Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 26 Number 18, August 2021 ISSN 1531-7714

Is it Suitable to Use the Same Categorization in Rating Scales When Applied to Students with Distinctive Levels of Achievement?

Mutasem M. Akour. *Department of Educational Psychology, Faculty of Educational Sciences, The Hashemite University, Zarqa, Jordan*
Hind Hammouri. *Department of Educational Psychology, Faculty of Educational Sciences, The Hashemite University, Zarqa, Jordan*
Saed, Sabah. *Department of Teaching and Curriculum*, *Faculty of Educational Sciences, The Hashemite University, Zarqa, Jordan*
Hassan Alomari, *University of Jordan, Jordan*

This study examined the efficiency of using the same rating scale categories in measuring affective constructs for students with distinctive levels of achievement. Data used in this study came from the Trends in Mathematics and Science Study (TIMSS) 2011, as a case, on the three scales that were designed to measure eighth graders' attitudes towards science. Data from the four higher and the four lower science performing countries were analyzed using Rasch model. Results revealed that the use of a four-point rating scale appears to be appropriate for some of the higher performing countries; however, it was not appropriate for the lower performing countries. In addition, category functioning and distances between threshold estimates differed by whether the country was a higher or a lower performing country; distances between threshold estimates were too close for the lower performing countries as compared to the higher performing countries. The findings of the current study question the utility of using these scales for a cross-national sample and deducing results concerning the samples' agreeability to the construct of attitudes towards science.

Likert-type rating scales are widely used in measuring several constructs in the educational and psychological sciences, such as attitudes, anxiety, personality traits, etc. These types of scales provide researchers with several features. For example, it enables researchers to assign several possible answers to each question. In addition, it requires all respondents to use the same stimuli when formulating their responses. Lopez (1996) asserted that respondents should be able to distinguish the response levels of each rating scale and to provide a clearly hierarchical ordering of the rating scale categories, so that it is possible to locate them at separate locations along the variable of interest. However, respondents may not use rating scales as intended by scale developers. Respondents may choose socially acceptable answers, misinterpret vague contents, or fall into a response set. Moreover, respondents may differ in interpreting a given rating scale in terms of their own understanding of the response labels (Smith, Wakely, de Kruif, & Swartz, 2003).

The number of response categories used in Likert scales usually affects the psychometric properties of the scale. Reliability and validity are two forms of evidence that can provide scale developers with some insight into

the optimal number of response options (Maitland, 2009). Reliability refers to the precision of scores obtained from a given scale. It quantifies scores' variations across replications of the scale at different points in time, or the consistency over multiple questions on a single occasion (Haertel, 2006). Several studies (e.g., Lozano, Garicai-Cueto, & Muniz, 2008; Muñiz, García-Cueto, & Lozano, 2005; Weng, 2004) attempted to explore the effect of assigning different numbers of response options on reliability. They found that increasing the number of response alternatives affected reliability positively. However, no significant gain resulted when the number increased beyond four options.

On the other hand, validity refers to the extent to which evidence and theory support the interpretation of scale scores for proposed uses (American Educational Research Association [AERA] et al., 2014). Validity studies in this regard were less common in the literature as compared to reliability studies. Some studies examined how changing the number of response categories would affect validity. For example, Lozano et al. (2008) found that increasing the number of response options from two to nine options resulted in higher values for Cronbach's alpha, and hence reliability was improved. Moreover, increasing the number of response options resulted in higher percentages of the explained variance, and hence factorial validity was improved. They concluded that the optimum number of options is between four and seven.

Likert data fall within the ordinal level of measurement; it is assumed that the value of each category is higher than the previous category but by an unspecified amount since intervals between values cannot be presumed equal (Jamieson, 2004 . Rasch model (Rasch, 1960) transforms Likert data into interval scales as logarithmic values of the odds (logits). Thus, differences between response choices become mathematically meaningful, as a necessary condition for computing statistics that assume interval data (Bond & Fox, 2015). Using Rasch model allows for the indication that a person endorsing a more extreme item in a scale should also endorse all fewer extreme items. Similarly, all respondents are expected to highly rate any easy-to-endorse item (Wright and Masters, 1982).

Several studies used Rasch model in optimizing rating scale categories in terms of collapsing response categories. For example, Smith et al. (2003) used Rasch measurement to optimize the number of points on a writing self-efficacy scale for students in the fourth and fifth grades. They found that collapsing the 10-point scale into a more meaningful 4-point scale best fitted the data. In another study, Royal et al. (2010) used Rasch model on a 5-point instrument that was administered to undergraduate students. They revealed that collapsing a 5-point rating scale into a 4-point scale improved measurement quality as compared to collapsing the scale into a 3-point one. On the other hand, when Daher, Ahmad, Winn, and Selamat (2015) applied Rasch model to data resulted from administering a spiritual well-being scale on a sample of adolescents, they found that using six categories resulted in better fit statistics and item reliability as compared to using three and four categories. Moreover, Colvin and Gorgun (2020) compared properties of scale categories when administering three forms of the Rosenberg Self-Esteem Scale that have 4, 6, and 8 response categories. They found that most of the psychometric properties were similar across the three variations. Based on these studies, it seemed that a minimum of a 4-point scale would be efficient to be used with school students.

Rating scales are commonly used in large-scale assessments, such as the Trends in Mathematics and Science Study (TIMSS). The typical administration of TIMSS produces a wealth of cognitive and noncognitive data in the fields of mathematics and science. Data on affective constructs collected by TIMSS are presented by means of attitudinal survey items in various formats to students, parents, and school personnel (Martin et al., 2012). Britton and Schneider (2007) emphasized that large scale assessments need to be fair for all students. The design of rating scales used in such assessments greatly affects the quality of the responses (Bond & Fox, 2015). Unless the rating scales that form the basis of data collection are functioning effectively, any conclusions based on those data will be insecure (Linacre, 2002). However, in large-scale assessments respondent samples are different. Royal, Ellis, Ensslen, and Homan (2010) asserted that in this case it is not possible to find one solution for choosing the ideal rating scale. Accordingly, an investigation into the efficiency of rating scale categories across samples is merited and needed.

In TIMSS 2011 and for the eighth grade, Singapore, Chinese Taipei, Korea, and Japan were the four highest achieving countries in science who showed a substantial difference in achievement as compared to the lowest

achieving countries, Ghana, Qatar, Oman, and Palestinian National Authority (Martin, Mullis, Foy, & Stanco, 2012). Therefore, the current study examined the efficiency of using the same number of categories in the rating scales in these two distinct and divergent groups of science performers.

The current study

In survey research, it is important to develop a survey that uses clear terminology and language so that each item transfer the same meaning to the respondents. Moreover, respondents should be able to clearly identify the ordered nature of the rating scale categories, and to distinguish the differences between each category. In practice, it is sometimes challenging to fulfill these requirements given that surveys have different shapes and sizes. For example, determining the number of response options would be problematic since using few response options is risky in that it could lead to inaccurate results, while using too many options could confuse the respondents. Introducing more alternatives that the respondents could not differentiate would introduce error variance in the model, and thus lead to lower accuracy.

On the other hand, participants may feel more comfortable with uneven and larger number of response categories since they do not have to expose themselves to a given choice. Therefore, it would be difficult to choose the ideal rating scale in large-scale assessments given the heterogeneity of the respondent samples (Royal et al., 2010). The current study used Rash measurement in exploring the functioning of rating scales used in collecting data for attitudinal surveys utilized in a large-scale assessment, i.e., TIMSS.

The current study was motivated by a finding from Sabah, Hammouri, and Akour (2013), where they validated a scale of attitudes toward science in TIMSS 2007 using Rasch model across different countries. Their study revealed that the attitudes toward science scale did not function as expected with the low achieving countries. Although great care has been taken to develop rating scales by TIMSS developers, the assumptions about both the quality of the measure and the utility of the rating scale in facilitating interpretable measures should be tested empirically (Bond & Fox, 2015). Therefore, the current study hope to provide scale developers with an insight into the functioning of using the same categorization schema in a scale that would be applied to students with heterogeneous achievement levels. This might help scale developers in selecting the more efficient categorization that would best fit all students, and that might elevate the reliability of the scores and the validity of the inferences made upon these scores. In addition, given that previous research (e.g., Weng, 2004; Lozano et al., 2008) suggested that using four to seven categories would optimize validity and reliability, the current study examined if this holds for international and large-scale assessments when applied to respondents with divergent levels of achievement.

Method

Participants

In TIMSS 2011, 63 countries and 14 regional benchmarking jurisdictions participated in the eighth-grade assessment, where 29 of them teach science as a general or an integrated subject. TIMSS was applied to ninth graders in three countries (known as "out of grade" countries) (Foy et al., 2013).

The sample of the present study consisted of 3200 students participated in TIMSS 2011. This sample was selected as follows. First, data for the four highest-achieving (HA) countries (Singapore, Chinese Taipei, Korea, and Japan) and the four lowest-achieving (LA) countries (Ghana, Qatar, Oman, and Palestinian National Authority) were selected. Second, wherever there are students with missing data on any of the 20 items of the "attitudes toward science" scale, their responses on all items were deleted. Third, Linacre (2002) mentioned that a sufficient sample size needed to provide stable item and person estimates could be as many as 100*(m+1) subjects, where (m+1) indicates the number of categories; accordingly, a random sample of 400 students (with no missing data) were selected from each country to form the data of the present study, resulting in 1600 students from the highest-performing countries and 1600 students from the lowest-performing countries.

Instruments

TIMSS 2011 used three scales to measure eighth graders' attitudes toward science (Martin & Mullis, 2012). These scales are Students Like Learning Science (SLS) scale, Students' Confident in Science ability (SCS) scale, and Students' Valuing Science (SVS) scale. These

scales have 20 items in total, resulting in 8000 pieces of item-level data within each country.

The first scale, SLS scale, consisted of five items: (1) I enjoy learning science; (2) I wish I did not have to study science; (3) Science is boring; (4) I learn many interesting things in science; and (5) I like Science. However, the SCS scale consisted of nine items: (1) I usually do well in science; (2) Science is more difficult for me than for many of my classmates; (3) Science is not one of my strengths; (4) I learn things quickly in science; (5) Science makes me confused and nervous; (6) I am good at working out difficult science problems; (7) My teacher thinks I can do well in science; (8) My teacher tells me I am good at science; and (9) Science is harder for me than any other subject.

The third scale, SVS scale, collected students' responses to six items: (1) I think learning science will help me in my daily life; (2) I need science to learn other school subjects; and (3) I need to do well in science to get into the University of my Choice; (4) I would like to do well to get the job I want; (5) I would like a job that involves using science; and (6) It is important to do well in science.

TIMSS scales utilize a four-point Likert response scale. This response type does not allow students to select a neutral response. The categories in this scale were (agree a lot=4, agree a little =3, disagree a little=2, and disagree a lot=1). Students were asked to indicate how much they agree or disagree with each item (statement) by filling the circle of one of these categories. In SLS scale, two out of five items were negatively worded; in SCS scale four out of nine items were negatively worded, while in SVS scale none of the items were negatively worded.

Data Analysis

Rasch Rating Scale Model (Andrich, 1978) analysis was performed using WINSTEPS computer program (Linacre, 2005b). Data for each country were analyzed separately. The following two preliminary steps were performed before analyzing data. The responses to the six negative worded items were reverse coded, and point–measure correlations were examined to ensure that all items were oriented in the same direction on the latent variable. It was assumed that the categories implement a clearly defined, conceptually exhaustive ordered sequence.

Assessing unidimensionality is another important assumption of Rasch model. In the present study, principal components analysis of residuals was used to examine whether the items within each scale measure one dimension. Each of the three scales was considered a unidimensional scale when the unexplained variance (after removal of the first latent variable) on any secondary dimension is less than 2 in eigenvalue units, and less than three items load on that dimension (Linacre, 2017).

To examine whether the responses provided by examinees to each of SLS, SVS, and SCS rating scales were functioning as intended by item developers, we followed the guidelines outlined in Linacre (2002) and Bond and Fox (2015). First, each rating category should contain a minimum of 10 observations to provide stable estimates. Second, the shape of the distribution of category frequencies is uniform, which is optimal for step calibration; other substantively meaningful distributions include unimodal. Third, average measures should advance monotonically with rating scale category values. Fourth, unweighted mean square fit statistics (outfit MNSQ) of each rating scale were less than 2.0. Values of MSNQ greater than 2 indicate that there is too much unexplained variance in the data. Higher values of MSNQ associated with a given response category suggest that the category has been used by respondents in unexpected contexts. Fifth, the thresholds indicate a hierarchical pattern to the rating scale, and magnitudes of the distances between adjacent category thresholds should be at least 1.4 logits and no more than 5 logits apart.

Moreover, to inspect the distinctions between thresholds visually, we plotted probability curves which show the probability of endorsing a given rating scale category for every agreeability-endorsability difference estimate, for each of the three scales and for each of the eight countries.

Furthermore, to determine if there are enough items in each scale that spread along the continuum and enough spread of ability among persons, person reliability and separation indices were computed for the scores on each scale. Moreover, item reliability and separation indices indicate whether item estimates would remain stable if other respondents were given the same items. It is important to note that Rasch model-based reliabilities underestimate classical reliability coefficients, because Rasch model treats data as discrete rather than

continuous (Bond and Fox, 2015). Person reliability greater than or equal 0.80, person separation index greater than or equal 2, item reliability greater than or equal 0.90 and item separation index greater than or equal 3, were considered adequate (Linacre, 2005a).

Results

The purpose of the present study was to investigate the effectiveness of rating scales utilized in large scale assessments in measuring affective constructs for students with distinctive levels of achievement. Data (1600 respondents from the HA countries and 1600 respondents from the LA countries) were evaluated by country individually. Following are the results of the analyses.

Dimensionality

Principal component analysis of the standardized residuals on each of the three scales that measure students' attitudes toward science in TIMSS 2011 showed that the unexplained variance in first contrast was less than 1.2 in eigenvalue units. This indicates that, for each scale, almost only one item loaded on that secondary dimension. Since Linacre (2017) asserted that at least three items should load on any secondary dimension to treat it as a meaningful one, it was inferred that all items on each scale fulfilled the assumption of unidimensionality.

Table 1 shows that all point measure correlations were all positive and ranged from 0.61 to 0.90 for HA countries and from 0.41 to 0.82 for LA countries. This result indicated item-level polarity, meaning that all items were oriented with related latent variables.

Table 1. Point measure correlations for each item across all countries

Scale	Item	High-achieving countries				Low-achieving countries			
		Singapore	Chinese Taipei	Korea	Japan	Ghana	Qatar	Oman	Palestine
SLS	1	0.82	0.89	0.90	0.87	0.48	0.75	0.58	0.71
	2	0.71	0.82	0.81	0.81	0.47	0.71	0.56	0.64
	3	0.85	0.89	0.89	0.89	0.52	0.79	0.61	0.74
	4	0.80	0.82	0.79	0.74	0.72	0.66	0.69	0.66
	5	0.84	0.84	0.84	0.79	0.73	0.78	0.71	0.73
SCS	1	0.79	0.84	0.82	0.78	0.44	0.55	0.41	0.52
	2	0.79	0.84	0.78	0.79	0.48	0.65	0.52	0.61
	3	0.74	0.81	0.76	0.76	0.47	0.60	0.52	0.66
	4	0.72	0.74	0.74	0.68	0.49	0.55	0.52	0.53
	5	0.70	0.80	0.77	0.63	0.49	0.62	0.51	0.55
	6	0.72	0.83	0.76	0.76	0.62	0.64	0.58	0.64
	7	0.75	0.82	0.81	0.82	0.60	0.63	0.56	0.60
	8	0.66	0.74	0.61	0.69	0.61	0.62	0.55	0.56
	9	0.73	0.75	0.72	0.72	0.63	0.61	0.54	0.59
SVS	1	0.71	0.78	0.77	0.76	0.54	0.70	0.53	0.58
	2	0.71	0.80	0.79	0.76	0.59	0.78	0.62	0.60
	3	0.79	0.84	0.80	0.78	0.67	0.78	0.72	0.69
	4	0.81	0.85	0.82	0.81	0.68	0.81	0.71	0.67
	5	0.78	0.80	0.80	0.82	0.70	0.82	0.72	0.70
	6	0.69	0.80	0.78	0.72	0.75	0.74	0.75	0.74

Moreover, the following calculations were performed: category frequencies, average measures, outfit MNSQ, thresholds calibration, and person and item separation indices and reliabilities for the three scales (SLS, SVS, and SCS) for each of the HA and LA countries. These results are demonstrated in Tables 2, 3, and 4 that present a summary of the diagnostic indicators of the rating scales functioning.

Table 2. Diagnostics for the SLS Rating Scale by Country (5 items)

	Country	Category Label[a]	Observed Count [b]	Average measure [c]	MNSQ Outfit	Threshold calibration	Person Separation Reliability	Item
Highest achieving countries	Singapore	1	830	-3.64	1.09	None	1.61	5.68
		2	864	-1.41	1.04	-3.35	0.72	0.97
		3	223	0.69	0.73	1.00		
		4	83	1.68	1.29	2.35		
	Chinese Taipei	1	433	-3.18	1.20	None	1.95	4.38
		2	814	-1.22	0.89	-3.57	0.79	0.95
		3	478	1.16	0.76	0.58		
		4	275	2.59	1.07	2.99		
	Korea	1	298	-4.45	1.26	None	2.25	4.18
		2	838	-1.55	0.88	-4.69	0.83	0.95
		3	687	1.53	0.81	0.33		
		4	177	3.91	0.95	4.37		
	Japan	1	350	-2.96	1.33	None	2.03	3.98
		2	838	-1.17	0.84	-3.34	0.81	0.94
		3	539	1.02	0.88	0.46		
		4	273	2.90	0.87	2.88		
Lowest achieving countries	Ghana	1	1311	-1.44	1,19	None	0.47	9.18
		2	317	-0.91	1.33	-0.32	0.18	0.99
		3	192	0.15	0.52	0.07		
		4	180	0.43	0.99	0.25		
	Qatar	1	967	-1.52	1.13	None	1.11	5.57
		2	496	-0.73	0.99	-1.04	0.55	0.97
		3	295	0.18	0.60	0.20		
		4	242	0.70	1.28	0.84		
	Oman	1	1277	-1.29	1.04	None	0.47	6.14
		2	397	-0.81	1.15	-0.66	0.18	0.97
		3	192	0.04	0.59	0.27		
		4	134	0.11	1.08	0.39		
	Palestine	1	1065	-1.27	1.09	None	0.98	4.50
		2	454	-0.62	1.09	-0.77	0.49	0.95
		3	269	-0.02	0.78	0.19		
		4	212	0.59	1.05	0.58		

[a] agree a lot=4, agree a little =3, disagree a little=2, and disagree a lot=1.

[b] Observed count for each response category on the scale (sum of observed count for each country=number of items*400).

[c] The average of measures across all observations in each category.

Table 3. Diagnostics for the SVS Rating Scale by Country (6 items)

	Country	Category Label [a]	Observed Count [b]	Average measure [c]	MNSQ Outfit	Threshold calibration	Person Separation Reliability	Item
Highest achieving countries	Singapore	1	1008	-3.58	0.98	None	1.60	9.32
		2	957	-1.43	0.86	-2.77	0.72	0.99
		3	331	0.27	0.98	0.45		
		4	104	1.41	1.65	2.32		
	Chinese Taipei	1	469	-3.15	1.26	None	2.16	11.78
		2	739	-1.14	0.86	-2.80	0.82	0.99
		3	780	0.91	0.95	-0.10		
		4	41	2.98	0.96	2.90		
	Korea	1	489	-3.11	1.02	None	1.85	8.42
		2	969	-1.14	0.97	-3.18	0.77	0.99
		3	784	0.77	0.87	0.0		
		4	158	2.37	1.31	3.18		
	Japan	1	414	-2.77	1.13	None	1.96	11.76
		2	709	-0.98	0.85	-2.51	0.79	0.99
		3	870	0.78	0.92	-0.31		
		4	407	2.85	1.07	2.82		
Lowest achieving countries	Ghana	1	1850	-2.25	1.02	None	0.79	4.44
		2	375	-1.21	0.75	-1.07	0.38	0.95
		3	115	-0.16	0.76	0.38		
		4	60	0.04	1.80	0.70		
	Qatar	1	1307	-2.30	0.98	None	1.48	5.80
		2	597	-0.97	1.00	-1.39	0.69	0.97
		3	281	0.13	0.83	0.27		
		4	215	0.99	1.47	1.13		
	Oman	1	164	-2.33	1.00	None	1.08	5.44
		2	496	-1.23	-.87	-1.22	0.54	0.97
		3	165	-0.04	0.64	0.36		
		4	91	0.55	1.75	0.86		
	Palestine	1	1454	-2.27	1.06	None	1.12	7.13
		2	595	-1.10	0.90	-1.24	0.65	0.98
		3	223	-0.11	0.92	0.35		
		4	128	0.64	1.12	0.89		

[a] agree a lot=4, agree a little =3, disagree a little=2, and disagree a lot=1.
[b] Observed count for each response category on the scale (sum of observed count for each country=number of items*400).
[c] The average of measures across all observations in each category.

Table 4. Diagnostics for the SCS Rating Scale (9 items) by country

	Country	Category Label [a]	Observed Count [b]	Average measure [c]	MNSQ Outfit	Threshold calibration	Person Separation Reliability	Item
Highest achieving countries	Singapore	1	745	-2.55	1.08	None	2.32	4.87
		2	1472	-0.90	0.86	-2.49	0.84	0.96
		3	1061	0.43	0.93	0.07		
		4	322	1.85	1.30	2.42		
	Chinese Taipei	1	423	-2.45	1.63	None	2.68	8.53
		2	970	-0.77	0.80	-2.67	0.87	0.99
		3	1335	1.17	0.76	-0.08		
		4	872	3.01	1.12	2.75		
	Korea	1	344	-3.14	1.47	None	2.54	10.58
		2	1184	-1.12	0.83	-3.54	0.87	0.99
		3	1642	1.28	0.89	-0.20		
		4	430	3.50	1.06	3.74		
	Japan	1	331	-2.51	1.30	None	2.49	12.71
		2	975	-0.79	0.77	-2.68	0.86	0.99
		3	1416	1.29	0.93	-0.11		
		4	878	3.26	1.17	2.79		
Lowest achieving countries	Ghana	1	1670	-1.17	1.04	None	1.24	7.38
		2	909	-0.64	1.13	-0.52	0.61	0.98
		3	585	0.03	0.64	0.13		
		4	436	0.23	1.25	0.39		
	Qatar	1	1482	-1.49	1.06	None	1.55	5.43
		2	1054	-0.66	0.90	-0.89	0.71	0.97
		3	675	0.01	0.81	0.10		
		4	389	0.33	1.31	0.79		
	Oman	1	1705	-1.45	0.98	None	1.22	5.67
		2	1026	-0.74	0.99	-0.73	0.60	0.97
		3	590	-0.10	0.63	0.07		
		4	279	-0.14	1.50	0.67		
	Palestine	1	1460	-1.43	1.10	None	1.49	6.99
		2	1045	-0.68	0.97	-0.88	0.69	0.98
		3	718	-0.02	0.81	0.02		
		4	377	0.43	1.16	0.86		

[a] agree a lot=4, agree a little =3, disagree a little=2, and disagree a lot=1.
[b] Observed count for each response category on the scale (sum of observed count for each country=number of items*400).
[c] The average of measures across all observations in each category.

At Least 10 Observations of Each Category

As Tables 2, 3, and 4 show, for the three scales (SLS, SVS, and SCS) in the selected countries each category frequency exceeds 10 responses that endorsed a particular category. That is, respondents endorsed each category with satisfactory frequency so that all rating scale categories were stable, and there were no need for the collapsing of any two adjacent categories into a single more-stable category.

Regular Observation Distribution

Tables 2, 3, and 4 revealed that the shape of distributions of category frequencies for the four HA countries met the guideline for effective rating scale; each distribution is unimodal (Linacre, 2002) suggesting that students in HA countries used the rating scale categories as intended. Although the distributions of category frequencies within each scale for the four LA countries were unimodal, these distributions were not symmetrical. Each distribution did not show smooth increases from a category to another. For example, in Ghana, the category frequencies for SLS scale decreases from 1311 to 317 (a difference of 994), then from 317 to 192 (a difference of 125), and then from 192 to 180 (a difference of 12). This implies that category frequencies are not nearly equal, for all three scales.

Average Measures Advance Monotonically with Category

Average measures were functioning as expected (increasing monotonically across the three rating scales) in all selected HA and LA countries except in Oman, it fails for the SCS scale as Table 4 shows; it is not ordered. It is ascending from category 1 to 3 then descending from category 3 to 4. The average measure for category 4 was recorded as (-0.14) noticeably less than (-0.10) for category 3. Meaning that, on average, Omani' students with more agreeability of science confidence endorsed the higher category. Saying it differently, students choosing category 4 are less agreeable, on average, than students choosing category 3.

OUTFIT Mean-Squares Less than 2.0

Tables 2 through 4 reveal that all outfit MNSQs associated with all categories for the three scales in all selected countries were less than 2. This suggests that there is a reasonable uniform level of randomness in the "attitudes toward science" data.

At Least 10 Observations of Each Category

As Tables 2, 3, and 4 show, all step calibrations advance with the categories for each of the three scales for all selected countries. However, for the HA countries, step calibrations advance a distance that ranged from 1.35 to 5.02 logits for SLS; from 1.87 to 3.22 logits for SVS; and from 2.35 to 3.94 logits for SCS. This advance met the guideline, it is more than or equal 1.4 and less than or equal to 5 logits (Linacre, 2002), except for Singapore and Korea in SLS scale. In Singapore, it is less than the lower limit (1.35); the redefining of these two (1 &2) categories to have wider substantive meaning or combining categories may be indicated. Whereas, in Korea, it exceeds the upper limit (5.02).

Figures 1 to 6 (see Appendix A) provide the visual method to examine the probability curves. The horizontal axis represents the difference between a person's "attitude" and the item's affective value. The probability that a respondent would endorse any one category is visually presented in these figures. These figures show that each category had a distinct peak in the probability curve graph, for HA countries, illustrating that each is indeed the most probable response category for some portion of the "attitudes toward science" variable. Figures 1 to 6 indicate that each step defines a distinct position on the variables (SLS, SVS, SCS) for each of the HA countries, indicating that each of the rating scales had been employed by students in a manner consistent with the intentions of the scales' developers. Whereas, for the LA countries, categories observed to be too close on the graph.

For the LA countries, step calibration advance a distance range from 0.18 to 1.24 logits for SLS scale; from 0.32 to 1.66 logits for SVS; and from 0.26 to 0.99 logits for SCS. Results for SLS and SCS did not meet the step advance limit for all LA countries (i.e., 1.4 logits ≤step advance≤ 5 logits: Linacre, 2002). However, for SVS scale, this guideline was met partially for these countries; the distances between the first and second step calibrations were more than or equal 1.4 logit. Whereas distances between the second and third thresholds were too close on the logit scale (less than 1.4 logit).

Person and Item Separation and Reliability

Results from Tables 2, 3, and 4 revealed that in HA countries, person separation indices ranged from 1.61 to 2.25 for SLS scale, from 1.60 to 2.16 for SVS scale, and from 2.32 to 2.68 for SCS scale. Person reliabilities

ranged from 0.72 to 0.83 for SLS scale, from 0.72 to 0.82 for SVS scale, and from 0.84 to 0.87 for SCS scale.

However, in LA countries, person separation indices range from 0.47 to 1.11 for SLS scale, from 0.79 to 1.48 for SVS scale, and from 1.22 to 1.55 for SCS scale. Person reliabilities range from 0.18 to 0.55 for SLS scale, from 0.38 to 0.69 for SVS scale, and from 0.60 to 0.71 for SCS scale. Both results indicate that the three scales were not reliable for LA countries, which indicates low variability of persons on the variables being measured in these countries (Green & Frantom, 2002).

Tables 2, 3, and 4 show that item separation and reliability met the criteria (i.e., separation > 3.0, reliability >.90; Linacre, 2002) for the three scales for all HA and LA countries. In HA countries, item separation ranged from 3.98 to 5.68 for SLS scale, from 8.42 to 11.78 for SVS scale, and from 4.87 to 12.71 for SCS scale. Whereas, in LA countries it ranged from 4.50 to 9.18 for SLS scale, from 4.44 to 7.13 for SVS scale, and from 5.43 to 7.38 for SCS scale. Item reliabilities were greater than 0.90 for the three scales in both HA and LA countries.

Discussion and Conclusions

This study evaluated the functioning of rating scales used in collecting data for attitudinal surveys utilized in a large-scale assessment (i.e., TIMSS) using Rash measurement. The examination of the functioning of the rating scales was done based on achievement; data from two groups of students who are distinctive in achievement were analyzed. The evaluation process was done based on the five guidelines outlined in Linacre (2002) and Bond and Fox (2015).

The findings revealed that the three scales met the requirement of having at least 10 observations in each response category for all countries with distinctive levels of achievement, meaning that respondents endorsed each category with satisfactory frequency. This indicates that locally stable estimates of the rating scale structure can be produced (Linacre, 2002).

The second requirement of the shape of the distribution of the category frequencies was met for the four HA countries. However, this guideline was not met for the four LA countries. This implies that each category is not contributing about equally to the measurement process (Linacre, 1999). This irregularity in observation frequencies across categories in LA countries may signal aberrant category usage (Linacre, 2002).

Regarding the third requirement, average measures were ordered and functioning as expected in almost all countries, even though the advances across categories for each of the three scales for all selected eight countries were uneven. Therefore, the increase in average measures with each successive rating point implies that higher attitudes toward science is associated with higher category labels. This requirement was not met in only one of the LA countries in one of the scales. The disordering occurred in the last category, which indicates that the fourth category did not represent more of the attitudes toward science than the third category. Accordingly, the meaning of the rating scale is uncertain for this data set, and consequently any derived measures are of doubtful utility (Linacre, 1999). This could be because the difference between a "disagree a little" and a "disagree a lot" may not be clear to these respondents.

Furthermore, the fourth requirement of idiosyncratic category use was met for all countries; outfit measures associated with all categories for the three scales in all selected countries, indicating that these categories were not used in unexpected contexts.

The requirement of ordered thresholds was met for all HA countries, except for two countries (Singapore and Korea) in one of the scales (SLS scale). In Singapore, it was less than the lower limit; this means that the redefining of the two categories to have wider substantive meaning or combining categories may be indicated. Whereas, in Korea, it exceeds the upper limit. Although, the difference is not large, it could be due to sampling error, or it could signal something in the scale, because when a category represents a very wide range of performance so that its category boundaries are far apart, then a "dead zone" develops in the middle of the category in which measurement loses its precision (Linacre, 2002). So, this scale may need a revision and thorough detection of the reasons why this result happens to assure that each step defines a distinct position on the variable and to avoid large gaps in the variable (Linacre, 1999). On the other hand, results for SLS and SCS did not meet the step advance limit for all LA countries, whereas it was met partially for SVS scale. The distances between the first and second step calibrations were more than or equal 1.4 logit, whereas distances between the second and third thresholds were too close on the logit scale. Accordingly, redefining the

categories for the LA countries to have wider substantive meaning or combining categories may be indicated (Linacre, 2002).

For some of the HA countries, person separation and person reliability for two scales, SLS and SVS, were lower than required. This indicates that both SLS when applied in Singapore and Chinese Taipei, and SVS when applied in Singapore, Korea, and Japan may not be sensitive enough to distinguish between high and low performers meaning that more items may be needed (Linacre, 2005a). Similarly, the three scales were not reliable for LA countries, which means that they are not sensitive enough to differentiate students into several different levels with respect to measured construct. On the other hand, item separation and reliability met the criteria for the three scales for all HA and LA countries. This result implies that the person sample is large enough, in each country, to confirm the hierarchy of item's difficulty to "agree with" (i.e., construct validity of the instrument; Linacre, 2005a). Saying it differently, item reliabilities indicate that the three scales cover a broad range of item endorsability along the construct continuum.

Based on the findings of the present study, the use of a 4-point rating scale appeared to be appropriate for some of HA countries; however, it was not appropriate for LA countries. In addition, category functioning and distances between threshold estimates differed by whether the country is a higher or a lower achieving one. Distances between threshold estimates were too close in LA countries; it is indicative of an issue with drawing distinctions between the rating scale categories. Moreover, SCS scale in Omani' sample lacked ordering in the "average measure" values from category 3 to category 4, this result comments on the functioning of the rating scale for this sample. Whether category disordering is due to a misspecification of the rating scale or to idiosyncrasies only found in the sample requires further investigation.

Person separation and reliability were not adequate for SLS and SVS when applied in Singapore, Chinese Taipei, and Korea. They were not acceptable for all LA countries. Step calibration did not meet Linacre's (2002) guideline when applied to Singapore, Korea and all LA countries. These two results shed light on the reliability and effectiveness of the scale categorization for these two HA and the four LA countries.

To sum up, the findings of the present study revealed that the three scales were functioning well for Japan only. These scales did not meet some guidelines for the other three HA countries. However, these scales were not functioning effectively for the LA countries. This result questions the utility of using these scales for international sample and deducing results concerning the samples' endorsability or "agreeability to" the construct of "attitudes toward science". It is expected that scales' developers exerted times for constructing, trying out, analyzing and re-analyzing these scales. However, it seems that some factors affecting students' responses to these scales were missed.

The findings of the current study did not align with the findings from previous research in that using four categories or using the same number of categories was efficient for all students in the sample (Daher et al., 2015; Royal et al. 2010; Smith et al., 2003). One reason for that might be related to the heterogeneity of the sample used in the current study. Previous research indicated several factors affecting students' responses, such as differential stimulus familiarity, social desirability, and response style (Smith et al., 2003). Given the large number of nations involved in TIMSS, and the fact that we do not have any qualitative data available to help explain the results of the study, researchers are invited to understand what constitutes a mindset for choosing one category over another. Furthermore, to study thoroughly factors influencing students' responses in their contexts in a step to tailor these scales to the intended populations. Although it is not easy to have an optimal scale for a national sample, even, it is more difficult when having international sample, findings from the analyses described above provided insight for revising the "attitudes toward science" scale with the goal of elevating reliability and validity.

When attempting to use rating scales across cultures, several challenges may arise. Diverse cultures may differ in their tendency in endorsing items on rating scales. Chen et al. (1995) reported that Japanese and Chinses students were more likely than North American students to use the midpoint in scales. In the present study, students in LA countries were more likely than students in HA countries to use the "disagree" part of the scales, while students in HA countries were more likely to use the "agree" part of each scale. This may reflect cultural differences rather than achievement-based differences, and thus implies that further

investigation is needed using data from students with comparable achievement levels but from diverse cultures.

Another factor that may have affected the efficiency of the categorization is the differentiation in reading level. Students in HA countries have better reading levels as compared to those in LA countries. Better reading levels might result in better comprehension of the scales' item content, and therefore resulted in differentially response patterns. This indicate the need to, probably, give more attention to the issue of items' wording.

TIMSS have used negatively worded items in two scales, which is widely recommended. However, Suárez-Álvarez et, al. (2018) recommended using direct items when the lecture skills of the respondents are low. It is probable that students of the four lower performing countries would have lower lecture skills, which might affect, negatively, their responses. Therefore, TIMSS developers might have to consider this effect and design procedures to control the effect of this factor.

The results of the present study need to be reviewed with two limitations in mind. First, it was assumed that respondents (8th graders) provided sincere responses to the "attitudes toward science" scale. Rewards and punishments were not administered based on student performances on the test. Therefore, students had to rely on intrinsic motivation to express their conceptions of the construct. Second, the sample of study had responses with complete or no missing data. Accordingly, it could be subjected to "non-response error" that results from participants' lack of response to some or all the items on the scale (Creswell, 2005; Cui, 2003; Fraenkel & Wallen, 2009). Non-response error becomes a problem when the participants who do not respond to some items of the "attitudes toward science" scale may differ on the scale's measures or on science achievement from those who do (Cui, 2003). Therefore, the findings of the present study may not be generalizable to other larger samples of students who may have different types of missing data.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association. https://www.testingstandards.net/uploads/7/6/6/4/76643089/9780935302356.pdf

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561-573. https://doi.rg/:10.1007/BF02293814

Bond, T. & Fox, C. (2015). *Applying the Rasch model: fundamental measurement in the human sciences* (3rd Ed.). Routledge.

Britton, E. & Schneider, S. (2007). Large-scale assessments in science education. In S. Abell and N. Lederman (Eds.) *Handbook of research on science education*. Lawrence Erlbaum associates.

Buckley, J. (2009). *Cross-National Response Styles in International Educational Assessments: Evidence from PISA 2006*. https://edsurveys.rti.org/PISA/documents/Buckley_PISAresponsestyle.pdf

Chen, C., Lee, S., & Stevenson, H. (1995). Response style and cross-cultural comparisons f rating scales among east Asian and north American students. Psychological Science, 6(3), 170-175.

Colvin, K. F., and Gorgun, G. (2020). Collapsing Scale Categories: Comparing the Psychometric Properties of Resulting Scales. *Practical Assessment, Research & Evaluation, 25*(6). https://scholarworks.umass.edu/pare/vol25/iss1/6/

Creswell, J. (2005). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (2nd ed.). Pearson Merrill Prentice Hall.

Cui, C. (2003). Reducing error in mail surveys. *Practical Assessment, Research, & Evaluation, 8(*18). http://pareonline.net/getvn.asp?v=8&n=18/

Daher, A. M., Ahmad, S. H., Winn, T., & Selamat, M. I. (2015). Impact of Rating Scale Categories on Reliability and Fit Statistics of the Malay Spiritual Well-Being Scale using Rasch Analysis. *Malays J Med Sci, 22*(3), 48-55.

Foy, P., Arora, A., &. Stanco, G. (editors) (2013). *TIMSS 2011 User Guide for the International Database*. TIMSS & PIRLS International Study Center.

Fraenkel, J., & Wallen, N. (2009). *How to design and evaluate research in education* (7th ed.). McGraw-Hill.

Green, K. E., & Frantom, C. G. (2002, November). *Survey development and validation with the Rasch model.* Paper presented at the International Conference on Questionnaire Development, Evaluation, and Testing.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 65-110). American Council on Education and Praeger Publishers.

Jamieson S. (2004). Likert scales: how to (ab)use them. *Medical education*, *38*(12), 1217–1218. https://doi.org/10.1111/j.1365-2929.2004.02012.x

Linacre JM. (1995). Misfit Statistics for Rating Scale Categories. *Rasch Measurement Transactions, 9*(3), 450.

Linacre, J. (1999). Investigating Rating Scale Category Utility. *Journal of Outcome Measurement, 3*(2), 103-122.

Linacre, J. (2002). Optimizing Rating Scale Category Effectiveness. *Journal of Applied Measurement, 3*(1), 85-106.

Linacre, J. (2005a). *Reliability and separation of measures. Help for Winsteps Rasch Measurement Software.* http://www.winsteps.com/winman/reliability.htm

Linacre, J. M. (2005b). *WINSTEPS Rasch measurement computer program*. Chicago, Winsteps.com.

Linacre, J. M. (2017). *A user's guide to Winsteps Ministeps Rasch-model computer programs [version 4.0.0].*

Lopez, W (1996/). Communication Validity and Rating Scales *Rasch Measurement Transactions, 10*(1), 482-483.

Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 4*(2), 73–79. https://doi.org/10.1027/1614-2241.4.2.73

Martin, M. & Mullis, I. (Eds.). (2012). *Methods and procedures in TIMSS and PIRLS 2011.* TIMSS & PIRLS International Study Center.

Martin, M., Mullis, I., Foy, P., & Stanco, G. (2012). *TIMSS 2011 International Results in Science.* https://timssandpirls.bc.edu/timss2011/downloads/T11_IR_Science_FullBook.pdf

Maitland, A. (2009). How Many Scale Points Should I Include for Attitudinal Questions? *Survey Practi*ce, 2 (5). https://doi.org/10.29115/SP-2009-0023

Muñiz, García-Cueto, & Lozano. (2005). Item format and the psychometric properties of the Eysenck Personality Questionnaire. *Personality and Individual Differences, 38*(1), 61-69. https://doi.org/10.1016/j.paid.2004.03.021

Rasch, G. (1960) *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhagen: Institute for Educational Research.

Royal, K. D., Ellis, A., Ensslen, A., & Homan, A. (2010). Rating Scale Optimization in Survey Research: An Application of the Rasch Rating Scale Model. *Journal of Applied Quantitative Methods, 5*(4), 607-617.

Sabah, S., Hammouri, H., & Akour, M. (2013). Validation of A Scale of Attitudes Toward Science Across Countries Using Rasch Model: Findings From TIMSS. *Journal of Baltic Science Education, 12(* 5), 692-702.

Smith, E. V., Wakely, M. B., de Kruif, R. E. L., & Swartz, C. W. (2003). Optimizing Rating Scales for Self-Efficacy (and Other) Research. *Educational and Psychological Measurement, 63*(3), 369-391. https://doi.org/10.1177/0013164403063003002

Suárez-Álvarez, J., Pedrosa, I., Lozano, L. M., García-Cueto, E., Cuesta, M., & Muñiz, J. (2018). Using reversed items in Likert scales: A questionable practice. *Psicothema*, *30*(2), 149 158. https://doi.org/10.7334/psicothema2018.33

Wang, L. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. Educational and Psychological Measurement, 64(6), 956-972. https://doi.org/10.1177/0013164404268674

Wright, B.D. & Masters, G.N. (1982) *Rating Scale Analysis.* MESA Press.

Citation:

Akour, M. M., Hammouri, H., Sabah, S., & Alomari, H. (2021). Is it suitable to use the same categorization in rating scales when applied to students with distinctive levels of achievement? *Practical Assessment, Research & Evaluation*, *26*(18). Available online: https://scholarworks.umass.edu/pare/vol26/iss1/18/

Corresponding Author

Mutasem M. Akour
The Hashemite University
Zarqa, Jordan
email: mutasem [at] hu.edu.jo

Appendix A

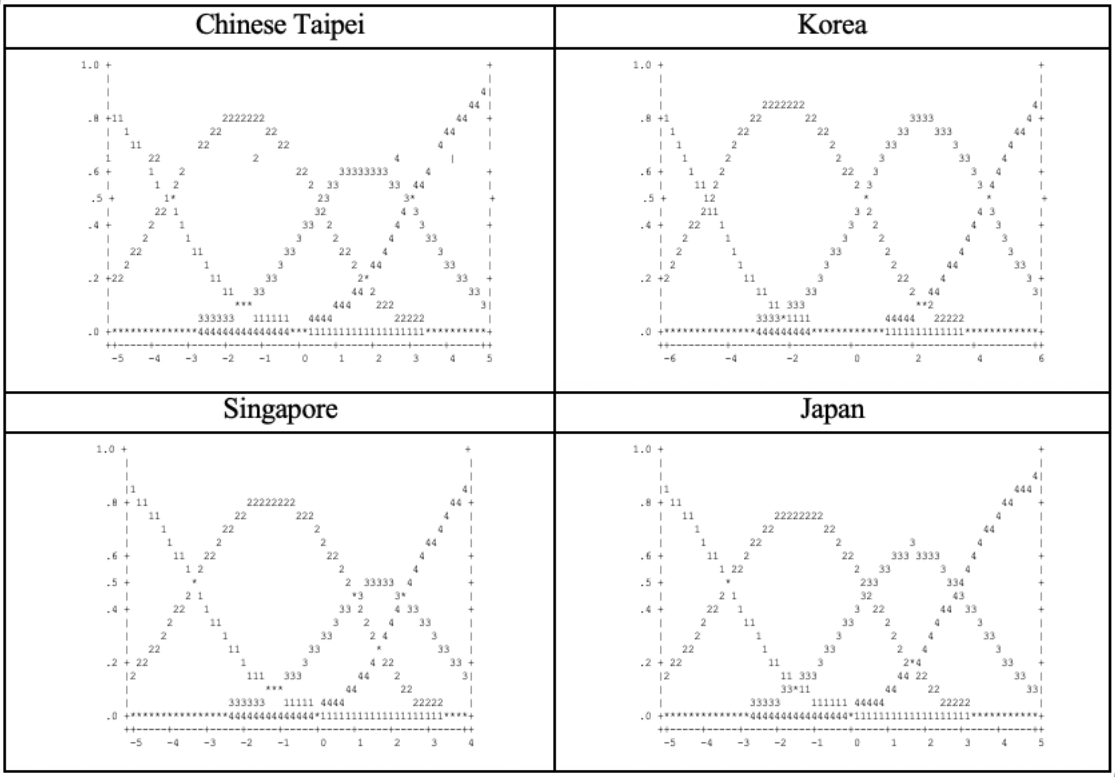

Figure1. Category Characteristic Curves for HA Countries for SLS Rating Scale

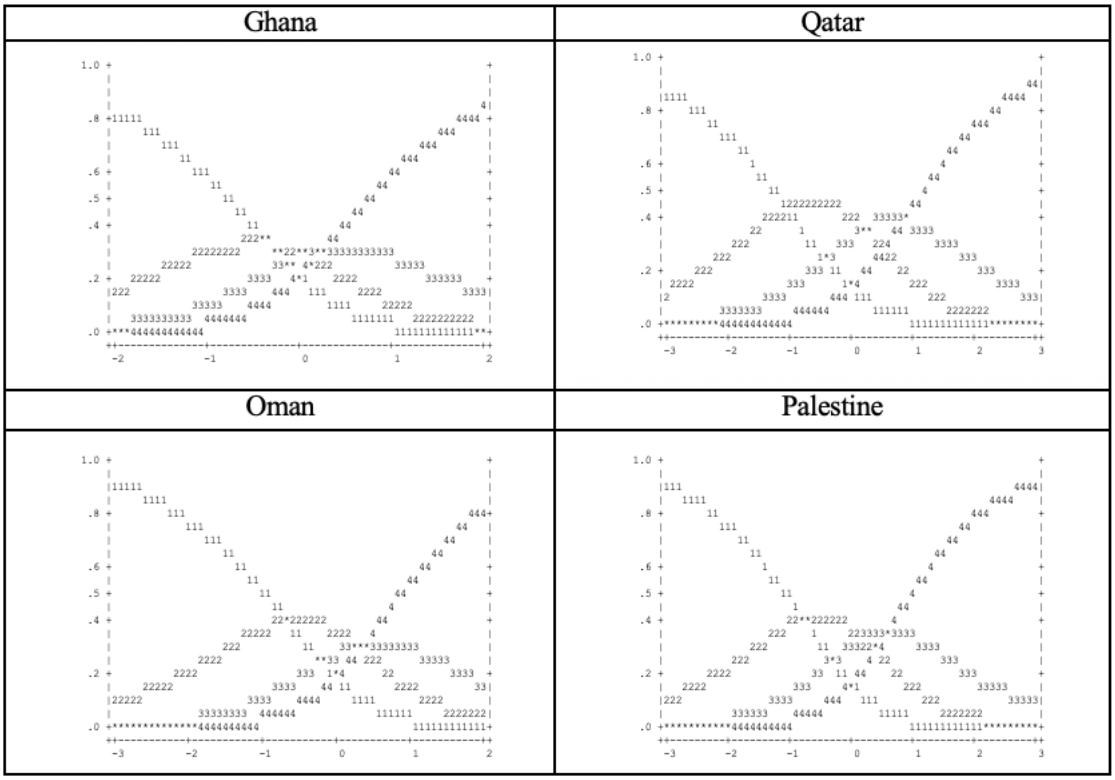

Figure 2. Category Characteristic Curves for LA Countries for SLS Rating Scale

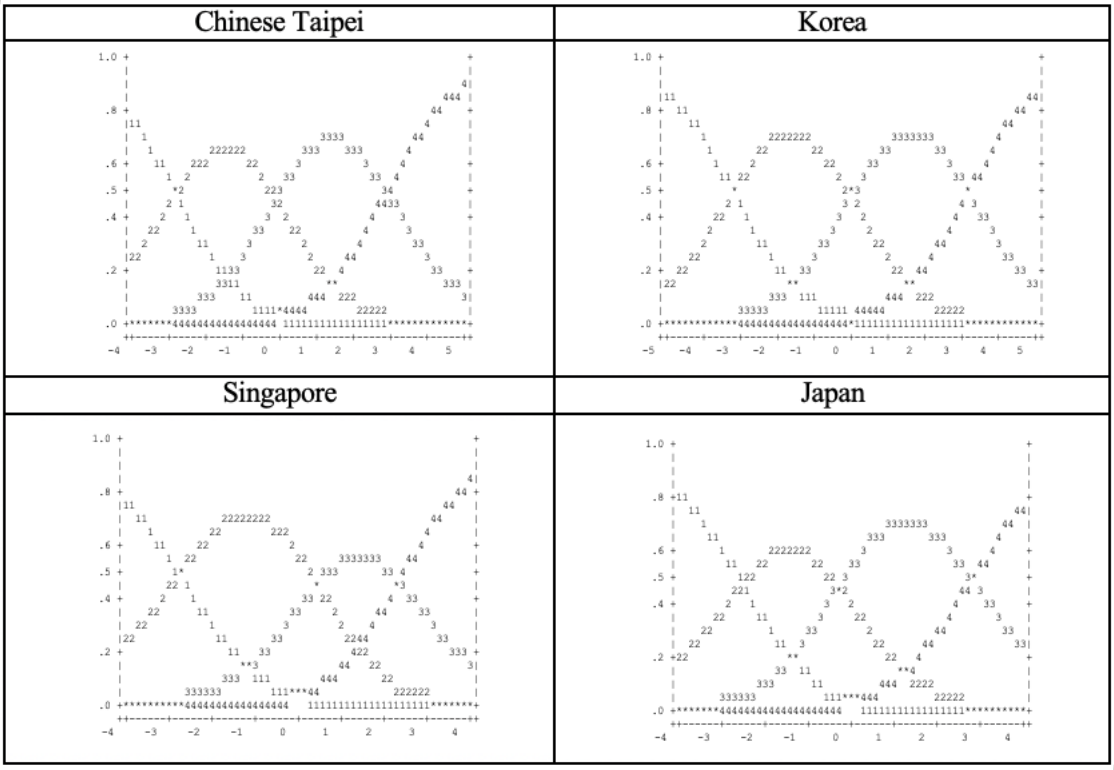

Figure 3: Category Characteristic Curves for HA Countries for SVS Rating Scale

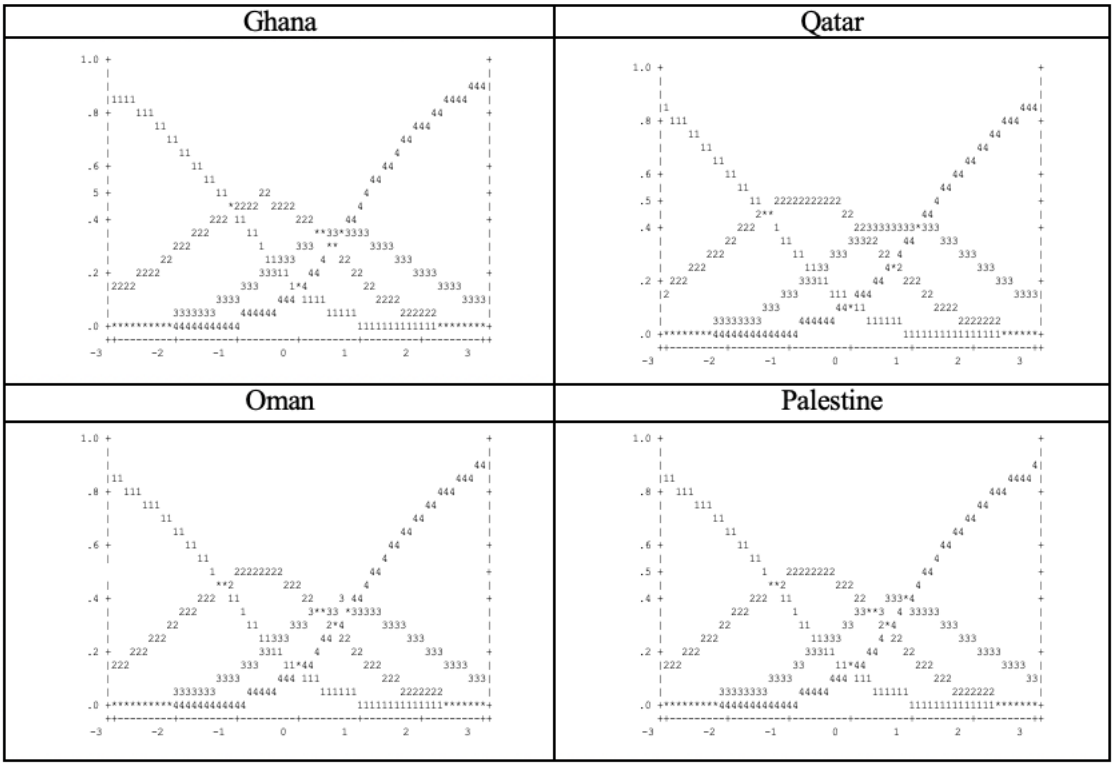

Figure 4. Category Characteristic Curves for LA Countries for SVS Rating Scale

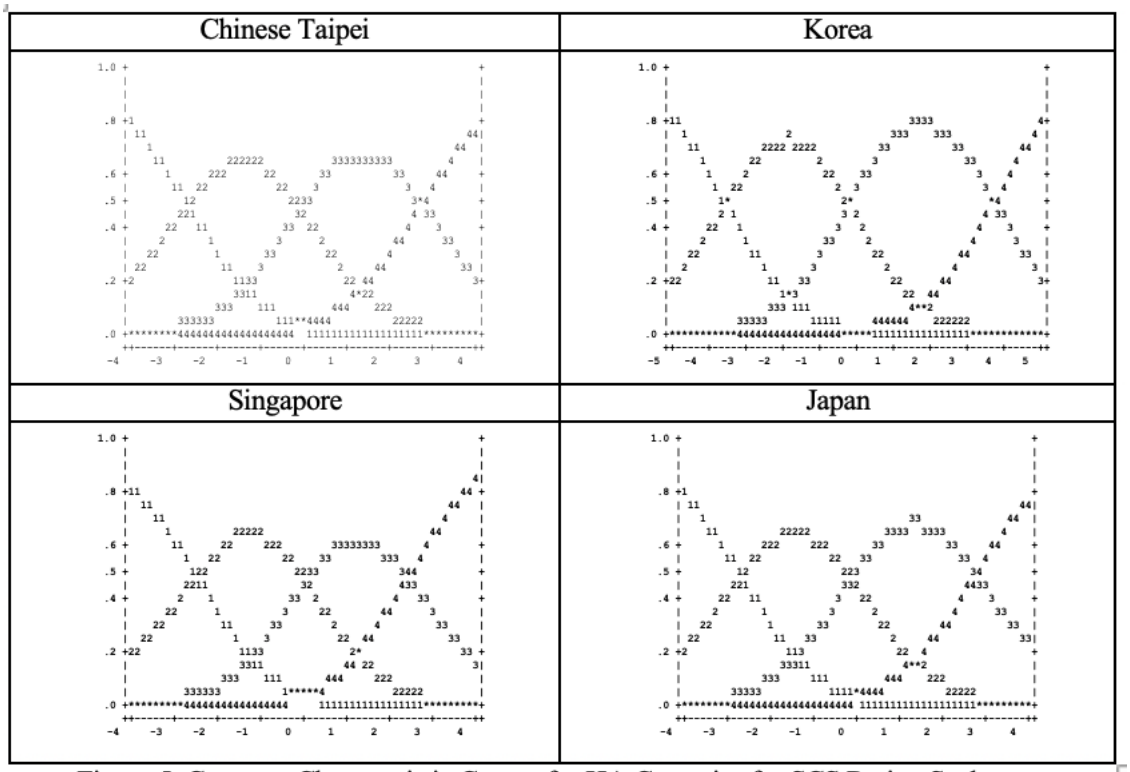

Figure 5. Category Characteristic Curves for HA Countries for SCS Rating Scale

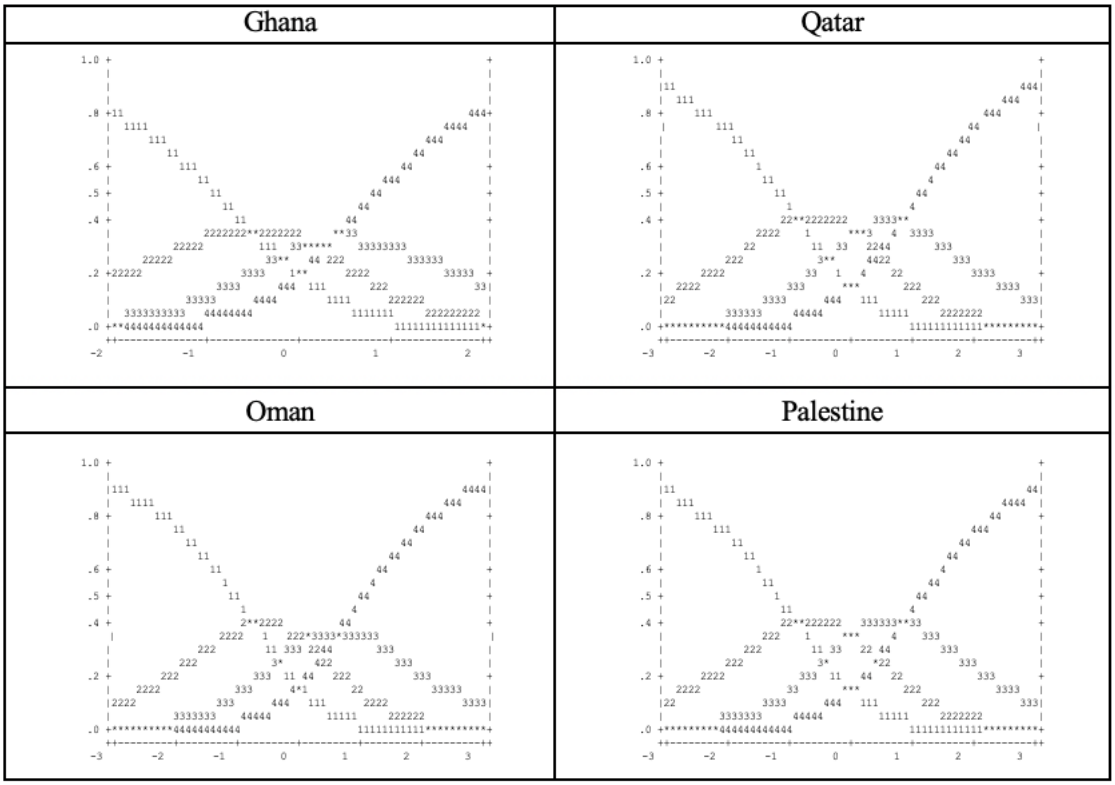

Figure 6. Category Characteristic Curves for LA Countries for SCS Rating Scale