cogent
education

*Corresponding author: Ghaleb H Alnahdi, Special Education, Prince Sattam Bin Abdulaziz University, Alkharj, Saudi Arabia
E-mail: ghalnahdi@gmail.com; g.alnahdi@psau.edu.sa

EDUCATIONAL ASSESSMENT & EVALUATION | RESEARCH ARTICLE

The Arabic version of the teacher efficacy for inclusive practices (TEIP-AR) scale: A construct validity study

Ghaleb H Alnahdi*

Abstract: This study aimed to examine the psychometric characteristics of the Arabic version of the Teacher Efficacy for Inclusive Practices (TEIP) scale. Data were collected from 432 participants in Saudi Arabia—185 in-service and 247 pre-service teachers. A statistical analysis was conducted using the following methods: Confirmatory Factor Analysis (CFA), Exploratory Factor Analysis (EFA), misfit analysis via Rasch modelling, and reliability Cronbach's alpha coefficients. Good internal consistency coefficients were obtained for the TEIP scale and each of its three subscales (>.8). Acceptable fit indices were obtained from the confirmatory factor analysis (CFA) for a scale with 18 items in three subscales: inclusive instructions, collaboration, and managing behaviour. In conclusion, the Arabic version of the TEIP scale is valid with Arabic samples and preserves the psychometric properties and the structure of the original scale in order to measure teachers' self-efficacy working in inclusive classrooms. Recommendations for rephrasing some items are also discussed.

Subjects: Test Development, Validity & Scaling Methods; Inclusion and Special Educational Needs; Disability Studies

Keywords: CFA; self-efficacy; Arabic; Saudi Arabia; inclusive education; TEIP; validation

1. Introduction

The international trend is now to include all students in the regular classroom and provide the necessary support for those who need additional services (UNESCO, 2009).This necessitates ensuring that teachers are ready to work with students of different capacities and needs. Saudi Arabia is

ABOUT THE AUTHOR

Ghaleb H Alnahdi is an associate professor in special education at Prince Sattam University. He earned his Ph.D. in Special Education from Ohio University. He also holds a master's degree in research and evaluation from Ohio University (2012) and another, in special education, from King Saud University (2007). His research focuses on intellectual disability, inclusive education, cross-cultural validation of scales, and teacher preparation. He is involved in several research projects with different research groups at the national and international levels.

PUBLIC INTEREST STATEMENT

Inclusion practices are becoming increasingly widespread throughout the world. With this change, expectations and attitudes from teachers have started to change as well. One of these new perspectives is regarding the ability to work in inclusive education. This study examines the suitability of the Arabic version of one of the most common questionnaires used to measure teachers' beliefs on their readiness to work with students with different abilities. The information from this study will help to confirm whether the Teacher Efficacy for Inclusive Practices (TEIP) scale is appropriate for use with teachers in the Arab world.

cogent · oa

one country that has begun in recent years to pay attention to the inclusion of students with disabilities in public schools. Since 2001, the movement was started to include students with disabilities in special classes in regular schools (Alnahdi, Saloviita, & Elhadi, 2019). In addition, students with learning disabilities are included in regular classrooms. According to Alhano (2006) that 80% of students with disabilities received their special educational services at regular schools. To support this ongoing process, it is important to have a measure of Saudi teachers' self-efficacy to work with students with disabilities in the classroom.

Teachers' self-efficacy for inclusive education refers to their degree of confidence in their skills and ability to work in inclusive education. According to Bandura (1997), self-efficacy is people's beliefs in their "capabilities to organize and execute the courses of action required to produce given attainments" (p. 3). Many studies have found that high teacher self-efficacy is related to student achievement (Denzine, Cooney, & McKenzie, 2005; Tschannen-Moran & Hoy, 2007; Zee & Koomen, 2016). In addition, self-efficacy is related to teachers' amount of effort needed to accomplish a task (Zee & Koomen, 2016).

Measures, in general, need to be verified and tested by multiple studies. The lack of such measures is especially dire in the Arab countries, where no reliable, tested Arabic-language scale to measure self-efficacy to work in inclusive classrooms in Arabic exists within the researcher's knowledge. One of the most common and used scales internationally to measure teacher self-efficacy is the Teacher Efficacy for Inclusive Practices (TEIP) scale (Sharma, Loreman, & Forlin, 2012), as it has been tested in different countries and regions—Canada, Australia, Hong Kong, India (Sharma et al., 2012), Japan, Finland (Yada & Savolainen, 2017), South Africa (Savolainen, Engelbrecht, Nel, & Malinen, 2012), China (Malinen, Savolainen, & Xu, 2012), the USA (Park, Dimitrov, Das, & Gichuru, 2016), and Bangladesh (Ahsan, Sharma, & Deppeler, 2012)—and showed acceptable psychometric characteristics. This is the first study with an Arabic sample and using an Arabic version of the scale, to the researcher's knowledge.

The TEIP scale contains 18 items within three domains: efficacy in managing behaviour, efficacy in inclusive instruction, and efficacy in collaboration. It is a 6-point Likert-type scale, on which a high score indicates a "high sense of perceived teaching efficacy for teaching in inclusive classrooms" (Sharma et al., 2012, p. 15).

This study aimed to examine the psychometric characteristics of the Arabic version of the TEIP scale using a sample from Saudi Arabia. The TEIP-AR is designed to measure in-service and pre-service teachers' self-efficacy to work in inclusive classrooms (Loreman, Sharma, & Forlin, 2013; Sharma et al., 2012), that is, classrooms containing students with different abilities and needs.

2. Methods

2.1. Sample
Questionnaires were distributed to two groups: pre-service teachers in a college of education at a Saudi public university, and teachers in schools in Riyadh. In all, 432 participants voluntarily answered the questionnaire, on paper; the researcher then transferred these records to electronic form. The sample consisted of 55% male participants and 45% female participants (see Table 1). Around 43% were in-service teachers, and the rest were pre-service teachers. Sixty per cent of them were majoring or had majored in special education.

2.2. Translation
First, the questionnaire was translated from English into Arabic by two bilingual researchers fluent in both languages. Then, the translated Arabic version was sent to another researcher who also specialised in English language to back-translate the Arabic version into English, and the back-translation was compared to the original scale. On the basis of the comparison, some minor changes were made to the Arabic version to ensure that the meaning of the original items was

cogent ·· education

Table 1. Participants' demographic information		Frequency	Percent
Job	*Teacher*	*185*	*43%*
	Pre-service teacher	*247*	*57%*
Major	*Special education*	*260*	*60%*
	Other major	*172*	*40%*
Gender	*Female*	*193*	*45%*
	Male	*239*	*55%*

preserved. Next, pilot testing (n = 30) was conducted before proceeding to distribute the questionnaire to the rest of the sample. In short, the recommended procedures for cross-cultural adaptation of self-report measures (Beaton, Bombardier, Guillemin, & Ferraz, 2000; Brislin, 1970) were implemented in the translation process.

3. Results

3.1. Reliability
The reliability of the TEIP scale and the three subscales were examined: Cronbach's alpha coefficients were .928 for the scale as a whole and for the EMB subscale, .825; EII subscale, .824, and EC subscale, .822. All of these values indicate good internal consistency (George & Mallery, 2003). See Table 2.

The results section is divided into three sections, corresponding to the different analyses. First, Rasch analysis was utilised to specify misfitting items and persons, in order to examine whether the models would better fit the data without them. Second, exploratory factor analysis (EFA) statistics were used to examine items loadings on the proposed factors. Third, confirmatory factor analysis (CFA) statistics were applied, with five different models.

3.2. Rasch analysis
In addition to common statistical analyses based on the classical test theory (CTT) approach, Rasch analysis was performed based on the item response theory approach (IRT). 'The ability of a scale to provide fundamental measurement should be established prior to the more commonly reported psychometric attributes. Rasch analysis offers a method of ensuring that key measurement assumptions are tested' (Tennant, McKenna, & Hagell, 2004, p. 24). Rasch analysis was conducted using R software with the eRm Package for Rasch modelling. Three different analyses were conducted in this Rasch analysis. First, items and person fitting analysis. second, the data and the proposed model were tested for fit with Erling Andersen's test. Third, item characteristics analysis was conducted for difficulty and information (explained later).

Since 'misfitting item and person performances degrade measures' (Bond & Fox, 2015, p. 365), item and person fit analyses were conducted using the eRm package in R. In the item fit analysis,

Table 2. Reliability statistics (Cronbach's alpha coefficient) (N = 432)	
	Cronbach's Alpha Coefficient
TEIP (18 items)	.928
Efficacy in managing behaviour (items 1, 2, 7, 8, 11, 17)	.825
Efficacy in inclusive instruction (items 5, 6, 10, 14, 15, 18)	.824
Efficacy in collaboration (items 3, 4, 9, 12, 13, 16)	.822

two items were misfitting based on the p-value (<.01) of chi-squared (items 1, 17). This was also confirmed by standardised outfit statistics for items 1 and 17, which were 2.89 and 2.3, respectively, and standardised infit statistics for those items, which were 2.70 and 2.49, respectively, which is larger than the threshold of 1.7 (Wright, Linacre, Gustafson, & Martin-Lof, 1994).

The person fit test indicated that 12 persons might be good to remove, as they had statistics larger than 1.5 in any of the four fit indices, that is, standardised/unstandardised infit and outfit statistics. Twelve persons were removed for further analysis (sample 2). In addition, 'Measurement invariance evaluation is an important aspect of test development' (Brown, 2014, p. 4), to ensure that items do not function differently with different subgroups. A global model test, Andersen's likelihood ratio test for goodness-of-fit with mean split criterion, is recommended (Futschek, 2014). The results were as follows: LR = 23.367 (df = 17, p = .138); they indicate that the null hypothesis that item parameters estimate will be similar across subgroups, that is, that no item bias is detected.

One important analysis that can be done based on IRT, such as by Rasch analysis, is to compute the item information function (Magis, 2013). This mathematical function helps evaluate how informative each item is for different levels of abilities—self-efficacy in this study (Magis, 2013; Zieba, 2013). 'Very easy items are usually more informative at low ability levels whereas highly difficult and discriminating items are more informative for larger ability levels' (Magis, 2013, p. 306). By reviewing the item information charts, we notice that the items would be suitable for people with very low self-efficacy because the highest point of the chart—where the item becomes more informative—was with low-ability persons for all items (see Appendix 1). Ability in the charts refers to participants' overall score on self-efficacy to work in inclusive education.

In sum, Rasch analysis on items showed that the item information function was high with people with low levels of self-efficacy; participants more familiar with students with disabilities might need items with stronger statements to yield more information for measuring their self-efficacy.

In addition, we review the difficulty parameters for the items—called 'item difficulty' because mostly used with achievement tests (Fox & Jones, 1998), but more accurately representing how difficult it is to endorse a certain item in a self-report measure (Fox & Jones, 1998). Rasch analysis of items showed that item 2 scored the highest on the difficulty parameters, which means that item 2 needs a person to have more self-efficacy to agree with it than other items. In contrast, item 17 was the lowest on the difficulty parameter, meaning even people with very low self-efficacy (as well as higher) could agree with it (see Figure 1).

3.3. Exploratory factor analysis (EFA)

EFA was applied using one of the recommended methods of factor extraction, principal factors (PF) (Brown, 2014). PF is robust to issues with normality assumptions, as it has no distributional assumptions (Brown, 2014). EFA using the principal factors (PF) extraction method was conducted, and attained three factors; for the EII factor, three items (14, 15, and 18) loaded as expected, while only items 5 and 6 loaded on the EMB. Item 10 loaded on the EC (similar to the Chinese version by Malinen et al., 2012). Item 3 loaded on the EMB. As for the first-factor EMB, all six items loaded on this factor as well as some items for other factors. In sum, items 3, 5, 6, and 10 did not function as expected for the three-factor scale.

However, two factors should be attained based on the Kaiser–Guttman rule for factors with eigenvalues >1 and the scree plot of eigenvalues (Brown, 2014). In the first unrotated solution, all items were loaded on the first factor with no less than .53. And the first factor alone explained 45% of the variance. For the rotated solution, first items in factor EMB loaded on the first scale with

Figure 1. Item-person chart.
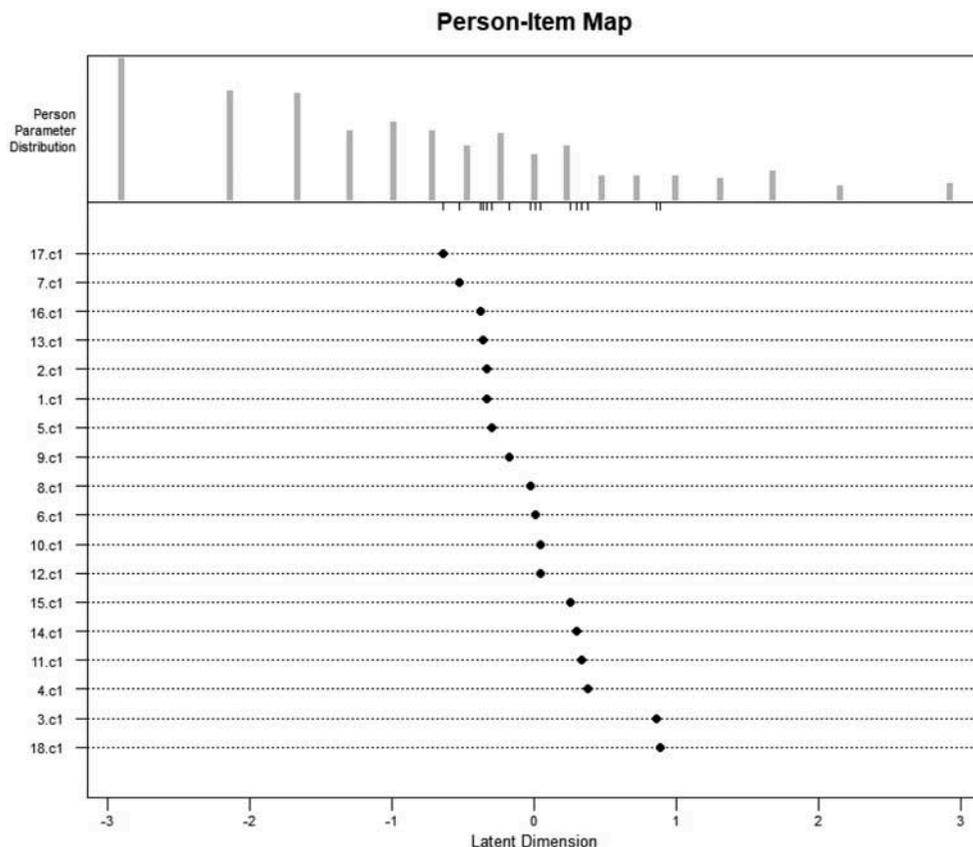
no less than .40, where this value of loading could be considered salient (Park et al., 2016); also, items in factor EC loaded on the second scale with no less than .365.

3.4. Confirmatory factor analysis (CFA)

In this study, we reported results for some often-used fit indices: normed fit index (NFI), the Tucker–Lewis coefficient (TLI), the comparative fit index (CFI), the goodness-of-fit index (GFI), and the adjusted goodness-of-fit index (AGFI) (Schermelleh-Engel, Moosbrugger, & Müller, 2003). RMSEA and SRMR were also calculated, as recommended by Hu and Bentler (1998). See Table 3.

First, CFA was conducted with the hypothesised model of three subscales with six items in each factor. The first model was the three-factor model with data from 432 participants. Fit indices were reasonably good (see Table 3); however, Amos software suggested a few modifications to improve the fit of the model. Therefore, in model 2, we covaried four item errors to another four item errors within the same factor only (see Figure 2). For example, in the EC factor, the software suggested to covary item 3 and 4 errors. Item 3 is *I can make parents feel comfortable coming to school* and item 4 is *I can assist families in helping their children do well in school*. That is, both items relate to similar abilities for dealing with families of students; therefore, it is logical to covary them. In the EMB factor, the software suggests to covary item 12 and 13 errors; item 12 is *I can collaborate with other professionals (e.g. itinerant teachers or speech pathologists) in designing educational plans for students with disabilities* and item 13 is *I am able to work jointly with other professionals and staff (e.g. aides, other teachers) to teach students with disabilities in the classroom*. In the EII factor, the software suggests to covary item 5 with item 6 errors, and item 14 with item 18 errors; item 5 is *I can accurately gauge student comprehension of what I have taught* and item 6 is *I can provide appropriate challenges for very capable students*; item 14 is *I am confident in my ability to get*

cogent ·· education

Model	SBS-χ^2	p	df	RMSEA	CFI	SRMR	GFI	AGFI	TLI	IFI	NFI
M1	451.227	.000	132	.075	.910	.0466	.889	.856	.895	.910	.878
M2	365.759	.000	128	.066	.933	.0424	.910	.880	.920	.933	.900
M3	308.792	.000	97	.071	.935	.0430	.914	.880	.919	.935	.908
M4	349.817	.000	128	.065	.921	.0471	.910	.880	.904	.921	.881
M5	299.209	.000	97	.071	.921	.0484	.913	.879	.902	.921	.888

Table 3. Goodness-of-fit indices for three-factor models of the TEIP scale

SBS-χ^2 = Satorra–Bentler scaled chi-squared; df = degrees of freedom; RMSEA = root-mean-square error of approximation; CFI = comparative fit index; SRMR = standardised root-mean-square residual; GFI = goodness-of-fit index; AGFI = adjusted goodness-of-fit index; TLI = Tucker–Lewis coefficient; NFI = normed fit index.
M1 = No control for error, M2 = Covariance for error (12–13, 14–18, 5–6, 3–4), M3 (n = 432, 16 items), M4 (n = 411, 18 items), M5 (n = 411, 16 items)

students to work together in pairs or in small groups and item 18 is *I can provide an alternate explanation or example when students are confused*.

In the third run of CFA (M3), two items were removed (items 1 and 17) based on the Rasch analysis for misfitting items with the whole sample of 432 participants. It shows a slight improvement in comparison to M1, with 18 items, but not compared to M2, where we covaried errors in two sets of four items (12–13, 14–18, 5–6, 3–4) within the same subscale. In the fourth run of CFA (M4), 21 participants were removed who showed misfit via Rasch analysis for misfitting persons and also through box plot screening. There was no improvement in the fit indices when the misfitting persons were dropped, which might mean their effect was limited. In the fifth run of CFA (M5), two items were removed (items 1 and 17) based on the Rasch analysis for misfitting items, and 21 participants were removed that showed misfit via Rasch analysis for misfitting persons and also through box plots screening. This model showed no noticeable improvement in the fit indices in comparison with M2, which contained all 18 items and all 432 participants.

In all models, the RMSEA value was below .5, which can be considered to indicate good fit (Hu & Bentler, 1999). TLI and NFI, with .9, also indicate an acceptable fit (Bentler & Bonett, 1980). In sum, the models show reasonable fit, except for chi-squared, which is o sensitive to large sample sizes even with reasonably fit data (Byrne, 2010). However, fitting indices did not improve significantly by removing misfitting items and persons (12 persons) suggested via Rasch analysis nor by removing nine persons after screening box plot. To conclude, the several steps of analysis support the 18-item scale with a three-factor structure, which was preserved in the Arabic version.

Model 2 shows slightly better fit indices than the other models and includes all 18 scale items and all 432 participants. Table 4 shows that all items have statistically significant loading to the latent factor at p < .001. The latent factor here is self-efficacy to work in inclusive education.

4. Discussion & recommendations

Several different types of analysis were employed in this study, were based on two different theories, CTT and IRT. Both theories are concerned with measurement 'quality'; one takes a holistic approach and the other starts with individual items and moves to the whole scale. CFA analysis was conducted five times. Of the three times using 18 items (all items), one analysis included only 411 participants after removing those who misfit (outliers) in the Rasch analysis and also through box plot screening. In addition, two CFA analyses were conducted with 16 items, after removing two items (1, 17) that showed misfit through Rasch analysis for misfitting items. However, these adjustments did not yield any significant improvement in the fit indices of the scale in general in comparison with M2, which included all 18 items and 432 participants. This result thus supports the original proposed structure of the scale (Sharma et al., 2012), and other studies validate the three-factor structure (Park et al., 2016).

cogent · education

Figure 2. CFA for all items with all participants (N = 432).
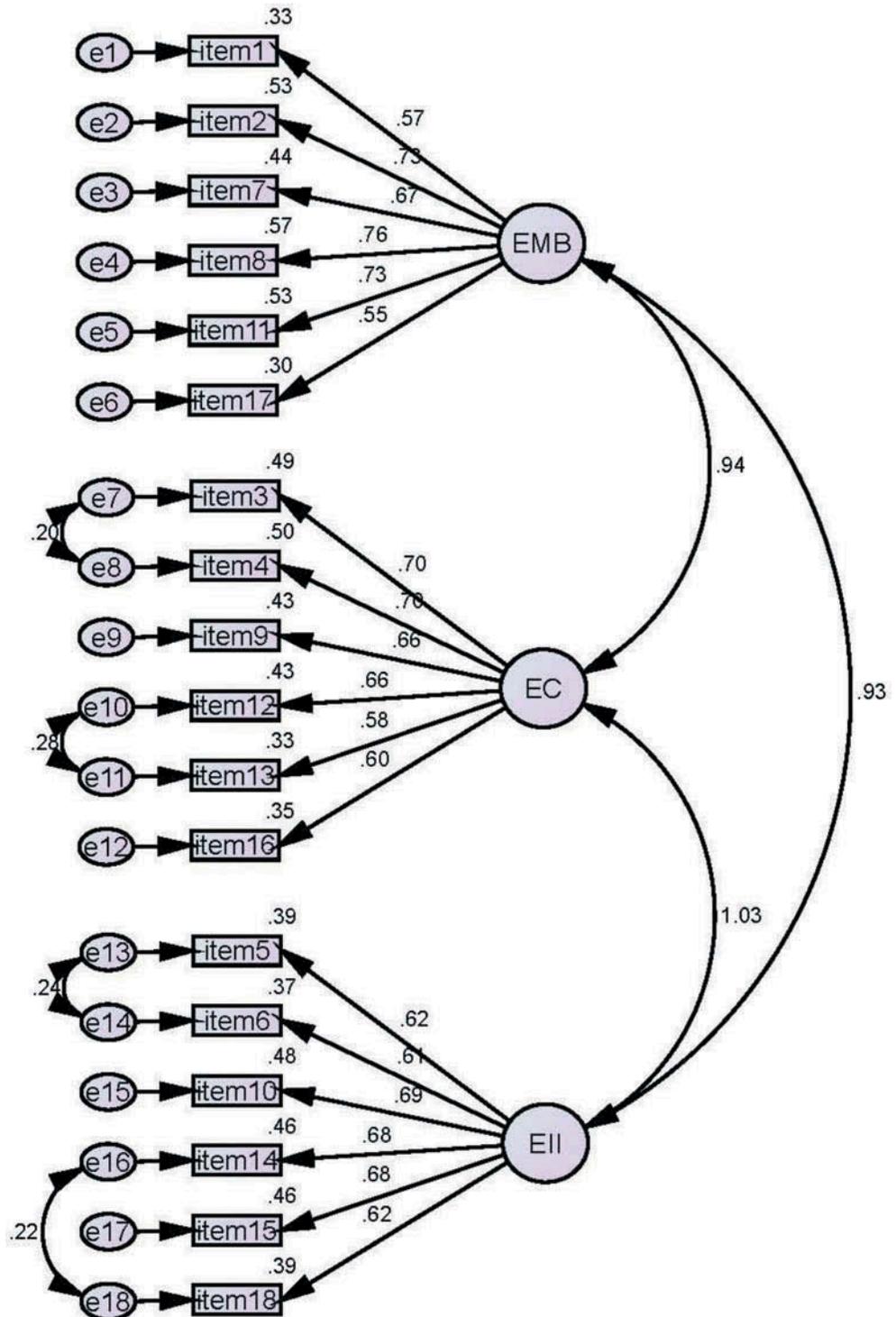
In the future replication of this study in Arabic contexts, we would recommend modifications to improve some items, including testing them through a pilot study. There were two items (5 and 6) loaded on the EMB factor because they were perceived as related to managing behaviour; therefore, it will help to rephrase these items to more clearly connect them to inclusive instruction and not behaviour, since the domain focus on instruction. For example, item 6, *I can provide*

Table 4. Parameter estimates for the CFA (N = 432), model 2

Estimate	LV	Standardised Loading	Unstandardised Loading	SE	CR	P
I am confident when dealing with students who are physically aggressive	EMB	0.55	1.00*	-	-	-
I am able to get children to follow classroom rules		0.73	1.19	0.11	10.96	<.001
I can control disruptive behavior in the classroom		0.76	1.30	0.12	11.20	<.001
I am confident in my ability to prevent disruptive behavior in the classroom before it occurs		0.67	1.19	0.12	10.36	<.001
I am able to calm a student who is disruptive or noisy		0.73	1.26	0.12	10.91	<.001
I can make my expectations clear about student behavior		0.57	1.03	0.11	9.32	<.001
I am able to provide an alternative explanation or example when students are confused	EII	0.62	1.00*	-	-	-
I can use a variety of assessment strategies (e.g. portfolio assessment, modified tests, performance-based assessment)		0.68	1.16	0.10	11.99	<.001
I am confident in my ability to get students to work together in pairs or in small groups		0.68	1.16	0.09	13.57	<.001
I am confident in designing learning tasks so that the individual needs of students with disabilities are accommodated		0.69	1.21	0.10	12.17	<.001
I can provide appropriate challenges for very capable students		0.61	1.08	0.10	10.91	<.001
I can accurately gauge student comprehension of what I have taught		0.62	1.11	0.10	11.15	<.001
I can make parents feel comfortable coming to school	EC	0.7	1.00*	-	-	-
I can assist families in helping their children do well in school		0.7	1.01	0.07	15.60	<.001
I am confident in my ability to get parents involved in the school activities of their children with disabilities		0.66	1.05	0.08	13.01	<.001
I can collaborate with other professionals (e.g. itinerant teachers or speech pathologists) in designing educational plans for students with disabilities		0.66	1.08	0.08	13.05	<.001
I am able to work jointly with other professionals and staff (e.g. aides, other teachers) to teach students with disabilities in the classroom		0.58	1.00	0.09	11.47	<.001
I am confident in informing others who know little about laws and policies related to the inclusion of students with disabilities		0.6	0.99	0.08	11.83	<.001

SE = standard error, CR = critical ratio, LV = latent variable
* fixed to 1 (no associated standard error, critical ratio or p-value).

appropriate challenges for very capable students, could be modified to *I can provide different levels of tasks for very capable student to learn something new*. Similarly, item 10 *I am confident in designing learning tasks so that the individual needs of students with disabilities are accommodated*, the word 'design' might make participants think about a complex level of design that might require them to work with others to accomplish, leading this item to load on the EC factor, which is about collaboration. Thus, for example, replacing *design learning tasks* with similar words like *preparing learning tasks* might improve this item. Indeed, in the Chinese version of the TEIP scale (Malinen et al., 2012), this item was excluded from the model for loading on all three factors during exploratory factor analysis, as was item 4, *I can assist families in helping their children do well in school.*

Rasch analysis of the items showed good item information function for people with low self-efficacy; participants familiar or competent with disabilities might need items with stronger statements for those items to be informative in measuring self-efficacy (see appendix 1).

5. Conclusion
The Arabic version of the TEIP scale showed acceptable fit for the proposed scale structure based on CFA analysis. The Arabic scale can thus be used to measure the latent variable of teachers' belief in their ability to work in inclusive classrooms. However, replication of this study will be informative, especially to confirm whether two items (1, 17) will continue to perform differently than other items in the scale, as Rasch analysis results showed them doing. In addition, rephrasing items 5, 6, and 10 is recommended, followed by testing to examine whether these items are improved. In any future update of the scale, including items with strong statements might yield more information for measuring self-efficacy, especially for people familiar with people with disabilities. Unless the target group of individuals has no prior knowledge of persons with disabilities, the scale as it currently (18 items) is appropriate. In addition, examining the dimensionality of the scale using the application of Rasch analysis will help us gain more understanding of the scale's psychometric properties.

Competing Interests
The authors declare no conflict of interest.

Author details
Ghaleb H Alnahdi
E-mail: ghalnahdi@gmail.com
ORCID ID: http://orcid.org/0000-0003-1692-9765
Department of Special Education, College of Education, Prince Sattam bin Abdulaziz University, Alkharj, Saudi Arabia.

Citation information
Cite this article as: The Arabic version of the teacher efficacy for inclusive practices (TEIP-AR) scale: A construct validity study, Ghaleb H Alnahdi, *Cogent Education* (2019), 6: 1618516.
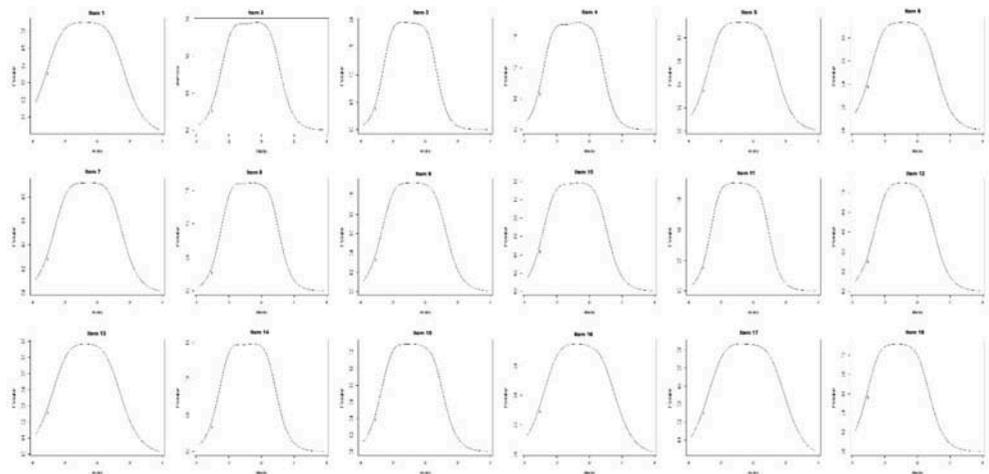
References
Ahsan, M. T., Sharma, U., & Deppeler, J. M. (2012). Exploring pre-service teachers' perceived teaching-efficacy, attitudes and concerns about inclusive education in Bangladesh. *International Journal of Whole Schooling, 8*(2), 1–20.

Alhano, I. (2006). *Representations of learning disabilities in Saudi Arabia elementary schools: A grounded theory study* (Doctoral dissertation). Retrieved from http://dr-banderalotaibi.com/new/research_dr.php

Alnahdi, G., ., Saloviita, T., ., & Elhadi, A., . (2019). Inclusive education in Saudi Arabia and Finland: Pre-service teachers' attitudes. *Support for Learning, 34*, 1. doi:10.1111/1467-9604.12239

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York, NY: Worth Publishers.

Beaton, D. E., Bombardier, C., Guillemin, F., & Ferraz, M. B. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine, 25*(24), 3186–3191. doi:10.1097/00007632-200012150-00014

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*(3), 588. doi:10.1037/0033-2909.88.3.588

Bond, T., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. New York, NY: Routledge.

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology, 1*(3), 185–216. doi:10.1177/135910457000100301

Brown, T. A. (2014). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Publications.

Byrne, B. M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. New York: Taylor & Francis.

Denzine, G. M., Cooney, J. B., & McKenzie, R. (2005). Confirmatory factor analysis of the Teacher Efficacy Scale for prospective teachers. *British Journal of Educational Psychology, 75*(4), 689–708. doi:10.1348/000709905X37253

Fox, C. M., & Jones, J. A. (1998). Uses of Rasch modeling in counseling psychology research. *Journal of*

Counseling Psychology, 45(1), 30. doi:10.1037/0022-0167.45.1.30

Futschek, K. (2014). Actual type-I-and type-II-risk of four different model tests of the Rasch model. *Psychological Test and Assessment Modeling, 56*(2), 168–177.

George, D., & Mallery, P. (2003). *SPSS for windows step by step: A simple guide and reference. 11.0 update* (4th ed.). Boston: Allyn & Bacon.

Hu, L. T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*(4), 424. doi:10.1037/1082-989X.3.4.424

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

Loreman, T., Sharma, U., & Forlin, C. (2013). Do pre-service teachers feel ready to teach in inclusive classrooms? *Australian Journal of Teacher Education, 38*, 27–44. doi:10.14221/ajte.2013v38n1.10

Magis, D. (2013). A note on the item information function of the four-parameter logistic model. *Applied Psychological Measurement, 37*(4), 304–315. doi:10.1177/0146621613475471

Malinen, O. P., Savolainen, H., & Xu, J. (2012). Beijing in-service teachers' self-efficacy and attitudes towards inclusive education. *Teaching and Teacher Education, 28*(4), 526–534. doi:10.1016/j.tate.2011.12.004

Park, M. H., Dimitrov, D. M., Das, A., & Gichuru, M. (2016). The teacher efficacy for inclusive practices (TEIP) scale: Dimensionality and factor structure. *Journal of Research in Special Educational Needs, 16*(1), 2–12. doi:10.1111/1471-3802.12047

Savolainen, H., Engelbrecht, P., Nel, M., & Malinen, O. P. (2012). *Understanding teachers' attitudes and self-efficacy in inclusive education: Implications for pre-service and in-service teacher education. European Journal of Special Needs Education, 27*(1), 51-68.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8*(2), 23–74.

Sharma, U., Loreman, T., & Forlin, C. (2012). Measuring teacher efficacy to implement inclusive practices. *Journal of Research in Special Educational Needs, 12*(1), 12–21. doi:10.1111/j.1471-3802.2011.01200.x

Tennant, A., McKenna, S. P., & Hagell, P. (2004). Application of Rasch analysis in the development and application of quality of life instruments. *Value in Health, 7*, S22–S26.

Tschannen-Moran, M., & Hoy, A. W. (2007). The differential antecedents of self-efficacy beliefs of novice and experienced teachers. *Teaching and Teacher Education, 23*(6), 944–956. doi:10.1016/j.tate.2006.05.003

UNESCO. (2009). *Inclusive education: The way of the future. Final Report of the International Conference of Education (48th Session)*, Paris: Author.

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

Yada, A., & Savolainen, H. (2017). Japanese in-service teachers' attitudes toward inclusive education and self-efficacy for inclusive practices. *Teaching and Teacher Education, 64*, 222–229. doi:10.1016/j.tate.2017.02.005

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

Zieba, A. (2013). The item information function in one and two-parameter logistic models–A comparison and use in the analysis of the results of school tests. *Didactics of Mathematics, 10*(14), 87–96.

Appendix 1.
Item information charts for all items.

Publishing with Cogent OA ensures:

- Immediate, universal access to your article on publication
- High visibility and discoverability via the Cogent OA website as well as Taylor & Francis Online
- Download and citation statistics for your article
- Rapid online publication
- Input from, and dialog with, expert editors and editorial boards
- Retention of full copyright of your article
- Guaranteed legacy preservation of your article
- Discounts and waivers for authors in developing regions

Submit your manuscript to a Cogent OA journal at www.CogentOA.com