Construct validation of pre-service science teacher efficacy beliefs instrument (STEBI-B): Rasch analysis technique

Jamal H. Abu-ALRUZ

Department of Curriculum and Instruction, The Hashemite University, Zarqa 13115, JORDAN

E-mail: jamalruz@hu.edu.jo

Received 20 Mar., 2018
Revised 13 Dec., 2018

Contents

Abstract

The purpose of the present study is to validate an Arabic version of STEBI-B using Rasch model techniques. The validated Arabic version of the instrument (A-STEBI-B) was administered to a sample of (168) Jordanian prospective elementary school science teachers enrolled in the Hashemite University classroom teacher preparation program. Rasch analysis program, WINSTEPS 3.69 was used to calculate the infit and outfit MNSQ and ZSTD statistics. The values of infit and outfit MNSQ and ZSTD of the 23 items in both scales (10 items of OE and 13 items of PE) fits the model well. This means that each item contributes to mesurement of only one construct in both scales, (i.e. outcome expectancy and personal teaching efficacy). In conclusion, the arabic version of STEBI-B is valid and reliable measure for use in jordan.

Keywords: Construct validation, Pre-service teachers, Efficacy beliefs, and Rasch model.

Introduction and Theoretical Framework

There is a consensus that beliefs are part of a group of affective state constructs describing the structure and content of a person's thinking and providing an understanding of his/her actions and practice (Bryan & Atwater, 2002; Klassen, tze & Betts, 2011). Along with various measures of teacher knowledge and skills, one component of teacher preparation and professional development program evaluation is assessing teacher self-efficacy and K-12 science teacher ability (Enochs & Riggs, 1990; Enochs, Smith & Huinker, 2000).

Understanding teachers' beliefs, specifically science teachers' self-efficacy beliefs, is essential to improving their professional quality and science teaching practices (Enochs, Scharmann, & Riggs, 1995; Pajares, 1992; Enochs & Riggs, 1990). Bandura (1986) found that the primary source of information for self-efficacy beliefs is through mastery of experiences. Research revealed that prospective elementary science school teachers' self-efficacy beliefs affected by their prior experiences with science teaching in prospective teacher preparation programs at their universities.

Ramey-Gassert, Shroyer and Staver (1996) also found that prospective teacher preparation programs play an important role in helping prospective teachers construct a strong science teaching self-efficacy, feel confident to prepare and teach science, and to use effective instructional methods to foster effective students'

science learning in their science classrooms. Examining prospective teachers' science teaching self-efficacy beliefs is significant in developing the quality of their science teaching, improving their science background, and evaluating and improving science teaching preparation programs.

Research on prospective science teachers' teaching self-efficacy beliefs has been examined using the Science Teaching Efficacy Beliefs Instrument for Prospective teachers (STEBI-B), developed by Enochs and Riggs (1990). The STEBI-B measures two components of prospective teachers' science teaching self-efficacy beliefs; which are personal science teaching self-efficacy beliefs, and science teaching outcome expectancy. The major concern in assessing the quality of instruments in science education research is reliability and validity of such instruments. Most science education instruments lack theoretical measurement framework. Moreover, ordinal-level attitudinal data routinely analyzed as if these data has equal intervals, thereby violating requirements of parametric tests (Boone, Townsend & Staver, 2011).

Applying Item Response Theory (IRT) models, such as Rasch models, provide significant advantages for the development and evaluation of Likert-type items and instruments (Liu, 2006). Rasch analysis converts ordinal data into a ratio scaled data and produces item parameters and person parameters that are of a ratio level of measurement (Bond & Fox, 2001; Boone & Scantlebury, 2006; Neumann & Nehm, 2011). Rasch-based analyses are also able to test whether item/scale comparability exists for a given sample by testing whether all items are answered in the same fashion or not. This allows empirical testing of Likert-type scale assumptions. It also allows for comparisons of students and items on quantitatively equivalent intervals.

Moreover, many simple Rasch diagnostic tools are available to allow one to evaluate the functioning of a scale with regard to reliability and validity. These tools go far beyond the simple calculation of an alpha coefficient. Researchers in a range of disciplines throughout the world now appreciate the necessity of this step before parametric tests carried out (Bond & Fox, 2007; Wright & Masters, 1982; Wright & Stone, 1979). The Rasch model is used to evaluate large data sets such as that of TIMSS (Trends of International Mathematics and Science Study), and many small data sets in many fields (Schulman& Wolfe, 2000). When using sets of items on a survey to determine a respondent's overall attitude, it must be transformed to a linear scale, which the Rasch model allows one to carry out. If a linear measurement scale is not used, this may compromises the validity of all subsequent statistical tests.

Statement of the Problem

Prospective elementary school teachers' science teaching self-efficacy beliefs depends on successful mastery of experiences in teacher preparation programs at the university level and influences their behaviors in teaching science when they become school teachers. Ultimately, prospective teachers' science teaching efficacy beliefs hypothesized to affect school students' effective learning of science. There is a growing literature investigating prospective teachers' science teaching efficacy beliefs, assessed using the science teaching efficacy beliefs instrument for prospective teachers (STEBI-B) developed by Enochs and Riggs (1990). However, there is little research about this issue in Jordan. This study comes as an attempt to develop, validate an Arabic version of Science Teaching Efficacy Beliefs Instrument (STEBI-B) for prospective elementary school teachers, using Rasch test analysis techniques.

Research Questions

- Is the STEBI-B valid in measuring prospective elementary school teachers' science teaching efficacy beliefs in Jordan?
- Is the STEBI-B reliable in measuring prospective elementary school teachers' science teaching self-efficacy beliefs in Jordan?

Methodology

Sample of the Study

The translated Arabic version of the instrument (A-STEBI-B) that passed the process of content validity (via back translation check) were administered to a representative sample of (168) Jordanian prospective elementary school science teachers enrolled in the Hashemite University classroom teacher preparation program.

Research Instrument (STEBI-B) and the Five Criteria Questionnaire

The STEBI-B, which is the source of the Arabic version of science teacher efficacy belief instrument (A-STEBI-B) measures two dimensions of prospective teachers' self-efficacy beliefs about science teaching: personal science teaching efficacy (13 items), and science teaching outcome expectancy (10 items).

A five-criteria questionnaire was used in this study to support the construct validity of the STEBI-B. The criteria were chosen on the basis of the literature review, specifically, the five-criteria questionnaire used by Enochs and Riggs (1990) in their validation study of the STEBI-B. These criteria were: the number of college science courses taken, the number of high school science courses taken, prospective teachers'

choice of teaching science, prospective teachers' use of activity-based or inquiry based instruction, and prospective teachers' self-rating of their science teaching.

Research Procedures

First Development and Validation Study:

- Translation of the STEB I -B and the five criteria questionnaire into Arabic.
- Assessing content validity: evaluation of translation by translators and evaluation of the ability of each item in the A-STEBI-B, and the five criteria questionnaire by Jordanian professionals in science education.
- Revision of translation of the A-STEBI-B, and the five criteria questionnaire.
- Administration of the revised A-STEBI-B, and the five criteria questionnaire to a sample of Jordanian prospective elementary school teachers, Assessing construct validity by Confirmatory factor analysis using IBM SPSS and IBM AMOS Version 20.
- Screening the items in the A-STEBI based on factor analysis results.

Second Development and validation study:

- Revision of the A-STEBI and the five criteria questionnaire based on the results of factor analysis, reliability and item-scale score correlation of the first study.
- Administration of the refined and revised A-STEBI and the five criteria questionnaire to a different sample of Jordanian prospective elementary school teachers.
- Assessing construct validity (uni-dimensionality) using Rasch analysis techniques with WINSTEPS Version 3.69 program, each one of the two subscales of the instrument treated separately.
- Assessing reliability and construct validity using Rasch techniques.

Results

Confirmatory Factor Analysis of the STEBI-B Instrument

Exploratory factor analysis using principal component analysis of the revised STEBI-B scale suggests that (23) items of the instrument define two separate constructs (Bliecher, 2004; Enochs & Riggs, 1990; Morrell and Carroll, 2003). The construct validity of the two-factor model instrument was checked by confirmatory factor analysis using IBM SPSS Amos version 20. The fit statistics for the two-factor

solution show that the measurement model yields reasonable fit indices (CFI = 0.98; RMSEA = 0.047; SRMR = 0.04) between the item response and the proposed measurement model (i.e. the two subscales: outcome expectancy and science teaching efficacy).

Overall Model Fit Information, Separation and Mean Logit for Outcome Expectancy items:

Table 1. Overall model fit information, separation and mean Logit for Outcome Expectancy items

Summary of 168 Measured Person								
					Infit		**Out fit**	
	Raw Score	**Count**	**Measure**	**Model Error**	**MNSQ**	**ZSTD**	**MNSQ**	**ZSTD**
Mean	37.7	10	0.72	0.41	**1.02**	**-0.1**	**1.02**	**-0.1**
S.D.	4.10	0.0	0.73	0.07	0.68	1.3	0.71	1.3
Max.	49.0	10	4.12	1.06	4.63	4.3	5.36	4.8
Min.	26.0	10	-0.92	0.35	0.10	-3.3	0.09	-3.1
Real RMSE	0.47	Adj. SD	0.55	Separation	**1.17**	Person	Reliability	**0.58**
Model	0.42	Adj.SD	0.59	Separation	**1.41**	Person	Reliability	**0.67**
S.E. Of Person Mean= 0.06								
Summary of 10 Measured Items								
					Infit		Out fit	
	Raw Score	Count	Measure	Model Error	MNSQ	ZSTD	MNSQ	ZSTD
Mean	633.0	168	0.00	0.10	1.00	0.1	1.02	0.2
S.D.	95.9	0.0	0. 54	0.01	0.12	1.0	0.15	1.2
Max.	49.0	10	4.12	1.06	4.63	4.3	5.36	4.8
Min.	26.0	10	-0.92	0.35	0.10	-3.3	0.09	-3.1
Real RMSE	0.10	Adj. SD	0.53	Separation	5.10	Item	Reliability	0.96
Model RMSE 0.10			Adj.SD 0.53				Separation 5.20	
Item Reliability 0.96								
S.E. Of items Mean= 0.18								

The separation index is an index of the spread of the person positions or item positions. For persons, the separation index is 1.17 for the data (real separation index) and the model separation is 1.41. The person reliability is 0.58 for the data and is 0.67 for the model, which is a moderate value, this value is expected due to person homogeneity in terms of the number of science method and science courses, gender, number of school science courses and self rating of their science teaching. For item

separation, it is 5.1 for the data and 5.2 for the model. The item reliability is 0.96, which is high value that supports the construct validity of the scale.

Table 2. The Infit and Outfit statistics and item measure for outcome expectancy items

Item number	measure	Model S.E.	Infit		Outfit		PT Corr.
			MNSQ	ZSTD	MNSQ	ZSTD	
1	-0.11	0.13	0.91	-0.6	0.90	-0.7	0.37
4	-0.36	0.11	0 .90	-0.7	0.85	-1.1	0.48
7	-0.52	0.10	1.12	1.0	1.24	1.8	0.33
9	-0.50	0.12	0.99	0.0	1.12	0.9	0.36
10	1.45	0.09	1.18	1.6	1.20	1.7	0.47
11	-0.03	0.09	0 .91	-0.8	0.91	-0.8	0.55
13	0.36	0.08	1.18	1.9	1.18	1.7	0.40
14	-0.11	0.09	0 .96	-0.4	0.93	-0.7	0.52
15	-0.02	0.11	0 .83	-1.4	0.81	-1.6	0.55
16	-0.16	0.10	1.02	0.20	1.07	0.6	0.40

The infit and outfit mean square values (MNSQ) for the ten outcome expectancy scale items are ranged between 0.81 and 1.24, and the standardized fit values between -1.6 to = 1.9. Although there is no rule of thumb for the acceptable values for infit and outfit statistics, some considerations were suggested by researchers (Bond & Fox, 2007; Frantom ,Green& Lam, 2002): mean square values of infit and outfit between 0.5 and 1.5, or 0.6-1.4, and 0.8-1.2; mean square values less than 1.3 for samples less than 500, 1.2 for samples 500-1000, and 1.1 for samples greater than 1000. Standardized (ZSTD) infit and outfit between -2 and +2, between -3 and +2, and less than +2. Consequently, all items of the outcome expectancy scale are within the acceptable range of fit statistics.

This result suggests the unidimensionality of the OE scale, which is a basic assumption of Rasch model. The point measure correlation for the OE scale items range between 0.33 and 0.55. This result revealed that each item contributed to define a common construct.

Moreover, Boone; Townsend and Staver (2011) suggested that instead of presenting both SE and OE items in the STEBI instrument to respondents, one is best served by presenting one set of items first (e.g., SE) and then a second set of items (OE). This would facilitate the possibility of presenting a SE rating scale that might be different from the OE rating scale, and presented in a manner that would not confuse.

Overall model fit information, separation and mean Logit for Personal science teaching efficacy items:

Table 3. The Infit and Outfit statistics and item measure for Personal science teaching efficacy items

Summary of 168 Measured Person								
					Infit		**Out fit**	
	Raw Score	**Count**	**Measure**	**Model Error**	**MNSQ**	**ZSTD**	**MNSQ**	**ZSTD**
Mean	46.5	13.0	0.62	0.36	1.02	-0.2	1.00	-0.2
S.D.	5.9	0.00	0.73	0.04	0.77	1.5	0.76	1.5
Max.	59.0	13.0	2.70	0.50	5.58	5.9	5.20	5.5
Min.	28.0	13.0	-1.33	0.31	0.11	-3.3	0.13	-3.2
Real RMSE	0.41	Adj. SD	0.60	Separation	1.45	Person	Reliability	0.68
Model	RMSE	0.36	Adj.SD	0.63	Separation	1.73	Reliability	0.75
S.E. Of Person Mean= 0.06								
Summary of 13 Measured Items								
					Infit		**Out fit**	
	Raw Score	**Count**	**Measure**	**Model Error**	**MNSQ**	**ZSTD**	**MNSQ**	**ZSTD**
Mean	600.8	168	0.00	0.10	1.00	0.0	1.00	0.1
S.D.	60.7	0.0	0. 40	0.01	0.15	1.3	0.16	1.4
Max.	730.0	168.0	0.79	0.13	1.25	2.3	1.24	2.1
Min.	462.0	168.0	-0.61	0.08	0.71	-2.6	0.70	-2.7
Real RMSE	0.10	Adj. SD	0.53	Separation	5.10	Item	Reliability	0.96
S.E. Of items Mean =0.12								

For persons, the separation index is 1.45 for the data(real separation index) and the model separation is 1.73. The person reliability is 0.68 for the data and is 0.75 for the model, which is a moderate value; this value is expected due to person homogenity, in terms of the number of science method and college science courses, gender, number of school science courses and self rating of their science teaching. For item separation, it is 3.69 for the data and 3.79 for the model. The item reliability is 0.93, which is high value that supports the construct validity of the scale.

Table 4. The Infit and Outfit statistics and item measure for personal efficacy items

Item number	measure	Model S.E.	Infit		Outfit		PT Corr.
			MNSQ	ZSTD	MNSQ	ZSTD	
2	-0.31	0.11	1.20	1.6	1.20	1.7	0.30
3	0.21	0.09	1.08	0.8	1.11	1.0	0.46
5	-0.19	0.11	1.10	1.0	1.11	0.9	0.38
6	0.00	0.11	1.04	0.4	1.04	0.4	0.53
8	-0.28	0.11	0.71	**-2.6**	0.70	**-2.7**	0.71
12	-0.28	0.11	1.00	0.0	1.00	0.1	0.47
17	0.58	0.09	0.86	-1.4	0.91	-0.9	0.61
18	-0.48	0.13	0.85	-1.2	0.87	-1.1	0.57
19	0.79	0.09	1.25	2.3	1.24	2.1	0.33
20	0.36	0.08	1.01	0.2	1.21	1.8	0.54
21	0.01	0.09	1.04	0.4	1.04	0.4	0.53
22	-0.61	0.11	0.93	-0.5	0.85	-1.2	0.51
23	0.22	0.09	0.87	-1.2	0.86	-1.2	0.60

The infit and out fit mean square values for the 13 personal efficacy scale items, as shown in table (4), are ranged between 0.71 and 1.25, and standardized fit values between -2.7 to 2.1. All items of the personal efficacy scale; except item 8 which has unacceptable in and outfit ZSTD, are within the acceptable range of fit statistics. This result revealed the unidimensionality of the PE scale, which is one of the basic assumptions of rasch model. Moreover; the point measure correlation for the PE scale items range between 0.30 and 0.71, this result suggests that each of the items in the personal efficacy scale contributed to define a common construct.

Discussion and Recommendations

The construct validity analysis of the Arabic version of STEBI-B instrument using Rasch analysis techniques revealed that the instrument is adequately valid and reliable. The item quality measures revealed that all items of both scales: OE and PE, fits with model expectations based on both MNSQ and ZSTD values. The person-item for both scales demonstrates that most persons tend to select options at or above the middle of the Likert scale. This may be due to language misinterpretation and differences between Arabic and English language, and the homogeneity of the study sample, in terms of some variables. For instance most of them were female students at the second year level in a four years preparation program, they completed successfully about 2-3 college courses in school science and methods of teaching science courses, with almost the same high school science courses etc.

Moreover, the results of response scale categories revealed that the sample responses of three items of the PE scale and seven items of the PE scale fails to increase by category value. This failure of the item-responses to map the underlying trait in an ordered fashion is problematic. Purski, Blanco, Riggs, Grimes, Fordtran, Barbola, Cornell & Lichtenst (2013) argues that the 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 negatives (e.g., not) and interpret the item as stated in the affirmative style.

(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 vice versa.

5. There is debate about 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 in-service teachers, being in the field, they should either indicate that they "can" or "cannot" do. Offering "undecided/uncertain" to some items in this case, may be, prompted respondents to decline indicating a lack of skill or commitment. In addition to that, there was another 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?"

In a review comparing efficacy research from 1986-1997 with that done from 1998 to 2009, Klassen, Tze, Betts & Gordon (2011, p 39–40) suggested four key areas for future directions in efficacy research; there is a need to: (a) conduct qualitative studies to determine the sources of teacher efficacy—how they "form, develop, and

change over time"—these have yet to be fully researched and may vary over the career span and across cultures; (b) offer valid measurements—there is a prevalence of invalid or ill-reported measurements in the research literature; (c) connect self-efficacy of science teachers to student outcomes; and (d) determine how teacher self-efficacy enhanced (e.g., through Teacher Professional Development, teacher researcher collaborations).

In conclusion, the Arabic version of the STEBI-B instrument is a valid and reliable measure, but some improvement and revisions is needed to improve the quality of instruments' items, such as using a positive statements, and using different rating scale for each one of the two scales of the STEBI-I. Lin and Gorrell (2001) argued that the concept of teacher efficacy might be culturally oriented, thus there is a need to examine the translated items carefully when applied in different cultures. So, language editing of the instruments' item statements of the Arabic version is highly needed, in order to make it consistent with the characteristics of the Arabic language and reflects the cultural differences.

References

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

Bleicher, R.E. (2004). Revisiting the STEBI-B: Measuring Self Efficacy in preservice elementary teachers. *School Science and Mathematics, 104*, 383-391.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model. Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Boone,W., Townsend, J.S. & Staver, J. (2011) . 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.

Boone, W. & Scantlebury, K. (2006). The Role of Rasch analysis when conducting science education research utilizing multiple-choice tests. *Science Education, 90*, 253-269.

Bryan, L., & Atwater, M. (2002), Teacher beliefs and cultural models: A challenge for science teacher preparation programs. *Science Education, 86*, 821-839.

Frantom, C. G., Green, K. E., & Lam, T. C. M. (2002). Item grouping effects on invariance of attitude items. *Journal of Applied Measurement, 3*, 38-49.

Enochs, L., & Riggs, I. (1990). Further development of an elementary science teaching efficacy belief instrument: A pre-service elementary scale. *School Science and Mathematics, 90*, 694-706.

Enochs, L. G., Scharmann, L. C. & Riggs, I. M. (1995). The relationship of pupil control to preservice elementary science teacher self-efficacy and outcome expectancy. *Science Education, 79*(1), 63-75.

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.

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*, 2143.

Lin, H., & Gorrell, J. (2001). Exploratory analysis of pre-service teacher efficacy in Taiwan. *Teaching and Teacher education, 17*(5), 623-635.

Liu, X., & Boone, W. J. (Eds.). (2006). *Applications of Rasch measurement in science education.* Maple Grove, MN: JAM Press.

Linacre, J. M. (2005). *WINSTEPS* (Version 3. 57 .2) [Computer software]. Chicago: Winsteps.

Morrell, P., & Carroll, J. (2003). An Extended examination of preservice elementary teachers' science teaching self-efficacy. *School Science & Mathematics, 103*(5), 246-251.

Neumann, I., Neumann, K., & Nehm, R. (2011). Evaluating instrument quality in science education: Rasch-based analyses of a nature of science test. *International Journal of Science Education, 33*(10), 1373-1405.

Pajares, M. F. (1992). Teachers' beliefs and educational research: cleaning up a messy construct. *Review of Educational Research, 62*, (3), 307-332.

Purski, L., Blanco, S., Riggs, R., Grimes, K., Fordtran, C. Barbola, G., Cornell, J., & Lichtenstein, M. (2013). Construct validation of the self-Efficacy teaching and knowledge instrument for the science teachers revised (SETAKIST-R): lesson learned. *Journal of Science Teacher Education, 24*(7), 1133-1156. DOI: 10.1007/s10972-013-9351-2.

Ramey-Gassert, L., Shroyer, M., & Staver, R. (1996).A Qualitative study of factors influencing science teaching self-efficacy of elementary level teachers. *Science education, 83*(8), 283-315.

Schulman, J.A., & Wolfe, E.W. (2000).Development of a nutrition self-efficacy scale for prospective physicians. *Journal of applied measurement, 1*(2),107-130.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.