAST) held in Houston, Texas, November 11–13, 2010. The CAST, the annual state-wide Science Teachers Association of Texas (STAT) meeting, is a well regarded large regional K-12 science education conference. In 2010, approximately 7,000 persons attended the CAST (STAT 2011).

Teachers who visited our booth were asked to participate in our study of self-efficacy scales; 334 K-12 teachers opted to do so. The study protocol, informational consent document, and scales were available to participants in paper documents and electronically. All aspects of this study were approved as an educational exemption by the University of Texas Health Science Center at San Antonio's institutional review board (Protocol Number HSC 20110090E) not requiring written informed consent. Teachers responded to the scale without personal identifiers.

Given the logistics and resources at the booth, some teachers took the SETAKIST-R on computer and others completed it on paper surveys. On-line entry was completed by 194 (58 %) teachers using laptop computers connected to a secure password-protected database. SETAKIST-R scale instructions and items were presented on screen in the same format as they were on paper. On-line users were assigned a unique identifier number. Paper forms were assigned random identifier numbers from an Excel generated random number list (2007 Version of Excel). All paper forms were bundled by participant number, secured, and later entered into the database by project staff. Use of the self-monitoring electronic database, combined with attentive staff data entry of paper forms, resulted in minimal missing data. For analyses, data were downloaded as reports into Excel spreadsheets, reviewed, then submitted for processing. We thanked the teachers for completing the instruments by giving them educational posters, bookmarks, pens, and a stress squeeze ball.

SETAKIST-R Statistical Analysis

Scaling

The response scale for the SETAKIST-R ranges from "Strongly Agree = 1" to "Strongly Disagree = 5." For the most part, items alternate between being positively worded and negatively worded. The responses to items 1, 3, 5, 7, 9, 11, 13, and 14 were reversed coded, so that higher values would indicate higher levels of self-efficacy. In addition, the responses were anchored at 0 to facilitate computation of the generalized linear models for our item-response analyses. Therefore, a 0 represented very low self-efficacy and a 4 high self-efficacy.

Rasch Analysis

Considerable discussion has been published regarding the scoring and statistical approaches to validating efficacy instruments (Boone et al. 2010; Desouza et al. 2004; Henson 2002, 2001; Henson et al. 2001; Roberts and Henson 2000). We opted

to apply IRT techniques, including Rasch item-response models to the data collected using the SETAKIST-R. Precedent for this is well-established (Boone et al. 2010; Desouza et al. 2004; Kyriakides and Creemers 2008).

Item-Response Theory (IRT) links item responses to levels of the latent trait. It assumes that the characteristic being measured is a single (unidimensional), continuous, and unobserved (latent) trait. The Rasch model maps the probability of agreement on a particular SETAKIST-R item to a person's underlying level of self-efficacy with respect to teaching science. We started by fitting a generalized linear model for an ordered categorical response, an ordered logistic regression model, with two sets of parameters: one set for items, one set for persons.

There are two basic models for ordered categorical responses: Rating Scale Model (RSM) and Partial Credit Model (PCM) (De Boeck and Wilson 2004). The RSM imposes strict criteria for the item responses represented in the data. These criteria include:

1. All values for the rating scale must be represented on all items; and
2. If a single item lacks a particular rating scale value, for example '5' on a 5 point scale, then the model fails to converge.

The PCM model relaxes the constraints. The constant scoring property is relaxed so that the items can have different numbers of categories.

We conducted our IRT analyses in three stages. First, we fit a PCM model with all 16 SETAKIST-R items to evaluate item-fit and unidimensionality. Second, we used principal components analysis on the person-item residuals to evaluate unidimensionality, and we used CFA to evaluate the fit of the component model to the data. Third, we re-fit the Rasch PCM model for the items associated with each component identified in our analysis of dimensions found among the person-item residuals.

We computed person and item reliability (separation) indices for each subscale. Person reliability is the Rasch equivalent for traditional measure of internal consistency, such as Cronbach's alpha. Low Person reliability (<0.80, separation index <2.0) suggests that an instrument lacks sufficient sensitivity to discriminate persons who are high from those who are low on the latent dimension. Low Item reliability (<0.90, separation index <3) suggest that the items are insufficiently distributed across the range of the latent dimensions to adequately measure the construct (low construct validity).

We also examined the person-item maps that map the responses on each item to the underlying latent dimension. Estimate of the person values on the latent dimension appear as a histogram at the top of the graph. Ideally, the distribution of the person self-efficacy measure should be symmetrically distributed across a sufficient range of the latent dimension to adequately evaluate an instrument. Skewed distributions indicate either that there is insufficient variation in self-efficacy appraisals among the sample participants or that the instrument itself lacks sufficient sensitivity. The mapping of the distribution of item responses to the latent dimension appears in the lower part of the graph. The items are ranked from the lowest to the highest along the vertical axis in terms of their mean value along the latent dimension, indicated by a solid black circle on the graph. Ideally, the mean responses should form a diagonal, from low to high, across the range the latent dimension. The model coefficients (differences in logits between each response level and the lowest

value (reference = 0)) for each individual response (1–4) is represented by open circles. The distribution of these individual responses is expected to follow an ordered pattern from low to high. Deviations from this pattern (disordered responses) suggest significant problems with the item wording, scaling, or both.

Results

Science Teacher Sample

Three hundred thirty-four K-12 science teachers completed the SETAKIST-R. The majority were female (84 %, N = 280) and white (83 %, N = 277). Twenty-five percent (N = 78) identified themselves of Hispanic or Latino origin and 8 % (N = 28) as African American. All grade levels were represented; with the majority of teachers instructing at the elementary school level (55 %, N = 185). The mean teacher age was 40 years (SD = 11; range: 21–67 years).

The average number of years of teaching experience was 11 (SD = 9; range: 0–42 years). About half of the teachers (51 %, N = 167) received their certification through 'traditional' university-based routes. The remainder obtained their certification through various alternative routes. Most respondents held only one certification (76 %, N = 251). Those holding more than one certificate typically had certification in multiple science areas, although some held certificates in English as a Second Language, special education, or mathematics. Teacher respondents were from 19 of the 20 Regional Educational Service Centers in the state of Texas; however, most of those responding to the SETAKIST-R were from Region 4 Houston (25 %, N = 83); Region 20 San Antonio (15 %, N = 51); Region 1 Edinburg (11 %, N = 35), and Region 10 Richardson (11 %, N = 35). The majority (61 %, N = 204) had more than five years teaching experience. Overall, our SETAKIST-R respondents were a representative sample of all CAST attendees. CAST organizers polled 1,000 participants and determined that the 2010 attendees were female (86 %), aged 25–54 (83 %), with more than 5 years teaching experience (75 %) (Science Teachers Association of Texas 2011).

Missing Data

Of our entire data set, only 11 forms were incomplete. Missing demographic data was left blank except for determining Educational Service Center locations based on school or district names provided. Eleven respondents chose not to list gender; two did not respond to the awards item. All respondents answered each of the 16 SETAKIST-R items, providing complete data.

SETAKIST-R Dimensionality

We fit a PCM model with all 16 SETAKIST-R items to evaluate item-fit and unidimensionality. The infit and outfit statistics for 16 items in the SETAKIST-R are displayed in Table 2. The statistics are organized according to the two a priori

expected constructs: Knowledge Self-efficacy and Teaching Self-efficacy. Except for item five, "improvising experiments," and item six, "inviting the principal to evaluate one's teaching," the infit and outfit statistics suggest that the items fit well.

The principal components analysis on the person-item residuals, however, suggests that the 16 items define two separate constructs. The mapping of the items to the constructs suggests that the structure of the SETAKIST-R is consistent with the proposed structure for the original SETAKIST. We used CFA to evaluate the fit of the two dimensions to the SETAKIST-R item-responses.

The fit statistics for the two-factor solution suggest that the hypothesized measurement model provides a reasonably close fit between the item-responses and the suggested measurement model: Knowledge Efficacy and Teaching Efficacy (Table 3).

Rasch PCM Analysis for the Knowledge Efficacy Items

The infit and outfit mean square residuals for the seven Knowledge Efficacy items all have MNSQ values within the 0.5–1.5 range (Table 4), except item 5 "improvising experiments" with the infit MNSQ = 1.958, indicating that this item fails to contribute much to the measurement of Knowledge Self-efficacy, and it easily could be deleted without affecting the scale. The person reliability, corrected for extreme values, was very low at 0.74, yielding a separation index of only 1.7. The corresponding item reliability is 0.80 (separation index = 2.0) are also quite low. This suggests that the Knowledge Self-efficacy scale lacks sufficient sensitivity to discriminate individuals with high levels from those with lower levels of Knowledge Self-efficacy. These low values could be a function of too few items

Table 2 Infit and outfit statistics for the 16 SETAKIST-R items

Content category	Item (verbiage abbreviated)	Chi square	*df*	*p* value	Outfit MNSQ	Infit MNSQ
Knowledge efficacy	1. Welcoming Questions	357.48	324	0.097	1.10	1.08
	3. Answering questions	313.33	324	0.655	0.96	0.85
	5. Improvising experiments	492.82	324	0.000	1.52	1.09
	7. Confident teaching	320.74	324	0.541	0.99	0.98
	9. Teach effectively	261.76	324	0.995	0.81	0.82
	11. Finding better ways	324.55	324	0.481	1.00	1.02
	13. Understand concepts	238.86	324	1.000	0.74	0.72
	14. Student interest	285.53	324	0.939	0.88	0.89
Teaching efficacy	2. Necessary skills	219.23	324	1.000	0.68	0.73
	4. Principal evaluate	633.31	324	0.000	1.95	1.22
	6. Teach well	305.07	324	0.768	0.94	0.89
	8. Difficult topic	257.25	324	0.997	0.79	0.83
	10. Difficult to explain	250.78	324	0.999	0.77	0.77
	12. Teach something else	394.94	324	0.004	1.22	0.94
	15. Anxious teaching	460.17	324	0.000	1.42	1.30
	16. Better understanding	359.78	324	0.083	1.11	1.03

Table 3 Fit statistics confirmatory factor analysis

Fit statistic	Value	Description
Population error		
RMSEA	0.063	Root mean squared error of approximation
90 % CI, lower bound	0.053	
Upper bound	0.074	
pclose	0.020	Probability RMSEA $\leq$ 0.05
Baseline comparison		
CFI	0.915	Comparative fit index
TLI	0.901	Tucker-Lewis index
Size of results		
SRMR	0.048	Standardized root mean squared residual
CD	0.956	Coefficient of determination

assessed on a sample of individuals failing to adequately represent an entire range of Knowledge Self-Efficacy.

The distribution of the estimated Knowledge Self-efficacy for each person displays a markedly negatively skewed distribution, with very few participants at the lower level of Knowledge Self-efficacy (Fig. 2, top). The person-item map for the Knowledge Efficacy items (Fig. 2, bottom) shows that the individual items tend to cluster at the lower levels of the latent trait. The person-item map also reveals some underlying problems with either the content or the scaling of the SETAKIST-R items. The item-responses on five of the eight items fail to map to the latent construct in the expected ordered way (indicated by a '*' along the right vertical axis).

Item one has only a spread from 1 to 2 (strongly agree to agree) indicating that this item may not be needed in the scale if all inservice science teachers are reporting that they "do" feel that they have the necessary skills to teach science. There appears to be some issue regarding the option of "undecided/uncertain" with numbers two (agree) and three (undecided/uncertain) flipping positions in Items 2, 5, and 11.

Rasch PCM analysis for the Teaching Efficacy Items

The infit and outfit mean square residuals for the eight Teaching Efficacy items have MNSQ values within the 0.5–1.5 range (Table 5), with the exception of item four, "invite the principal to evaluate my teaching," which falls outside of the outfit MNSQ = 1.592, indicating that this item fails to contribute much to the measurement of Teaching Self-efficacy.

The person-item map shows that the Teaching Efficacy items (Fig. 3) are reasonably distributed across the range of the latent-trait. The person-item map also reveals that there are some underlying problems with either the content or scaling of the SETAKIST-R Teaching Efficacy items. The person-item map also reveals that there are some underlying problems with either the content or scaling of the

Table 4 Infit and outfit statistics for the knowledge efficacy items

Item (abbreviated)	Chi square	*df*	*p* value	Outfit MNSQ	Infit MNSQ
1. Welcoming questions	359.426	298	0.008	1.202	1.099
3. Answering questions	218.172	298	1.000	0.730	0.732
5. Improvising experiments	319.082	298	0.192	1.067	1.958
7. Confident teaching	300.105	298	0.455	1.004	0.892
9. Teach effectively	186.251	298	1.000	0.623	0.659
11. Finding better ways	318.442	298	0.199	1.065	0.891
13. Understand concepts	191.073	298	1.000	0.639	0.707
14. Student interest	263.277	298	0.927	0.881	0.874

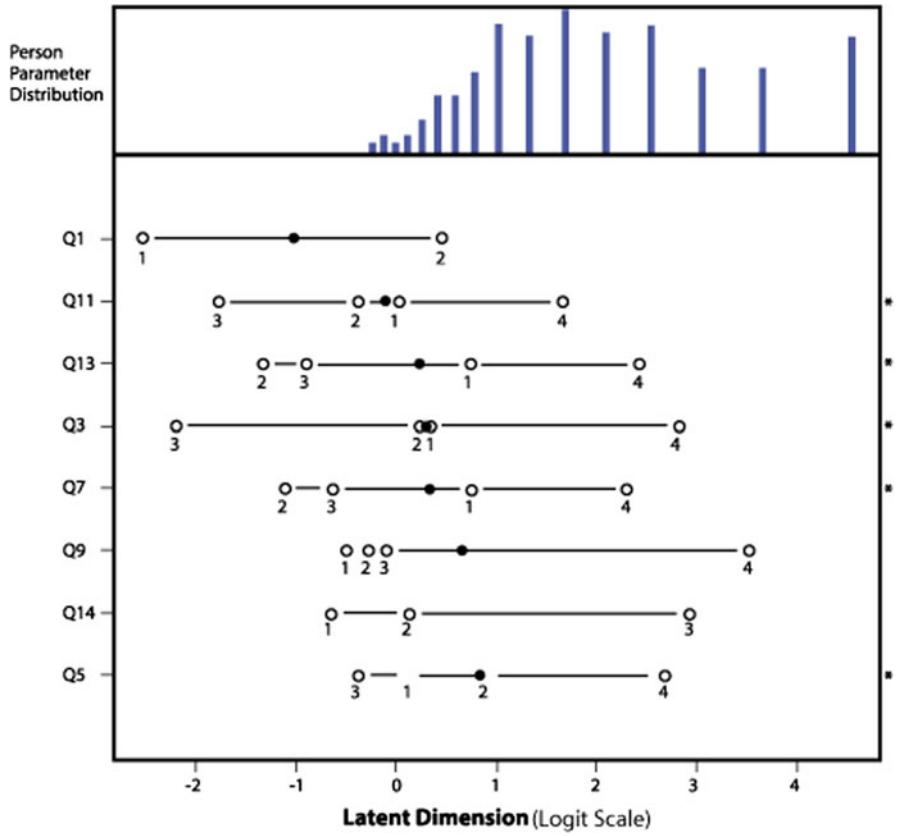

Fig. 2 Person-item *map* for the knowledge efficacy items

SETAKIST-R Teaching Efficacy items. The person reliability, corrected for extreme values, was very low at 0.68, yielding a separation index of only 1.5. The corresponding item reliability is just as poor, with an item reliability of 0.72 and

Table 5 Infit and outfit statistics for the teaching efficacy items

Item (abbreviated)	Chi square	*df*	*p* value	Outfit MSQ	Infit MNSQ
2. Necessary skills	215.085	321	1.000	0.668	0.793
4. Principal evaluate	512.637	321	0.000	1.592	1.175
6. Teach well	279.426	321	0.955	0.868	0.838
8. Difficult topic	249.390	321	0.999	0.775	0.843
10. Difficult to explain	229.158	321	1.000	0.712	0.724
12. Teach something else	309.810	321	0.663	0.962	0.865
15. Anxious teaching	393.856	321	0.003	1.223	1.181
16. Better understanding	284.893	321	0.927	0.885	0.870

separation index of only 1.6, far below the 0.90 and 3.0 needed to support the construct validity of the scale. This suggests that the Teaching Self-efficacy scale lacks sufficient sensitivity to discriminate individuals with high levels from those with lower levels of Teaching Self-efficacy. These low values could be a function of too few items assessed on a sample of individuals that fails to adequately represent the entire range of Knowledge Self-Efficacy.

The distribution of the latent dimension associated with Teaching Self-efficacy is less skewed than Knowledge Self-efficacy and shows a wider range of values for Teaching Self-efficacy within the sample (Fig. 3, top). The item responses are also distributed across a wider range of the latent dimension, but still tend to cluster at the lower end of the latent dimension. Seven of the eight items have disordered mapping of the item responses to the underlying trait. Only one item, 12, "I might be better at teaching something other than science," maps in the expected ordered fashion. Items 2, 4, 6, 8, 10, 15, and 16 all flip on the agree (2) and undecided/uncertain (3) option. Interestingly, all these items use negative terms, like "not." This may contribute to inservice teachers apparent confusion about assigning "undecided" rather than "agree" to a negatively stated item.

Discussion

We have illustrated the academic lineage of an elementary science teacher self-efficacy scale, the SETAKIST (Roberts and Henson 2000), heretofore unpublished in peer-reviewed literature, and its transition (with slight modifications) to the SETAKIST-R for use as one measure of a K-12 science TPD program. We administered the SETAKIST-R in a manner similar to Roberts and Henson (2000) to a convenience sample of 334 K-12 teachers attending a state-wide science teacher conference.

We described the validation data of the SETAKIST-R. It was hoped that the SETAKIST-R data would show it to be an adequate K-12 science teacher self-efficacy measure—it is short, can be administered in less than 10 min, and can be given repeatedly over the course of TPD with little imposition upon respondents.

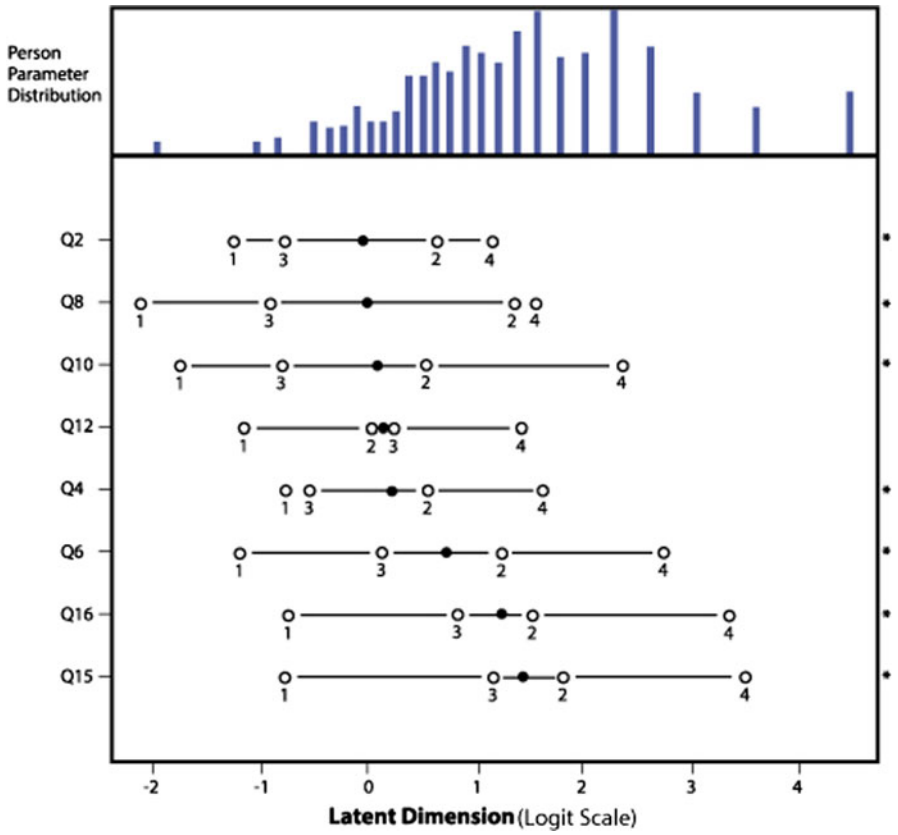

Fig. 3 Person-item *map* for the teaching efficacy items

In validating the SETAKIST-R, the initial Rasch PCM analysis suggested that there are potentially two dimensions to the SETAKIST-R consistent with Science Knowledge and Science Teaching constructs. This was modestly confirmed by the CFA. *However*, detailed analysis of the fit statistics, person and item reliabilities, and the person-item maps for the items within each dimension identified a number of problems with either the items, the scaling, or both.

Low person and item reliability (separation) are typically a function of too few items that fail to capture the entire range of the latent dimension and/or too few participants, with too little variation along the latent dimension in the sample to adequately assess either Knowledge or Teaching Self-efficacy. The person-item map reveals a number of issues that account for the poor performance of the two subscales with respect to their ability to adequately discriminate teachers with higher Teaching or Knowledge Self-efficacies from those with lower levels of self-efficacy. The failure of the item-responses to map to the underlying trait in an ordered fashion is particularly problematic. Potential sources for the observed disordered responses include:

1. The use of verbal negatives and phrases in the composition of the items.
   (a) Respondents often read through negatives (e.g., not) and interpret the item as stated in the affirmative.
   (b) The use of negative sounding phrases such as "difficult to teach," "might be better at," "anxious when," "wish I understood better", may invite a negative mindset or may impart a defensive posture in respondents.
2. The choice of scaling in which Strongly Agree is assigned 1 and Strongly Disagree assigned 5 is a bit counter-intuitive for many respondents.
3. The above conditions combined can create overlapping problems for the respondents.
4. Another problem is introduced with the alteration of positively and negatively worded items; from one item to another, one has a situation in which respondents may inadvertently circle a "4" (Agree) rather than a "2" (Disagree) or visa versa.
5. There is debate as to whether inclusion of the option "undecided/uncertain" is appropriate for persons in the field. For pre-service teachers, it makes sense that they might feel unprepared on any or all of these items; but for inservice teachers, being in the field, they should either indicate that they "can" or "cannot" do (Bandura 1997, 2006). Possibly offering "undecided/uncertain" to some items in this case prompted respondents to decline indicating a lack of skill or commitment. Or there was some other type of confusion between choosing "agree" (2) or "undecided/uncertain" (3).
6. Use of modifiers such as those in item three "typically able" and 11 "continually improvising" could add to respondent confusion—what is "typical" and who does anything "continually?"

The Teaching Efficacy subscale contains eight items; given the wording and scoring, *higher* subscale scores imply increased Teaching Efficacy. The Knowledge Efficacy subscale contains seven items; given the negative wording of the items, scoring was reversed; hence, *higher* subscale scores imply increased Knowledge Efficacy.

To avoid having to "reverse score" items, they should be rewritten from their "negatively" stated aspect to one more "positively" stated. For example, Item 2, "I do not feel I have the necessary skills to teach science," could be re-written as "I have the necessary skills to teach science." This would avoid a "double negative" of sorts—I agree or disagree that I do not have skills. However, for purposes of this manuscript, substantial changes to item wording and scale presentation are beyond its scope of work. It does set into motion additional work to be completed on this scale before it is used as one of our TPD project evaluation measures. This work would need to be completed before we would offer the tool to the broader educational research community.

Item-responses lack sufficient spread across the Knowledge Efficacy latent trait indicating that the items themselves may not be good representations of the extent of Knowledge Efficacy. While the items appear to be better distributed across the Teaching Efficacy constructs, addition of more items that map to higher levels of Teaching Efficacy would improve the scale. One item arrayed only on "strongly agree" (1) and "agree" (2) indicating that inservice K-12 science teachers in this study already agree that they "welcome student questions;" hence, this item is of no consequence in the scale

for this group. In earlier work done with this data, exploratory factor analysis removed this item from the scale when it failed to load adequately on either factor.

Limitations of the Study

The study sample was a *select* group of science teachers—those who chose to attend an intense weekend science TPD conference. Results from the sample may not generalize to more representative groups of K-12 science teachers. It may be that results were disarrayed because of the cross-sectional study design and the multiple variables associated with the demographics. There were no efforts to examine differences among demographic indicators using the IRT techniques, or determination of causal relationships between Science Teaching Efficacy and Science Knowledge Efficacy.

Instructions to respondents could have been reinforced by stressing that they report their "capabilities as of now" (Bandura 1997, p. 44), not their potential capabilities or their past performance. Simple definitions of self-efficacy, setting it apart from feelings, and aligning it with beliefs could have been offered to respondents to better acquaint them with this construct.

Recommendations

In prior statistical work utilizing our same data, our team performed factor analysis as was done by Roberts and Henson (2000). This classical approach yielded the same two-factor structure with one of the 16 items not loading on either subscale. Things looked promising. We applied multivariate analyses methods and observed that the SETAKIST-R showed no subject matter or grade level differences in our study sample, thus, we believed it could be used with our mixed grade level teacher programs. This analysis also disclosed interesting associations related to teachers receiving recognitions and teachers with greater number of years experience. We thought we might have a robust scale with two distinct constructs to use as a measure of our TPD program over an extended period of time—thus, being able to contribute to the longitudinal self-efficacy literature—and explain some demographic phenomena. But, our excitement was short-lived.

IRT techniques, including Rasch Analyses, gave us pause. In this more appropriate analysis, we confirmed the presence of the two factors; however, the item responses did not scale in order. A closer look at the item disarray in each subscale and the actual verbiage within items drew our attention to several issues: the configuration of negatively and positively stated items, the possible impact of negative terms/phrases and ambiguous modifiers, the likely effect of offering an "out" (uncertain/undecided) for inservice teachers, and disclosing the actual "values" of the response choices (1 for agree, 5 for disagree, being counterintuitive to some respondents).

Among the various measures that can be used to evaluate the benefits and impacts of TPD, self-efficacy holds a position (Moreno and Tharp 2006). However, given what our data show, we cannot recommend the SETAKIST-R to be used in assessing K-12 science teacher self-efficacy with confidence.

While we have found several revisions of the STEBI as a standard of practice, such as the Mathmatics Teaching Efficacy Beliefs Instrument (MTEBI) (Enochs et al. 2000), and even modifications of it (e.g., Bursal and Paznokas 2006), we offer this lesson learned. Rather than attempting to revamp a scale, it might be better to go back to the "master" and begin again. Bandura (1997, 2006) exhorts us to create scales that first invite us to indicate that we "can" or "cannot" do something; then, if indicating that we "can," we are to estimate the level to which we believe we can. He carefully outlines procedural instructions in the design of self-efficacy scales, starting with information gathered from other resources. He suggests that our items be phrased in terms of now—what we "can do" rather than what we "will do" (Bandura 1997, p. 43) and cautions against creating scales with fewer than 10-unit intervals as they may be less sensitive (Bandura 1997, p. 44). He notes that beliefs will differ in level (from simple to complex), generality (situation dependent or perception of importance), and strength (some entrance level assurance is needed to attempt a task, then individuals or collectives can persevere and increase mastery) (Bandura 1997, p. 43).

While building upon initial work done by pioneers in the field, teacher self-efficacy scale development has become fraught with copying and/or modifying existing items from other scales. By trying to improve or revamp each others' scales, we, the research community, may have created a type of "in-breeding" that clouds better thinking about efficacy item construction. It has certainly stymied growth in this promising field (Bong 2006; Guskey and Passaro 1994; Klassen et al. 2011) and caused us to speak of scales as being reliable rather than scores (Henson et al. 2001). In addition, there are some basic scale construction rules being ignored with regards to item development, response formats, identifying the latent variable (DeVellis 2003), and in the way we treat raw data (Schumacker and Linacre 1996). Further, there are issues in applying conventional factor analysis (good for identifying underlying variables) where Rasch analysis would be more informative in confirming the existence of a factor and in illustrating "item and person location on the variable" (Schumacker and Linacre 1996).

In a review comparing efficacy research from 1986-1997 with that done from 1998 to 2009, Klassen et al. (2011, p 39–40), we learn that there are four key areas suggested for future directions in efficacy research; there is a need to:

1. Conduct qualitative studies to determine the sources of teacher efficacy—how they "form, develop, and change over time" (Klassen et al. 2011, p. 39)—these have yet to be fully researched and may vary over the career span and across cultures. We should also apply elements from the systematic research model used in student self-efficacy studies reported in Usher (2009) and Usher and Pajares (2009);
2. Offer valid measurements—there is a prevalence of invalid or ill-reported measurements in the research literature;
3. Connect teacher self-efficacy with student outcomes; and
4. Determine how teacher self-efficacy can be enhanced (e.g., TPD, teacher-researcher collaborations).

We further extend the conversation by suggesting that if science is viewed in its broadest most generalized terms, there may yet be a way to offer one scale as an

outcome measure for mixed grade-level pre-collegiate science TPD programs. The National Science Education Standards (National Research Council 1996), as well as the newer Framework for K-12 Science Education (National Research Council 2011) which have influenced the Next Generation Science Standards (2012; http://www.nextgenscience.org/), speak of overarching themes, themes that cut across grade level and discipline—perhaps these are the areas to which science teacher self-efficacy measures could be pointing.

Conclusions

Our goal was to identify and validate a brief measure of science teacher self-efficacy. We summarized the academic lineage and evolution of the SETAKIST, heretofore unpublished in peer-reviewed literature. We modified three items of the SETAKIST, hence creating the SETAKIST-R, and collected empiric validation data. It was hoped that the SETAKIST-R, distilled from the work of others in the field, would perform well as a science teacher self-efficacy measure. The SETAKIST-R could be self-administered in less than 10 min and our data verified the presence of two underlying subscales—'Knowledge Efficacy' and 'Teaching Efficacy.' However, Rasch analyses indicated problems with item wording and scaling. Given that the SETAKIST-R is not robust, we do not recommend its use as a measure of science teacher self-efficacy.

Acknowledgments This project was supported by (a) the National Center for Research Resources and the Division of Program Coordination, Planning, and Strategic Initiatives of the National Institutes of Health through a Science Education Partnership Award, Grant Number, R25-OD025122; (b) a Science Education Drug Abuse Partnership Award, R25-DA025578, from the National Institutes on Drug Abuse (NIDA); and (c) support from the Max and Minnie Tomerlin Voelcker Fund, which established the Voelcker Biosciences Teacher Academy. We gratefully acknowledge the academy teachers who assisted in data collection: Dabs Hollimon, Wanda Pagonis, and Raul Ramirez. We thank Cynthia Ortiz and William Sanns of the UT Health Science Center Department of Epidemiology and Biostatistics for their assistance with data management. We appreciate the cooperation of the CAST conference organizers, Frank Butcher and Lauren Swetland. Finally, we extend our heartfelt thanks to the 344 teachers who took time out of their conference schedule to complete the SETAKIST-R. They made this validation study possible.

Appendix: Self-Efficacy Teaching and Knowledge Instrument for Science Teachers-Revised (SETAKIST-R)

To improve our program, we are trying to find out more about what science teachers think. There are no correct responses to the following statements. You are simply offering your opinion. Indicate your true feelings, not what you think may be a response that is expected. For each of the 16 items, circle the single best response according to the scale below. Your answers are strictly confidential and will be combined with others' responses so that no individual can be identified. **It is important that you respond to all statements circling only one answer:**

1. Strongly Agree
2. Agree
3. Undecided/Uncertain
4. Disagree
5. Strongly Disagree

	Strongly agree	Agree	Undecided/ Uncertain	Disagree	Strongly Disagree
1. When teaching science, I usually welcome student questions.	1	2	3	4	5
2. I do not feel I have the necessary skills to teach science.	1	2	3	4	5
3. I am typically able to answer students' science questions.	1	2	3	4	5
4. Given a choice, I would not invite the principal to evaluate my science teaching.	1	2	3	4	5
5. I feel comfortable improvising during science lab experiments.	1	2	3	4	5
6. Even when I try very hard, I do not teach science as well as I would like.	1	2	3	4	5
7. After I have taught a science concept once, I feel confident teaching it again.	1	2	3	4	5
8. I find science a difficult topic to teach.	1	2	3	4	5
9. I know how to teach important science concepts effectively.	1	2	3	4	5
10. I find it difficult to explain to students why science experiments work.	1	2	3	4	5

Please turn page over to complete SETAKIST-R

	Strongly agree	Agree	Undecided/ Uncertain	Disagree	Strongly Disagree
11. I am continually finding better ways to teach science.	1	2	3	4	5
12. I might be better teaching something other than science.	1	2	3	4	5
13. I understand science concepts well enough to teach science effectively.	1	2	3	4	5
14. I know how to make students interested in science.	1	2	3	4	5
15. I feel anxious when teaching science content that I have not taught before.	1	2	3	4	5
16. I wish I had a better understanding of the science concepts I teach.	1	2	3	4	5

SETAKIST-R
As used in a validation study held at the 2010 Conference for the Advancement of Science, Houston, Texas.

Original SETAKIST reported in:
Roberts, J.K. & Henson, R.K. (2000, November 16). *Self-Efficacy Teaching and Knowledge Instrument for Science Teachers (SETAKIST): A Proposal for a New Efficacy Instrument.* Paper presented at the 28th Annual Meeting of the Mid-South Educational Research Association, Bowling Green, Kentucky. (ERIC Document Reproduction Service No. ED 448 208, Columbia University, NY.)

See also:
Roberts JK, Henson RK, Tharp BZ, Moreno NP. (2001). An examination of change in teacher self-efficacy beliefs in science education based on the duration of inservice activities. *Journal of Science Teacher Education 12* (3):199-213

References

Armor, D., Conry-Oseguera, P., Cox, M., King, N., McDonnell, L., Pascal, A., Pauley, E., & Zellman, G. (1976). *Analysis of the School Preferred Reading Program in selected Los Angeles minority schools*. Santa Monica, California: The Rand Corporation. (ERIC Document Reproduction Service No. ED 130 243, Columbia University, NY).

Ashton, P. & Webb, R. (1982, March). Teachers' sense of efficacy: Toward an ecological model of teacher motivation. Paper presented at the annual meeting of the American Educational Research Association, New York. (Copy made available from the author).

Bandura, A. (1977a). *Social learning theory*. Englewood Cliffs, New Jersey: Prentice-Hall.

Bandura, A. (1977b). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*, 191–215.

Bandura, A. (1978). The self system in reciprocal determinism. *American Psychologist, 33*(4), 344–358.

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist, 37*(2), 122–147.

Bandura, A. (1983). Self-efficacy determinants of anticipated fears and calamities. *Journal of Personality and Social Psychology, 45*, 464–469.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, New Jersey: Prentice-Hall.

Bandura, A. (1993). Perceived self-efficacy in cognitive development and functioning. *Educational Psychologist, 28*(2), 117–148.

Bandura, A. (1996). *Self-efficacy in changing societies*. New York: Cambridge University Press.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman and Company.

Banudra, A. (2006). Guide for constructing self-efficacy scales. In F. Pajares & T. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (pp. 307–337). Greenwich, CT: Information Age.

Bong, M. (2006). Asking the right question: How confident are you that you could successfully perform these tasks? In F. Pajares & T. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (pp. 287–305). Greenwich, CT: Information Age.

Boone, W. J., Townsend, J. S., & Staver, J. (2010). Using Rasch theory to guide the practice of survey development and survey data analysis in science education and to inform science reform efforts: An exemplar utilizing STEBI self-efficacy data. *Science Education, 95*, 258–280.

Bursal, M., & Paznokas, L. (2006). Mathematics anxiety and preservice elementary teachers' confidence to teach mathematics and science. *School Science and Mathematics, 106*(4), 173–280.

Caprara, G. V., Barbaranelli, C., Steca, P., & Malone, P. S. (2006). Teachers' self-efficacy beliefs as determinants of job satisfaction and students' academic achievement: A study at the school level. *Journal of School Psychology, 44*, 473–490.

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach (Statistics for Social Science and Public Policy)*. New York: Springer.

Dembo, M. H., & Gibson, S. (1985). Teachers' sense of efficacy: An important factor in school improvement. *The Elementary School Journal, 86*(2), 173–184.

Desouza, J. M. S., Boone, W. J., & Yilmaz, O. (2004). A study of science teaching self-efficacy and outcome expectancy beliefs of teachers in India. *Science Education, 88*, 837–854.

Devellis, R. F. (2003). *Scale development theory and applications*. Applied Social Research Methods Series (2nd ed., Vol. 26). Thousand Oaks, CA: Sage Publications.

Enochs, L. G. & Riggs, I. M. (1990, April 8–11). *Further development of an elementary science teaching efficacy belief instrument: A preservice elementary scale*. Paper presented at the 63rd Annual Meeting of the National Association for Research in Science Teaching, Atlanta, Georgia. (ERIC Document Reproduction Service No. ED 319 601, Columbia University, NY).

Enochs, L. G., Smith, P. L., & Huinker, D. (2000). Establishing factorial validity of the mathematics teaching efficacy beliefs instrument. *School Science and Mathematics, 100*(4), 194–202.

Gibson, S., & Dembo, M. H. (1984). Teacher efficacy: A construct validation. *Journal of Educational Psychology, 76*(4), 569–582.

Guskey, T. R. (1988). Teacher efficacy, self-concept, and attitudes toward the implementation of instructional innovation. *Teaching and Teacher Education, 4*, 63–69.

Guskey, T. R. (2000). *Evaluating professional development*. Thousand Oaks, California: Corwin Press.

Guskey, T. R., & Passaro, P. D. (1994). Teacher efficacy: A study of construct dimensions. *American Educational Research Journal, 31*, 627–643.

Guskey, T. R., & Yoon, K. S. (2009). What works in professional development? *Phi Delta Kappan, 90*(7), 495–500.

Henson, R. K. (2001 January). *Teacher self-efficacy: Substantive implications and measurement dilemmas*. Keynote address given at the annual meeting of the Educational Research Exchange. College Station, Texas: Texas A&M University.

Henson, R. K. (2002). From adolescent angst to adulthood: Substantive implications and measurement dilemmas in the development of teacher efficacy research. *Educational Psychologist, 37*(2), 137–150.

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.

Hutchings, P. & Shulman, L. (1999, September/October). The scholarship of teaching: New elaborations, new developments. *Change, 31*(5), 10–15.

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 Psychological Review, 23*, 21–43. doi:10.1007/s10648-010-9141-8.

Koballa, T. R., & Crawley, F. E. (1985). The influence of attitude on science teaching and learning. *School Science and Mathematics, 85*, 223–232.

Kyriakides, L., & Creemers, B. P. M. (2008). A longitudinal study on the stability over time of school and teacher effects on student outcomes. *Oxford Review of Education, 34*(5), 521–545.

Moreno, N. P., & Tharp, B. Z. (2006). How do students learn science? In J. Rhoton & P. Shane (Eds.), *Teaching science in the 21st century* (pp. 291–305). Arlington: NSTA Press.

National Research Council. (1996). *National science education standards*. Washington, DC: National Academies Press.

National Research Council. (2011). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: National Academies Press.

Bandura, A. (n.d.) Teacher self-efficacy scale. Retrieved from http://people.ehe.osu.edu/ahoy/.

Next Generation Science Standards. (2012). Retrieved from http://www.nextgenscience.org/. Washington, DC: Achieve Incorporated.

Riggs, I. M. (1995, April). *The characteristics of high and low efficacy elementary teachers.* Paper presented at the annual meeting of the National Association for Research in Science Teaching, San Francisco, California.

Riggs, I. M. & Enochs, L. G. (1989, March 30–April 1). *Toward the development of an elementary teacher's science teaching efficacy belief instrument.* Paper presented at the annual meeting of the National Association for Research in Science Teaching, San Francisco, California.

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.

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

Roberts, J. K. & Henson, R. K. (2000, November, 16). *Self-efficacy teaching and knowledge instrument for science teachers (SETAKIST): A proposal for a new efficacy instrument*. Paper presented at the 28th Annual Meeting of the Mid-South Educational Research Association, Bowling Green, Kentucky (ERIC Document Reproduction Service No. ED 448 208, Columbia University, NY).

Roberts, J. K., Henson, R. K., Tharp, B. Z., and Moreno, N. P. (2000, January 28). *An examination of change in teacher self-efficacy beliefs in science education based on duration of inservice activities*. Paper presented at the annual meeting of the Southwest Educational Research Association, Dallas, Texas (ERIC Document Reproduction Service No. ED 438 359, Columbia University, NY).

Roberts, J. K., Henson, R. K., Tharp, B. Z., & Moreno, N. P. (2001). An examination of change in teacher self-efficacy beliefs in science education based on duration of inservice activities. *Journal of Science Teacher Education, 12*(3), 199–213.

Ross, J. A. (1992). Teacher efficacy and the effect of coaching on student achievement. *Canadian Journal of Education, 17*, 51–65.

Ross, J. A., & Gray, P. (2006). Transformational leadership and teacher commitment to organizational values: The mediating effects of collective teacher efficacy. *School Effectiveness and School Improvement, 17*(2), 179–199.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs, 80*, 1–28.

Scher, L., & O'Reilly, F. (2009). Professional development for K-12 math and science teachers: What do we really know? *Journal of Research on Educational Effectiveness, 2*(3), 209–249.

Schumacker, R. E. & Linacre, J. M. (1996). *Factor analysis and Rasch analysis*. Retrieved from http://www.rasch.org/rmt/rmt94k.htm (April, 26, 2013).

Science Teachers Association of Texas (STAT). (2011). *STAT presents the art of science CAST 2011 sponsorship packages*. Austin, Texas: STAT Publication.

Shidler, L. (2009). The impact of time spent coaching for teacher efficacy on student achievement. *Early Childhood Education, 36*, 453–460.

Shulman, L. S. (1986). Those who understand: Knowledge and growth in teaching. *Educational Researcher, 15*(2), 4–14.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*(1), 1–22.

Shulman, L. (1998, October). *Fostering a scholarship of teaching and learning*. Paper presented at the 10th Annual Louise McBee Lecture, Athens, Georgia.

Stein, M. K., & Wang, M. C. (1988). Teacher development and school improvement: The process of teacher change. *Teaching and Teacher Education, 4*, 171–187.

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

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

Usher, E. L. (2009). Sources of middle school students' self-efficacy in mathematics: A qualitative investigation of student, teacher, and parent perspectives. *American Educational Research Journal, 46*, 275–314.

Usher, E. L., & Pajares, F. (2009). Sources of self-efficacy in mathematics: A validation study. *Contemporary Educational Psychology, 34*, 89–101.

Zielinski, E. J., Dana, T. M., & Courson, S. K. (2000, April). *Effects of a long-term biotechnology professional development program on stages of concern, levels of use, self-efficacy, and classroom implementation.* Paper presented at the annual national conference of the National Association for Research in Science Teaching (NARST), New Orleans, LA.