Research Article
Open Access

Item response theory-based evaluation of psychometric properties of the safety attitudes questionnaire—korean version (saq-k)

Abstract

Years have passed since patient safety culture began to be measured. Several measurement instruments have been developed and used in the field, including the Safety Attitudes Questionnaire (SAQ). Validation is an essential step when administering these instruments to health care professionals. The problem is that confirmatory factory analysis (CFA) has mainly been used for the validation, despite the fact that CFA only works well with continuous responses. Yet SAQ and its variants, including the Korean version, use a 5-point Likert scale, which is by definition not a continuous, but a categorical measure. To resolve this conflict, we used item response theory (IRT) graded response model (GRM), an extension of CFA to categorical variables, to revalidate the SAQ-K and unravel the properties of each item and the SAQ-K instrument as a whole. We calculated difficulty and discriminating parameters for all items first. Then, based on the results, we estimated the expected score of each domain along the latent trait continuum, showing how the current SAQ behaves. We also used the item-level and domain-level information function, which can enlighten us on how to improve the precision of the current instrument. Most importantly, we obtained IRT-based empirical Bayes estimates of the latent trait level of each person and compared them with the traditional 0 to 100 SAQ arithmetic mean domain score scale. The correlation coefficients were very high, but when plotted we observed a considerable discrepancy between them. This finding led to the validity of the traditional scoring method in question. We expect the results and methodology of this study to bring about active discussions on the use of such safety culture instruments and help future studies on improving patient safety.

Keywords: safety culture, safety attitude questionnaire, patient safety, item response theory

Volume 3 Issue 5 - 2016

Heon-Jae Jeong,[1] Wui-Chiang Lee[2]

[1]The Care Quality Research Group, Chuncheon, Korea
[2]Department of Medical Affairs and Planning, Taipei Veterans General Hospital, Taipei, Taiwan

Correspondence: Wui-Chiang Lee, Director, Department of Medical Affairs and Planning, Taipei Veterans General Hospital, 201, Section 2, Shihpai Rd, Taipei City, Taiwan, Tel +886-2-28757120, Fax +886-2-28757200, Email leewuichiang@gmail.com

Received: May 29, 2016 | **Published:** June 14, 2016

Introduction

As cultural issues have been acknowledged as one of the most important aspects in improving patient safety,[1-4] many resources have been invested to tap into the safety culture from a small clinic setting to even a nationwide health care setting. The first step is, of course, accurately measuring safety culture, which plays an important role not only in understanding the topography of safety culture, but also in evaluating the effectiveness of endeavors for safety improvement. Consequently, various instruments to measure safety culture have been developed and utilized.[5-8] Among them, the Safety Attitudes Questionnaire (SAQ) is one of the most frequently used instruments[5,6,9] and has already been translated into several languages for various countries.[10-15] Jeong et al. developed the 34-item Korean version of SAQ (SAQ-K)[16] and devised novel approaches for analyzing its results using empirical Bayes methods.[9,17,18] Like the original version, SAQ-K consists of six distinct domains, the definitions of which are described in Table 1.[8] SAQ-K has been thoroughly validated and widely used in health care organizations in Korea.

Despite the successful application of such instruments in gauging patient safety culture in health care, we have to admit that those safety culture instruments, including SAQ-K, should be revalidated for several reasons. In most cases, the critical part of the survey questionnaire validation process has been conducted using confirmatory factor analysis (CFA).[19] The problem here is that CFA can and should only be used for continuous responses, not for categorical responses such as a 5-point Likert scale, because CFA is basically a linear regression model. Therefore, it cannot appropriately deal with the errors in a measurement model for dichotomous or categorical responses—errors that can neither be normally distributed nor have constant variance. In addition, in the CFA paradigm, the probability of choosing an option for an item can go out of bounds.[20] Simply put, validating a Likert scale-based instrument with CFA is theoretically incorrect. Unfortunately, however, the original SAQ and all its variants, including SAQ-K, measure responses on a 5-point Likert scale (1= disagree strongly, 2= disagree slightly, 3= neutral, 4= agree slightly, 5= agree strongly) and then convert them to a scale from 0 to 100, with 25-point intervals. In all honesty, we might have misused the continuous response targeted method for categorical variables based on the unfounded assumption that the Likert scale would behave just like continuous response.

Another issue in the everyday use of SAQ thus far is that we use the simple mean domain score (arithmetic mean of the item scores in a domain), from 0 to 100, which is the method that the original SAQ rubric suggests. Although it is easy to calculate, this approach implicitly assumes that items are τ ("tau") equivalent, which means all the items in a domain are equally important; in CFA terminology, they should share similar factor loadings. If this assumption does not hold, then we should use a factor score that places more weight on the items with larger load in gand less weight on items with smaller

Submit Manuscript | http://medcraveonline.com

©2016 Jeong et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and build upon your work non-commercially.

loading to prevent overvaluing or undervaluing items. In this way, we can obtain more reliable estimates.[21] Now, we are trapped in a catch-22 situation: When we try to prove that τ equivalence is satisfied or to use the factor score approach, weneed to runa CFAfirst to calculate the loadings. However, CFA is not an appropriate method for categorical responses, at least not by definition. It seems like there is no way out. How can we escape this dilemma?

Table 1 SAQ Domain Definitions and Number of Items

SAQ Domain	Definition	Number of Items
Teamwork Climate (TC)	Perceived quality of collaboration between personnel	5
Safety Climate (SC)	Perception of a strong and proactive organizational commitment to safety	6
Job Satisfaction (JS)	Positivity about the work experience	5
Stress Recognition (SR)	Acknowledgment of how performance is influenced by stressors	4
Perception of Management (PM)	Approval of managerial action	10
Working Conditions (WC)	Perceived quality of the work environment and logistical support	4

We propose using the item response theory (IRT)—specifically, the graded response model (GRM)—to understand SAQ's item properties and obtain adjusted domain scores that are logically more sound. Developed by Same jima,[22] the idea of GRM is also known as the cumulative logit model and can effectively handle the categorical responses as well as enable us to understand the psychometric properties of both the whole instrument and each item.

Fayers defined IRT as "a model-based measurement, also known as latent theory, in which trait level estimates depend on both persons' responses and on the properties of the items that were administered".[23] For most readers, this definition is a bit complicated, and fully understanding the behind-the-scenes logic is even more complicated. Therefore, we do not use jargon like phi-gamma psychophysics,[24] nor do we describe detailed mathematical formulas for IRT parameter estimation here. Rather, we show how to interpret the key element of IRT GRM into plain language through a simple illustrative example, which will help the readers understand the rest of this article.

An Illustrative Example of IRT GRM

Assume we are analyzing the property of an item called K to estimate the level of a certain latent trait, which is an unobservable characteristic of a person, such as perception of teamwork climate (TC) on SAQ. Item K was measured on a 5-point Likert scale (1 = disagree strongly; 5 = agree strongly). We describe the IRT GRM perspective in Figures 1 & 2, where the level of the latent trait is denoted on the x axis in a scale with a mean of 0 and a standard deviation (SD) of 1. At this point, it should be understood that, unlike the scale in the classic test theory (CTT) paradigm, this scale with a mean and SD in the IRT paradigm is not sample specific; in other words, even when the instrument is administered to other groups, the items behave the same way with the same item properties, yielding comparable scores for the groups. This is called parameter invariance, and it is one of the most important and most frequently misunderstood assumptions of IRT.

Returning to the example, we first add meanings to Figure 1 that clearly shows us the essential nature of GRM, the cumulative logit model: i) y axis is the probability of choosing a certain response category (option) equal to or higher than the category versus lower than the category (e.g., category 2, 3, 4, or 5 versus category 1), and ii) there are multiple S-shaped curves due to multiple categories as an option of an item, contrary to the single line (not curved) for an item of the CFA paradigm.

Next, we interpret the boundary characteristic curves in Figure 1. A person with a latent trait level of -2.33 has a 50% chance of answering 1 versus 2, 3, 4, or 5 (We denote the curve for this as Pr ($K\geq2$), and the remaining curves are named according to the same rule). A person with a latent trait -1.26 has a 50% possibility of answering 1 or 2 versus 3, 4, or 5. A person with a latent trait 0.27 has a 50% probability of choosing 1, 2, or 3 versus 4 or 5. A person with a latent trait of 1.95 has a 50% chance of answering 1, 2, 3, or 4 versus 5. These latent trait values are traditionally called difficulty parameters (denoted as "b"), as a reflection of history that IRT was frequently used to assess the properties of a test item to measure at what level a subject can give a correct answer—in other words, how difficult an item is. As you may have already noticed, an item with n categories (options) to choose have n-1 boundary characteristic curves. We have four curves for each item of SAQ-K as it uses a 5-point Likert scale.

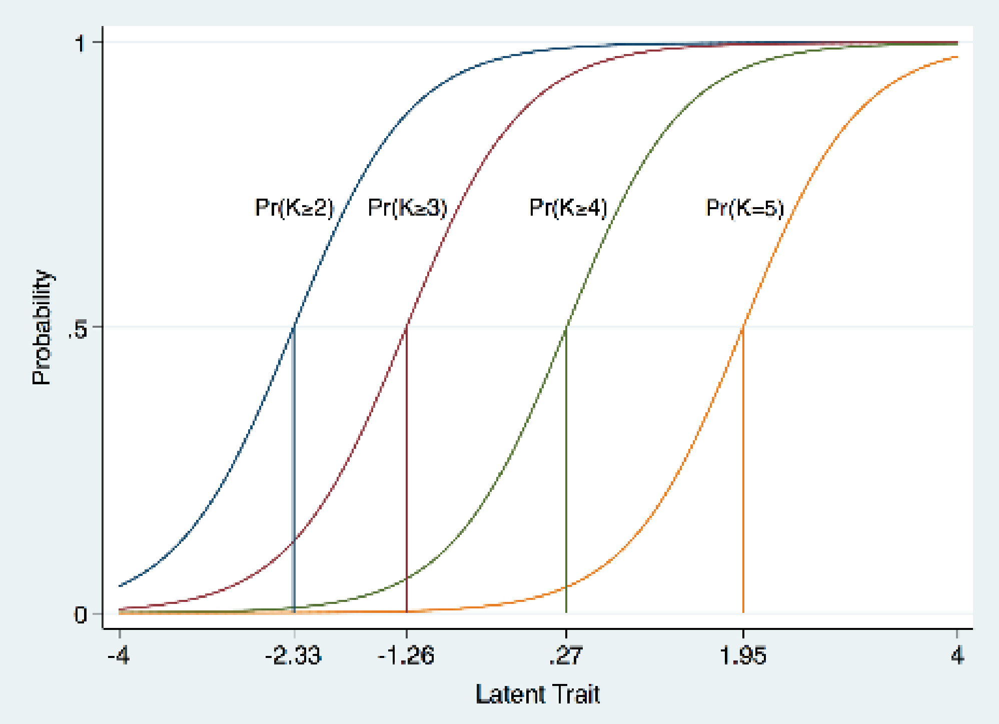

Figure 1 Boundary characteristic curves (BCC) for item K.

Although not explicitly, an important characteristic of IRT is embedded in Figure 1, which is called the discrimination parameter (denoted as "a"). It refers to how fast the probability of choosing an option category is changing near the value of the difficulty parameter. We can understand this as a differential coefficient of the curves at the 50% probability; the higher this parameter is, the more sensitively an item can detect a difference in a latent trait.

When we run the IRT GRM, what is usually returned from most statistical software is a form as shown in Table 2 (information on statistical significance is omitted here). Note that only one discrimination parameter exists, unlike multiple difficulty parameters. This is a convention of GRM.

Table 2A Example of a GRM Result in Text Form.

Discrimination (a)	1.80
Difficulty (b)	
≥2	-2.33
≥3	-1.26
≥4	0.27
=5	1.95

In addition, we can find the probability of which category (option) will be answered by a person with a certain level of latent trait. Such information can be depicted in the form of a graph called category characteristic curves (see Figure 2).The interpretation is rather intuitive: A respondent with a latent trait lower than -2.15 is likely to respond 1, a respondent with a latent trait ranging between -2.15 and -1.37 is likely to respond 2, and so on.

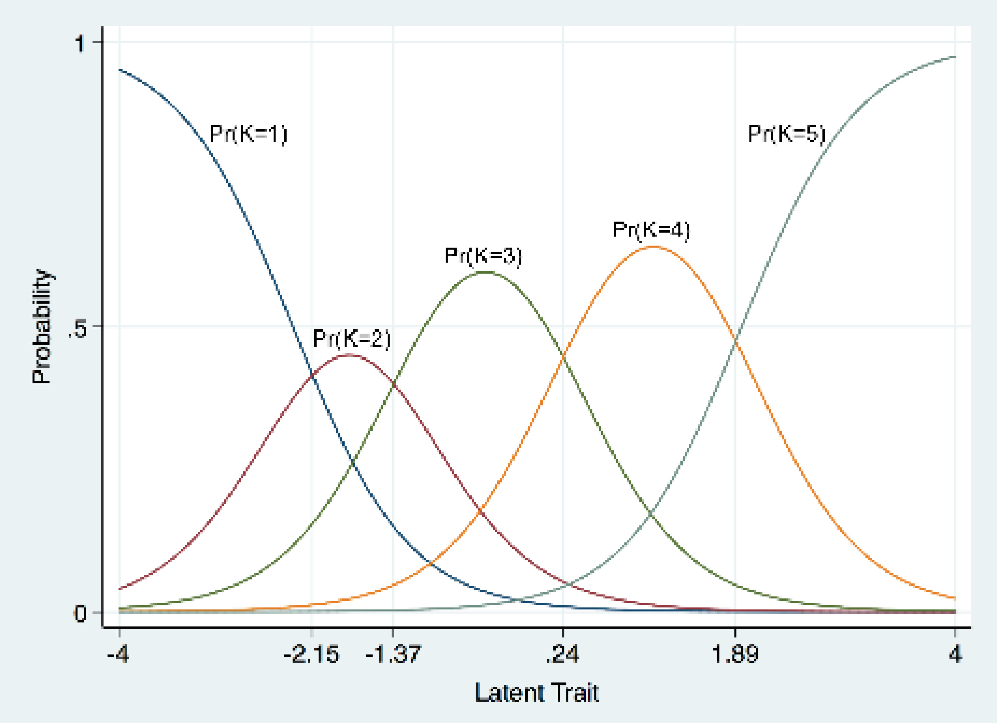

Figure 2 Category characteristic curves for item K.

Because GRM is for ordered responses, the order of curves in the BCC or CCC style plots is always arranged from category one (left) to five (right) for a 5-point Likert scale; therefore, the category for which a curve stands will no longer be indicated in this article, if no specific reason to do so exists.

This example with multiple curves and uneven intervals between them, provides empirical evidence that the continuous response assumption of CFA may not work well for categorical response; it also shows why we actually can and should take advantage of the IRT framework for a non-continuous response. We use many other features of IRT in this article, but the discussion thus far offers a sufficient background for understanding the content of this study, even if the reader is not accustomed to an IRT paradigm.

Based on the background provided, this study aims to accomplish the following:

i. to re-validate SAQ-K with IRT GRM, unraveling both the item-level and instrument-level properties that have not yet been thoroughly investigated; and

ii. to calculate new estimates of SAQ-K scores for each domain based on the IRT approach and compare them with the scores based on the simple arithmetic mean approach that the original SAQ rubric suggested.

Although we followed an IMRaD structure (Introduction, Methods, Results, and Discussion), the contents are somewhat blended; in particular, some paragraphs on methods and discussions are located in the results section. This is intentional. We tried to help readers who are not used to the IRT framework to fully understand the contents of this article.

Methods

We used the same dataset analyzed to validate SAQ-K; the data were collected in a large metropolitan hospital in Seoul from October through November 2013.Detailed information as to survey administration can be found in our previous articles.[16,18]

To examine the psychometric properties of SAQ-K with IRT GRM, we first obtained discrimination (a) and difficulty (b) parameters for each of the SAQ-K items. To help understand how the SAQ-K items may behave, we roughly classified the items into four distinct patterns and plotted their category characteristic curves (CCC) in Figure 3, which shows the probabilities of choosing a certain response category of the Likert scale through the latent trait continuum (please refer to Figure 2 for a better understanding).

Next, we tried to reveal the characteristics of the instrument as a whole; the expected domain score against latent trait was calculated by summing up the expected score of each item at every point of the latent trait level of each domain. By depicting this information in the form of test characteristic curve (TCC) in the traditional 0 to 100 SAQ scoring system, we tried to understand which score we could expect from SAQ-K for a person with a certain trait level on the current scoring scheme.

We also visited the amount of information that each item possesses. First, we plotted item information along the latent continuum in a format called item information function (IIF). Then, by adding up the IIF of items in each domain, we obtained a domain-level test information function (TIF), which shows how precisely SAQ-K can estimate the level of a respondent's latent trait. More specifically, TIF helps us decide which region on the latent trait continuum can be estimated most precisely or most poorly—in other words, where the holes are.

Lastly, we predicted a respondent's latent trait level of each domain by using empirical Bayes (EB) method in the IRT sense and calculated the correlation coefficient between the EB estimates and traditional arithmetic mean SAQ domain scores that did not take into account the τ equivalence issue. We also visualized the correlation with a scatter plot with linear regression and locally weighted scatter plot smoothing (LOWESS) lines for each domain.

Running IRT GRM, we used mean-variance adaptive Gauss-Hermite quadrature for numerical integration, with seven integration points that were computationally efficient and provided results with enough precision. For all analyses, Stata 14.1 (Stata Corp., College Station, Texas) was used.

Results

Characteristics of respondents

Of the 1,381 questionnaires returned, we analyzed 1,142 questionnaires that contained no missing values. We did this list-wise deletion to compare the two different approaches of CFA and IRT fairly; that is to say, IRT models can bear missing values, but CFA cannot. Table 3 shows the characteristics of the respondents, including gender, work year, and job type. The majority of respondents were female nurses—the composition that reflects the general situation of the Korean health care system.

Table 2B *Characteristics of Respondents*

Characteristics	**N**	**%**
Gender		
Male	300	26.3
Female	842	73.7
Work Years		
Less than 6 months	77	6.7
7–11 months	122	10.7
1–2 years	193	16.9
3–4 years	249	21.8
5–10 years	290	25.4
11–20 years	150	13.1
More than 21 years	61	5.3
Job Type		
Physician	378	33.1
Nurse	609	53.3
Pharmacist	10	0.9
Supporting Staff	132	11.6
Administration	9	0.8
Other	4	0.4
Total	1,142	100.0

Psychometric properties of SAQ-K items

The properties of all 34 items of SAQ-K, difficulty (b), and discrimination (a) are described in Table 4. The "Info rank" column will be explained in a later section. From now on, the item ID is used to describe the results.

For discrimination, JS3 showed the highest value (4.22) while WC1 showed the lowest (1.12), yielding a greater than three-times difference. Although we do not describe each value of difficulty parameters, it is worth noting that not every item shows that the four difficulty parameters were evenly distributed around zero. In addition, items within the same domain do not share the same pattern in parameter distribution. In order to help readers better understand, we classified the CCC of all 34 items into four distinct patterns (Figure 3). This classification was done by the authors' subjective analysis, not based on any statistical pattern analysis, which is beyond the scope of this study.

Table 3 *IRT Parameters of Each SAQ-K Item and Its Rank of Information Amount*

ID	**Items**	**a**	**Difficulty (b)**				**Info Rank**
			≥2	**≥3**	**≥4**	**=5**	
Teamwork Climate							
TC1	Nurse input is well received in this clinical area	1.94	-3.08	-1.90	-.11	1.45	5
TC2	Disagreements in this clinical area are resolved appropriately (i.e., not *who* is right, but *what* is best for the patient)	2.48	-2.51	-1.47	-.11	1.33	1
TC3	I have the support I need from other personnel to care for patients	2.41	-2.69	-1.72	-.32	1.04	2
TC4	It is easy for personnel here to ask questions when there is something that they do not understand	2.08	-3.08	-1.78	-.37	1.11	3
TC5	The physicians and nurses here work together as a well-coordinated team	1.93	-2.46	-1.45	-.05	1.51	4
Safety Climate							
SC1	I would feel safe being treated here as a patient	1.89	-3.02	-1.83	-.29	1.42	5
SC2	Medical errors are handled appropriately in this clinical area	2.37	-2.92	-1.94	-.46	1.11	2

Citation: Jeong HJ, Lee WC. Item response theory-based evaluation of psychometric properties of the safety attitudes questionnaire—korean version (saq-k). *Biom Biostat Int J.* 2016;3(5):174–187. DOI: 10.15406/bbij.2016.03.00079

Table Continued

ID	Items	a	Difficulty (b)				Info Rank
			≥2	≥3	≥4	=5	
SC3	I know the proper channels to direct questions regarding patient safety in this clinical area	2.04	-2.26	-1.41	-.07	1.45	3
SC4	I receive appropriate feedback about my performance	2.53	-2.56	-1.76	-.32	1.30	1
SC5	I am encouraged by my colleagues to report any patient safety concerns I may have	1.95	-2.69	-1.78	-.17	1.18	4
SC6	The culture in this clinical area makes it easy to learn from the errors of others	1.68	-3.44	-2.07	-.20	1.63	6
Job Satisfaction							
JS1	I like my job	2.51	-2.06	-1.32	-.08	1.11	4
JS2	Working here is like being part of a family	3.06	-1.63	-.89	.19	1.45	3
JS3	This is a good place to work	4.22	-1.67	-.92	.18	1.41	1
JS4	I am proud to work in this clinical area	3.52	-2.11	-1.27	-.03	1.14	2
JS5	Morale in this clinical area is high	2.46	-2.03	-.99	.47	1.82	5
Stress Recognition							
SR1	When my workload becomes excessive, my performance is impaired	1.52	-3.49	-2.21	-.72	.87	4
SR2	I am less effective at work when fatigued	1.74	-2.59	-1.97	-.81	.82	3
SR3	I am more likely to make errors in tense or hostile situations	2.17	-2.22	-1.43	-.29	1.17	1
SR4	Fatigue impairs my performance during emergency situations (e.g., emergency resuscitation, seizure)	2.12	-2.31	-1.48	-.45	1.01	2
Perception of Management							
PM1	Unit management supports my daily efforts	2.43	-2.25	-1.30	.16	1.64	8
PM2	Hospital management supports my daily efforts	1.39	-3.74	-2.65	-.48	1.04	10
PM3	Unit management doesn't knowingly compromise patient safety	2.91	-2.44	-1.59	-.14	1.09	5
PM4	Hospital management doesn't knowingly compromise patient safety	3.06	-2.28	-1.34	.04	1.34	3
PM5	Unit management is doing a good job	2.96	-2.44	-1.55	-.07	1.41	4
PM6	Hospital management is doing a good job	2.84	-1.99	-1.13	.31	1.72	6
PM7	Problem personnel are dealt with constructively by our unit management	1.79	-3.30	-2.15	-.23	1.15	9
PM8	Problem personnel are dealt with constructively by our hospital management	3.40	-1.90	-1.22	.25	1.53	1
PM9	I get adequate, timely info about events that might affect my work from unit management	3.12	-2.02	-1.10	.34	1.69	2
PM10	I get adequate, timely info about events that might affect my work from hospital management	2.73	-2.21	-1.21	.29	1.66	7
Working Condition							
WC1	The levels of staffing in this clinical area are sufficient to handle the number of patients	1.12	-1.94	-.58	1.26	3.31	4
WC2	This hospital does a good job of training new personnel	2.19	-2.27	-1.19	.38	1.99	3
WC3	All the necessary information for diagnostic and therapeutic decisions is routinely available to me	2.43	-2.96	-1.59	.15	1.81	2
WC4	Trainees in my discipline are adequately supervised	2.99	-2.94	-1.57	.04	1.70	1

Note: a: discriminating parameter

Citation: Jeong HJ, Lee WC. Item response theory-based evaluation of psychometric properties of the safety attitudes questionnaire—korean version (saq-k). *Biom Biostat Int J.* 2016;3(5):174–187. DOI: 10.15406/bbij.2016.03.00079

In Figure 3 (also refer to Table 5), the first pattern shows well-balanced items: The probability curves for categories of an item are almost evenly distributed, and the intervals were almost the same. Several items, including WC4, were included in this category. The top right pane shows left-shifted curves, so that the crossing point of categories four and five is around 1 on the latent trait. The bottom left is a unique pattern where the region of answering category 2 is very narrow; therefore, it is relatively less likely that category 2, "disagree slightly," is chosen by respondents. An extreme case of this was the last pattern, albeit there was only one item, SR2, where the probability of answering category 2 did not take the lead at all along the latent trait continuum. In this case, statistically speaking, choosing category 2 might be very unlikely, which led us to ask whether this item is properly designed and is really worth including in the instrument. This inquiry will be answered in a later section.

Table 4 *Item List by Domain of Four Category Characteristic Curve Patterns*

	TC	SC	JS	SR	PM	WC
Balanced	TC2, TC3		JS2, JS3, JS4, JS5		PM1, PM3, PM4, PM5, PM6, PM9, PM10	WC2, WC3, WC4
Left Shifted	TC1, TC4	SC1, SC2, SC6		SR1	PM2, PM7	
Narrow Category 2	TC5	SC3, SC4, SC5	JS1	SR3, SR4	PM8	WC1
Low Category 2				SR2		

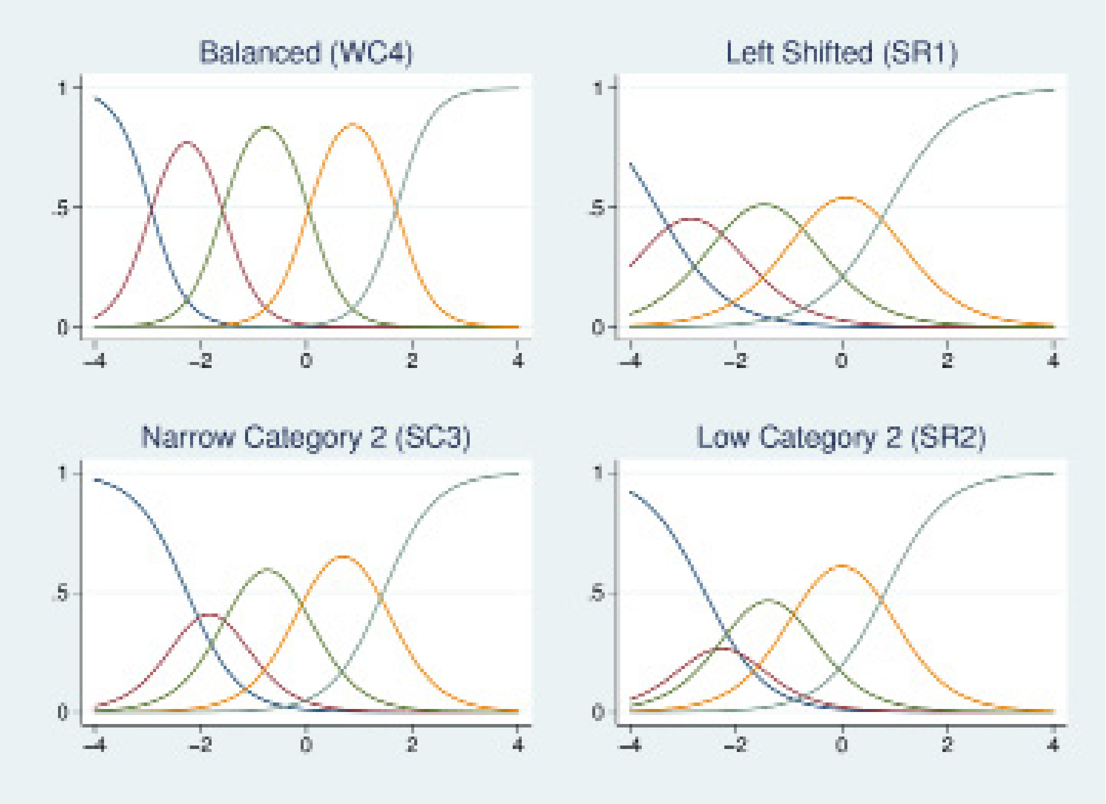

Figure 3 Graphical display of the four patterns of category characteristic curves with exemplary items

ote: x-axis: latent trait; y-axis: probability of answering a certain category.

Test characteristics of SAQ-K by domain

Figure 4 shows the scores we can expect from respondents who possess different levels of the latent trait by each domain. To illustrate, for the TC domain, respondents with an above average latent trait level are likely to score higher than 65.8. For this expected score for individuals with an average latent trait, we can see significant differences across domains: 66.67 (SC), 59.3 (JS), 70.63 (SR), 61.5 (PM), and 55.5 (WC).

By extension, we can apply the 95% interval to this graph. Because the scale of latent trait is a mean of 0 and SD of 1, the 95% interval means the latent trait level ranges from -1.96 to 1.96. Again, using the TC domain as an example, we can expect 95% of randomly selected respondents to score between 30.2 and 94.8. Instead of providing those numbers here, we present a graphical summary of the mean and 95% expectation score (Figure 5). We can easily find a considerable amount of discrepancy across domains.

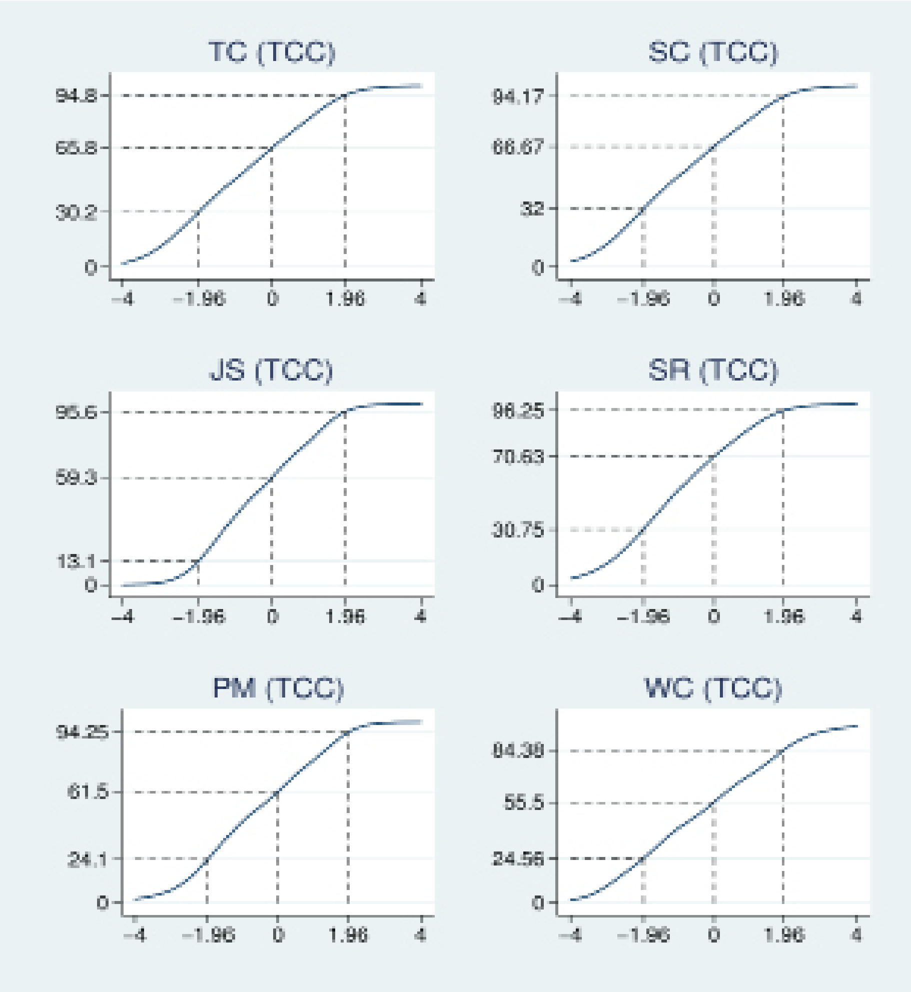

Figure 4 Test characteristic curves by SAQ-K domain.

Note: x-axis: latent trait; y-axis: expected SAQ-K domain score in the traditional 0 to 100 SAQ scoring system.

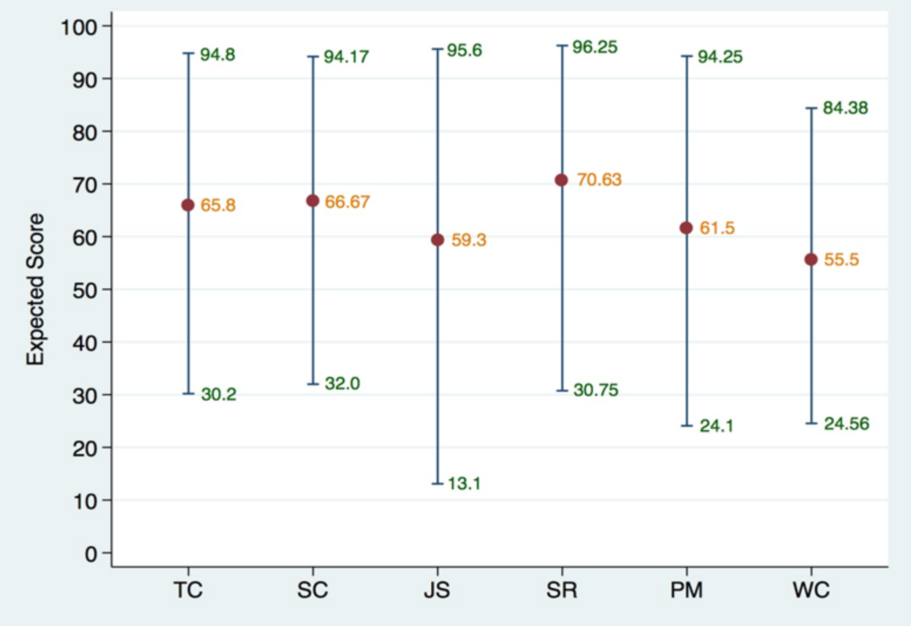

Figure 5 Expected average scores and 95% intervals by SAQ-K domain.

Amount of information by item and domain

Figure 6 shows the amount of information an item and domain convey. In the IRT realm, the information is the conceptualization of the precision and reliability of the item and the instrument. That is to say, the high information region (where a curve stays high) has more capacity to measure an individual's latent trait level with more precision and with smaller error. The left pane shows the information amount of each item against the latent trait level. There are too many items and crisscrossing, so we did not indicate the item ID on the plot. Instead, we arbitrarily chose a trait level X at which item information curves were not too crowded and numbered the curves from top to bottom. This rank is listed in the last column (Info rank) of Table 3.

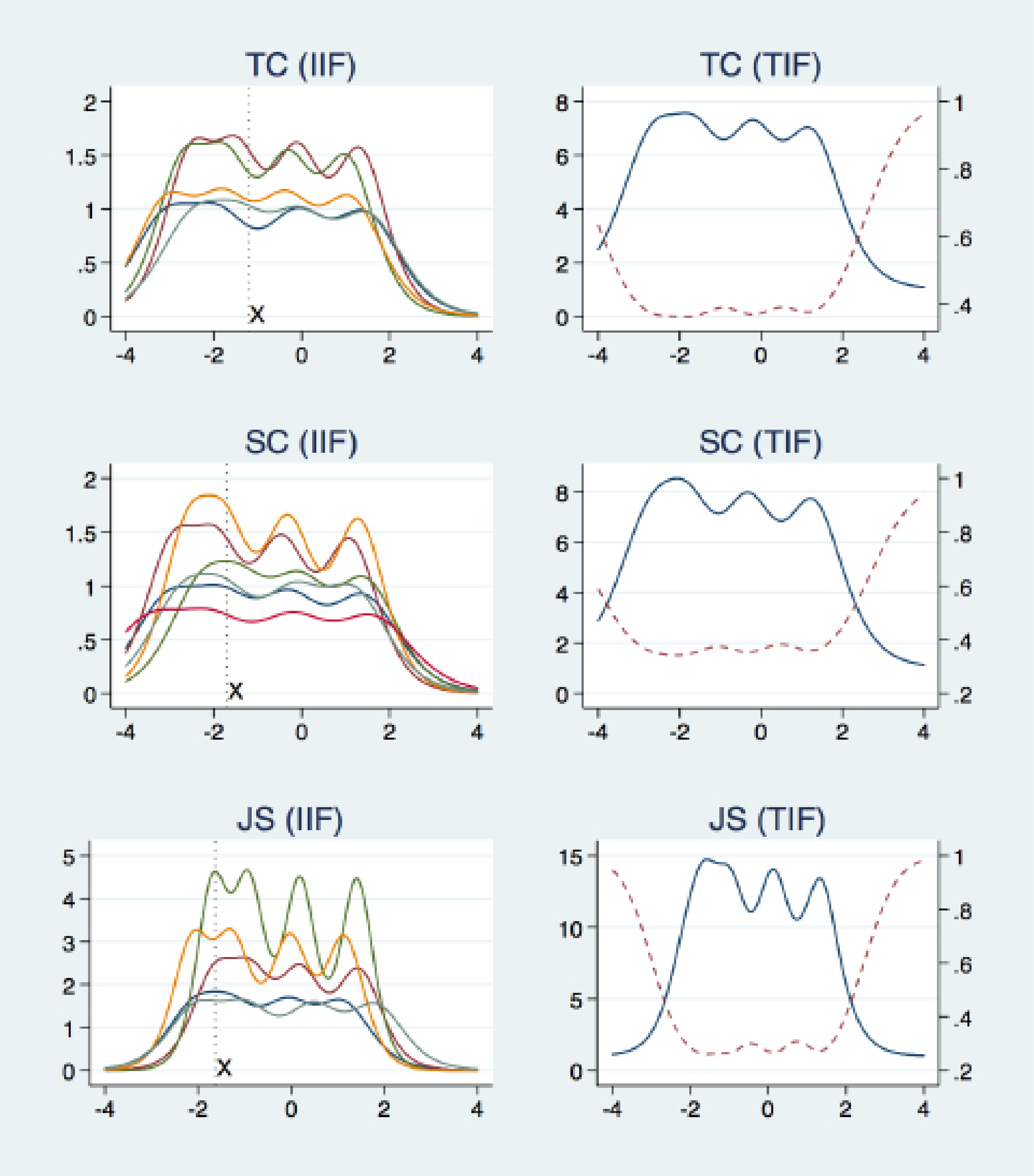

Figure 6 Item information function (IIF) and test information function (TIF) by SAQ-K domain.

Note: x-axis: latent trait; y-axis (left): amount of information; y-axis (right): standard error; solid line: information amount; dashed line: standard error.

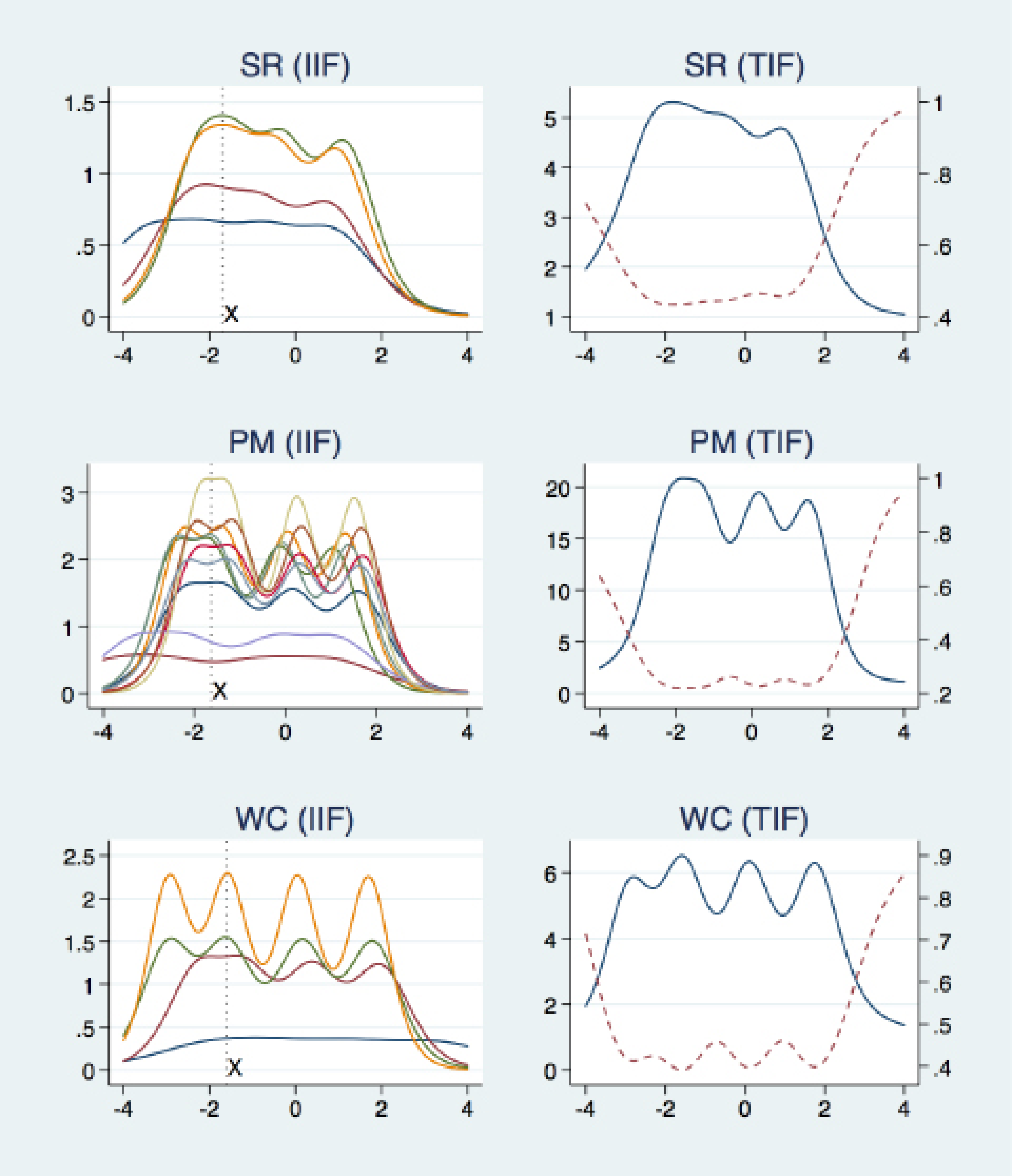

Figure 6 Item information function (IIF) and test information function (TIF) by SAQ-K domain (continued).

Note: x-axis: latent trait; y-axis (left): amount of information; y-axis (right): standard error; solid line: information amount; dashed line: standard error.

Because SAQ-K consists of a 5-point Likert scale, we do not have a unimodal or symmetric item information function (IIF); instead, each category of an item provides its own information, leading to the multimodal curves with four peaks for each (the use of five categories means that there are four thresholds between category pairs in a cumulative setting). Note that for most domains except for JS, IIFs stayed high compared to latent trait levels even lower than -2, but in the opposite direction, IIFs generally began to drop even before 2, suggesting that the items provide less information in the high latent trait region than in the lower latent trait region.

It is also worth noting that the information amount varies significantly across different items. SR, PM, and WC domains are great examples. Two items in SR (SR1 and SR2), two items in PM (PM2 and PM7), and WC1 showed much lower IIFs than other items in each domain, suggesting that those items did not add much information to the whole domain in which they were included. The Info rank column in Table 3 provides details on which item showed low or high IIF.

In Figure 6, the plots of the right pane are called test information function (TIF) and depict the information amount of six domains (solid line), each of which was obtained by summing up the IIFs on the left pane. Thus, generally a domain with more items shows a higher TIF. The dashed lines are the standard error of the latent trait estimates and provide almost mirror images of information curves in a vertical direction, which makes sense in that a precise measurement means a smaller error. TIFs of TC and SC showed their peaks from around -3 to 1.8; outside these regions, the curves dropped. JS, SR, and PM showed a similar peak at the upper value, around 1.8, but the lower value of the peak was a bit higher, around-2, which means the precision is high in a narrower range. WC showed a relatively wide plateau in TIF, extending the upper peak to around 2. Note that not a single domain showed a flat and high TIF, and outside a certain range the TIF curve dropped while the standard error curve surged, suggesting that the amount of information that the items and test can provide varied considerably along the latent trait.

Comparing IRT approach and traditional arithmetic mean approach in obtaining SAQ-K domain estimates

As described in an earlier section, the six SAQ domain scores were obtained by computing the arithmetic mean of items of each domain without considering the weighting on the items. On the other hand, IRT GRM calculated empirical Bayes (EB) estimates of latent trait level by taking into account parameters of each item. We applied both approaches and calculated the correlation coefficient of each domain (Table 5). The correlation coefficients for all six domains were very high and statistically significant.

Table 5 *Correlation Coefficients between Traditional SAQ Scoring and IRT-based EB Estimates of Latent Traits*

Domain	TC	SC	JS	SR	PM	WC
Correlation	0.993	0.9914	0.9921	0.9807	0.9909	0.9689

Note: All correlations coefficients were statistically significant.

We then plotted the estimates from the two approaches in a two-way scatter graph format, which tells us more information that is a bit different from the very high correlation coefficients discussed earlier. Generally, all domains showed a linear relationship between the two approaches, but unlike the EB estimates, which showed continuous characteristics (see y values of the observations), traditional arithmetic means showed a discontinuous nature due to the categorical characteristics of the Likert scale that SAQ uses. More importantly, a certain domain score from the traditional approach corresponded to many different EB estimates from the IRT approach. This phenomenon might arise from the issue of τ equivalence. For example, with the arithmetic mean approach, a person who chose answers 2, 3, 3, 3, and 4 for the first through fifth items in the TC domain might yield a mean score of 3; another person who chose 4, 3, 3, 3, and 2 for the same items would have the same mean score, 3. However, if the amount of information or importance of the first and the last items differed, and such a difference was legitimately compensated, those two people's domain scores should be different. This is exactly what the graphs shows: There are many combinations of answers of the items, and those combinations may yield the same arithmetic mean, but once the importance of each item is taken into consideration, the EB estimates of the latent trait can and should vary across the combinations that have the same arithmetic mean, as shown in Figure 7.

If we observe the plots horizontally, this phenomenon is even more clearly observed. For example, for SR on the latent trait level (y-value), 0, the traditional arithmetic score expands from almost 50 to 75, which means that people with the same SR trait level can be considered very differently in the arithmetic mean paradigm.

Discussion

The primary purpose of applying IRT, besides handling the CFA's issues in dealing with categorical responses, is to develop scale or metric that assigns values to the latent trait. Readers who are relatively new to the IRT paradigm might find this process a bit hard to admit, but actually it is not. Scale development begins by choosing an anchor point and then deciding the size of the unit, indicating the distance from the anchor point. Such a process may seem arbitrary, but it has always been around us. For example, the freezing point of water is 32° and the boiling point is 212° on the Fahrenheit scale, with 180 intervals between the two points, which serve as the anchor points while the intervals are the measurement unit. In the Celsius scale, the same freezing point and boiling point are 0° and 100°, respectively, resulting in 100 equal intervals between them.[23,25] The point here is that we arbitrarily but systematically assign values to the anchor points and then decide how many intervals should occur—in other words, the resolution of the unit in measuring the target phenomenon. As such, although we used a scale with a mean of 0 and a standard deviation of 1 for the latent trait in this article, we can always transform the scale to a traditional 0 to 100 scale of SAQ, maintaining the strengths of the IRT approach.

Along the same lines, one of the most fascinating properties of the IRT scale is that the latent traits and item parameters are expressed in the same metric: the so-called conjoint scale. To illustrate, the difficulty parameters (b) of the items are referenced to the very same scale for respondents' latent trait (e.g., if b is -2, then b is located at -2 S Don the latent trait continuum).Thus, any transformation of a scale into another one does not cause a problem as long as the transformation is applied to both latent trait and item parameters simultaneously. Naturally, unlike the classic test theory, we do not need a comparison to the sample who took the test together to estimate a subject's latent trait level. In the IRT paradigm, a person's information and item properties do not depend on each other; therefore, we just use the item parameters to obtain the latent trait level of each person because the probability of how an item behaves at each point of the latent trait has already been on the shelf since the development and testing of the item.[26,27]

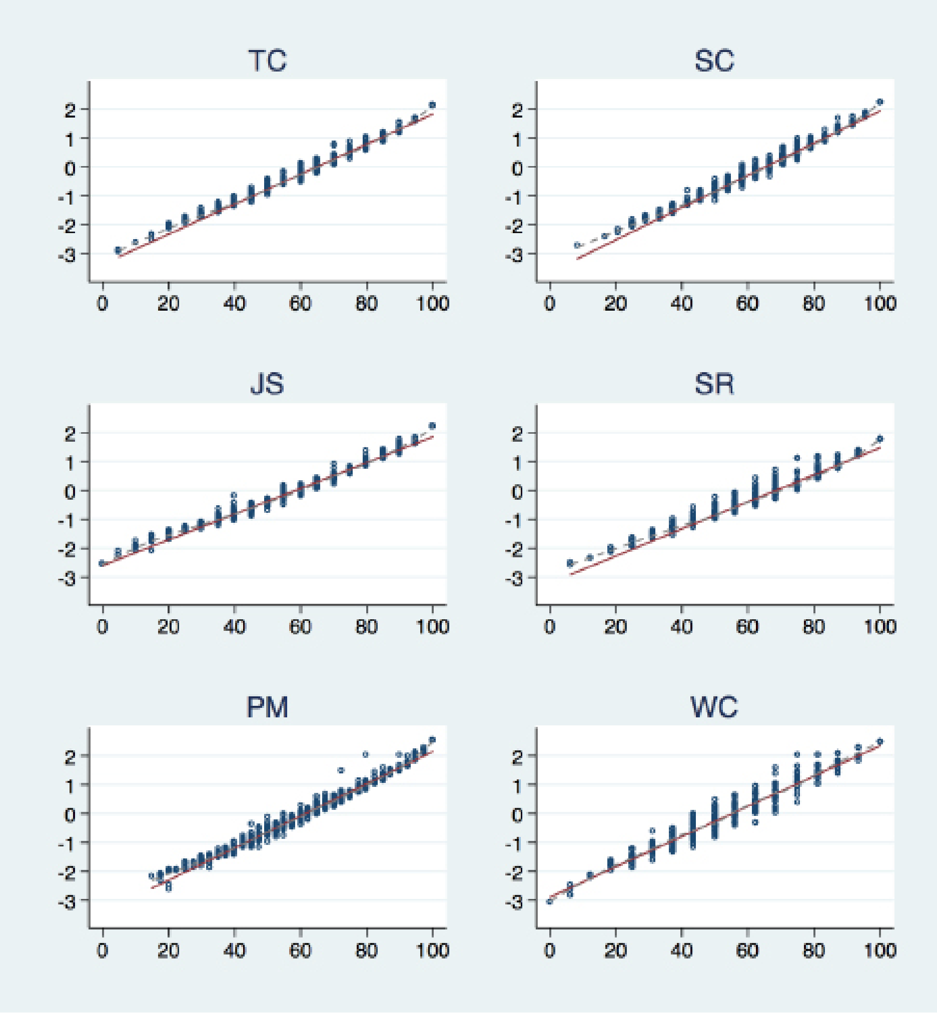

Figure 7 Scatter plot of traditional SAQ scores and IRT-based estimates of latent traits and fitted lines.

Note: x-axis: traditional arithmetic mean of SAQ domain scores; y-axis: latent trait level from IRT; solid line: linear regression line; dashed-line: LOWESS smothering line.

These ideas boil down to the following crucial characteristic of IRT, called parameter invariance, which is described as "when an IRT model fits the data exactly in the population, then the estimated item parameters should be the same, within sampling error, regardless of what sample the data were derived from, and the estimated person latent traits should be the same regardless of what items they are based on".[28] This parameter invariance provides us with the utmost practical usefulness. Simply put, we can use the items with any groups regardless of the difference among groups; if needed, even different instruments can be administered to different respondents. For some of the readers, this parameter invariance might not make much sense, but they probably have already gone through such test instruments. Many computer-adaptive tests, such as Test of English as a Foreign Language (TOEFL) and Graduate Record Examinations (GRE), are great examples.[29,30] Test takers of those exams can choose the date and time of their test, which means their scores are independent of other test takers (no comparison with others). Each test taker is presented with different item sets, which means as long as we know the item properties, we can measure the subject's ability; we do not need to administer the same items to all subjects.

Although there are other aspects of IRT, such as uni-dimensionality or local independence, describing all of them are beyond the scope of this article. Let us return to the results of this study.

Picking up the above-mentioned scale issue, we revisit the problem of the current arithmetic mean-based scoring system for each domain used in almost all cases of SAQ use. Most frontline personnel who use SAQ think of the current scoring as an absolute value—that is to say, they believe scores that were generated in the traditional way can be directly compared across different domains. Therefore, if mean scores of subjects of a certain domain are clumped around a certain value and another domain shows a wide spread, people interpret this different distribution patterns in domain scores as the cultural characteristics of the target population who took the survey, which is the phenomenon we want to measure. However, by reviewing the test characteristic curves (TCC) in Figure 4, we can see that this approach may have a serious flaw. In the current arithmetic scoring system, each domain has its own expected score distribution as emphasized in Figure 5. Therefore, what we thought of as hospitals' safety culture profile may actually come from instrument characteristics, not from respondents' characteristics. In addition, many hospitals including Johns Hopkins, trace the temporal changes in SAQ scores.[31] They frequently set a certain threshold value of change, such as 20, and if a certain domain score drops more than that, it is interpreted as a red flag. However, as TCC effectively demonstrates, the sensitivity of the traditional domain score on the change in latent trait varies significantly. For example, the 95% range of expected score in the JS domain spans from 13.1 to 95.6, which was much wider than that of WC (from 24.56 to 85.38). This means that the same amount of change in the latent trait level of JS is reflected as a larger change in SAQ score than that of WC in traditional SAQ scoring (as Figure 4 demonstrates, TCC of JS is steeper). Again, if this holds, what we have regarded as safety culture scores might have been the mixture of what we really wanted to measure and the test characteristics should not have been intervened or at least should have been canceled out.

In Figure 6, the IIF and TIF provide rich information in designing and modifying the test instrument. Let us discuss TIF first. As the sum of individual item information in IIF, TIF allows us to understand which region on the latent trait can or cannot be measured precisely—in other words, the size of error in the region. If we want an instrument targeting hospitals with a high safety culture, we might not need much information in the lower latent trait level. In this case, we probably want to remove some items focusing on the low trait level and add items for the high trait level to improve the precision at the high latent trait level. On the other hand, if the instrument is supposed to be generic and will be administered to all hospitals, then we prefer a relatively flat TIF to measure the latent trait along the entire range equally well. Therefore, if there is a so-called hole, the low information region along the latent trait continuum, we can add an item or two to make up for the hole and improve precision of the specific region. Therefore, the IIF's information can serve as the basis for this instrument modification. In addition, more often than not, we need to remove some items due to, for example, respondents' time constraints. The decision about which item should be taken out first can be helped by IIF; as in the left pane of Figure 6, some items clearly showed a very low IIF curve compared to other items in the domain, which means they do not contribute much to the information of the domain. Taking out those with low information function items and observing the change in TIF would probably be a logical process in controlling the number of items of the instrument.

We already raised concerns from the comparison between empirical Bayes (EB) estimates based on IRT and the traditional domain scores by arithmetic mean. We think a more important issue is not just the compatibility between them, but rather how SAQ scores are currently being used. In order to focus on the psychometric properties of SAQ, we mainly used scores, whether in the form of EB estimates or arithmetic mean. In many cases, however, SAQ scores are interpreted in the concept of "percent agreement"—that is, the percentage of respondents whose mean domain score is equal to or higher than 75 (the level of "agree slightly" in a Likert scale of SAQ). This approach is very meaningful, showing how the safety culture is spread among the respondents, health care professionals in a certain clinical area, or even a hospital. However, if we recall, a single latent trait score from IRT corresponds to a range of arithmetic mean scores; for example, for all six domains, the latent trait level 0 corresponded to around 60–80 of the traditional arithmetic mean score (refer to the discussion around Figure 7). Considering this, the problem becomes obvious. Certainly, for a group of respondents who share the same latent trait level, some of them might be classified as "those who agree," whereas others might not. This phenomenon could be thought of as misclassification bias, which seriously undermines the validity of the survey.

In addition, this percent agreement approach basically reduces the resolution of the data to a dichotomized form; thus, the variance of scores cannot be addressed, and information on the score distribution, such as score change in a certain trait region (e.g., score change from 20 to 70), can easily be ignored. Regardless of which approach is used between traditional mean or IRT, this issue from percent agreement-based interpretation does exist. Therefore, more debate on this is required.

Finally, we also see the potential of IRT in national and international collaboration to improve patient safety. SAQ is certainly not the only safety measurement instrument; several other instruments, some of which have been developed by the hospitals themselves, are being used. Unfortunately, there was no way to combine those instruments and compare the results across hospitals or countries. However, using the IRT approach, items' innate properties are revealed, and if those multiple instruments share a few core items, then we can merge the instruments to yield all the item properties on a single latent trait scale, which is a sharp contrast to the classic test theory (CTT). Indeed, this idea has been used in various fields, such as developing a new instrument for dementia patients by combining a few different instruments.[32] Although the method is much more complicated, multi-group studies even at the country level have also been tried in other industries.[33] The safety culture instrument deserves such efforts, as such an endeavor will save lives globally.

Conclusion

Please do not get us wrong. We do not have any intention to nullify the results from previous studies using SAQ and its variants based on CFA and the traditional arithmetic mean approach; nor are we sales representatives of IRT. We conducted this study only to re-validate SAQ using a methodologically correct approach, thereby enabling future studies to be performed on the solid theoretical foundation of the instrument. IRT happened to provide such a framework, and some features of it, such as amount of information through IIF or TIF and test characteristic curves, extend the potential of SAQ much further. In addition, it is important to note that we see a possibility for updating the survey instrument by adding or subtracting some items, but we can still use the instruments to compare results temporally or across different groups with the help of IRT. We do sincerely hope that not only the results, but also the methodology used in this study can play a role in laying the groundwork for improving patient safety.

By the way, just in case somebody asks why we should bother to use IRT instead of CFA based on the assumption that the data are approximately continuous, simply say "If you build a regression model with dichotomized or ordinal data, are you so reckless to use linear regression? That is why we bother."

Acknowledgement

None.

Conflict of interest

None.

References

1. Jeong HJ, Pham JC, Kim M, et al. Major cultural-compatibility complex: Considerations on cross-cultural dissemination of patient safety programmes. *BMJ Qual Saf*. 2012;21(7):612–615.
2. Chassin MR, Loeb JM. High-Reliability Health Care: Getting There from here. *Milbank Q*. 2013;91(3):459–490.
3. Nance JJ. *Why Hospitals Should Fly: The Ultimate Flight Plan to Patient Safety and Quality Care*. 2008.
4. Dekker S. *Just culture: Balancing safety and accountability*. 2012.
5. Colla JB, Bracken AC, Kinney LM, et al. Measuring patient safety climate: a review of surveys. *Qual Saf Health Care*. 2005;14(5):364–366.
6. Etchegaray JM, Thomas EJ. Comparing two safety culture surveys: safety attitudes questionnaire and hospital survey on patient safety. *BMJ Qual Saf*. 2012;21(6):490–498.
7. Morello RT, Lowthian JA, Barker AL, et al. Strategies for improving patient safety culture in hospitals: a systematic review. *BMJ Qual Saf*. 2013;22(1):11–18.
8. Sexton JB, Helmreich RL, Neilands TB, et al. The Safety Attitudes Questionnaire: psychometric properties, benchmarking data, and emerging research. *BMC Health Serv Res*. 2006;6(1):44.
9. Lee GS, Mi-jin Park, Hae-ran Na. A Strategy for Administration and Application of a Patient Safety Culture Survey. *J of Quality Improvement in Health Care*. 2015;21(1):80–95.
10. Nordén-Hägg A, Sexton JB, Kälvemark-Sporrong S, et al. Assessing Safety Culture in Pharmacies: The psychometric validation of the Safety Attitudes Questionnaire (SAQ) in a national sample of community pharmacies in Sweden. *BMC Clin Pharmacol*. 2010;10(1):8.
11. Carvalho REFLD, Cassiani SHDB. Cross-cultural adaptation of the Safety Attitudes Questionnaire-Short Form 2006 for Brazil. *Revista Latino-Americana de Enfermagem*. 2012;20(3):575–582.
12. Deilkås ET, Hofoss D. Psychometric properties of the Norwegian version of the Safety Attitudes Questionnaire (SAQ), generic version (short form 2006). *BMC Health Services Research*. 2008;8(1):191.
13. Lee WC, Wung HY, Liao HH, et al. Hospital safety culture in Taiwan: a nationwide survey using Chinese version Safety Attitude Questionnaire. *BMC Health Serv Res*. 2010;10(1):234.
14. Raftopoulos V, Pavlakis A. Safety climate in 5 intensive care units: A nationwide hospital survey using the Greek-Cypriot version of the Safety Attitudes Questionnaire. *J Crit Care*. 2013;28(1):51–61.
15. Natalie Z, Kaspar Küng, Susan MS, et al. Assessing the safety attitudes questionnaire (SAQ), German language version in Swiss university hospitals-a validation study. *BMC Health Services Research*. 2013;13(1):347.
16. Heon-Jae Jeong, Su Mi Jung, Eun Ae An, et al. Development of the Safety Attitudes Questionnaire-Korean Version (SAQ-K) and Its Novel Analysis Methods for Safety Managers. *Biometrics & Biostatistics International Journal*. 2015;2(1):1–20.
17. Heon-Jae Jeong, Su Mi Jung, Eun Ae An, et al. Combinational Effects of Clinical Area and Healthcare Workers' Job Type on the Safety Culture in Hospitals. *Biometrics & Biostatistics International Journal*. 2015;2(2):1–24.
18. Heon-Jae Jeong, Su Mi Jung, Eun Ae An, et al. A Strategy to Develop Tailored Patient Safety Culture Improvement Programs with Latent Class Analysis Method. *Biometics & Biostatistics International Journal*. 2015;2(2):1–27.
19. Cole DA. Utility of confirmatory factor analysis in test validation research. *J Consult Clin Psychol*. 1987;55(4):584–594.
20. Embretson SE, Reise SP. *Item response theory*. 2013.
21. Acock AC. *Discovering structural equation modeling using Stata*. 2015.
22. Samejima F. *Estimation of latent ability using a response pattern of graded scores*. 1969.
23. Fayers P. Item response theory for psychologists. *Quality of Life Research*. 2004;13(3):715–716.
24. Herrick RM. Psychophysical methodology: Deductions from the phi-gamma hypothesis and related hypotheses. *Perception & Psychophysics*. 1970;7(2):73–78.
25. Glas CA, P de Boeck, Wilson M. Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach. *Journal of Educational Measurement*. 2005;42(3):303–307.
26. Weiss DJ, Yoes ME. *Item response theory, in Advances in educational and psychological testing: Theory and applications*. 1991;69–95.
27. Thissen D, Steinberg L. Item response theory. *The Sage Handbook of Quantitative Methods in Psychology*. 2009;148–177.
28. Stata Corp. *Stata 14-Item Response Theory Reference Manual*. 2015.
29. Wainer H, Wang X. Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement*. 2000;37(3):203–220.
30. Mills CN. Development and introduction of a computer adaptive Graduate Record Examinations General Test. *Innovations in computerized assessment*. 1999;117–135.
31. Pronovost PJ, Goeschel CA, Marsteller JA, et al. Framework for patient safety research and improvement. *Circulation*. 2009;119(2):330–337.
32. Mungas D, Reed BR. Application of item response theory for development of a global functioning measure of dementia with linear measurement properties. *Stat Med*. 2000;19(11-12):1631–1644.
33. Muthén B, Asparouhov T. IRT studies of many groups: the alignment method. *Front Psychol*. 2014;5(978):10.