International Journal of Science Education

ISSN: 0950-0693 (Print) 1464-5289 (Online) Journal homepage: http://www.tandfonline.com/loi/tsed20

Rasch analysis for psychometric improvement of science attitude rating scales

Pey-Tee Oon & Xitao Fan

To cite this article: Pey-Tee Oon & Xitao Fan (2017): Rasch analysis for psychometric improvement of science attitude rating scales, International Journal of Science Education, DOI: 10.1080/09500693.2017.1299951

To link to this article: http://dx.doi.org/10.1080/09500693.2017.1299951

Published online: 10 Apr 2017.

INTERNATIONAL JOURNAL OF SCIENCE EDUCATION, 2017
http://dx.doi.org/10.1080/09500693.2017.1299951

Rasch analysis for psychometric improvement of science attitude rating scales

Pey-Tee Oon and Xitao Fan

Faculty of Education, University of Macau, Macao, People's Republic of China

ABSTRACT

Students' attitude towards science (SAS) is often a subject of investigation in science education research. Survey of rating scale is commonly used in the study of SAS. The present study illustrates how Rasch analysis can be used to provide psychometric information of SAS rating scales. The analyses were conducted on a 20-item SAS scale used in an existing dataset of *The Trends in International Mathematics and Science Study* (TIMSS) (2011). Data of all the eight-grade participants from Hong Kong and Singapore ($N = 9942$) were retrieved for analyses. Additional insights from Rasch analysis that are not commonly available from conventional test and item analyses were discussed, such as invariance measurement of SAS, unidimensionality of SAS construct, optimum utilization of SAS rating categories, and item difficulty hierarchy in the SAS scale. Recommendations on how TIMSS items on the measurement of SAS can be better designed were discussed. The study also highlights the importance of using Rasch estimates for statistical parametric tests (e.g. ANOVA, *t*-test) that are common in science education research for group comparisons.

ARTICLE HISTORY
Received 20 October 2016
Accepted 22 February 2017

KEYWORDS
Psychometric information; test analysis; item analysis; Rasch analysis; attitude toward science; rating scale; science education research

Background

Students' attitudes towards science (SAS) are a concern for many science educators and researchers globally (Osborne, Simon, & Collins, 2003; Potvin & Hasni, 2014), because those attitudes have serious implications for both science education and for future career choices. The importance of SAS is reflected in some important and influential international assessments, such as *The Trends in International Mathematics and Science Study* (TIMSS) and *Programme for International Student Assessment* (PISA), both of which incorporated assessment of SAS.

Survey of rating scale is commonly employed in SAS scales (Boone, Staver, & Yale, 2014; Liu, 2010; Sondergeld & Johnson, 2014). Potvin and Hasni (2014) discussed that Likert-type rating scale was 'easy to use', 'quite statistically reliable', and 'allow quantitative comparisons' (p. 111). Of the 216 published articles involving SAS reviewed by Potvin and Hasni (2014), 189 involved surveys that were based on rating scale of some kind.

CONTACT Pey-Tee Oon peyteeoon@umac.mo Faculty of Education, University of Macau, Avenida da Universidade, Taipa, Macao, People's Republic of China

In their critical review of 150 published articles that covered 66 SAS instruments, Blalock et al. (2008) concluded that most articles showed a lack of psychometric evidence for the rating scales used. Noll (1935) made a plea to the field to measure attitude in a 'scientific' (objective) way. Despite this plea made many decades ago, a significant number of studies were still found to lack psychometric evidence (Blalock et al., 2008). A recent study by Boone, Townsend, and Staver (2011) further raised questions concerning the scarcity of psychometric evidence and the misconceptions concerning some psychometric concepts in the published research articles involving rating scale data. Issues highlighted included failure to demonstrate internal consistency, misconception or confusion about internal consistency and unidimensionality (Kind, Jones, & Barmby, 2007). Other criticisms of SAS scales used in science education research include lack of validity evidence as well as lack of understanding about validity (Kind et al., 2007).

A common issue for many SAS scales in science education is that, researchers often assume that the ordinal-scale score (raw score) is linear on which parametric statistical tests can readily be performed on. However, the raw scores of the rating scales are not expressed on a linear scale, and many science education researchers are not aware that even the basic statistics (e.g. means, standard deviations) assumes linearity (Wright & Master, 1982). Wright (1999) further discussed that many social scientists were not aware of the harmful effects of misusing the non-linear raw scores. A more difficult issue in conventional measurement is the dependency between persons' scores and the difficulty levels of items, which makes it very difficult or even impossible to compare scores from different rating scales. Furthermore, as item statistics depend on the persons who answered the items, a change of respondent sample will change the item statistics, too.

In science education research, conventional psychometric investigation of SAS rating scales typically involves test level analysis (e.g. item mean, item-total correlations), internal consistency estimation (e.g. Cronbach's coefficient alpha), exploratory factor analysis for scale structure, (e.g. Jocz, Zhai, & Tan, 2014; Swarat, Ortony, & Revelle, 2012; Wang & Berlin, 2010), and some other correlational analyses with external criteria for validation purpose (e.g. Tuan, Chin, & Shieh, 2005; Velayutham & Aldridge, 2013; Velayutham, Aldridge, & Fraser, 2011). Several scholars (e.g. Boone et al., 2014; Liu, 2010; Neumann, Neumann, & Nehm, 2011) have recently made the call that science education researchers may use Rasch model for developing psychometrically better assessment instruments in science education research. A quick look at the science education literature involving SAS rating scales reveals that Rasch analysis is not something that science education researchers are very familiar with, and it has been rarely used for SAS rating scales in science education research. In our view, although Rasch model has been receiving some more attention in science education, many researchers in this field are probably still unaware of what Rasch analysis can offer, and of how Rasch analysis can help in improving the psychometric quality of assessment in science education research.

Aim of the paper

The purpose of this study is to illustrate how Rasch analysis can be used to provide psychometric information about SAS rating scales, and how such information may help science education researchers to improve the psychometric quality of the scales. It should be noted that the paper is neither intended to be an evaluation of a particular

instrument, nor is it intended to be a critique of the conventional item and test analyses. Instead, the paper intends to illustrate how Rasch analysis may readily provide insightful information for evidence-based psychometric improvement of SAS rating scales.

As an example of using Rasch analysis for the purpose of providing information about the psychometric characteristics of a rating scale, we used an existing dataset and conducted the relevant analyses, as described below.

Methods

Data and participants

Data for the present study were retrieved from *TIMSS* (2011) (Michael, Mullis, Foy, & Stance, 2012). Data of all the eighth-grade participants from Hong Kong and Singapore ($N = 9942$) were included for analyses. Of these students, 4015 were from Hong Kong and 5927 were from Singapore.

Instrument

The rating scale items used in the present study were retrieved from *TIMSS* (2011) data. Twenty questions were extracted from the eighth-graders' science context questionnaire measuring student attitudes towards science (Martin & Mullis, 2012), and the items were shown in Table 1. The items used a four-step scale (1 = *Disagree A Lot*; 2 = *Disagree*

Table 1. Model-data fit statistics.

Item	Total score	Total count	Measure (*logit*)	Infit MNSQ	Outfit MNSQ	PTMEASURE correlation
A1. Enjoy learning science	31382	9889	−.68	.68	.64	.73
A2. Wish have not to study science	29623	9884	−.23	1.16	1.16	.64
A3. Science is boring	29325	9867	−.17	1.06	1.05	.66
A4. Learn interesting things	32871	9880	−1.12	.90	.82	.63
A5. Like science	30671	9881	−.50	.65	.62	.75
B1. Science will help me	33016	9877	−1.16	.95	.97	.59
B2. Need science to learn other things	28976	9876	−.07	1.28	1.52	.54
B3. Need science to get into university	31204	9866	−.65	1.15	1.14	.60
B4. Need Science to get the job I want	29525	9878	−.21	1.24	1.26	.60
B5. Job involving science	26194	9877	.59	1.19	1.21	.66
B6. Important to do well in science	33053	9891	−1.16	1.07	1.03	.58
C1. Usually do well in science	28112	9883	.14	.68	.69	.73
C2. Science is more difficult	27533	9883	.28	1.09	1.16	.62
C3. Science not my strength	25407	9870	.76	1.09	1.18	.67
C4. Learn quickly in science	26978	9871	.40	.69	.70	.73
C5. Science makes confused and nervous	27599	9873	.25	1.28	1.36	.56
C6. Good at working out problems	23651	9863	1.16	.81	.83	.71
C7. I can do well in science	24896	9862	.87	.92	.96	.66
C8. I am good at science	22857	9863	1.34	.98	1.00	.66
C9. Science is harder for me	28019	9875	.16	1.23	1.28	.62

A Little; 3 = *Agree A Little*; 4 = *Agree A Lot*). Items A2, A3, C2, C3, C5, and C9 were negatively worded items, and their scores were reverse-coded prior to the analyses.

TIMSS included three motivational constructs, namely, intrinsic value (interest), utility value, and ability beliefs in assessing students' attitudes towards science (Martin & Mullis, 2012). The first construct consisting of 5 items is labelled as 'Student like learning science (SLS)' scale. The second construct consisting of 6 items is labelled as 'Student value science (SVS)' scale. The third construct consisting of 9 items is labelled as 'Student confidence in science (SCS)' scale (Martin & Mullis, 2012). The 5-SLS items were indexed from A1–A5, the 6-SVS items were indexed from B1–B6, and the 9-SCS items were indexed from C1–C9.

Analyses

The Rasch analyses were conducted using *Winsteps* software (Version 3.81.0). Our focus was on Rasch analysis results and interpretations, and how the Rasch modelling analysis findings could be used for considerations of improving the psychometric quality of the SAS rating scale.

Rasch analysis has some basic concepts and statistics that are unfamiliar for most science education researchers, even for those who may be familiar with the statistics and concepts in classical test theory (e.g. item difficulty, item discrimination, reliability). First, instead of using raw score (or some form of linear transformation of raw score), Rasch model uses *logit,* which is a statistical concept defined as the log (odds), that is, for a given item and the respondent's trait/ability level, the ratio of probability of correct response to its complement (1 – the probability). *Logit* is defined as a unit in Rasch measurement (Linacre & Wright, 1989). When data fit the Rasch model, the item and person estimates are estimated independently of one another (Rasch, 1960, 1980). Conceptually, this is very different from the classical test theory, where a person's score depends on the easiness/difficulty levels of a given set of items, while item difficulty estimates, in turn, depend on the ability/trait levels a given group of respondents, making the item and person estimates inter-dependent.

There are many statistical programs to perform Rasch analysis. The present study utilised the *Winsteps* program developed by Mike Linacre. It is a user-friendly program with detailed manual for all analyses. The software developer, Mike Linacre (2014) provides almost instantaneous online consultation for enquiries related to Rasch analysis.

Rasch analysis provides important information on different aspects of measurement that can be readily used to facilitate our efforts in improving the psychometric quality of rating scales, and these are elaborated below.

Model-data fit

In Rasch modelling analysis, fit statistics are provided to assess model fit (Bond & Fox, 2015), that is, whether the actual data are close to the Rasch model's expectation. When data fit the Rasch model, the item and person estimates are estimated independently of one another (Rasch, 1960, 1980). Data within the threshold of fit are assumed to be unidimensional (Bond & Fox, 2015). This property is important, and the item and person estimates are only meaningful if all the items of the scale contribute to the measure of a single latent trait (Bond & Fox, 2015). The estimates are degraded if other attributes are

also being measured by the scale. To a greater extent, the results from the scales are not interpretable if the latent trait is not manifested in the data (Bond & Fox, 2015).

The data fit is typically assessed by *infit* and *outfit* mean square (MNSQ). *Infit* is sensitive to inlier misfit, while *outfit* is sensitive to outlier misfit. Infit and outfit are quantified by MNSQ, and values of *infit*/*outfi*t range of .50 to 1.5 are regarded as fitting the Rasch model (Romine & Walter, 2014; Wright & Linacre, 1996). Items with fit statistics beyond the range limits should be interpreted with caution, as these fit statistic values suggest the misbehaviors of items. MNSQ fit indices above 1.5 and below .50 indicates misbehaving of items. For the former, it signalled that students with high agreeability endorse low endorsability items and vice versa. For the latter, it signalled that items behave suspiciously well and there might be secondary dimension that positively correlates with the latent trait (Masters, 1988; Romine & Walter, 2014).

Dimensionality map

Rasch's Principal Component Analysis (PCA) of residuals attempts to identify a potential secondary dimension (i.e. noises) (Linacre, 2014). The 'noises' threaten the measurement of latent trait (e.g. attitudes towards science). If a variance of 50% is explained by Rasch measures and the first contrast of unexplained variance reports a strength of less than two items (in eigenvalues), these indicate that the data can be assumed to be unidimensional (Linacre, 2014; Romine & Walter, 2014). In other words, it means that the secondary dimension or noises are not substantial enough to distort the measurement of latent trait.

Reliability and targeting

Rasch model produces two reliability indices to help researchers to determine whether the person and item estimates of Rasch are reliably calibrated on an interval scale. The person separation index indicates replicability of person ordering on an interval scale if they were given a parallel set of items measuring the same latent trait (Wright & Master, 1982). The item separation index indicates replicability of item placements on an interval scale if these same items were answered by another sample of the same population (Bond & Fox, 2015). The two reliability indices inform the researchers about the confidence level of Rasch's person and item estimates. The commonly accepted threshold for the separation index is 3.0 (Bond & Fox, 2007) or at least 2.0 (Lee, Grossman, & Krishnan, 2008).

In common Rasch model practice, careful inspection of the Wright map, which puts the item estimates (indicated by the number item) and the person estimates (labeld as '#') simultaneously on an interval scale in unit *logit* (Figure 2), will inform the researchers about how well the items are functioning. The map provides *prima facie* evidence on how well the items have targeted the sample to which they are intended for. A good targeting will show clusters of item and person stay opposite to each other where the spread of items is covered by the corresponding spread of persons (Bond & Fox, 2015). This feature of Rasch is of distinct advantage as compared to the conventional analyses because test developer can tell immediately from the map whether the items are targeting the respondents well. If the person distribution is clustered heavily on top in comparison with the item distribution near the bottom of the map, the items do not provide much information about the respondents. For more precise estimation

of latent trait (e.g. attitudes towards science), more endorsable items should be added into the scale.

Invariance

This is considered as an important property of scientific measurement (Bond & Fox, 2015). 'Invariance' means that the person and item estimates discussed above should remain the same regardless of measurement conditions. More specifically, person estimates should remain identical regardless of which appropriate measures were used, and item estimates should remain stable from different but relevant samples. Differential item functioning (DIF) can be examined to obtain information on whether the item and person statistics remained invariant. Lack invariance is signalled by a DIF contrast of greater than .50 logit (Bond & Fox, 2015; Linacre, 2014).

Category function

The diagnosis of appropriate response categories enhances reliability of a SAS scale (Boone et al., 2014). The diagnosis focuses on whether the optimal response categories have been used. The Andrich's Rating Scale Model (RSM) can be employed to investigate the usage of response categories for a SAS scale (Wright & Master, 1982). A set of criteria can be used to verify the functioning of each response category (Linacre, 2004). These criteria include (a) a minimum of 10 observations for each category ($N > 10$); (b) average category measures increase monotonically with categories ($N(q)_{\text{increase}}$); (c) outfit mean square statistics less than 2.00 (MNSQ < 2); (d) The category threshold increases monotonically with categories (τ_{increase}); (e) category thresholds are at least 1.4 to 5 logits apart ($\tau_{1.4\text{–}5\ \text{logits}}$); and (f) there are distinct peaks for every probability curves.

Illustrative results and discussions

Rasch analysis results

Infit-outfit statistics

Table 1 presents the model-data fit statistics. All items are within the acceptable limits of *infit* and *outfit* MNSQ statistics, indicating that the data met the expectations of Rasch model. All items had positive point-biserial (PTMEASURE) correlations, indicating that the items were properly scored and they functioned as they were expected to (Linacre, 2014). However, a point to be noted is that Item B2 ('I need science to learn other school subjects.') had marginal MnSq *outfit* for the Rasch model.

PCA of residuals

The variance explained by Rasch measures was 49.5%, and the first three eigenvalues for the unexplained variance were 3.6, 2.6 and 2.1 in the PCA of residuals analysis (not reported in the table). We further examined which items contributed most to the noise by examining the plot of residual loadings for the SAS data (Figure 1).

The plot shows that Items B4 (*Need Science to get the job I want*), B3 (*Need science to get into university*), B2 (*Need science to learn other things*), B5 (*Job involving science*), B1 (*Science will help me*), and B6 (*Important to do well in science*), corresponding to Items

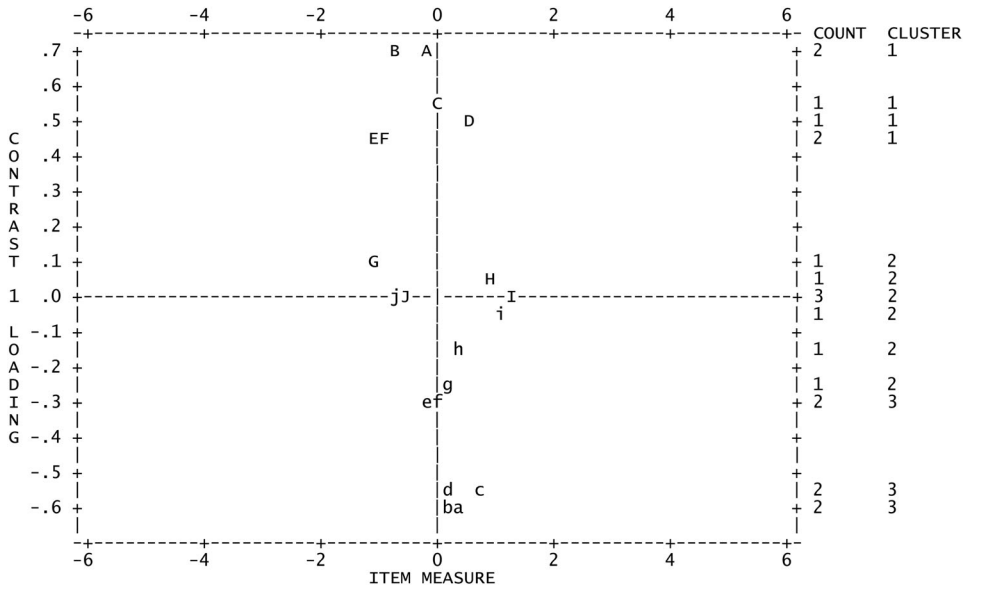

Figure 1. Plot of residual loadings for SAS data showing contrasts between items at the top versus those at the bottom (Linacre, 2014). Items A, B, C, D, E, F, G, H, I, J, a, b, c, d, e, f, g, h, i, j are corresponding to items B4, B3, B2, B5, B1, B6, A4, C7, C8, A5, C2, C9, C3, C5, A2, A3, C1, C4, C6, and A1.

A, B, C, D, E, and F in Figure 1, had factor loadings greater than.40. On the other extreme, Items C2 (*Science is more difficult*), C9, (*Science is harder for me*), C3 (*Science not my strength*), C5 (*Science makes confused and nervous*), corresponding to Items a, b, c, d in Figure 1, had factor loadings < −.60. These two groups of contrasting items clustered to each other more than they did to the other SAS items. The former are all positively worded items while the latter are negatively worded items.

We further examined the dimensionality of the items by removing all the negatively worded items: A2 (*Wish have not to study science*), A3 (*Science is boring*), C2 (*Science is more difficult*), C3 (*Science not my strength*), C5 (*Science makes confused and nervous*), and C9 (*Science is harder for me*). After removing these items, the variance explained by Rasch measures increased from 49.5% to 55.7%, approximately 6% increment, and the first three eigenvalues for the unexplained variance decreased to 3.2, 2.1, and 1.3, respectively.

SVS and SCS showed strongest contrast to each other – Items B1 (*Science will help me*), B2 (*Need science to learn other things*), B3 (*Need science to get into university*), B4 (*Need Science to get the job I want*), B5 (*Job involving science*), and B6 (*Important to do well in science*) were positively loaded (Table 2) while Items C1 (*Usually do well in science*), C4 (*Learn quickly in science*), C6 (*Good at working out problems*), C7 (*I can do well in science*), and C8 (*I am good at science*) were negatively loaded (Table 2). The disattenuated correlation between SVS and SCS items is.68, indicating that the two sets of measures are not correlated too highly (Linacre, 2014; Schumacker, 1996). On the contrary, SLS Items, A1 (*Enjoy learning science*), A5 (*Like science*), and A4 (*Learn interesting things*), showed similar loadings with those of SCS items, and they also had high correlation (.93) with SCS items (Table 2). These results strongly suggest that SLS and SCS measures are consistent

Table 2. Factor loadings of all the positively worded items.

Contrast	Loading	Measure (*logit*)	Infit MNSQ	Outfit MNSQ	Entre number	
1	.71	−.16	1.20	1.18	A	B4
1	.69	−.68	1.10	1.07	B	B3
1	.47	.00	1.28	1.44	C	B2
1	.42	.78	1.21	1.21	D	B5
1	.40	−1.28	.96	1.00	E	B1
1	.37	−1.27	1.07	1.04	F	B6
3	−.54	.25	.87	.90	a	C1
3	−.54	.56	.84	.86	b	C4
3	−.53	1.46	.89	.92	c	C6
3	−.52	1.67	1.03	1.04	d	C8
3	−.50	1.12	.98	1.01	e	C7
2	−.27	−.72	.81	.77	f	A1
2	−.25	−.50	.79	.75	g	A5
2	−.10	−1.22	1.00	.95	G	A4

with each other, and they measure very similar or nearly the same latent trait (Linacre, 2014).

Person and item separation indices

Rasch analysis provides assessment of reliability in the form of item and person separation indices. The item separation index was 43.42 and the person separation index was 3.06. These indices indicate the spread of items and persons reliably calibrated along the latent trait measured by the scale. From the results, we may infer that the items had reasonable separation, that is, some were more endorsable, while others were less endorsable, similar to a cognitive test with some items being more difficult while others being easier. Similarly, some respondents were more agreeable while others were less so.

Wright map

Wright map (Figure 2) provides information about how the item endorsability (or difficulty) levels matched respondents' trait (or ability) levels, and such information can be very useful for understanding where additional items may be needed for future improvement of the scale. Figure 2 lays out the locations of the 9942 students and the 20 items on a common scale. The first column is the logit scale and Columns 2 and 3 graphically described the locations of the respondents and the 20 items, respectively. This map transformed the student scores and item scores on a common interval scale in *logit* unit. For the present study, the student and item *logit* scale runs from −5 to +6 logits. Students closer to the top of the figure were more in favour of science than those near the bottom. Items near the top are less endorsable items, and these appeared to be stronger SAS statements (more difficult or less likely for respondents to agree with). Students showing greater extent of positive towards science were more likely to agree with these items. In contrast, students who did not perceive science positively were less likely to agree with these items.

Items probing student confidence in science (Items C1–C9) appeared to be least likely to be endorsed by these Asian students (Figure 2). The results corroborated with other findings reported elsewhere that Asian students who outperformed their counterparts in international science and mathematics assessment tended to show lower confidence in the subjects (e.g. Leung, 2002), most likely due to cultural disposition (Chang &

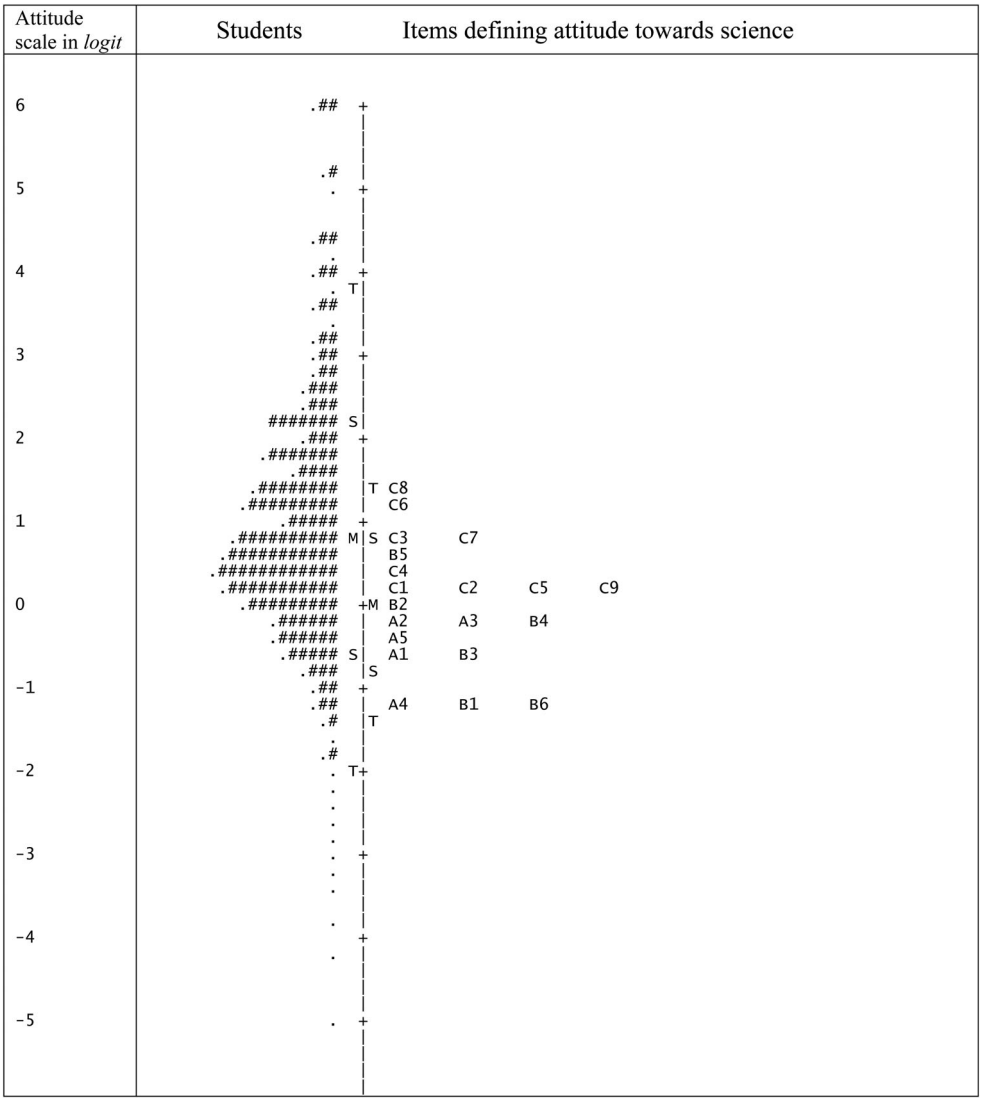

Figure 2. Wright map with item and person estimates calibrated on a linear scale. The first column is a logit scale, the second and third columns described locations of persons and items, respectively. Note: EACH '#' IS 75: EACH '.' IS 1 TO 74.

Cheng, 2008). Among the SVS items (B1–B5), Item B5 ('I would like a job that involves using science'), as indicated by its position relative to other SVS items on the Wright map, appeared to be the most difficult to endorse by these students (Figure 2). Oon and Subramaniam (2013) found that Singapore school students, despite of their interest and good grades in science, were generally not inclined towards a science-related careers. But in general, the SVS items appeared to be the least difficult to endorse (Figure 2). It is very likely that students in Asia need no reminder about the importance of doing well in science for its utilitarian value (Oon & Subramaniam, 2013). The orderings for the SAS items appeared to be logical.

The content sufficiency and content validity can be assessed from the distribution and ordering of the items on the map (Wright & Master, 1982). Visual inspection revealed that there was no obvious gap between the items (Figure 2). However, the map indicated that more less-agreeable items could be built into the scale in order to measure the SAS construct better, because no items appeared to have targeted the more agreeable persons (>1.5 logit) (Figure 2).

Differential item functioning

Rasch's DIF statistics were computed for the two cohorts of students from Singapore and Hong Kong, respectively, to examine whether the items functioned in the same way for the two geographical samples (Table 3). Item B3 (0.5x), B6 (.81x) and Item C5 (.6x) showed obviously larger country DIF statistic values than others. A closer look at the *t* test statistic of these three DIF values showed $|t| > 2$, with $p < .05$, suggesting that the observed DIF values were statistically significant. As a result, these three items would be considered

Table 3. DIF statistics for Hong Kong and Singapore students.

Item	Country	Measure	DIF Contrast
A1. Enjoy learning science	Hong Kong	−.64	.08
	Singapore	−.72	
A2. Wish have not to study science	Hong Kong	−.34	−.19
	Singapore	−.15	
A3. Science is boring	Hong Kong	−.29	−.22
	Singapore	−.08	
A4. Learn interesting things	Hong Kong	−1.00	.21
	Singapore	−1.21	
A5. Like science	Hong Kong	−.50	.00
	Singapore	−.50	
B1. Science will help me	Hong Kong	−1.20	−.06
	Singapore	−1.14	
B2. Need science to learn other things	Hong Kong	−.11	−.07
	Singapore	−.04	
B3. Need science to get into university	Hong Kong	−.36	.53
	Singapore	−.88	
B4. Need Science to get the job I want	Hong Kong	.01	.38
	Singapore	−.37	
B5. Job involving science	Hong Kong	−59	.00
	Singapore	−59	
B6. Important to do well in science	Hong Kong	−.72	.81
	Singapore	−1.54	
C1. Usually do well in science	Hong Kong	.03	−.18
	Singapore	.22	
C2. Science is more difficult	Hong Kong	.10	−.30
	Singapore	.40	
C3. Science not my strength	Hong Kong	.66	−.17
	Singapore	.83	
C4. Learn quickly in science	Hong Kong	.31	−.16
	Singapore	.46	
C5. Science makes confused and nervous	Hong Kong	−.08	−.57
	Singapore	.49	
C6. Good at working out problems	Hong Kong	1.12	−.06
	Singapore	1.18	
C7. I can do well in science	Hong Kong	1.03	.27
	Singapore	.76	
C8. I am good at science	Hong Kong	1.46	.20
	Singapore	1.26	
C9. Science is harder for me	Hong Kong	.02	−.23
	Singapore	.25	

as not invariant across the two geographical samples. In other words, these three items appeared to have different levels of endorsability for the two samples.

Functioning of response categories

A minimum of 10 observations was observed for each category ($N > 10$) and the *outfit* MNSQ for each category reported values below 2.00 (Table 4).

The average measure increased monotonically from categories 1 (*Disagree A Lot*) to 4 (*Agree A Lot*) [$N(q)$increase]. Besides, the threshold calibrations increased monotonically with the distance between the range of 1.4 and 5 logits apart ($\tau_{1.4-5\ \text{logits}}$) (Table 4). Each category also has its distinct peak (Figure 3). These results lend support to the use of the 4-step response categories of this measure (Bond & Fox, 2015).

The category curves (Figure 3) provide information about the appropriateness of the response categories for this sample. The *y*-axis (0–1) is the expected probability of each response category to be endorsed by the respondents. The *x*-axis represents the item difficulty (i.e. item endorsability in this example) for the respondents, with positive values indicating higher level of liking for science. Figure 3 indicates that those who exhibited higher positive attitude towards science (i.e. those with high positive values on the *x*-axis) tended to endorse Category 4 (*Agree A Lot*), while those who did not perceive science favourably (i.e. those with low values on the *x*-axis) tended to endorse Category 1 (*Disagree A Lot*). In other words, this graph suggested that the response categories in this SAS scale functioned as intended.

Additional insights from Rasch analysis for improving SAS rating scales

In addition to the illustrative discussion about the major aspects of information from Rasch analysis, we may consider how Rasch analysis results could be used for the purpose of improving the psychometric quality of a SAS rating scale, as discussed below.

Invariance measurement of SAS

Zenisky, Hambleton, and Robin (2004) stated that DIF analyses were common for many large-scale assessments. In science education research, only very few (e.g. Wagler & Wagler, 2013) considered this issue. Invariance property of a scale is important for a measure. If items were found to function differently for different groups (e.g. favouring one group while disadvantaging the other), the measurement results could be misleading. In our example above, Item B3 (*I need to do well in science to get into the university of my choice*), B6 (*Important to do well in* science) and Item C5 (*Science makes me confused and nervous*) were flagged in Rasch analysis as being not invariant. This suggests that the

Table 4. Summary of the 4-point category response.

Category	Observed count, N (%)	Average measure, N (q) (*logit*)	Outfit MNSQ	Threshold Calibration (*logit*)
1. Disagree A Lot	15660 (8)	−1.20	1.32	None
2. Disagree A Little	47318 (24)	−.17	.95	−1.81
3. Agree A Little	77528 (39)	.88	.84	−.12
4. Agree A Lot	57003 (29)	2.37	1.08	1.93

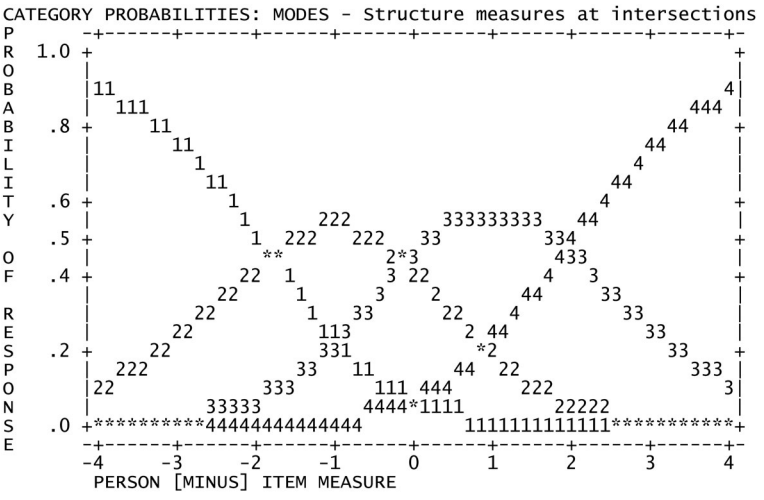

Figure 3. The probability curves showing how probable is the observation of each of the four categories (1 = *Disagree a lot*; 2 = *Disagree a little*; 3 = *Agree a little*; 4 = *Agree a lot*) for measure relative to the item measure (Linacre, 2014).

ratings of these three items need to be interpreted with caution, as they functioned differently across the samples from Hong Kong and Singapore.

Unidimensionality of SAS construct

The Rasch's model-fit statistics and PCA analysis of residuals provide insight for the unidimensionality of the latent trait (e.g. attitudes towards science) being measured. The Rasch PCA analysis of residuals is different from the conventional factor analysis. The former focuses on analysing the item residuals as representing random 'noise' in measurement, while the latter focuses on the commonalities of the items, which theoretically represents the latent trait that a measure is trying to measure, as opposed to the measurement 'noise' in the PCA analysis of residuals. Because of this, the Rasch residual-based PCA, as compared to the conventional factor analysis, can be used to identify secondary dimensions that may exist in the data (e.g. the negatively worded items as evident in the present study). The secondary dimension is undesirable for a unidimensional scale, as it measures something different from the latent trait. This issue is particularly important if item scores are to be summed for a total score/mean score. When such undesirable secondary dimension is identified, some remedies may be needed (e.g. to word all items positively to avoid the potential noise introduced by negatively worded items).

Optimum utilization of SAS rating categories

Appropriate rating categories on a scale contribute to the reliability of the instrument (Boone et al., 2011), but the importance of this has been overlooked in science education research. It is suggested that underutilised categories (e.g. very few or no respondents tended to choose a given category of a rating scale) should be removed or collapsed (Wright & Linacre, 1992) in order to enhance optimum usage of the categories.

Often, the use of rating categories is the result of the subjective judgment of a researcher. Rasch modelling analysis can help a researcher to decide if the rating categories used are reasonable, or may need adjustment. The results shown in the example above suggest that rating categories used are reasonable for the intended measurement purpose.

Item difficulty hierarchy in the SAS scale

Rasch analysis indicated that the scale covered a range of item difficulty (i.e. endorsability) that reflected different levels of attitudes towards science. Nevertheless, the results from the Wright map suggested that the items on the scale tended to be easy to endorse, because most items clustered near the lower part of the map. In other words, it indicates that most items targeted the students with less positive attitudes towards science. A few more difficult items targeting students who perceived science more favourably can be added into the scale. The addition of these items could enhance the validity of the SAS measurement (Boone et al., 2014; Chang & Engelhard, 2016).

Parametric test based on Rasch estimates

In science education, it is common to conduct parametric statistical tests (e.g. *t*-test, ANOVA) on raw scores of rating scales (Boone et al., 2014). Many researchers in science education may not be aware that the label of rating categories (e.g. '1' to '4' for an item to be rated) does not reflect the exact distance between the rating categories. As highlighted by Bond and Fox (2007), the distance is an 'unspecified amount' (p. 106). Succinctly illustrated by Boone et al. (2014), a coding scheme of (6, 5, 4, 3, 2, 1) and (5, 4, 3, 2, 1, 0) will arrive at different ratio of agreements between the same set of items. They called this equal-distance assumption as 'unfortunate leap' (p. 24). Parametric statistical tests are often not appropriate for such data of categorical rating scale, especially for students with very low level or very high level of liking for science (Boone et al., 2011). It is possible that a statistical test based on raw scores of categorical rating scales could suggest statistically significant differences where none actually exists, or vice versa.

Examples were presented in Table 5. Data of 282 students ($n = 138$ from Hong Kong and $n = 144$ from Singapore) who reported the lowest level of liking for science were used for this illustration. Both raw scores and Rasch person estimates were used in comparing the two groups. The analysis based on the raw scores reported statistical significant difference ($p < .05$) for the two groups of students on liking for science. In contrast, analysis based on the interval data of Rasch's estimates indicated no significant difference between the two groups ($p > .05$).

Table 5. Parametric *t*-test results using Rasch estimates and raw scores ($n_{\text{Hong Kong}} = 138$; $n_{\text{Singapore}} = 144$).

	Respondent	Mean (*logit*)	SD	t	df	*p*
Rasch estimate	Hong Kong	−2.39	.98	−1.73	280	.09
	Singapore	−2.19	.94			
Raw score	Hong Kong	9.06	1.53	−1.98	280	.04
	Singapore	9.42	1.52			

Recommendation on TIMSS SAS items improvement

The PCA of residuals suggested that the negatively framed items in TIMSS may not measure the same underlying construct as that measured by the positively framed items, as discussed by Smith (1996). All the negatively framed items closely clustered to each other, much more so than they did to the other positive items. A better fit was achieved by removing all the negative items, as shown by the significant increase of the variance explained by Rasch, and by the decrease of noise in the data. Bainer and Smith (1999) cautioned,

> … be careful when introducing reverse coded or negatively worded items into the instrument. Although this practice has been recommended as a means of offsetting response set biases, there are clear indications in a variety of settings that the responses to the negative worded items do not measure the same underlying construct as the positively worded items. There may be a substantial correlation between the two variables, as there was in this case, but the combination of the positively and negatively worded items in the same calibration often causes the item fit statistics to have an unexpectedly high proportion of misfitting items. (p. 263)

Based on the findings presented previously, we recommend that negatively worded SAS items of TIMSS be replaced by positively worded items. However, if their inclusion is inevitable, we suggest the creation and scoring of a sub-scale for the negatively framed items.

Osborne et al. (2003) did an extensive review of SAS studies, and they reported that SAS has pluralistic connotations as SAS might consist of several constructs, including (a) *the perception of the science teacher,* (b) *anxiety towards science,* (c) *the value of science,* (d) *self-esteem at science,* (e) *motivation towards science,* (f) *enjoyment of science,* (g) *attitudes of peers and friends towards science,* (h) *attitudes of parents towards science,* (i) *the nature of the classroom environment,* (j) *achievement in science,* and (k) *fear of failure on course* (p. 1054). The selection of SAS constructs very much depends on the subjective judgment of researchers. In assessing attitudes towards science, much research has incorporated student liking for science (SLS), value of science (SVS), and confident in science (SCS) (e.g. Murphy & Beggs, 2003; Wang & Berlin, 2010), similar to what is included in TIMSS. However, there has been a dearth of studies that have empirically examined the psychometric characteristics of these three components. Wang and Berlin (2010), through PCA of raw scores, identified seven components that explained 60% of variance for a SAS instrument designed to have included SLS, SVS, and SCS constructs. The present study conducted similar analyses for TIMSS data as Wang and Berlin's (2010), and identified one strongest component that explained 46% of variance of the SAS total scale consisting of the SLS, SVS, and SCS constructs. Such conventional factor analysis results often led to the conclusion that the 'instrument can justifiably be used as a single measure of general attitudes towards science class' (Wang & Berlin, 2010, p. 2422). This conclusion, however, failed to consider the *unexplained variance*, which were 67% and 54% in Wang and Berlin (2010) and the TIMSS SAS scale shown in this study, respectively. Using Rasch's PCA of residuals, the present study examined the unexplained variance, and found that the two sets of measures, the SVS and the SCS, are not well correlated as they showed strong contrasts with each other (Linacre, 2014; Schumacker, 1996). The statistical evidences suggested that these two measures were distinct from each other. On the

other hand, the results also showed that SLS and SCS were highly correlated, and might have measured very similar construct.

Theoretically, 'confidence in science' is defined as 'the extent to which a student is confident and feels successful in science class' (Wang & Berlin, 2010, p. 2418). On the other hand, the 'value of learning science' is defined as 'the degree of the alignment between science courses and future goals, such as college or career' (Andersen, & Chen, 2016, p. 7). For students to learn science, the former is an intrinsic aspect of motivation, while the latter is an extrinsic aspect of motivation. These two aspects can be conceptually different. For example, the grades that students received, or their success/failure to comprehend scientific understanding, could affect the intrinsic motivation, but not necessarily so for the extrinsic motivation. The different motivational orientations might have posed inconsistencies between them, but such inconsistencies may have often gone unnoticed. On the other hand, student 'liking for science' is defined as 'doing something as it is inherently interesting and enjoyable' (Ryan & Deci, 2000, p. 55, as reported in Plamer, 2005, p. 1858), and this has often been referred to as intrinsic motivation, similar to 'confidence in science'.

To examine whether the contrast loadings between the SCS and SVS remain invariant across sub-samples, we randomly split the data of the 9942 students into 10 sub-samples without repeated cases, and run the PCA of residuals on these 10 sub-samples. The contrast loadings between SCS and SVS for the 10 sub-samples were scatter-plotted against one another. A total of 45 possible pairwise graph plots were produced (e.g. Group 1 versus Group 2, Group 2 versus Group 3, Group 9 versus Group 10, etc.), and Figure 4 illustrates one out of the 45 graphs showing the contrast loadings between the SCS and SVS measures. The two independent sub-samples produced loadings that overlapped by 99.4% (e.g. 99.4% shared variance between the two independent samples). The shared variances across 45 possible subgroup sample pairs ranged from 97.3% to 99.7%. In other words, the results remained linearly invariant across the independent sub-samples, and this provided indication that SCS and SVS are two measures that have measured SAS

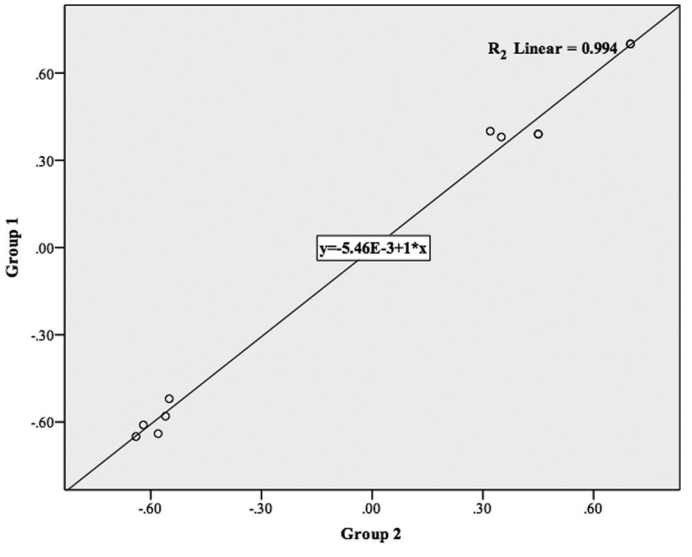

Figure 4. Scatter-plot for contrast loadings between Group 1 and Group 2 randomly split students.

differently, as evident in their contrast loadings that highlighted the distinction between the two sub-measures.

The above statistical findings and theoretical justifications concerning the three constructs prompted us to suggest that, for the study of SAS, SLS and SCS be combined into one internal factor, while SVS be treated as an external factor. As noted by Hidi and Renninger (2006), the internal factor is underpinned by affect and emotion, while the external factor is situational that relates to environmental factors. The results for the two factors therefore should be interpreted separately.

Conclusion

This article provides a practical guide to science education researchers on how to use Rasch analysis to improve psychometric quality of SAS rating scales, and to encourage science education researchers to apply Rasch analysis to better assess psychometric quality of SAS rating scales as suggested by Boone et al. (2014), Liu (2010), and Neumann et al. (2011). As illustrated above, Rasch analysis can provide useful information that is typically unavailable from the conventional psychometric analysis, and such information can be very helpful for improving the psychometric quality of a rating scale.

It is also noted that science education researchers could be better off in using Rasch estimates for parametric statistical tests for greater accuracy of research finding (Harwell & Gatti, 2001). As explained by Boone et al. (2011),

> Across the broad landscape of science education, research plays a limited role in national, state, and local policies, programs, and practices because it lacks the credibility of strong explanatory and predictive power that result from strong connections between theory and research practice. If science education research is to gain credibility in the eyes of policy makers, program developers, and practitioners, research practice must tighten its connection to sound theory. (p. 259)

We echo their admonition and urge science education researchers to adopt better analytical practice, such as using Rasch estimates for statistical parametric tests and Rasch analysis for objective psychometric assessment.

Acknowledgement

We express our gratitude to Dr. William P. Fisher for his assistance in addressing the reviewers' comments and we appreciate the suggestions from the two anonymous reviewers that greatly improved the quality of the manuscript.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Martin, M. O., & Mullis, I. V. S. (Eds.). (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Andersen, L., & Chen, J. A. (2016). Do high-ability students disidentify with science? A descriptive study of U.S. ninth graders in 2009. *Science Education*, *100*(1), 57–77.

Bainer, D. L., & Smith, R. M. (1999). Developing a unidimensional instrument to measure the effectiveness of school-based partnerships. *Journal of Outcome Measurement*, *3*(3), 248–265.
Blalock, C. L., Lichtenstein, M. J., Owen, S., Pruski, L., Marshall, C., & Toepperwein, M. (2008). In pursuit of validity: A comprehensive review of science attitude instruments 1935–2005. *International Journal of Science Education*, *30*(7), 961–977.
Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.
Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model* (3rd ed.). New York, NY: Routledge, Taylor & Francis.
Boone, W. J., Staver, J. S., & Yale, M. S. (2014). *Rasch analysis in human sciences*. Dordrecht: Springer.
Boone, W. J., 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*(2), 258–280.
Chang, C.-Y., & Cheng, W.-Y. (2008). Science achievement and students' self-confidence and interest in science: A Taiwanese representative sample study. *International Journal of Science Education*, *30*(9), 1183–1200.
Chang, M. L., & Engelhard, G. (2016). Examining teachers' sense of efficacy scale at the item level with Rasch measurement model. *Journal of Psychoeducational Assessment*, *34*(2), 177–191.
Harwell, M. R., & Gatti, G. G. (2001). Rescaling ordinal data to interval data in educational research. *Review of Educational Research*, *71*, 105–131.
Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist*, *41*(2), 111–127.
Jocz, J. A., Zhai, J., & Tan, A. L. (2014). Inquiry learning in the Singaporean context: Factors affecting student interest in school science. *International Journal of Science Education*, *36*(15), 2596–2618.
Kind, P. M., Jones, K., & Barmby, P. (2007). Developing attitudes towards science measures. *International Journal of Science Education*, *29*(7), 871–893.
Lee, Y. S., Grossman, J., & Krishnan, A. (2008). Cultural relevance of adult attachment: Rasch modeling of the revised experiences in close relationships in a Korean sample. *Educational and Psychological Measurement*, *68*(5), 824–844.
Leung, F. K. S. (2002). Behind the high achievement of East Asian students. *Educational Research and Evaluation*, *8*(1), 87–108.
Linacre, J. M. (2004). Optimal rating scale category effectiveness. In E. V. Smith, Jr. and R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 258–278). Maple Grove, MN: JAM Press.
Linacre, J. M. (2014). *WINSTEPS (Version 3.81.0)* [Computer Software]. Chicago, IL: Winsteps.com.
Linacre, J. M., & Wright, B. D. (1989). The 'length' of a logit. *Rasch Measurement Transactions*, *3*(2), 54–55.
Liu, X. (2010). *Using and developing measurement instruments in science education: A Rasch Modeling approach*. Charlotte, NC: Information Age.
Masters, G. N. (1988). Item discrimination: When more is worse. *Journal of Educational Measurement*, *25*, 15–29.
Michael, O. M., Mullis, I. V. S., Foy, P., & Stance, G. M. (2012). *TIMSS 2011 International results in science*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.
Murphy, C., & Beggs, J. (2003). Children's perceptions of school science. *School Science Review*, *84* (308), 109–116.
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.
Noll, V. H. (1935). Measuring the scientific attitude. *Journal of Abnormal and Social Psychology*, *30*, 145–154.

Oon, P. T., & Subramaniam, R. (2013). Singapore school students' views about physics according to whether they intend to choose this subject as a tertiary field of study: A Rasch Analysis. *International Journal of Science Education*, *35*(1), 86–118.
Osborne, J., Simon, S., & Collins, S. (2003). Attitudes towards science: A review of the literature and its implications. *International Journal of Science Education*, *25*(9), 1049–1079.
Potvin, P., & Hasni, A. (2014). Interest, motivation and attitude towards science and technology at K-12 levels: A systematic review of 12 years of educational research. *Studies in Science Education*, *50*(1), 85–129.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Paedagogiske Institut.
Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.
Romine, W. L., & Walter, E. M. (2014). Assessing the efficacy of the measure of understanding of macroevolution as a valid toold for undergraduate non-science majors. *International Journal of Science Education*, *36*(17), 2872–2891.
Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, *25*, 54–67.
Schumacker, R. E. (1996). Disattenuating correlation coefficients. *Rasch Measurement Transactions*, *10*(1), 479.
Smith, R. M. (1996). A comparison of methods for determining dimensionality in Rasch measurement. *Structural Equation Modeling: A Multidisciplinary Journal*, *3*, 25–40.
Sondergeld, T. A., & Johnson, C. C. (2014). Using Rasch measurement for the development and use of affective assessments in science education research. *Science Education*, *98*(4), 581–613.
Swarat, S., Ortony, A., & Revelle, W. (2012). Activity matters: Understanding student interest in school science. *Journal of Research in Science Teaching*, *49*(4), 515–537.
Tuan, H. L., Chin, C. C., & Shieh, S. H. (2005). The development of a questionnaire to measure students' motivation towards science learning. *International Journal of Science Education*, *27* (6), 639–654.
Velayutham, S., & Aldridge, J. M. (2013). Influence of psychosocial classroom environment on students' motivation and self-regulation in science learning: A structural equation modeling approach. *Research in Science Education*, *43*(2), 507–527.
Velayutham, S., Aldridge, J., & Fraser, B. (2011). Development and validation of an instrument to measure students' motivation and self-regulation in science learning. *International Journal of Science Education*, *33*(15), 2159–2179.
Wagler, A., & Wagler, R. (2013). Assessing the lack of measurement invariance for the measure of acceptance of the theory of evolution. *International Journal of Science Education*, *35*(13), 2278–2298.
Wang, T. L., & Berlin, D. (2010). Construction and validation of an instrument to measure Taiwanese elementary students' attitudes toward their science class. *International Journal of Science Education*, *32*(18), 2413–2428.
Wright, B. D. (1999). Fundamental measurement for psychology. In S. E. Embretson& S. L. Hershberger (Eds.), *The new rules of measurement: What every educator and psychologist should know* (pp. 65–104). Hillsdale, NJ: Lawrence Erlbaum.
Wright, B. D., & Linacre, J. M. (1992). Combining and splitting categories. *Rasch Measurement Transactions*, *6*, 233–235.
Wright, B. D., & Linacre, J. M. (1996). Reasonable mean-square fit values, Part 2. In J. M. Linacre (Ed.), *Rasch Measurement Transactions* (p. 370). Chicago, IL: Mesa Press.
Wright, B. D., & Master, G. N. (1982). *Rating scale analysis*. Chicago, IL: Mesa Press.
Zenisky, A., Hambleton, R. K., & Robin, F. (2004). DIF detection and interpretation in large-scale science assessments: Informing item writing practices. *Educational Assessment*, *9*(1–2), 61–78.