Research Article
Open Access
CrossMark

The level of collapse we are allowed: comparison of different response scales in safety attitudes questionnaire

Abstract

Various survey instruments have been used to measure patient safety culture. Many of these instruments use a (usually5-point) Likert scale. This studyused the Safety Attitudes Questionnaire-Korean version (SAQ-K), consisting of 34 items in six domains, to examine whether other scales, such as dichotomized and trichotomized scales, present equivalent estimates to the currently used 5-point Likert scale (1=disagree strongly, 2=disagree slightly, 3=neutral, 4=agree slightly, 5=agree strongly). For each item, we generated a 3-point scale by collapsing responsesfor 1 and 2 into one category and 4 and 5 into another category, yielding a scale of 1=disagree, 2=neutral, and 3=agree. A dichotomized scale was generated by collapsing responses for 1 through 3 from the original scale to 0=disagree and 4 and 5 to 1=agree. Correlations among the results from the five measurement scales for each respondent, as the unit of analysis, for each of the six domains were estimated: currently used simple mean of item scoreson a 5-point scale, empirical Bayes (EB) estimate from a 5-point graded response model (GRM), EB estimate from a 3-point GRM, EB estimate of a 2-parameter (2PL) item response theory (IRT) model, and EB mean of a 1-parameter (1PL) IRT model. All correlation coefficients were statistically significant ($p<.01$) and mostly exceeded 0.9 between the currently used simple mean and 3-point GRM estimates, although for dichotomized scales most coefficients were between 0.8 and 0.9. When we aggregated the responses to the clinical area level, the correlation became much higher, exceeding 0.9, except for those involving dichotomous scales in the stress recognition domain. This study found that dichotomous or trichotomous scales performed well compared to the current 5-point scale, suggesting such collapsing could replace the original scale at least in the analysis phase of collected data. Further study is needed to examine whether such simpler scales can be used in the survey-administering phase with sufficient validity.

Keywords: patient safety, safety culture, safety attitudes questionnaire, measurement scale

Volume 4 Issue 4 - 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 & National Yang-Ming University School of Medicine, Taipei, Taiwan

Correspondence: Wui-Chiang Lee, Department of Medical Affairs and Planning, Taipei Veterans General Hospital & National Yang-Ming University School of Medicine, Taipei, Taiwan, Tel +886-2-28757120, Fax +886-2-28757200, Email leewuichiang@gmail.com

Received: August 29, 2016 | **Published:** September 12, 2016

Introduction

As culture among healthcare professionals is agreed to be one of the most important factors in ensuring safe care,[1] many resources have been invested in measuring and portraying the topography of a safety culture. Thus, various measurement instruments have been developed and used around the world.Some are globally accepted more generic ones, such as the Safety Attitudes Questionnaire (SAQ) and Hospital Survey on Patient Safety (HSOPS);[2] others are country- or even organization-specific instruments.

These instruments seem to haveplayed their role well, and we have no intention of casting a suspicious eye on their validity. Yet their efficiency draws our attention. By efficiency, we mean whether they consume too much of healthcare professionals' time to complete the survey or researchers' time to analyze the collected data.

To illustrate, under the premise that instruments show equivalent validity, it is obvious that the one with a smaller number of items is more efficient. Therefore, taking out less useful items could be one way to increase efficiency. Indeed, Jeong et al. (2016), using item response theory (IRT), proposed a method to reveal the amount of impact of each item on the safety culture estimatesin order to selectless influential items that can be removed from an instrument.[3] In this way, the survey can be administered to healthcare professionals with minimum burden.

Another way, which has not yet been explored,is to consider trying different levels of measurement, especially scales with fewer response options. Thus far, most safety culture instruments, including SAQ and HSOPS, have relied on a 5-point Likert scale. Typically, the responses from such scalesare converted to a simple mean score and analyzed. Yet the possibility exists that a 3-point or even a simple yes/no type of format could be valid enough compared to the traditional 5-point scale. In particular, when combined with the IRT approach, which provides much more precise estimates than simple mean scores, reduced response options may function well enough.

Aclear strength of using fewer-option scales is evident in the analysis phase. Assume that we have 10 items with a 5-point Likert scale. Theoretically, the number of possible response combinations is 5^{10}=9,765,625. Then, the contingency table of the responses naturally becomes too sparse; thus,we cannot obtain reliable estimates. Calculating simple average scores does not suffer from such ranging waves of data, but more sophisticated analyses dealing with multiple dimensions and their covariance structure almost always face challenges. Indeed, the safety culture instrument that we analyzed in this study, Safety Attitudes Questionnaire-Korean version (SAQ-K), consists of 34 items in a 5-point Likert scale across six domains, and we failed when running several analyses because the amount of computation exceeded the capacity of the computer. Quite often, research is broken down from the boundless running time of analysis and ends up reporting simpler analysis results. However, if we collapse the collected data down to a simpler scale, like a dichotomous scale, the computations become exponentially simple: In the above 10-item example, if we use a dichotomous scale, the number of combinations in item responses is only 2^{10}=1,024. For this magnitude of data, computing resources are no longer a significant hindrance.

To enjoy the merit of simpler measurement scales, we should first ensure that the simpler scales provide at least similar, preferably indistinguishable measurement estimates as those from the currently used 5-point Likert scale. Therefore, we examined the correlations among the currently used 5-point measurement scale and others, such as dichotomized and trichotomized scales. The analysis was conducted at two levels. First, each survey respondent's score was a unit of analysis. Then, we took into account how the SAQ-K is actually used in hospital settings—namely, most times, clinical-area level scores are the main interest of administering a safety culture survey, so we also calculated clinical area-specific safety scores and plugged them into correlation analyses.

Methods

We used the SAQ-K dataset collected from 1,142 respondents working in a large metropolitan hospital with 72 clinical areas in Seoul, Korea, in 2013. As it is not the focus of this study, the details of participants' characteristics are not described here, but they can be found in our previous articles.[4,5] The definitions of SAQ-K domains and number of items for each domain are described in Table 1.[6]

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

Table 1 SAQ-K domain definitions and number of items

For data preparation (see Figure 1), we first trichotomized the original 5-point scale of SAQ-K responses (1=disagree strongly, 2=disagree slightly, 3=neutral, 4=agree slightly, 5=agree strongly) by collapsing responsesfor 1 and 2 into a disagree category and 4 and 5 into an agree category, yielding a 3-point scale: 1=disagree, 2=neutral, and 3=agree. Then, dichotomized scale responses were generated by collapsing responses for 1 through 3 from the original scale to 0=disagree and 4 and 5t o 1=agree. The rationale of dichotomization between 3 and 4 from the original scale was that the rubric of SAQ regards people who answered higher than or equal to 4 as those who agreed with the statement in an item.

With the item responses in different levels of scale, we calculated the SAQ score for each of the six SAQ-K domains forevery 1,142 respondent. To achieve more accurate estimates, we used different versions of IRT according to the response structure. Eventually, we obtained five different types of scores for each domain: i) currently used simple average item scores from the 5-point Likert scale (5 Mean),[7] ii) empirical Bayes (EB) estimates from an IRT graded response model(GRM) with the original 5-point scale responses (5GRM), iii) EB estimates from GRM with a collapsed 3-point scale (3 GRM), iv) EB estimates of a 2-parameter IRT model with dichotomous responses (2PL), and v) EB estimates of a 1-parameter IRT model with dichotomized responses (1PL). The 2PL has an additional discrimination parameter, which provides more information on item characteristics, than the 1PL model.[8] The correlations among those five different scores for each of the six SAQ-K domains were calculated with each respondent as the unit of analysis.

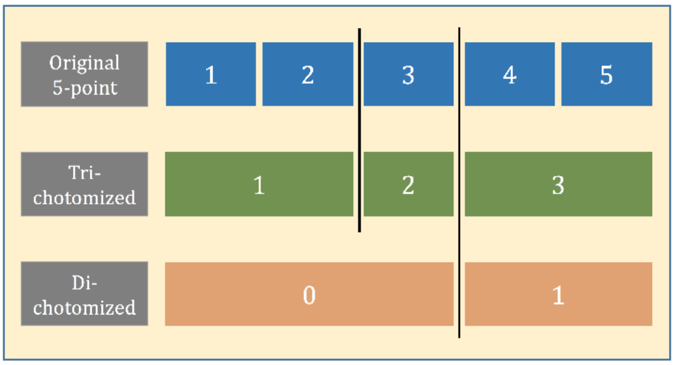


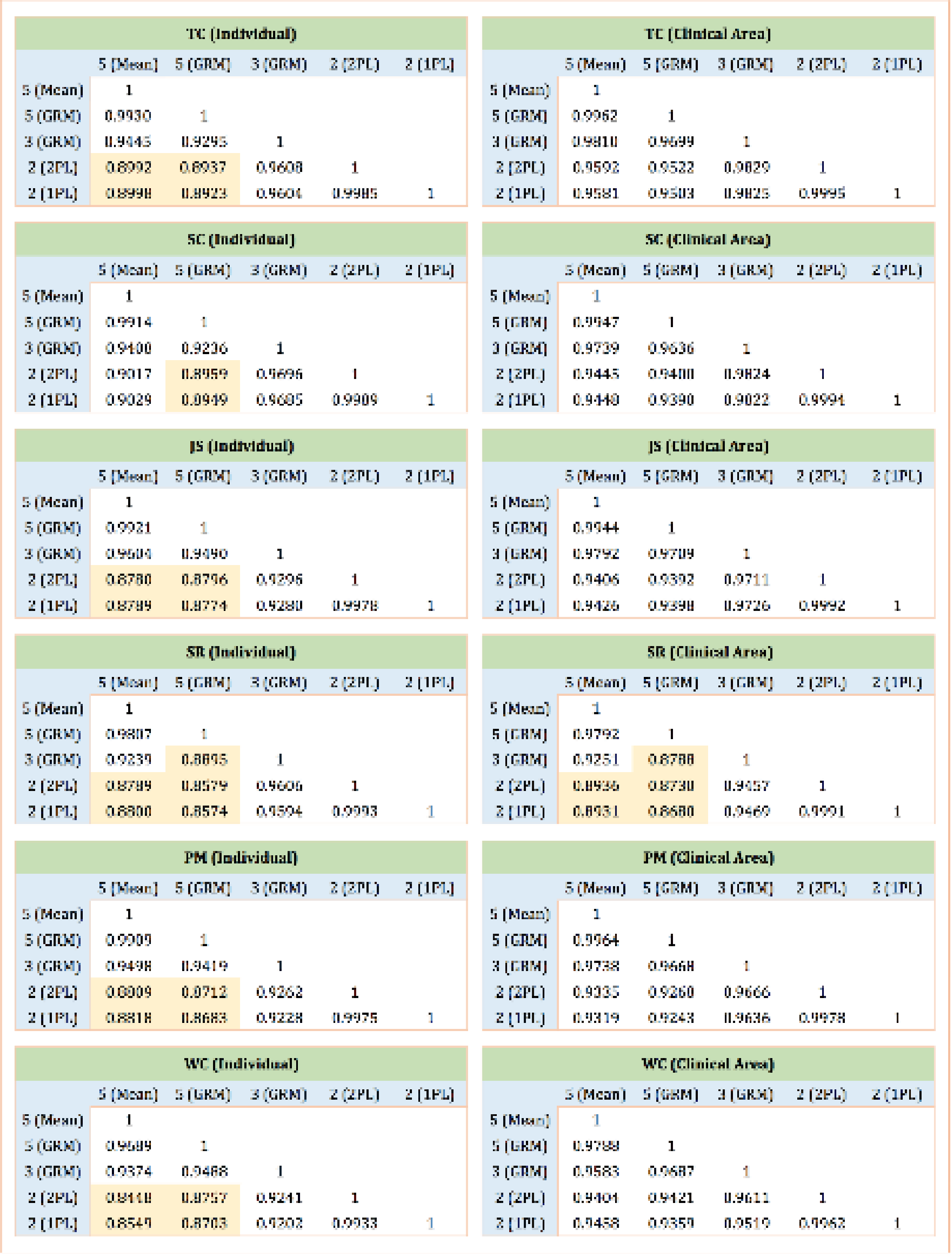
Figure 1 Scale collapsing scheme to generate trichotomized and dichotomized response data.

Next, we addressed the clinical area—there were 72 areas in this dataset—as the unit of analysis, and thus calculated domain scores for each clinical area by averaging individual respondents' scores obtained from the above step, yielding five scores from different measurement scales. Then, the correlations among them were calculated by domain, in the same way as done with the individual subject-level data structure.

Results

Table 2 shows the correlation coefficients among the domain scores from each of the five measurement scales. Let us describe the left pane, where an individual respondent was the unit of analysis. The left-most column—our primary interest—describes the comparison between the original scale (5 Mean) and the others. All correlation coefficients were very high, easily exceeding 0.9, and all correlations showed statistical significance (p<.01). Overall, the 3-point scale seemed to perform well, showinghigh correlation coefficients compared to the original simple 5-point Likert scale score. Correlation coefficients betweenthe dichotomized scale (1PL and 2PL) and the original scale were relatively smaller, although they were still higher than 0.8. Note that yellow shaded values mean the coefficient was lower than 0.9, the majority of which were involved with 1PL or 2PL. Across all combinations, the smallest correlation coefficient was 0.8448 between the original scale and the dichotomous (2PL) scale in the WC domain.

These high correlation coefficients certainly suggest a strong linear relationship among different levels of scale.[9] Yet the matrix plotstell a somewhat different story. As the number of response options decreased, the graphs were more scattered, although the linearity remained (Figure 1, left pane). This increased discrepancy was prominent when the response options were dichotomized.



TC (Individual)

	5 (Mean)	5 (GRM)	3 (GRM)	2 (2PL)	2 (1PL)
5 (Mean)	1				
5 (GRM)	0.9930	1			
3 (GRM)	0.9445	0.9295	1		
2 (2PL)	0.8992	0.8937	0.9608	1	
2 (1PL)	0.8998	0.8923	0.9604	0.9985	1

TC (Clinical Area)

	5 (Mean)	5 (GRM)	3 (GRM)	2 (2PL)	2 (1PL)
5 (Mean)	1				
5 (GRM)	0.9962	1			
3 (GRM)	0.9810	0.9699	1		
2 (2PL)	0.9592	0.9522	0.9829	1	
2 (1PL)	0.9581	0.9503	0.9825	0.9995	1

SC (Individual)

	5 (Mean)	5 (GRM)	3 (GRM)	2 (2PL)	2 (1PL)
5 (Mean)	1				
5 (GRM)	0.9914	1			
3 (GRM)	0.9400	0.9236	1		
2 (2PL)	0.9017	0.8959	0.9696	1	
2 (1PL)	0.9029	0.8949	0.9605	0.9909	1

SC (Clinical Area)

	5 (Mean)	5 (GRM)	3 (GRM)	2 (2PL)	2 (1PL)
5 (Mean)	1				
5 (GRM)	0.9947	1			
3 (GRM)	0.9739	0.9636	1		
2 (2PL)	0.9445	0.9400	0.9824	1	
2 (1PL)	0.9440	0.9390	0.9822	0.9994	1

JS (Individual)

	5 (Mean)	5 (GRM)	3 (GRM)	2 (2PL)	2 (1PL)
5 (Mean)	1				
5 (GRM)	0.9921	1			
3 (GRM)	0.9604	0.9490	1		
2 (2PL)	0.8780	0.8796	0.9296	1	
2 (1PL)	0.8789	0.8774	0.9280	0.9978	1

JS (Clinical Area)

	5 (Mean)	5 (GRM)	3 (GRM)	2 (2PL)	2 (1PL)
5 (Mean)	1				
5 (GRM)	0.9944	1			
3 (GRM)	0.9792	0.9709	1		
2 (2PL)	0.9406	0.9392	0.9711	1	
2 (1PL)	0.9426	0.9398	0.9726	0.9992	1

SR (Individual)

	5 (Mean)	5 (GRM)	3 (GRM)	2 (2PL)	2 (1PL)
5 (Mean)	1				
5 (GRM)	0.9807	1			
3 (GRM)	0.9239	0.8895	1		
2 (2PL)	0.8789	0.8579	0.9606	1	
2 (1PL)	0.8800	0.8574	0.9594	0.9993	1

SR (Clinical Area)

	5 (Mean)	5 (GRM)	3 (GRM)	2 (2PL)	2 (1PL)
5 (Mean)	1				
5 (GRM)	0.9792	1			
3 (GRM)	0.9251	0.8788	1		
2 (2PL)	0.8936	0.8730	0.9457	1	
2 (1PL)	0.8931	0.8680	0.9469	0.9991	1

PM (Individual)

	5 (Mean)	5 (GRM)	3 (GRM)	2 (2PL)	2 (1PL)
5 (Mean)	1				
5 (GRM)	0.9909	1			
3 (GRM)	0.9498	0.9419	1		
2 (2PL)	0.8809	0.8712	0.9262	1	
2 (1PL)	0.8818	0.8683	0.9228	0.9975	1

PM (Clinical Area)

	5 (Mean)	5 (GRM)	3 (GRM)	2 (2PL)	2 (1PL)
5 (Mean)	1				
5 (GRM)	0.9964	1			
3 (GRM)	0.9738	0.9668	1		
2 (2PL)	0.9335	0.9260	0.9666	1	
2 (1PL)	0.9319	0.9243	0.9636	0.9978	1

WC (Individual)

	5 (Mean)	5 (GRM)	3 (GRM)	2 (2PL)	2 (1PL)
5 (Mean)	1				
5 (GRM)	0.9689	1			
3 (GRM)	0.9374	0.9468	1		
2 (2PL)	0.8448	0.8757	0.9241	1	
2 (1PL)	0.8549	0.8703	0.9202	0.9933	1

WC (Clinical Area)

	5 (Mean)	5 (GRM)	3 (GRM)	2 (2PL)	2 (1PL)
5 (Mean)	1				
5 (GRM)	0.9788	1			
3 (GRM)	0.9583	0.9687	1		
2 (2PL)	0.9404	0.9421	0.9611	1	
2 (1PL)	0.9458	0.9359	0.9519	0.9962	1

Table 2 Correlation coefficients among different measurement scales

Note: 5 (Mean) isthe average item scores from the original 5-point scale; GRM and PL depict the IRT-based empirical Bayes estimates. All correlation coefficients were significant (p<.01); yellow shadingmeans coefficients were smaller than 0.9; in the left pane, individual respondentand clinical area were units of analysis in the left and right panes, respectively.

When we turn to the clinical level results, the estimates from different scales became much more similar. The correlation coefficients were higher (Table 2, right pane) than those from the individual respondent level (left pane). All values were higher than 0.9 except for those in the SR domain. In addition, matrix plots showed much improved linearity (Figure 2, right pane). For the 3-point scale, the most dots lie in an almost single line, and even 1PL and 2PL dichotomous scales showed good—tight—linearity with the original scale. As seen in the individual level analysis, the SR domain was an exception of such improvement, still showinga scattered pattern.

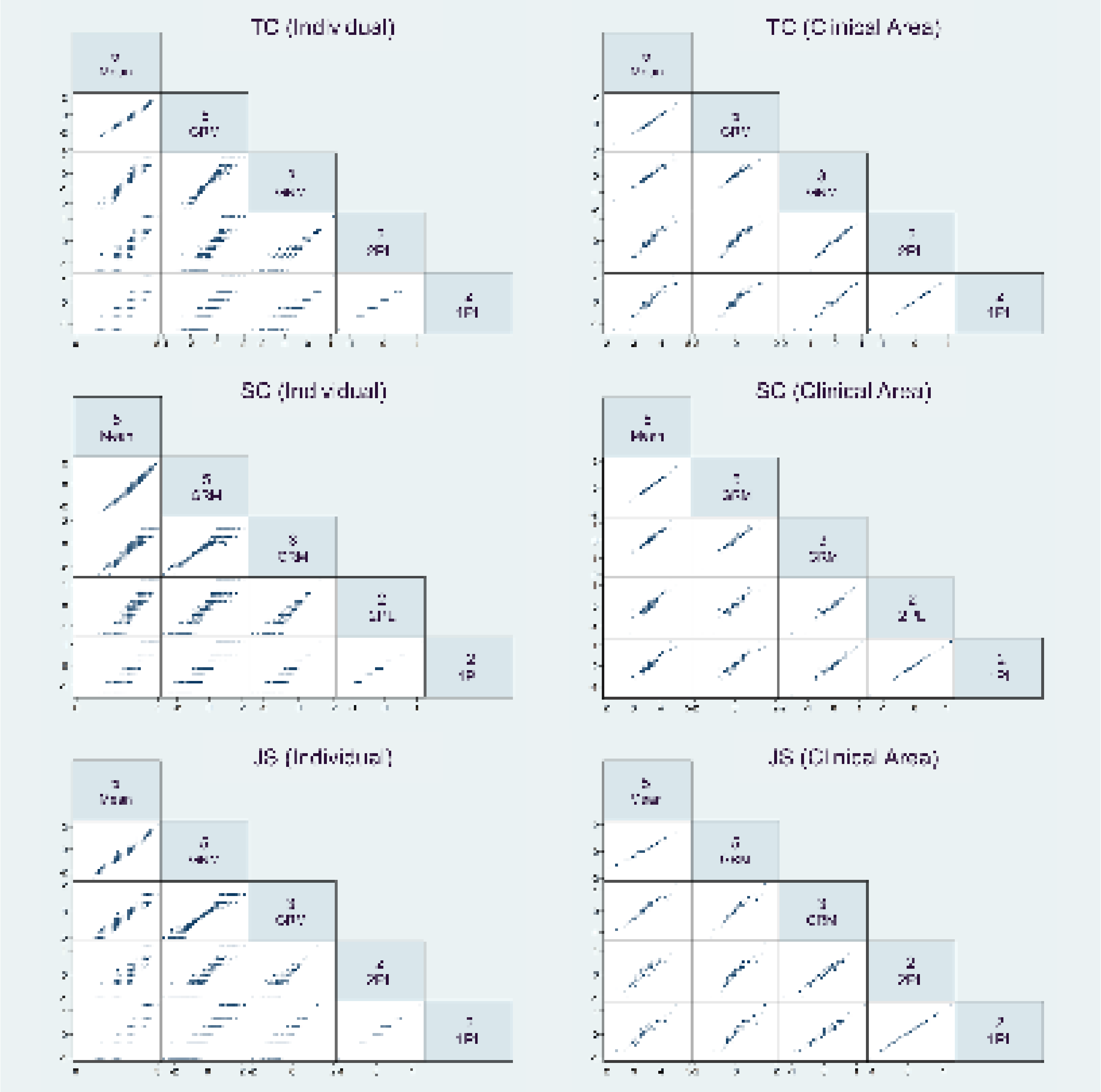


Figure 2 Matrix plots among different measurement scales.

Note: The left pane indicates the individual respondent level analysis; the right pane is the clinical area level analysis results.

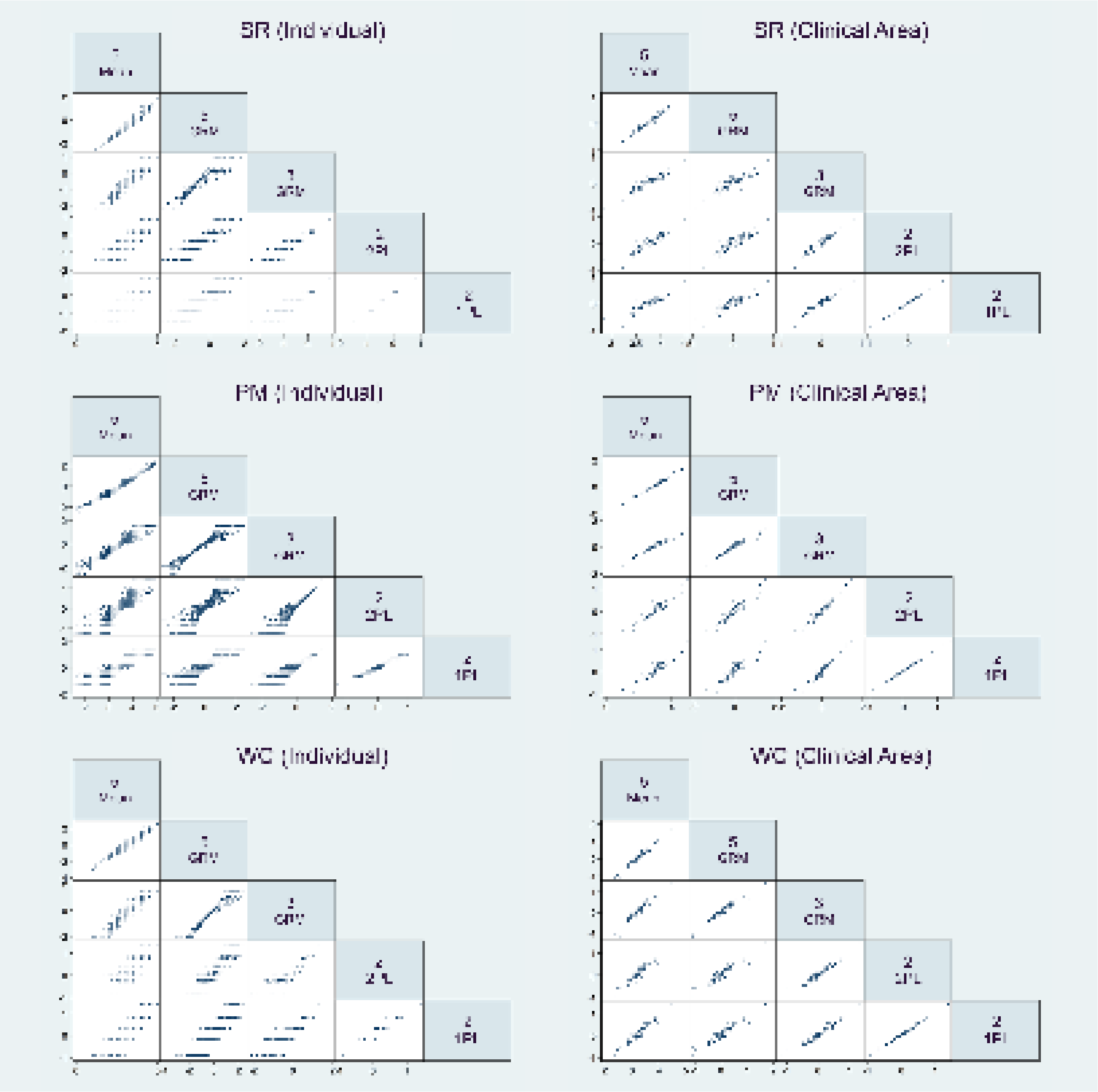


Figure 2 Matrix plots among different measurement scales. (continued).
Note: The left pane indicates the individual respondent level analysis; the right pane is the clinical area level analysis results.

Discussion

Culture is an intrinsically abstract concept; thus, measuring safety culture is not a cakewalk. To gauge such a notional idea, we use various measurement scales. Probably the most commonly used is a 5-point Likert scale. Responding to items like 'Nurse input is well received in this clinical area'(an item from SAQ)[6] by using the Likert scale allows us to quantify one's attitude to the statement. By measuring responses from multiple items, we can estimate safety culture with a reliable resolution.

The above approach seems impeccable, but in materializing it, we often overlook a fundamental principle: A very common misunderstanding is to think of a Likert scale as an interval scale, where the difference between adjoining response options are the same as seen in measuring temperature—namely, the difference between 50 and 51 degrees and 60 and 61 degrees is the same.Thus, with an interval scale, we can measure central tendency such as mean, median, and mode and obtain a variation of the data, like a standard deviation. Yet, technically, a Likert scale is an ordinal scale, where the amount of difference between response options is not the same or at least unknown. We cannot say that the difference between 'disagree strongly' and 'disagree slightly' and between 'agree slightly' and 'agree strongly' is the same.Assigning values of 1 to 5 to the response options is acceptable, but calculating any statistics like mean or even applying a regression model to the scores are, therefore, technically inappropriate.[10]

However, more often than not, we take a leap of faith, believing—or hoping—that the Likert scale works just as an interval scale or even ratio scale does.[11] On the strict definition of levels of scale, the SAQ and its variants, including the SAQ-K, are not free of the potential flaw at the theoretical level as they use a 5-point Likert scale as an interval scale. Despite this issue, the 5-point Likert scale and analyses of the responses as an interval scale are by far the *de facto* standard method used for SAQs and many other safety culture survey instruments, regardless of whether they are the gold standard or not. Reality is reality. Thus, we intentionally set aside the flawand confined ourselves to comparing the performance of other scales to the original *de facto* standard scale. The estimates we calculated for the other scales (dichotomized and trichotomized) were obtained through IRT (GRM for polychromous; 1 and 2PL for dichotomous), which completely recognizes the responses as ordinal scale data and analyzes them accordingly.

The results of this study were quite straight forward. Different measurement scales showed very high correlations with each other. However, for the individual respondent level, as the number of response options decreased, the graphs became a bit scattered, although the direction of the linearity remained well (Figure 2). This finding suggests that, for each respondent, measuring one's safety attitudes with an instrument based on collapsed response options might not serve as a direct substitute for the original 5-point scale-based measurement. However, with regard to the clinical area-specific scores obtained by aggregating scores from all healthcare professionals in a certain clinical area, simpler measurement scales like 3-point or even dichotomous scales provided safety culture estimates that are highly equivalent to those from the original 5-point scale. This phenomenon might be explained by the multiple respondents in an area offsetting the varying deviations from each individual. If so, in analyzing safety culture at the clinical area or higher level, such as hospital, region, and country, we can expect valid results from an instrument with a dichotomous or trichotomous measurement scale.

When estimating safety culture (called latent trait in psychological terminology) with collapsed scales, we used IRT. Thus, in order to better understand the rationale of such collapsing, we briefly introduce how IRT works behind the scenes. The premise of IRT is that each item has a certain range on the latent trait (safety culture)continuum where the item performs best—in other words, where an item can measure the trait most precisely. Theoretically, this region is where the probability that a person chose an option is 50%; the response is switched from not selecting an option to selecting that option.[12] To illustrate, if students are facing a very difficult question on a math exam, only a few students will give the correct answer. Such a question can distinguish whether a person has a very high capability in math because that switching point lies at a very high location along the trait continuum; therefore, the question does not provide much information about students with ordinary or low proficiency in math. That high trait region is where the question performs best.

Thus, naturally, items with multiple response options, like a 5-point Likert scale, can provide more information because there are four such switching points: between disagree strongly and disagree slightly, disagree slightly and neutral, and so forth. Those high information regionsare usually spread out along the latent trait continuum, although the level of spread varies a lot item by item. On the contrary, for a dichotomous response, only one best-performing region exists between disagree and agree, meaning the amount of information a single item can provide is concentrated in a relatively narrower region along the latent trait continuum.

As such, we might ask how we can collapse response options. For a single item, more response options definitely provide more information across the latent trait level. However, if we have multiple items for a safety culture (e.g., each SAQ-K domain contains four to ten items), even in a dichotomous response, the abovementioned switching points of those items can be spread out along a wide range of the latent trait continuum. Thus, by analyzing responses from items with different switching points, we still can measure the culture quite precisely. This approach is actually used in many fields. On the Graduate Record Examination (GRE), only a couple of dozen questions are asked, and the responsesare essentially treated as a dichotomized format (correct or incorrect), but still we receive scores with high granularity.[13] Again, the SAQ and its variants were not developed based on the IRT approach; thus, further study is required for each item's properties and performance when asked using different scales. Yet it is certain that, with such information, we can develop a more concise instrument in terms of both number of items and measurement scales.

We admit that the current study dealt with already collected data in a 5-point Likert scale and, therefore, our results might not have taken into account the potential bias that collapsed response options could cause in the administering phase. To illustrate, dichotomized scales (e.g., yes/no and disagree/agree) tend to force a respondent to choose a side. A person who does not have a concrete position when completing the survey might be troubled because there is no 'neutral' or 'neither disagree nor agree' option. A bigger issue is that once a 'neutral' person chooses a side, either a gree or disagree, the person may tend to answer the rest of the items as if she has already been in the position—that is to say, a kind of confirmation bias can be induced.[14] By including multiple seemingly unrelated items in an instrument,combined with the use of the IRT approach, we might be able to reduce such phenomena to a certain degree. Yet bias is bias. The ideal way is always to obtain the most unbiased possible data in the first place, and future study is certainly needed in this regard.

Maybe what we did in this study was a much simpler version of a non-inferiority test in clinical practice, which examines whether the effectiveness of a newly developed treatment is equivalent to that of the old treatment.[15] At least we began in that direction, just reporting whether trichotomized or dichotomized responses show equivalent safety culture estimates as the original scale so that they can replace the old one. Yes, we showed splendid correlation coefficients and scatter matrix plots that suggest such simplified scales might be an appropriate substitute for the 5-point scale. Indeed, those simpler scales would set us free from the zillions of combinations from contingency tables, meaning the fans of our computers no longer need to overheat. Yet as we moved forward, we struggled with a much more fundamental question: What is the real gold standard measurement scale for a safety culture survey instrument?

We cannot answer this question yet. For now, our best choice might be to ensure consistency by using the current SAQ measurement scale until the validity of a new scale is widely agreed upon and the scale is ready to roll out. That way, we can at least secure validity in the comparison of the safety culture across different time points and different organizations or regions from the data we collected so far. Still, that does not mean we should stay with the current tools forever. We can get closer to completely safe care only by improving the safety culture, and to improve the culture, we first need the best tool to measure it. Somebody once said "if we cannot measure it, we cannot manage it"; our interpretation is that, "if we can measure it better, we can improve it further." We have no right to settle for the present; rather, we are obligated to improve our safety culture measurement instruments and save millions of lives.

Conclusion

This article is the sixth in our SAQ-K series. As we drill down deeper and deeper, we see ourselves touching the more fundamental issues of the patient safety culture measurement process. This study is a kind of proposal to our colleagues and safety enthusiasts around world to overhaul the currently usedsafety culture measurement instrument sand find a more efficient way to measure and analyze the culture. We hope that this study can serve as a solid foundation for not only improving or remodeling the current instruments, but also developing new instruments in the future.

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. 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.
3. Jeong HJ, Lee WC. Item Response Theory-Based Evaluation of Psychometric Properties of the Safety Attitudes Questionnaire-Korean Version (SAQ-K). *Biometrics & Biostatistics International Journal*. 2016;3(5):1–15.
4. Jeong HJ, Jung SM, An EA, 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):1–8.
5. Jeong HJ, Jung SM, An EA, 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):1–11.
6. 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.
7. Orlando M, Sherbourne, Thissen D. Summed-score linking using item response theory: Application to depression measurement. *Psychological Assessment*. 2000;12(3):354–359.
8. Fayers P. Item response theory for psychologists. *Quality of Life Research*. 2004;13(3):715–716.
9. Lee RJ, Nicewander WA. Thirteen ways to look at the correlation coefficient. *The American Statistician*. 1988;42(1):59–66.
10. Agresti A, Kateri M. Categorical data analysis. 2011.
11. Jeong HJ, Lee WC. Ignorance or Negligence: Uncomfortable Truth Regarding isuse of Confirmatory Factor Analysis. *Journal of Biometrics & Biostatistics*. 2016;7(3):298.
12. Drasgow F, Hulin CL. Item response theory. *Handbook of Industrial and Organizational Psychology*. 1990;1:577–636.
13. Mislevy RJ. Book Review: A Review of Computerized Multistage Testing. *Journal of Educational and Behavioral Statistics*. 2015;40(4):425–431.
14. Nickerson RS. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*. 1998;2(2):175–220.
15. Kong L, Kohberger RC, Koch GG. Type I error and power in noninferiority/equivalence trials with correlated multiple endpoints: an example from vaccine development trials. *Journal of Biopharmaceutical*. 2004;14(4): 893–907.