Heliyon 7 (2021) e07882

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

The validity of the science teacher efficacy belief instrument (STEBI-B) for postgraduate, pre-service, primary teachers

Eileen V. Slater [a,*], Christina Maria Norris [b], Julia E. Morris [a]

[a] Edith Cowan University, School of Education, 2 Bradford St, Mt Lawley 6050, Western Australia, Australia
[b] Edith Cowan University. School of Education, 270 Joondalup Drive, Joondalup 6027, Western Australia, Australia

ARTICLE INFO

Keywords:
STEBI
Self-efficacy
Science teacher education

ABSTRACT

The STEBI-B (Enoch and Riggs, 1990) has been widely used as a measure of undergraduate primary pre-service teacher self-efficacy since its creation. However, the publication of its use within postgraduate teaching courses has been limited. The postgraduate pre-service teachers (Graduate Diploma and Master of Education students) are a very different population, presenting with more life experience and importantly, more experience in Science. This brings the generalizability of the STEBI-B to this population into question. The validity of the STEBI-B for use with a postgraduate, pre-service teacher population was investigated using a Rasch model analysis. Results support the two-factor structure presented by the original authors, the rewording proposed by Bleicher (2004), and additional modifications to the Likert scale and wording to improve targeting for this specific population. With simple, justified modifications the STEBI-B can be used as a tool to positively influence course design in postgraduate, pre-service teacher, science education courses.

1. Introduction

In line with research-informed practice, the authors sought an appropriate tool to measure self-efficacy in the student population (Kazempour and Sadler, 2015) to support the review of a postgraduate, pre-service, primary science education course at their Australian university. Various publications document the importance of self-efficacy for future science teaching (Dembo and Gibson, 1985; Mansfield and Woods-McConney, 2012; Palmer, 2006), which is lacking in many pre-service primary teachers who are often generalist teachers working across all subject areas (Gibson and Dembo, 1984; Gunning and Mensah, 2011; Knaggs and Sondergeld, 2015; McKinnon and Lamberts, 2013). The scale, chosen for its clear factor structure and reported reliability, was the Science Teaching Efficacy Belief Instrument (STEBI-B) which was purposefully designed for undergraduate pre-service primary teachers (Enoch and Riggs, 1990).

Postgraduate pre-service teachers enter education courses with different characteristics from undergraduate teachers. They have existing content knowledge in one or more subject areas due to their past study, and a more diverse range of both professional and life experiences. This includes their prior formal and informal science learning experiences, which may shape their self-efficacy for primary science teaching (Norris et al., 2018). Consequently, when the STEBI-B was administered to the cohort of postgraduate, pre-service primary teachers in the science education course, it was evident that the factor structure and reliability of the STEBI-B would need to be established for this specific population. Deehan (2017) summarised the methodologies and interventions of 140 papers reporting results using the STEBI-B, confirming that the postgraduate pre-service cohort used for this study was a new sample. Therefore, the revalidation of the instrument is imperative to ensure a similar factor structure and reliability are present in a cohort that is different from previously reported samples.

1.1. History of the STEBI

The Science Teaching Efficacy Belief Instrument (STEBI-A) was developed by Enochs and Riggs (1990) to measure in-service primary school generalist teachers' self-efficacy to teach science. This original instrument used a five-point, 25-item Likert scale format from strongly agree through strongly disagree with 'uncertain' at the center of the scale. The STEBI-A consisted of two factors, a personal science teaching efficacy belief scale (PSTE) and a science teaching outcome expectancy scale (STOE) with reliability reported to be 0.92 and 0.77, respectively (Enochs and Riggs, 1990).

* Corresponding author.
E-mail address: e.slater@ecu.edu.au (E.V. Slater).

https://doi.org/10.1016/j.heliyon.2021.e07882
Received 8 March 2021; Received in revised form 23 May 2021; Accepted 25 August 2021

The STEBI-B is a modified version of the STEBI-A for use with pre-service primary teachers. Questions were reworded to reflect future tense and a sample of 212 pre-service students was used for the validation study. Two items were removed for cross loading on the otherwise homogenous factors, resulting in a final instrument containing 23 items with a five-point Likert scale response format. The STEBI-B contained 13 items reflecting the PSTE and 10 items reflecting the STOE (Enochs and Riggs, 1990), with coefficient alpha reported as 0.90 and 0.76 respectively. Bleicher (2004) confirmed the factor structure and reported similar coefficients for the PSTE ($\alpha = 0.87$) and STOE ($\alpha = 0.72$). Bleicher (2004) also suggested that refining the wording of two items, 10 and 13, on the STOE scale to remove the term 'some' would improve the factor structure and reliability as this was used to qualify 'some students' rather than keeping consistent with other items that referred only to 'students'. The reported coefficient alpha for the STOE subscale is consistently lower than that of the PSTE, with a minimum reported alpha of 0.56 (Velthuis et al., 2014; Deehan, 2017). This may be due to there being 13 items in the PSTE and only 10 in the STOE.

The five-point Likert scale used in the STEBI-B could be problematic. Previous research on the use of a central category which does not linguistically fit with the progressive language used in the other categories, such as 'unsure', 'not sure' and 'not decided', has negative effects on the ordering of the category thresholds. Categories such as 'unsure', while presented as a midpoint between agreement and disagreement, do not function as a reflection of a level of agreement, but rather attract respondents who do not have enough information or knowledge to make a judgment (Andrich et al., 1997).

In previous studies, statistical methods such as Factor Analysis, which are based on classical test theory, have been applied to assess the construct validity of assessment instruments. Even though factor analysis can provide an indication of the dimensionality of an instrument, it has been shown to be affected by item difficulty. That is, it can result in factors which are effectively difficulty factors rather than dimensionally independent factors. In addition, in the assessment of the properties of items, difficult items that may be critical for a given construct, are excluded because they do not discriminate. A Rasch model analysis takes item difficulty into account when the factor structure of a set of items is investigated (Nunnaly and Bernstein, 1994).

1.2. Choice of a Rasch model analysis

Because of its properties, the Rasch model has increasingly been used to determine the psychometric properties of instruments in the fields of psychology, education and health (Cano et al., 2011; Hagquist et al., 2009; Tennant and Connaghan, 2007). The simplest Rasch model is the dichotomous Rasch model for analysing responses that are scored correct or incorrect. The polytomous Rasch model is used in cases where there are more than two response categories, for example in Likert scale format questionnaires (e.g. Andrich, 1978; Masters, 1982). There are a number of software packages available for performing Rasch model analyses, for example Winsteps (Linacre, 2018) and RUMM2030 (Andrich et al., 2014).

A Rasch model analysis places the magnitude of the property of the students and the magnitude of the same property of the items on a scale with interval level measurement and therefore expands the range of statistical analysis techniques available for investigating the data (Bond and Fox, 2015). The Rasch models are based on the property of invariance of comparisons, that is, that the comparison between any pair of students is invariant with respect to which items, from within the relevant class of items, are used for the assessment, and likewise the comparison between any pair of items is invariant with respect to which students are assessed, from within the relevant class of students. This formulation is relevant for both dichotomous and polytomous items such as those used for rating scales or partial credit (Andrich, 2011). It also results in the successive categories of a polytomous item being scored with successive integers beginning with 0 (Andrich, 1978). A benefit of the model in an educational context is that when a participant's integer scores are summed across the items, as they are with the STEBI-B, then the total score is the sufficient statistic for estimating the participant's location on the scale. The total score is transformed non-linearly to provide the interval level measurements, termed locations or logits, on the scale.

Cronbach's alpha is commonly reported as a measure of scale reliability. In the case of no missing responses, the software RUMM2030, which is used for the analyses of data reported in this paper, calculates Cronbach's alpha. In addition it calculates a similarly defined index, the person separation index (PSI) based on the Rasch model estimates and its standard errors. The PSI is an estimate of the proportion of the true variance of the distribution of person estimates relative to the sum of this variance and the error variance in the estimates (Rumm laboratory, 2018). When the distribution of responses does not have floor or ceiling effects, Cronbach's alpha and the PSI have almost identical values. In addition, the PSI is calculated readily when there are some missing responses in profiles.

For comparisons among defined groups, it is required that the items function invariantly among the groups, that is, that their parameter estimates are statistically invariant. Lack of invariance is referred to as Differential Item Functioning (DIF) and it is a problem for assessment validity and fairness. A central assumption of the Rasch model is that there is no DIF and therefore the model is powerful in identifying when DIF is present in the data (Hagquist et al., 2009). DIF is identified using RUMM2030 by comparing the means of observed responses of raw scores for class intervals against the item characteristic curve (ICC) for each item. If DIF is present for a subgroup (i.e. gender) on an item, then the plotted means will deviate from the ICC. The software also calculates an analysis of variance (ANOVA) of residuals from the ICC by group and provides a statistical analysis of DIF which complements the graphical analysis.

2. Aim

This paper presents a Rasch model analysis used to evaluate the psychometric properties, including the factor structure (construct validity) and reliability, of the STEBI-B for use with a postgraduate, pre-service, primary teacher population. Only when the scale is validated for use with this population can it be used to meaningfully guide course design through accurate measurement of individuals and the targeting of interventions reflective of their self-efficacy needs.

3. Methods

3.1. Research design

The broader research aimed to measure changes in the self-efficacy, in terms of teaching primary science, of students enrolled in a primary education, postgraduate, science unit. Prior to using the STEBI-B to determine the effect of the science unit on the postgraduate student's self-efficacy for teaching primary science, an analysis of its validity and reliability was required. When administering an instrument to a sample which differs from those previously reported, it is necessary to establish that the factor structure and reliability are consistent with those reported for the previous samples. Approval to conduct this analysis fit within the broader study, with ethics approval granted by the University's Human Research Ethics Committee (application 12776) and all participants providing informed consent prior to completion of the questionnaires.

In investigating the factor structure of the STEBI-B for this sample, first the existence of two separate factors, the STOE and PSTE, needed to be established. The research design therefore involved an initial analysis of all 23 items as a single scale, before separating the items into the reported existing subscales, if evidenced. As the Rasch model assumes all items represent the same construct (Sick, 2011, p.15), an analysis of all 23 items should show misfit to the model. In addition, the factor structure

needed to remain consistent for both pre and post measurement points, in order to accurately measure change at the two time points in the broader research. It was expected that students would move up (or down) the scale systematically, while the factor structure would remain the same.

3.2. The sample

The STEBI-B was administered to postgraduate pre-service teachers at an Australian university as part of a broader research study (Norris, 2017). The postgraduate pre-service teachers were enrolled in a Graduate Diploma of Education (Primary) and self-selected into the research while enrolled in the Science Education unit within their course. The pre-intervention STEBI–B was completed by 361 pre-service teachers, with 275 of the participants also completing the post-intervention STEBI-B. Table 1 summarises the gender and science education backgrounds of the 270 participants who completed the items in the questionnaire relating to this combination of demographic information.

Of note in Table 1 are the 33% of students who have a bachelor's degree or higher in an area of Science. This is very different from the type of learner who presents in the undergraduate primary science units of study, with implications for self-efficacy. Subsequently, the STEBI-B was analysed to confirm its suitability for use as a pre/post analysis tool with this cohort of students.

3.3. Data analysis

The psychometric properties of the STEBI-B were investigated using the Rasch model (Rasch, 1961). The software RUMM2030 (Andrich et al., 2014) was used to analyse responses to the instrument. Data were entered with negatively worded items reverse coded as per the original questionnaire (Enochs and Riggs, 1990). Both pre and post assessment data were entered using Time One (T1, pre-intervention) and Time Two (T2, post-intervention) as person variables. Therefore, the analysis represented 636 responses, which could be separated into comparison groups for individual analysis based on Time (T1, pre-intervention N = 361; T2 post intervention N = 275). The purpose of the T1/T2 grouping in this Rasch model analysis is to consider the validity and reliability of the questionnaire for the sample at both time points.

3.3.1. Person/item alignment and reliability

Cronbach's coefficient alpha as well as the Person Separation Index (PSI) (Andrich, 1982), were examined. In general and under ideal circumstances, the PSI and Cronbach's coefficient alpha have equivalent values. However, in the case of floor or ceiling effects, Cronbach's coefficient alpha is artificially inflated, and the PSI is a more accurate indication of reliability. Tennant and Conaghan (2007) specify a minimum value of 0.7 for PSI is required when making decisions about groups use and 0.85 for use with individuals. In this analysis, excellent PSI and Cronbach alpha was deemed to be >.80, with .70 the accepted minimum. In the STEBI-B questionnaire the Rasch model analysis provides an estimate of the pre-service teachers 'difficulty to endorse' estimate for each item, that is, a sense of how hard the items were for participants to agree with. As the item and person estimates are on the same scale, positive values indicate items that were more difficult to endorse and persons with higher efficacy. This analysis was also used to identify 'gaps' in the distribution of items, where there are no items aligned to persons levels of the latent construct (see Figure 2). This can guide item development to fill particular 'gaps' in a scale.

Table 1. Science learning background and gender of participants (adapted from Norris, 2017).

Highest Level of Science Education	Female % (N = 204)	Male % (N = 66)	Total % (N = 270)
Year 9 high school	2	0	1
Year 10 high school	15	12	14
Year 11 high school	7	8	7
Year 12 (final year) high school	45	42	44
Undergraduate Degree	25	30	26
Postgraduate Degree	6	8	7

3.3.2. Data fit to the model

The summed responses of more than one item are thought to be more valid and reliable than a response to one item only, assuming each item measures the same latent trait. Rasch model analysis identifies items that do not 'fit' with other items, that is, they don't measure the same construct. Fit was assessed statistically through two fit statistics, the fit residual and chi-square fit statistics, as well as visually through the Item Characteristic Curves (ICC) when assessing STOE and PSTE item fit. If the mean (*M*) and standard deviation (*SD*) of the fit residual statistic is close to 0 and 1 respectively it indicates that the data fit the model. If an individual item fits the model, the item fit residuals have values between -2.5 and 2.5. These cut-offs are supported by Andrich and Marais (2012), with values outside of these ranges indicating the items either over or under discriminate. The chi-square fit statistic compares observed mean responses with what is expected according to the Rasch model, indicating misfit when the divergence is statistically significant. To avoid a type one error, the Bonferroni adjustment (Bland and Altman, 1995) can be used to determine significance levels, and was applied in this study due to the multiple comparisons being tested.

3.3.3. Violations of local independence

Multidimensionality (more than one factor or construct) was diagnosed through three methods. First, through a principal component analysis (PCA) of residuals where the absence of any meaningful pattern supports the assumption of unidimensionality. Second, by forming two subtests of items based on evidence from the PCA and determining the statistical equivalence of person estimates based on these two subtests of items through t-tests (Smith, 2002). If less than a chance level (5%) of persons has different person estimates (logits) for the two subtests of items, the assumption of unidimensionality is supported.

Finally, the 'subtest' option in RUMM2030 allows the reliability estimates of two sets of data to be compared. The first set of data uses the original items and estimates and assumes they are independent. The second set of data combines items which are hypothesised to be dependent into a single polytomous item. If the second set of data is shown to have lower reliability values, then the case for multidimensionality is supported (Marais and Andrich, 2008).

3.3.4. Differential item functioning (DIF)

Comparisons between the summed scores on an instrument for members of subgroups can be misleading if, for the same level of the trait being measured, members of the different subgroups respond differently to individual items (Looveer and Mulligan, 2009). A differential item functioning (DIF) analysis identifies item bias, by determining items that function differently for subgroups of the population. DIF was assessed graphically through an inspection of the Item Characteristic Curve (ICC) for each item and confirmed statistically through an ANOVA of the residuals.

3.3.5. Response category functioning

This analysis determines if participants' categoric responses to the items (i.e., strongly disagree – strongly agree) are consistent with the metric estimate of the construct (Tennant and Conaghan, 2007). The criteria for evaluating the response categories of the Likert scale used in the STEBI-B were: a minimum number of responses per category of 10 (Linacre, 1999), that the category thresholds progress monotonically (category thresholds represent the point at which the difficulty of endorsement of two adjacent categories has 50:50 probability), that the

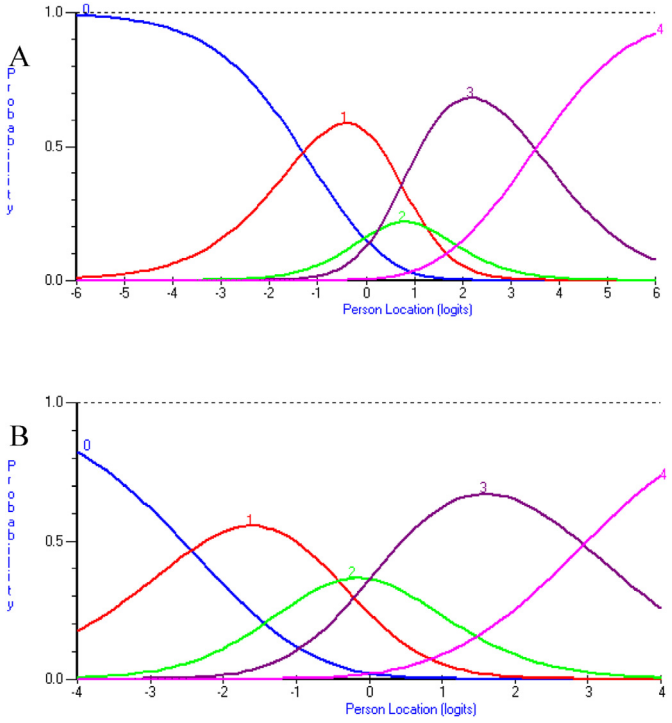

Figure 1. Example category probability graphs A) disordered categories and B) ordered categories.

category probability graphs indicate monotonic progression of categories (Figure 1), and in the case of a five point scale, the distance between category thresholds is at least 1.0 logits and no more than 5 logits. The acceptable distributions for response categories are uniform, normal, bi-modal or slightly skewed (Bond and Fox, 2015).

4. Results

4.1. Item/person alignment and reliability

Figure 2 shows histograms of the Rasch person estimates (top histogram in red) and threshold distribution estimates (bottom histogram in blue) for the STEBI-B on the same scale. The mean of the person estimates was 1.09 relative to the mean of items which is constrained to be 0, suggesting that many participants found the items easy to endorse.

There is an absence of item thresholds to separate the persons between 1 and 2 logits, suggesting further investigation of the targeting of the instrument is warranted as there may be a 'gap' where there are not thresholds that will separate the persons. The person separation index was 0.85 and coefficient alpha was 0.83, indicating good reliability, and an excellent ability of item fit residual fit statistics and chi-square tests to detect misfit when analysing data fit to the model.

4.2. Data fit to the model

Table 2 shows the summary fit statistics for all samples (pre, post and stacked) for the 23 item STEBI-B, providing evidence of misfit to a unidimensional model in all samples.

In the stacked sample, the mean of the item fit residual was 0.34 and the *SD* was 1.58, indicating the data deviate from the unidimensional model. The chi-square probability was also significant. These results suggested multidimensionality needed to be investigated and was expected based on previously reported constructs.

4.2.1. Violations of local independence

In the PCA of the residuals, the principal component loadings showed a relatively large difference between the first two components, with Eigen values of 3.596 and 1.683 respectively. An analysis of the first principle component (PC1) showed two distinct groupings of items. The STEBI-B purports to measure two constructs and these groups were represented in the PC1 loadings. A t-test was run using the two sets of items identified by the PC1 loadings and labelled STOE:PSTE. The STOE was represented by items 1, 4, 7, 9, 10, 11, 13, 14, 15 and 16 with the remaining items comprising the PSTE. The percentage of students with statistically significant differences in person estimates based on these two groupings of items was 20.44%, supporting the proposition that the STEBI-B is not unidimensional. In addition, a subtest analysis based on the items which constituted the STOE and PSTE resulted in a difference in the PSI from 0.85 to 0.66, while the alpha dropped to 0.49, supporting a multidimensional structure (Marais and Andrich, 2008).

Since the reported multiple factor structure (Enochs &Riggs, 1990; Bleicher, 2004) was supported by the Rasch model analysis of the STEBI-B for the pre, post and stacked samples for this population, two separate Rasch analyses, one of the STOE scale and one of the PSTE scale were conducted.

4.2.2. Item/person alignment and reliability – STOE and PSTE

Figure 3 shows histograms of the Rasch person estimates (top histogram in red) and item threshold estimates (bottom histogram in blue) for the STOE on the same scale, for the stacked sample. The mean of the person estimates was 1.19 relative to the mean of items which is constrained to be 0, supporting the hypothesis that many participants found the STOE items easy to endorse. Figure 4 shows histograms of the Rasch person estimates (top histogram in red) and item threshold estimates

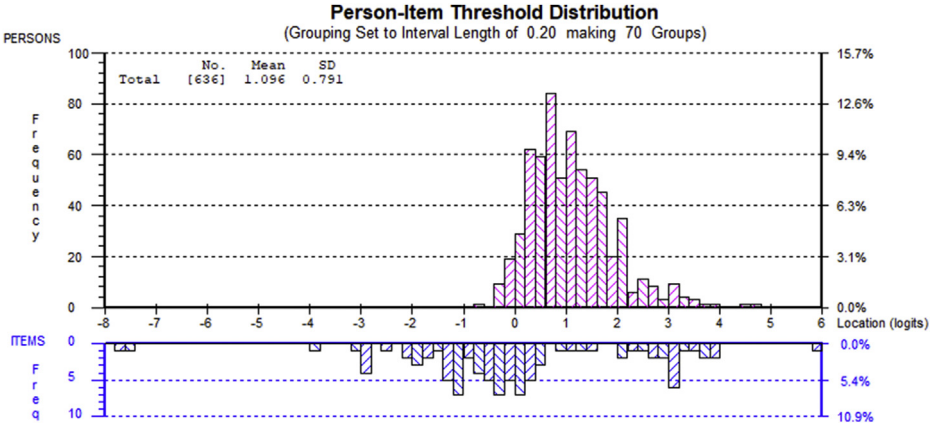

Figure 2. Person-item Threshold Distribution Graph for the STEBI-B.

Table 2. Summary Statistics Output for 23 item STEBI-B.

	Pre		Post		Stacked	
	ITEMS		ITEMS		ITEMS	
	Location	Fit Residual	Location	Fit Residual	Location	Fit Residual
Mean	0.0000	0.3378	0.0000	0.2921	0.0000	0.3454
Std. Dev.	1.0156	1.7517	0.7577	1.5212	0.6852	1.5810
	PERSONS		**PERSONS**		**PERSONS**	
	Location	Fit Residual	Location	Fit Residual	Location	Fit Residual
Mean	1.2362	-0.2592	1.1754	-0.3894	1.0961	-0.3490
Std. Dev.	0.7334	1.4987	0.8402	1.6312	0.7914	1.5810
	RELIABILITY		**RELIABILITY**		**RELIABILITY**	
PSI	0.84		0.85		0.85	
Cronbach's Alpha	0.81		0.83		0.83	
	FIT		**FIT**		**FIT**	
Total Item Chi-Square	192.7759		147.6079		269.7112	
Degrees of Freedom	69		69		69	
Chi-Square Probability	0.000000		0.000000		0.000000	

(bottom histogram in blue) for the PSTE on the same scale, for the stacked sample. The mean of the person estimates was 1.28 relative to the mean of items which is constrained to be 0, supporting the hypothesis that many participants found the PSTE items easy to endorse.

There is an absence of item thresholds to separate the persons between 1.0 and 2.0 logits for both the STOE and PSTE scales, suggesting further investigation of the targeting of both scales is warranted. For the STOE, the person separation index was 0.73 and coefficient alpha was 0.69; while these were 0.85 and 0.84 respectively for the PSTE.

4.2.3. Data fit to the model – STOE and PSTE

Table 3 shows the fit statistics for the STOE for all three samples. The fit statistics for the pre-intervention sample showed the better fit of the items to a unidimensional Rasch model, while post and stacked models suggested multidimensionality.

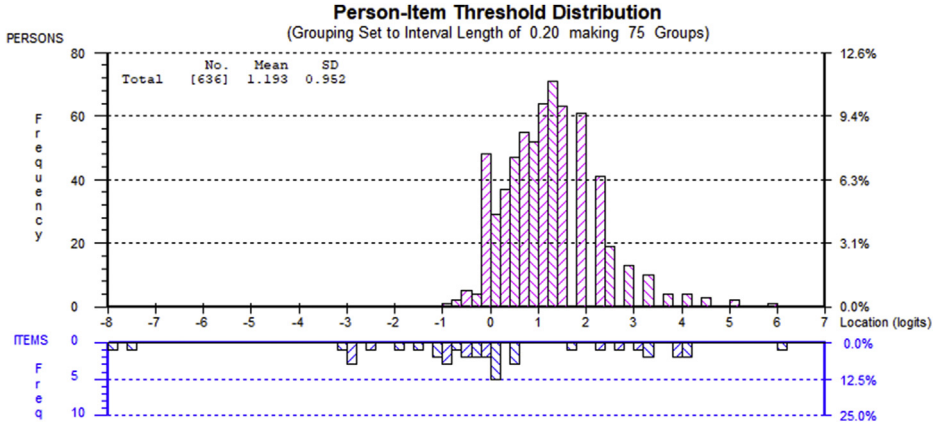

Figure 3. Person-item threshold distribution for STOE.

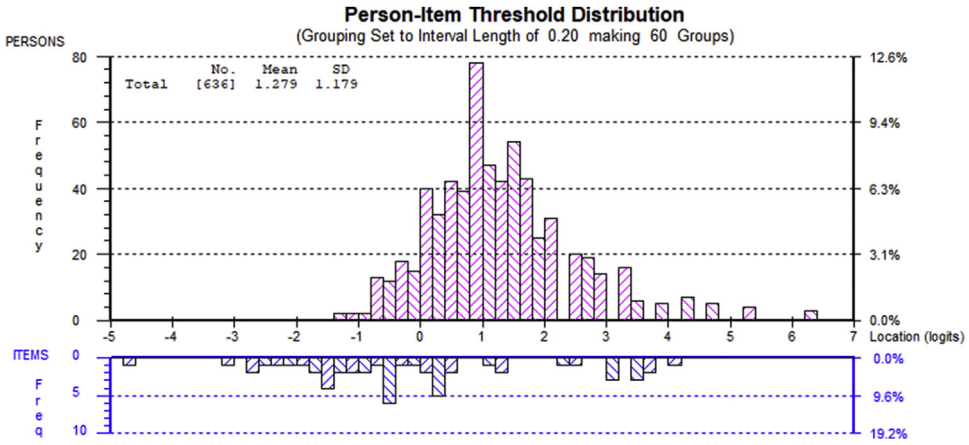

Figure 4. Person-item threshold distribution for PSTE.

The mean of the item fit residual for the STOE stacked data was 0.18 and the *SD* was 2.00, indicating deviation of the items from the model. The mean of the person fit residuals was -0.42 and the *SD* was 1.32, indicating a better fit of the persons to the model. It is therefore more likely that misfit to the Rasch model is being influenced by the items, rather than by individual persons whose response patterns do not fit with the predicted Rasch model. The STOE PSI and Cronbach alpha scores indicated a good ability of these tests to detect misfit. The chi-square probability suggested the pre-intervention sample for the STOE fit the unidimensional model, while the post-intervention and stacked samples did not.

Table 4 summarises model fit statistics for the PSTE for all three samples. In the PSTE analysis, three persons were removed from the data automatically by Rumm2030 for item calibrations due to extreme scores, recording the highest level of agreement with every item. Curtis (2001) advises the removal of these types of extreme persons due to significant effects on item calibrations.

The mean of the item fit residuals for the stacked sample was 0.31 and the *SD* was 2.83, indicating the data diverge from the unidimensional model. The person statistics showed a better fit of the persons to the model. The chi-square probability suggests that the post-intervention data fit a unidimensional model at the 0.0007 (Bonferroni adjusted) level while the pre-intervention and stacked data do not. The PSI and Cronbach alpha $>.80$ indicated an excellent ability of these tests to detect misfit.

To further investigate the source of misfit, individual item fit was also analysed using individual item fit residuals and chi-square probabilities. Viewing the Item Characteristic Curve (ICC) graphs can direct the researcher toward items that may show misfit. Figure 5 shows an example of an ICC graph for item 13, where divergence from the predicated curve is evident for the first and third class intervals.

For the STOE stacked sample, items 10 (2.83) and 13 (4.62) had high fit residuals and item 13 also had a significant chi-square probability (0.00). Item 10 also showed misfit in the pre-intervention data with a high fit residual (4.03) while item 13 (4.77, 0.00) showed misfit against both measures in the post-intervention data.

In the PSTE stacked sample, items 6 (4.233), 8 (4.884), 17 (-3.319), 18 (-3.29) and 21 (-3.103) had high fit residuals and significant chi-square probabilities at the Bonferroni level of adjustment (0.000769), indicating misfit to the model. Items 6, 8 and 18 also showed misfit in the pre-intervention sample, while no items showed misfit in the post-intervention sample.

4.2.4. Violations of local independence – STOE and PSTE

In the PCA of the residuals for the STOE stacked sample, the principal component loadings showed no relatively large difference between the first two components, with Eigenvalues of 1.5 and 1.2 respectively. An analysis by PC1 showed no distinct grouping of items. In a subtest analysis of two groups, hypothetically formed based on positive and negative loadings on PC1, the PSI dropped slightly to 0.68 and the error adjusted correlation between the two tests was 0.93. This supports the notion that the STOE is a unidimensional scale. Residual correlations were inspected with no item correlating with any other item above 0.3. Therefore, no violations of trait or response dependence were evident for the STOE construct.

For the PSTE stacked sample, in the PCA of the residuals the principal component loadings showed no relatively large difference between the first two components, with Eigenvalues of 1.5 and 1.2 respectively. An analysis by PC1 showed no distinct grouping of items. In a subtest analysis of two groups, formed based on positive and negative loadings on PC1, the PSI dropped slightly to 0.77 and the error adjusted correlation between the two tests was 0.90. This further supports the PSTE as a unidimensional scale. Items 6 and 8 had a residual correlation above 0.3 which indicates a possible violation of independence of response. These two items are part of the cluster of items around 0 logits (see Figure 5).

4.2.5. Differential item functioning - STOE and PSTE

The demographic factor that could be tested for DIF was gender. In analysing the ICC graphs for gender in the STOE stacked sample, no items showed potential uniform DIF. The ANOVA results supported this assumption, with an F ratio probability less than 0.002 being indicative of a statistically significant difference in group means on an item. This held true for the pre and post intervention samples on the STOE construct.

In analysing the ICC graphs for gender in the PSTE stacked sample, items 8, 18 and 19 showed potential uniform DIF; this is exemplified for item 18 (F ratio probability 0.000484) in Figure 6. The ANOVA results supported the graphical analysis, with an F ratio probability less than 0.001 being indicative of a statistically significant difference in group means on an item. Items 18 ('I will typically be able to answer students' Science questions') and 19 ('I wonder if I will have the necessary skills to teach Science') were uniformly endorsed more positively by males while item 8 ('I will generally teach Science ineffectively') was more easily positively endorsed by females. This held true for the pre-intervention data, while the post intervention data indicated no gender difference. Further differential analysis based on science education background was not analysed due to unequal and small group sizes.

4.2.6. Response category functioning – STOE and PSTE

Despite the large STOE stacked sample, exceeding 20 persons per item, the category 'strongly agree' was rarely used and did not meet the

Table 3. Summary statistics output for STOE.

	Pre		Post		Stacked	
	ITEMS		ITEMS		ITEMS	
	Location	Fit Residual	Location	Fit Residual	Location	Fit Residual
Mean	0.0000	0.2991	0.0000	0.1198	0.0000	0.1879
Std. Dev.	1.2058	1.4778	0.7381	1.8203	0.7124	2.0044
	PERSONS		**PERSONS**		**PERSONS**	
	Location	Fit Residual	Location	Fit Residual	Location	Fit Residual
Mean	1.4970	-0.3404	1.1896	-0.5012	1.1931	-0.4262
Std. Dev.	0.9123	1.2591	1.0056	1.3978	0.9522	1.3257
	RELIABILITY		**RELIABILITY**		**RELIABILITY**	
PSI	0.70		0.75		0.73	
Cronbach's Alpha	0.67		0.71		0.69	
	FIT		**FIT**		**FIT**	
Total Item Chi-Square	47.3372		85.9010		98.2754	
Degrees of Freedom	40		40		40	
Chi-Square Probability	0.198086		0.000034		0.000001	

Table 4. Summary statistics output for PSTE.

	Pre		Post		Stacked	
	ITEMS		ITEMS		ITEMS	
	Location	Fit Residual	Location	Fit Residual	Location	Fit Residual
Mean	0.0000	0.4600	0.0000	0.1085	0.0000	0.3104
Std. Dev.	1.0031	2.7713	0.8807	1.5007	0.8054	2.8312
	PERSONS		**PERSONS**		**PERSONS**	
	Location	Fit Residual	Location	Fit Residual	Location	Fit Residual
Mean	1.2424	-0.2801	1.4983	-0.4139	1.2787	-0.3740
Std. Dev.	1.0507	1.3658	1.3410	1.3268	1.1794	1.3746
	RELIABILITY		**RELIABILITY**		**RELIABILITY**	
PSI	0.85		0.85		0.85	
Cronbach's Alpha	0.84		0.84		0.84	
	FIT		**FIT**		**FIT**	
Total Item Chi-Square	223.6505		68.2929		214.7027	
Degrees of Freedom	39		39		39	
Chi-Square Probability	0.000000		0.00255		0.000000	

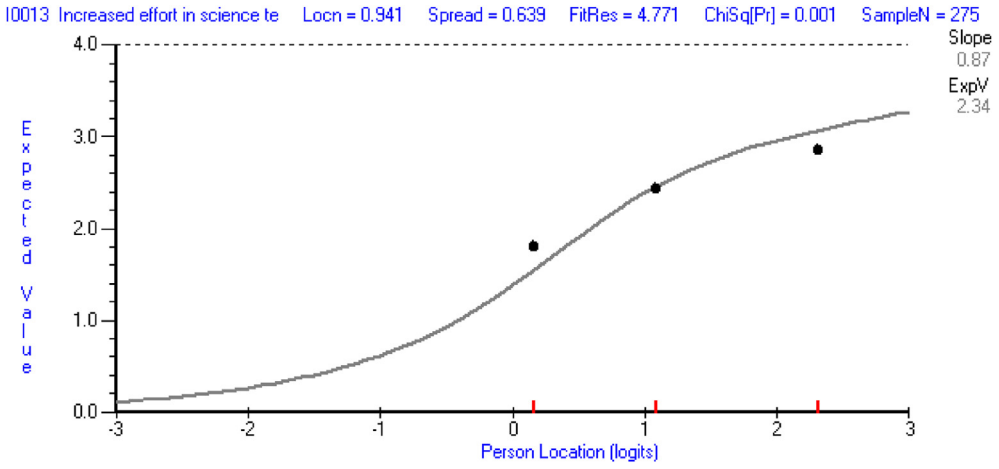

Figure 5. Sample ICC graph for item 13.

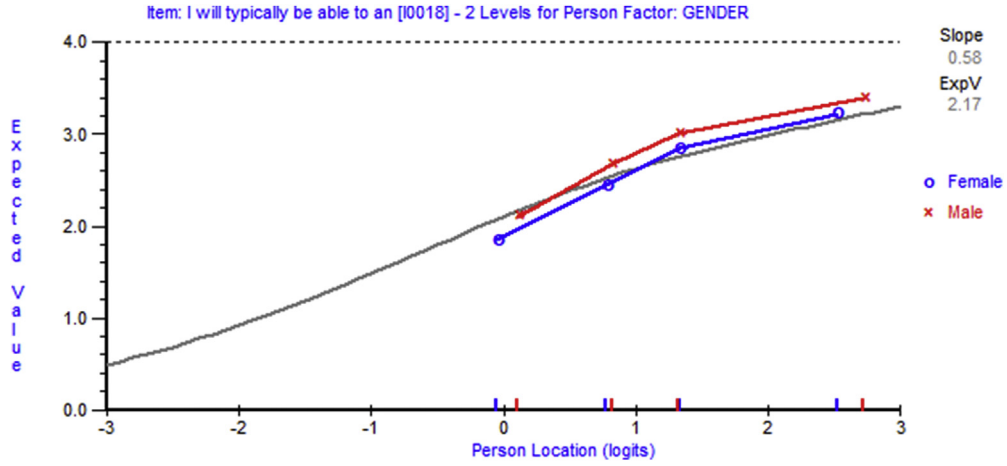

Figure 6. ICC graph for item 18 showing uniform DIF for gender.

recommended minimum of 10 respondents for every item (Linacre, 1999). Table 5 summarises the category use, reflecting a usage pattern which held true for both the pre-intervention and post-intervention STOE samples.

Similarly, for the PSTE stacked sample, the category 'strongly agree' was rarely used and does not meet the recommended minimum of 10 respondents for 10 of the 13 items. None of the respondents used this category for item 2, while the highest use of the category was 4.5% for item 20. This pattern of response category use held true for both the pre-intervention and post-intervention PSTE samples (Table 6).

The second criterion for assessing response category functioning is category thresholds, which should increase monotonically (Bond and Fox, 2015). For STOE items 4, 9 and 13 the thresholds were disordered in the stacked sample as well as in the post-intervention sample. Inspection of the category probability curves also indicated category disorder for these items, where at least one category was never the most probable category at any value along the *x* axis (Bond and Fox, 2015). The pre-intervention sample had no disordered thresholds (see Figure 1).

Category thresholds for PSTE items 8 and 19 were disordered in the stacked and pre-intervention samples. This is also reflected in the category probability graphs for these items, where category 'unsure' was never the most probable category at any value along the *x* axis. The post-intervention sample had a large number of items with disordered thresholds: 2, 3, 6, 8, 18, 19, 21, 22 and 23. When viewing the category probability curves for this sample, item 6 was the most problematic. Neither category 'agree' nor 'unsure' were ever the most probable category along the *x* axis for item 6, as 'disagree' dominated the responses along a large proportion of person abilities.

The recommended minimum distance of 1.0 logits (Bond and Fox, 2015) was not met consistently by the Likert scale for any item in the stacked sample nor the post-intervention sample for both the STOE and the PSTE constructs, with distances between thresholds two and three consistently problematic across items. While the pre-intervention STOE sample had more evident distances between thresholds, movement from threshold two to threshold three for all items was still less than the desirable 1.0 logits. The pre-intervention PSTE sample had three items, 3, 21 and 23 which progressed monotonically and met the minimum distance criteria.

5. Discussion

5.1. STOE scale

The STOE scale had sufficient reliability (PSI = .73), no trait or response dependence and no item displayed DIF for gender. Items 10 and 13 could be reworded (Bleicher, 2004) or removed, as they show misfit to the model. The response category 'unsure' is used by approximately 40% of respondents to question 10 in the stacked sample, suggesting the question is problematic in its current form. Item 13 also exhibits disordered thresholds, which will influence its misfit to the model. If item 10 and 13 are removed from the STOE scale the PSI drops to .70 and the alpha to .68, with all remaining items and persons fitting the model. This does not resolve the disordered category thresholds for items 4 and 9.

Collapsing categories 'unsure' and 'disagree' resolves the category threshold disorder, the PSI becomes .72 and the alpha increases to .73. This also provides for a minimum distance between thresholds of 1.0 logits for all items. If only items 4, 9 and 13 have categories unsure and disagree collapsed, the PSI is .73 and alpha is 0.71. This does not resolve the distance between all item thresholds; however, the data do fit the model. The problem with collapsing these two categories, under either solution, is that they have qualitatively different meanings. Collapsing agree and strongly agree, for example, is a matter of the intensity of your agreement. However, being 'unsure' is not a level of disagreement.

Removing items 10 and 13 and collapsing the scale to 4 points resulting in an eight item scale is one option for improving the STOE scale; it fit the unidimensional Rasch model and has a scale which progresses monotonically with acceptable distances between category thresholds for all items. The PSI becomes .70 and the alpha becomes 0.71.

Noting the aforementioned qualitative implications of collapsing the scale, a second possible resolution would be to reword items 10 and 13 and remove the central category in the Likert scale. This forces a choice between agreement and disagreement. Middle categories in Likert scales that are not worded to clearly place the response on a continuum with other response categories, such as 'uncertain', have been shown to be problematic (Andich, De Jong & Sheridan, 1997). An analysis with the current data set cannot model this option without losing substantial amounts of data (for example, 40% of responses for item 10). Nevertheless, a revised version could be piloted with a postgraduate, pre-service teacher population to determine if the desired effect is achieved in relation to category use, monotonic progression, distance between threshold categories and model fit.

5.2. PSTE scale

The PSTE scale had sufficient reliability (PSI = .85), and no trait dependence; however, item 6 'I will not be very effective in monitoring science experiments' and item 8 'I will generally teach science ineffectively' both displayed response dependence. Comparing the wording of these items and their proximity on the scale adds weight to this statistical outcome. Items 6 and 8 were identified as showing misfit to the Rasch model analysis in both the stacked and pre-intervention samples. In addition, item 8 showed gender DIF and had disordered thresholds in all three samples. When item 8 is removed the PSI remains at 0.85.

The use of negatively worded items, while historically common, has been shown to be unnecessary in scale development. Negatively wording

Table 5. Category frequencies stacked STOE.

ITEM	SA	A	U	D	SD
1	2	46	142	372	74
4	2	5	60	424	145
7	0	50	188	363	35
9	2	14	66	438	116
10	5	171	254	202	4
11	0	38	150	404	44
13	12	89	123	353	59
14	2	52	126	415	41
15	2	63	143	383	45
16	1	46	154	366	69

Table 6. Category frequencies stacked PSTE.

ITEM	SA	A	U	D	SD
2	0	1	26	285	321
3	6	34	150	302	141
5	2	26	101	436	68
6	6	40	115	384	88
8	8	48	83	341	153
12	10	59	154	317	93
17	9	58	146	346	74
18	4	49	168	341	71
19	27	175	114	273	44
20	28	73	138	310	84
21	4	29	145	390	65
22	1	7	36	291	298
23	6	63	175	330	59

old	new		SD	D	A	SA
1	1	When a student does better in science, it is often because the teacher exerted a little extra effort.				
4	2	When the science grades of students improve, it is often due to their teacher having found a more effective teaching approach.				
7	3	If students are underachieving in science, it is most likely due to ineffective science teaching.				
9	4	The inadequacy of a student's science background can be overcome by good teaching.				
11	5	When a low-achieving student in science progresses, it is usually due to extra attention given by the teacher.				
14	6	The teacher is generally responsible for the achievement of students in science.				
15	7	Students' achievement in science is directly related to their teacher's effectiveness in teaching science.				
16	8	If parents comment that their child is showing more interest in science, it is probably due to the child's teacher.				
2	9	I will continually find better ways to teach science.				
3	10	Even if I try very hard, I will not teach science as well as I will most subjects.				
12	11	I understand science concepts well enough to be effective in teaching primary school science.				
17	12	I will find it difficult to explain to students why science experiments work.				
19	13	I wonder if I will have the necessary skills to teach science.				
20	14	Given a choice, I will not invite the principal to evaluate my science teaching.				
22	15	When teaching science I will usually welcome student questions.				
23	16	I do not know what to do to turn students on to science.				

Figure 7. Modified STEBI-B for postgraduate pre-service primary teachers.

items places more cognitive load on respondents and has a subsequent effect on their responses due to attention and confusion, without the desired impact on response bias (van Sonderen et al., 2013). Retaining negatively worded items is therefore not a consideration when optimising the STEBI-B for use with this sample.

The further systematic removal of items which show misfit to the model, in the order 18, 21, 6 and 5 results in an eight item scale with a PSI of 0.8 and alpha of 0.8. Collapsing the categories 'unsure' and 'disagree', as for the STOE and inclusive of the same qualitative interpretation concern, removes all threshold disorder and results in the desirable distance between thresholds.

When comparing this study's results for gender DIF with other studies reporting on the STEBI regarding gender bias, the following was noted. In other reported studies gender analysis has been completed using the cumulative score for each construct, with the number of female participants being far greater (>75%) than that of their male counterparts. Research has shown there to be no statistically significant difference for the STOE construct (Norris, 2017; Bleicher, 2004; Cantrell et al., 2003; Enoch and Riggs, 1990). When analysing the PSTE construct, research has shown a statistically significant difference in favour of males (Bleicher, 2004; Cantrell et al., 2003; Enoch and Riggs, 1990). This analysis identified similar results to the previously conducted ANOVAs for these constructs, while further identifying items 18, 19 and 8 in the PSTE as the problematic items that are likely resulting in males scoring significantly higher on the PSTE scale than females.

5.3. Targeting the STEBI-B

It would be desirable to use cognitive interviewing (Willis, 2005) with prospective respondents in order to determine the effect of removing qualifiers for this sample; for example 'often' in item 1, 'When a student does better in science, it is *often* because the teacher exerted a little extra effort.' Removing some or all of the 'qualifiers' would theoretically make these items more difficult to endorse by making them more definitive. Shifting the targeting of these items is particularly important where the goal is to measure growth through pre and post intervention administration of the instrument. This change, as well as removing the 'unsure' option of the Likert scale, forces participants to be more decisive when responding to the items.

Previous studies have reported that pre-service teachers find the items on the STOE scale difficult to positively endorse; attributed to less experience in school settings (Norris, 2017; Cantrell et al., 2003; Knaggs and Sondergeld, 2015; Menon and Sadler, 2016). This was not the case for the post graduate sample. Therefore, past experience, rather than future prediction, may be one factor affecting the scale, and this could also be determined through cognitive interviewing.

In making the identified items more difficult to endorse, it is proposed that the histogram of persons will move to the left along the logits scale, bringing the mean of the persons closer to zero. Theoretically bumping items such as items 7,11 and 22, which are some of the most easily endorsed items, into the 'gap'. Whether this completely resolves the gap in item thresholds occurring along the scale or whether additional items to replace those removed are required, remains to be tested.

5.4. A modified STEBI-B

The data fit analysis of the STOE and PSTE scales suggested that multidimensionality could be addressed through fixing specific items. By applying the preceding suggested modifications to the STEBI-B, a parsimonious model exists with eight items in each scale, measured on a four-point Likert scale. The STOE scale retains items 1, 4, 7, 9, 11, 14, 15 and 16. The PSTE scale retains items 2, 3, 12, 17, 19, 20, 22 and 23. Both scales fit the Rasch model and have ordered response category thresholds with distances above 1.0 logits. The new numbering is represented in Figure 7, where items 1 through 8 represent the STOE and items 9 through 16 represent the PSTE. Highlighted in Figure 7 are items where qualifiers can be dropped if cognitive interviewing indicates that this would help to better target the items to a postgraduate, pre-service sample.

6. Conclusion

Due to the prevalence of the STEBI as a measure of teacher self-efficacy, it is desirable that the instrument can be adapted for use with a wide range of in-service and pre-service teachers. The Rasch model analysis conducted for this study begins to explore how the STEBI-B can be applied within a postgraduate pre-service teacher education cohort. The analysis has implications for the use of the STEBI-B

within this cohort, and echoes issues cited in other studies (Deehan, 2017).

Given the interest in reviewing pre-service teacher education to address growing international concerns about students' performance in science (Appleton, 2003; Hackling, 2014; Velthuis et al., 2014), it is essential that instruments such as the STEBI-B are valid and reliable for use to inform evidence-based decision making in initial teacher education. The measurement of self-efficacy in the primary science context is specifically important due to the high number of generalist teachers working in both Australia (77% of Year 4 teachers) and internationally (44% of Year 4 teachers of science) (Martin et al., 2016; Thomson, Wernert, O'Grady and Rodrigues, 2017). While we acknowledge that there are limitations to this research, such as small group sizes limiting DIF analysis, and a single university providing the sample, the findings of the Rasch analysis show the potential for the STEBI-B to be a reliable and useful tool for examining pre-service teacher efficacy within a post-graduate teacher education context; as it has been shown to be in other populations.

Declarations

Author contribution statement

Eileen V. Slater: Performed the experiments; Analyzed and interpreted the data; Wrote the paper.

Christina M. Norris: Conceived and designed the experiments; Performed the experiments; Wrote the paper.

Julia E. Morris: Performed the experiments; Wrote the paper.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability statement

The authors do not have permission to share data.

Declaration of interests statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

References

Andrich, D., 1978. A rating formulation for ordered response categories. Psychometrika 43, 561–573.

Andrich, D., 1982. An index of person separation in latent trait theory, the traditional kr.20 index, and the Guttman scale response pattern. Educ. Res. Perspect. 9 (1), 95–104.

Andrich, D., 2011. Rating scales and Rasch measurement. Expert Rev. Pharmacoecon. Outcomes Res. 11 (5), 571–585.

Andrich, D., Marais, I., 2012. A Course in Rasch Measurement Theory: Measuring in the Educational, Social and Health Sciences. Springer, Singapore.

Andrich, D., De Jong, J., Sheridan, B., 1997. Diagnostic opportunities with the Rasch model for ordered response categories. In: Rost, J., Langeheine, R. (Eds.), Applications of the Latent Trait and Latent Class Models in the Social Sciences. Waxman Verlag, Muster, Germany, pp. 58–68.

Andrich, D., Sheridan, B., Luo, G., 2014. RUMM2030 [Computer Software]. RUMM Laboratory, Perth, Western Australia, Australia.

Appleton, K., 2003. How do beginning primary school teachers cope with science? Toward an understanding of science teaching practice. Res. Sci. Educ. 33 (1), 1–25.

Bland, J.M., Altman, D.G., 1995. Multiple significance tests: the Bonferroni method. Br. Med. J. 310, 170.

Bleicher, R.E., 2004. Revisiting the STEBI-B: measuring self-efficacy in preservice elementary teachers. Sch. Sci. Math. 104 (8), 383–391.

Bond, T.G., Fox, C.M., 2015. Applying the Rasch Model: Fundamental Measurement in the Human Sciences, third ed. Taylor and Francis, Hoboken.

Cano, S.J., Barrett, L.E., Zajiceck, J.P., Hobart, J.C., 2011. Beyond the reach of traditional analyses: using Rasch to evaluate the DASH in people with multiple sclerosis. Multiple Sclerosis J. 17 (2), 214–222.

Cantrell, P., Young, S., Moore, A., 2003. Factors affecting science teaching efficacy of preservice elementary teachers. J. Sci. Teach. Educ. 14 (3), 177–192.

Curtis, D., 2001. Misfits: people and their problems. What might it all mean? Int. Educ. J. 2 (4), 91–99.

Deehan, J., 2017. The Science Teaching Efficacy Belief Instruments (STEBI A and B): A Comprehensive Review of Methods and Findings from 25 Years of Science Education Research (Springer briefs in education). Springer, Cham. https://link-springer-com.ezproxy.ecu.edu.au/book/10.1007%2F978-3-319-42465-1. (Accessed 18 February 2019).

Dembo, M.H., Gibson, S., 1985. Teachers' sense of efficacy: an important factor in school improvement. Elem. Sch. J. 86 (2), 173–184.

Enochs, L., Riggs, I., 1990. Further development of an elementary science teaching efficacy belief instrument: a preservice elementary scale. Sch. Sci. Math. 90w, 694–706.

Gibson, S., Dembo, M.H., 1984. Teacher efficacy: a construct validation. J. Educ. Psychol. 76 (4), 569–582.

Gunning, A., Mensah, F., 2011. Preservice elementary teachers' development of self-efficacy and confidence to teach science: a case study. J. Sci. Teach. Educ. 22 (2), 171–185.

Hackling, M.W., 2014. Challenges and opportunities for Australian science education. Prof. Educat. 13 (5), 4–7.

Hagquist, C., Bruce, M., Gustavsson, J.P., 2009. Using the Rasch model in nursing research: an introduction and illustrative example. Int. J. Nurs. Stud. 46 (3), 380–393.

Kazempour, M., Sadler, T.D., 2015. Pre-service teachers' beliefs, attitudes, and self-efficacy: a multi-case study. Teach. Educ. 26, 247–271.

Knaggs, C.M., Sondergeld, T.A., 2015. Science as a learner and as a teacher: measuring science self-efficacy of elementary preservice teachers. Sch. Sci. Math. 115 (3), 117–128.

Linacre, J.M., 1999. Investigating rating scale category utility. J. Outcome Meas. 3 (2), 103–122.

Linacre, J.M., 2018. Winsteps® (Version 4.3.1) [Computer Software]. Beaverton, Oregon: Winsteps.Com. Retrieved January 1, 2018. Available from. https://www.winsteps.com/.

Looveer, J., Mulligan, J., 2009. The efficacy of link items in the construction of a numeracy achievement scale—from kindergarten to year 6. J. Appl. Meas. 10 (3).

Mansfield, C.F., Woods-McConney, A., 2012. "I didn't always perceive myself as a science person": examining efficacy for primary science teaching. Aust. J. Teacher Edu. 37 (10), 37–52.

Marais & Andrich, 2008. Formalizing dimension and response violations of local independence in the unidimensional Rasch model. J. Appl. Meas. 9 (3), 200–215.

Martin, M.O., Mullis, I.V.S., Foy, P., Hooper, M., 2016. TIMMS 2015 International Results in Science. TIMMS & PIRLS International Study Centre, Boston, MA. Retrieved from. timss2015.org/download-center.

Masters, G.N., 1982. A Rasch model for partial credit scoring. Psychometrika 47, 149–174.

McKinnon, M., Lamberts, R., 2013. Influencing science teaching self-efficacy beliefs of primary school teachers: a longitudinal case study. Int. J. Sci. Educ., Part B 4 (2), 172–194.

Menon, D., Sadler, T.D., 2016. Preservice elementary teachers' science self-efficacy beliefs and science content knowledge. J. Sci. Teach. Educ.

Norris, C.M., 2017. Exploring the impact of postgraduate preservice primary science education on students' self-efficacy. Doctoral dissertation. Edith Cowan University. Research Online. https://ro.ecu.edu.au/theses/2040.

Norris, C.M., Morris, J.E., Lummis, G.W., 2018. Preservice teachers' self-efficacy to teach primary science based on 'science learner' typology. Int. J. Sci. Educ. 40 (18), 2292–2308. In press.

Nunnally, J.C., Bernstein, I.H., 1994. Psychometric Theory. McGraw-Hill, New York, NY.

Palmer, D., 2006. Sources of self-efficacy in a science methods course for primary teacher education students. Res. Sci. Educ. 36 (4), 337–353.

Rasch, G., 1961. Probabilistic Models for Some Intelligence and Attainment Tests (Reprinted 1980, Expanded Ed. With Forward and Afterward by B. D. Wright). The University of Chicago Press, Chicago.

Rumm Laboratory, 2018. Cronbach's Alpha and the Person Separation Index (PSI). Retrieved from. http://www.rummlab.com.au/rmrelidx2030.pdf.

Sick, J., 2011. March). Rasch measurement and factor Analysis *SHIKEN*. JALT Testing & Evaluation SIG Newsletter 15 (1), 15–17.

Smith, E.V., 2002. Detecting and evaluating the impact of multi-dimensionality using item fit statistics and principal component analysis of residuals. J. Appl. Meas. 3, 205–231.

Tennant, A., Conaghan, P.G., 2007. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis Rheum. 57, 1358–1362.

Thomson, S., Wernert, N., O'Grady, E., Rodrigues, S., 2017. TIMSS 2015: Reporting Australia's Results. Camberwell, Victoria. Retrieved from. www.acer.edu.au/timss.

van Sonderen, E., Sanderman, R., Coyne, J.C., 2013. Correction: ineffectiveness of reverse wording of questionnaire items: let's learn from cows in the rain. PloS One 8 (9).

Velthuis, C., Fisser, P., Pieters, J., 2014. Teacher training and pre-service primary teachers' self-efficacy for science teaching. J. Sci. Teach. Educ. 25, 445–464.

Willis, G.B., 2005. Cognitive Interviewing. A Tool for Improving Questionnaire Design. Sage Publications, Thousand Oaks, CA.