MedCrave
Step into the World of Research

Biometrics & Biostatistics International Journal

Research Article

Open Access

CrossMark

Does differential item functioning occur across respondents' characteristics in safety attitudes questionnaire?

Abstract

Statistically, yes. Practically, maybe. A complete overhaul is suggested.

Keywords: safety culture, safety attitudes questionnaire, patient safety, item response theory, differential item functioning

Volume 4 Issue 3 - 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 01, 2016 | **Published:** August 11, 2016

Introduction

When we administer a survey questionnaire to a population, we implicitly assume that people with the same level of attributes being measured will give the same response to a certain item designed to measure the attributes.[1,2] Otherwise, the survey questionnaire is suspected of having limited value as a measurement instrument for the attributes. Yet this assumption is frequently challenged. For a realistic example, it is common that respondents from a country where humility is much encouraged give a lower score to an item about self-confidence than those from a country where more self-assured people are well-respected, even when their underground trait levels of self-confidence are the same. In more psychometric terminology, this phenomenon—namely, unequal responding patterns among groups—is called differential item functioning (DIF), which is a profound bias threatening survey-based research.[3] If DIF is in doubt, we naturally question whether a difference in survey scores between two groups stems from the real difference in the trait that we want to measure or DIF between the groups, at least to a certain degree.[4] Thus, it is essential to ensure equivalence in the responding pattern for survey items among groups before moving forward to any group-to-group comparison of survey scores and more sophisticated analysis. Unfortunately, however, more often than not, this step is omitted in survey-based studies.[5]

In this series of Safety Attitudes Questionnaire–Korean Version (SAQ-K) articles, we have intentionally postponed the discussion on DIF[6-9] because we planned to utilize item response theory (IRT) for DIF detection. Using IRT is known to be a superior method given its conditional invariance property, which enables better decisions on DIF than traditional sum scores of a questionnaire.[10] We waited for the successful application of IRT to SAQ-K, which we achieved in our most recently published article.[11] Thus, we can no longer put off this DIF investigation on SAQ-K.

Various approaches can be used to examine DIF, such as the Mantel-Haenszel (MH) method and logistic regression (LR)-based techniques.[12,13] For our SAQ-K data, we chose the LR approach—more specifically, an iterative hybrid ordinal logistic regression—because it can effectively handle the polytomous property of SAQ-K items (5-point Likert scale). The use of the MH method is somewhat limited to dichotomous variables.[3,14,15] In addition, the LR method has a higher power than the MH method in detecting items with DIF, albeit a downside of the power does exist (as discussed in a later section).[2,16,17]

Although various group criteria can be tested with DIF, we focused on a single criterion: job type. Specifically, we analyzed DIF between physicians and nurses, the groups that constitute the majority of healthcare professionals in a healthcare organization. This study is not a complete overhaul of DIF for SAQ-K; rather, we hope the findings of this study will guide or at least ignite further studies in the item functioning of patient safety culture survey instruments like SAQ. For readers not familiar with the approaches introduced in this article, we provide a brief overview.

A brief introduction to DIF and its detection

For easier understanding, let us begin with an item with a dichotomous response: correct or incorrect, coded as 1 and 0, respectively. In Figure 1, the two lines in each graph indicate the two groups: A (solid) and B (dashed). The x-axis is the latent trait level that the item is supposed to measure, and the y-axis is the log odds of the correct answer. The graphs depict different types of DIF (graphs are purposefully simplified to the level of linear function).

If the two groups respond to the item in the same manner (no DIF), the relationship between the log odds of the correct answer (y-axis) and latent trait level shows the pattern in graph on the left. Because the lines from groups A and B are superimposed, they appear as one single line.

If one group shows higher log odds scores over the entire range of latent traitlevel as shown in the middle graph, the phenomenon is called uniform DIF. On the contrary, if the slopes of the graphs are significantly different (meaning the lines might eventually cross somewhere in the latent trait continuum—maybe even outside the graph), and as such each group is favored over the different latent trait region, the condition is called non-uniform DIF (as in the graph on the right).[13, 17] In epidemiologic terminology, uniform DIF corresponds to confounding, and non-uniform DIF can be said to be effect modification.[3]

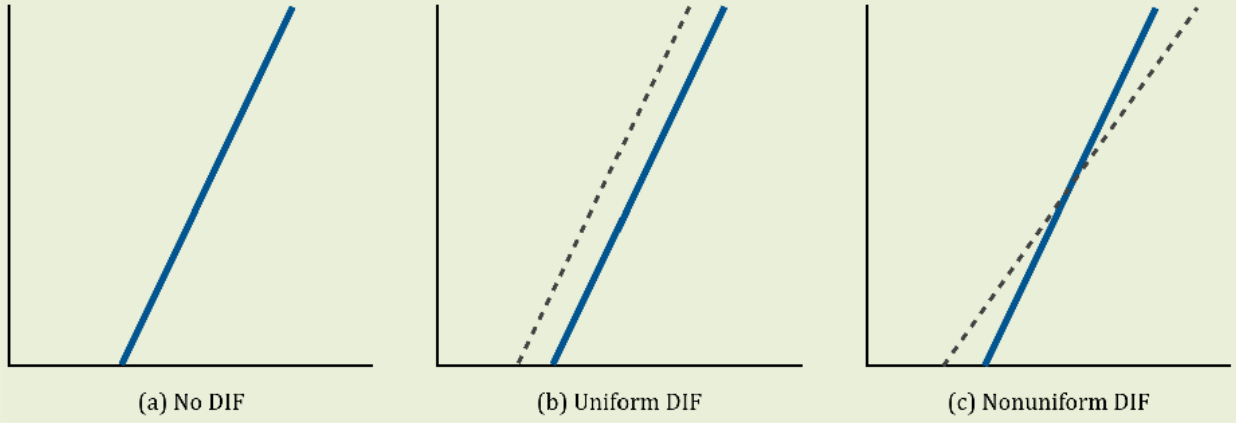

Figure 1 Three types of differential item functioning.

Note x-axis: latent trait level; y-axis:log odds of correct response; solid line: group A; dashed line: group B.

Here, we describe the discussed DIF types in the form of logistic regression models (Table 1). The essence of the LR-based approach for DIF detection is to examine whether a model has a better fit than the nested model, which can be tested with a likelihood ratio χ^2 test. Uniform DIF is investigated by comparing the log likelihood of model 1 with 2 (degree of freedom (*df*)=1) and non-uniform DIF by comparing model 2 with 3 (*df*=1). The comparison between model 1 and model 3 (*df*=2) is supposed to detect the total DIF effect—both uniform and non-uniform DIF.[2,3,15,17] For all three models, the number of response options for a particular item is the same; therefore, the *df* is determined solely by the number of regression coefficients in the models compared.[2]

Table 1 Logistic models for DIF detection

Logistic Model	DIF Type*
Model 1	No DIF (a)
Model 2	Uniform DIF (b)
Model 3	Non-uniform DIF (c)

Note DIF Type* is valid when the model has a better fit than the immediate nested model

This LR approach has been reported to show a good power for detecting DIF, but it has also raised the issue of inflated type I error rates, especially on large samples.A large sample may make the analysis so sensitive that items with a very small amount of DIF that might have been ignored are stigmatized as DIF items. Combined with some researchers' tradition of always dropping DIF-suspected items (even when the items still measure the intended latent trait), such inflated type I error creates inefficiency in instrument development and analysis.

Therefore, it is highly recommended to check the effect size in the model comparison process, such as through R^2 statistics. In the realm of logistic regression, R^2 calculation and interpretation have never been straightforward; thus, several pseudo R^2 statistics have been applied, such as Coxand Snell's, Nagelkerke's, and McFadden's R^2s. Observing these values suggests the magnitude of DIF for each item, but universal agreement on this is ideal, and knowing how to interpret it has not been defined yet.

In addition, proportional change in the regression coefficient between models can be used as both one of the DIF detection criteria and an effect size measure. To illustrate, uniform DIF is strongly suspected when a change in β_1 equal to or larger than 10 percent occurs between model 1 and model 2, albeit this 10 percent threshold is subjective. In simple language, this test criterion asks whether including a group term (β_2) influences the relationship between latent trait level and response.[2]

Other methods for DIF detection are available, such as testing the significance of regression coefficients in the models, but obviously no silver bullet that can be applied to every dataset has yet been invented. As such, varying methods have been utilized across studies, and the choice of method is left to the researcher's discretion. One thing for sure is that we should take advantage of a statistical significance-based approach like the likelihood ratio χ^2 test in conjunction with effect size measures like pseudo R^2 statistics or proportional β_1 change.

Methods

We used the same dataset from our previous studies on SAQ-K; the survey was conducted in a large metropolitan hospital in Seoul from October through November 2013. Detailed information as to survey process and participants can be found in our previous articles.[6-8] Note that this study used questionnaires collected only from doctors and nurses as these were groups to be compared.

SAQ-K consists of 34 items in six domains. The definition and number of items of each domain are summarized in Table 2.[6,21]

Table 2 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

DIF was tested within each domain, meaning six sets of tests were conducted. Items that revealed significance in any of the three likelihood ratio χ^2 tests with an alpha of 0.01 (Model 1 versus 2, Model 2 versus 3, and Model 1 versus 3) were flagged as having DIF.

In what follows, we describe the magnitude of DIF. For individual items, we calculated McFadden's pseudo R^2 statistics for the defined model comparisons. Among the various pseudo R^2 statistics, we chose McFadden's because, despite being a subjective decision, McFadden's pseudo R^2 statistic satisfies most of Kvalseth's eight criteria for a reliable R^2,[22,23] suggesting that it can serve as a decent measure. We then applied Zumbo et al.'s guideline to evaluate R^2: Below 0.13 is negligible, between 0.13 and 0.26 is moderate, and above 0.26 is large DIF.[17,24]

In addition, proportional change was obtained for additional information about the amount of uniform DIF: A 10% change was regarded as a meaningful size of uniform DIF.[2,25]

For individual participants, we calculated the difference between the DIF-adjusted (purified) SAQ-K score and the initial unadjusted score for each domain. Finally, we drew test characteristic curves to show the impact of DIF for each group.

In sum, we began from DIF detection and magnitude evaluation for each item, then moved to investigate the aggregate impact of DIF items on domain score for each individual participant and each group (i.e., doctors and nurses).

Results

Characteristics of respondents

Of the 1,381 questionnaires returned, we analyzed 987 questionnaires collected from 378 doctors and 609 nurses. Table 3 shows the characteristics of the respondents, inclusive of gender, work year, and job type.

Table 3 Characteristics of Respondents

Characteristics	N	%
Gender		
Male	230	23.3
Female	757	76.7
Work years		
Less than 6 months	68	6.9
7–11 months	109	11.0
1–2 years	177	17.9
3–4 years	225	22.8
5–10 years	250	25.3
11–20 years	117	11.9
More than 21 years	41	4.2
Job type		
Physician	378	38.3
Nurse	609	61.7
Total	987	100

DIF detection and effect size analysis

Table 4 shows the results of the logistic regression approach for DIF detection. DIF items revealed from the likelihood ratio χ^2 tests are highlighted in bold. Among the 34 SAQ-K items, 15 were flagged as DIF items; they all showed statistical significance from the total DIF effect test comparing Models 1 and 3: Pr(χ^2_{13}, 1) was smaller than 0.01. Ten items showed significance only in Pr(χ^2_{12}, 1) statistic from the comparison of Models 1 and 2, suggesting the typical uniform DIF manner. One item (SC6) showed significance only in Pr(χ^2_{23}, 1), suggesting it has an archetypal nonuniform DIF nature. Four items (TC1, JS1, JS4, and WC3) were significant for both Pr(χ^2_{12}, 1) and Pr(χ^2_{23}, 1) statistics.

Unlike the likelihood ratio χ^2 tests that raised the DIF flag for approximately half of the items, McFadden's pseudo R^2 statistics were negligible for all items (<0.13);[17,24] the largest was only 0.0651 in pseudo R^2 between Models 1 and 2 for TC5. To verify this result, we calculated other pseudo R^2 statistics, such Cox and Snell's R^2 and Nagelkerke's R^2, but they all yielded negligible values. For a proportional β_1 change, two items were higher than 10%: TC6 (10.78%) and SC1 (11.87%). These two items had statistical significance only from the χ^2_{12}, not the χ^2_{23}, suggesting an apparent uniform DIF.

Citation: Jeong HJ, Lee WC. Does differential item functioning occur across respondents' characteristics in safety attitudes questionnaire? *Biom Biostat Int J.* 2016;4(3):103–111. DOI: 10.15406/bbij.2016.04.00097

Table 4 Likelihood Ratio χ^2 Test Results, McFadden's Pseudo R^2, and Percent Change of β_1 Change

ID	Items	Statistics			McFadden's pseudo R^2			$\Delta\beta_1$ (%)
		Chi12	Chi23	Chi13	R12	R23	R13	
Teamwork climate								
TC1	**Nurse input is well received in this clinical area**	**.0000**	**.0000**	**.0000**	.0071	.0084	.0155	.6600
TC2	Disagreements in this clinical area are resolved appropriately (i.e., not *who* is right, but *what* is best for the patient)	.2171	.9322	.4651	.0006	.0000	.0006	.0300
TC3	I have the support I need from other personnel to care for patients	.9734	.1957	.4327	.0000	.0007	.0007	.0000
TC4	It is easy for personnel here to ask questions when there is something that they do not understand	.2740	.5781	.4709	.0005	.0001	.0006	.2200
TC5	**The physicians and nurses here work together as a well-coordinated team**	**.0000**	.0125	**.0000**	.0628	.0023	.0651	**10.7800**
Safety Climate								
SC1	**I would feel safe being treated here as a patient**	**.0000**	.0789	**.0000**	.0461	.0013	.0473	**11.8700**
SC2	**Medical errors are handled appropriately in this clinical area**	**.0018**	.8938	**.0074**	.0041	.0000	.0041	.3100
SC3	I know the proper channels to direct questions regarding patient safety in this clinical area	.0508	.0783	.0315	.0014	.0011	.0025	.1500
SC4	**I receive appropriate feedback about my performance**	**.0000**	.6333	**.0000**	.0131	.0001	.0132	1.2700
SC5	**I am encouraged by my colleagues to report any patient safety concerns I may have**	**.0010**	.5548	**.0038**	.0042	.0001	.0043	1.5800
SC6	**The culture in this clinical area makes it easy to learn from the errors of others**	.7935	**.0011**	**.0046**	.0000	.0044	.0044	.0700
Job Satisfaction								
JS1	**I like my job**	**.0000**	**.0085**	**.0000**	.0138	.0025	.0163	1.3300
JS2	Working here is like being part of a family	.1776	.8104	.3916	.0006	.0000	.0007	.7100
JS3	This is a good place to work	.0372	.7643	.1091	.0016	.0000	.0016	.1000
JS4	**I am proud to work in this clinical area**	**.0075**	**.0004**	**.0000**	.0027	.0048	.0075	2.0600
JS5	Morale in this clinical area is high	.1621	.0923	.0912	.0007	.0011	.0018	.6400
Stress Recognition								
SR1	**When my workload becomes excessive, my performance is impaired**	**.0004**	.0607	**.0003**	.0000	.0014	.0064	.4400
SR2	I am less effective at work when fatigued	.9492	.8186	.9721	.0000	.0000	.0000	.0300
SR3	I am more likely to make errors in tense or hostile situations	.5271	.9680	.8181	.0000	.0000	.0001	.1600

Citation: Jeong HJ, Lee WC. Does differential item functioning occur across respondents' characteristics in safety attitudes questionnaire? *Biom Biostat Int J.* 2016;4(3):103–111. DOI: 10.15406/bbij.2016.04.00097

Table Continued

ID	Items	Statistics			McFadden's pseudo R^2			$\Delta\beta_1$ (%)
		Chi12	Chi23	Chi13	R12	R23	R13	
SR4	Fatigue impairs my performance during emergency situations (e.g., emergency resuscitation, seizure)	.6489	.1608	.3372	.0000	.0007	.0008	.1500
Perception of Management								
PM1	Unit management supports my daily efforts	.5967	.3684	.5801	.0001	.0003	.0004	.1400
PM2	**Hospital management supports my daily efforts**	**.0033**	.0915	**.0032**	0035	.0012	.0046	.3900
PM3	Unit management doesn't knowingly compromise patient safety	.1514	.5124	.2883	.0008	.0002	.0010	.1000
PM4	Hospital management doesn't knowingly compromise patient safety	.8516	.1809	.4015	.0000	.0007	.0007	.0300
PM5	Unit management is doing a good job	.8760	.9452	.9856	.0000	.0000	.0000	.0300
PM6	**Hospital management is doing a good job**	**.0000**	.7495	**.0002**	.0066	.0000	.0066	2.1100
PM7	Problem personnel are dealt with constructively by our unit management	.0762	.7073	.1935	.0013	.0001	.0014	.2200
PM8	Problem personnel are dealt with constructively by our hospital management	.2687	.2487	.2789	.0005	.0005	.0010	.3100
PM9	I get adequate, timely info about events that might affect my work from unit management	.3319	.0901	.1485	.0004	.0011	.0015	.2600
PM10	I get adequate, timely info about events that might affect my work from hospital management	.6260	.8889	.8794	.0001	.0000	.0001	.0900
Working Condition								
WC1	**The levels of staffing in this clinical area are sufficient to handle the number of patients**	**.0000**	.9814	**.0000**	.0167	.0000	.0167	1.5700
WC2	**This hospital does a good job of training new personnel**	**.0000**	.8747	**.0000**	.0080	.0000	.0080	2.1600
WC3	**All the necessary information for diagnostic and therapeutic decisions is routinely available to me**	**.0000**	**.0061**	**.0000**	.0084	.0033	.0117	.8100
WC4	Trainees in my discipline are adequately supervised	.6223	.9509	.8840	.0001	.0000	.0001	.0900

Note Chi12, Chi23, and Chi13 denote the likelihood ratio χ^2 statistic between Models 1 and 2, Models 2 and 3, and Models 1 and 3, respectively; R12, R23, R13 indicate McFadden's pseudo R^2 from a comparison Models 1 and 2, Models 2 and 3, and Models 1 and 3, respectively; $\Delta\beta_1$ is proportional change in β_1 and shown as a percentage.

Graphical analyses of DIF effects

From the above analyses, we generated various graphs. Due to space constraints, we present only those graphs with the most distinctive results here. Again, the purpose of this study is to test DIF in SAQ-K; what follows mainly explains how to interpret such graphical results and perform various diagnostic analyses for readers.

depicts DIF pattern by item; we selected two items (SC1 and SC6) as examples. The upper-left graph (SC1) shows a typical uniform DIF pattern; the item score (y-axis) for doctors (solid line) is higher than that for nurses (dashed line) over the entire range of latent trait level (x-axis). On the other hand, in the upper-right graph (SC6), two lines cross at around zero in the latent trait continuum; in the higher latent trait region, nurses' score was larger than doctors' and the lower latent trait level region doctor's score was larger than nurses', suggesting typical nonuniform DIF pattern.

We also depicted item response functions to deeply understand how DIF arises in the SAQ-K items with categorical response. In the graphs for item response function, there should be five curves (lower-right graph, SC6), each of which corresponds to each of the response options of the 5-point Likert scale (1 through 5 in our case). The reason why there are only four curves for SC1 (lower-left graph) was that too few respondents chose the first category; thus, it was collapsed into the second category. Eventually, the analysis for SC1

was done with four response categories (1 and 2 together, 3, 4, and 5 in the Likert scale). The x-axis is respondents' latent trait as always, and the y-axis is the probability of a category being chosen.

In the item response function of SC1, for every response category, the solid line (doctors) was located to the left of the corresponding dashed line (nurses), suggesting that over the entire range of the latent trait level, doctors had a greater propensity to choose a higher response option than nurses. On the contrary, for the item response function of SC6, in the latent trait region higher than zero, dashed lines (nurses) are located to the left of the solid lines (doctors); in the lower latent trait region, the solid lines (doctors) are on the left. These item response functions for each category clearly explain how different DIF patterns could be generated behind the curtain. Other DIF items basically share similar patterns of the above two items, albeit the magnitudes varied.

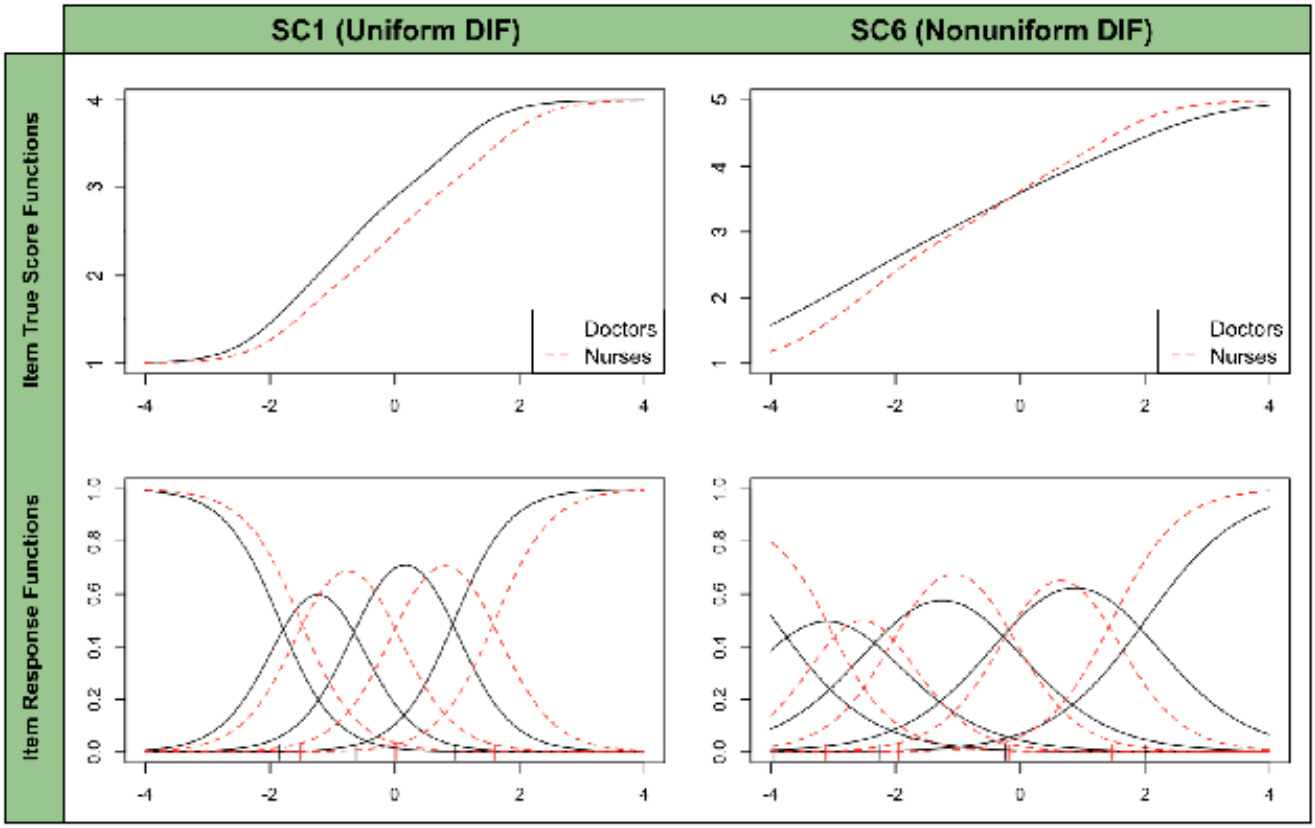

Figure 2 Item True Score Functions and Item Response Functions of DIF Items.

Now, we turn to DIF impact on individual respondents.Figure 3 shows the difference in latent trait level between the initial IRT-based trait estimate (DIF was ignored) and the purified trait estimate (DIF was accounted for). Unlike Figure 2, where item characteristics from the entire population were presented, Figure 3 shows an individual respondent's information: Each circle and triangle stands for the latent trait estimate of a domain for a single participant. In a word, Figure 3 shows the expected amount of change in latent trait estimates when DIF is accountedfor. In both graphs, the y-axis is the difference (initial – purified) and the x-axis of the right graph is the initial latent trait level.Here, we present the TC domain as it shows the most overt pattern.

The box plot on the left shows that the median difference was around -0.03, and the interquartile range(i.e., the middle 50% of respondents, depicted as the box)ranged from approximately -0.06 to 0.08. This is a typical right-skewed distribution with outliers at high values. The right graph shows the initial – purified difference against the initial latent trait level that ignored DIF. The interpretation of this graph is as follows: Across the entire latent trait continuum, doctors (circle) show a positive difference, suggesting that accounting for DIF leads to lower scores than the initial scores; for nurses (triangle), the pattern is reversed.

What is delineated above corresponds well with the test characteristic curves (TCC) in Figure 4. The y-axis of TCC is the possible score from a domain.As there are five items with the 5-point Likert scale ranging from 1 to 5, the minimum value is 5 and the maximum is 25 for the y-axis (left graph). Given the two DIF items (TC1 and TC5), the y-axis of the right graph ranges from 2 to 10. The point here is that, for DIF items, we can observea clear difference in TCC curves between doctors and nurses, although the absolute magnitude was not that much. On the other hand, the left graph for all items displayeda much smaller difference than the right graph. This can be explained as the DIF impact being diluted by the non-DIF items, resulting in the overall score difference between the two groups becoming minimal.

Citation: Jeong HJ, Lee WC. Does differential item functioning occur across respondents' characteristics in safety attitudes questionnaire? *Biom Biostat Int J.* 2016;4(3):103–111. DOI: 10.15406/bbij.2016.04.00097

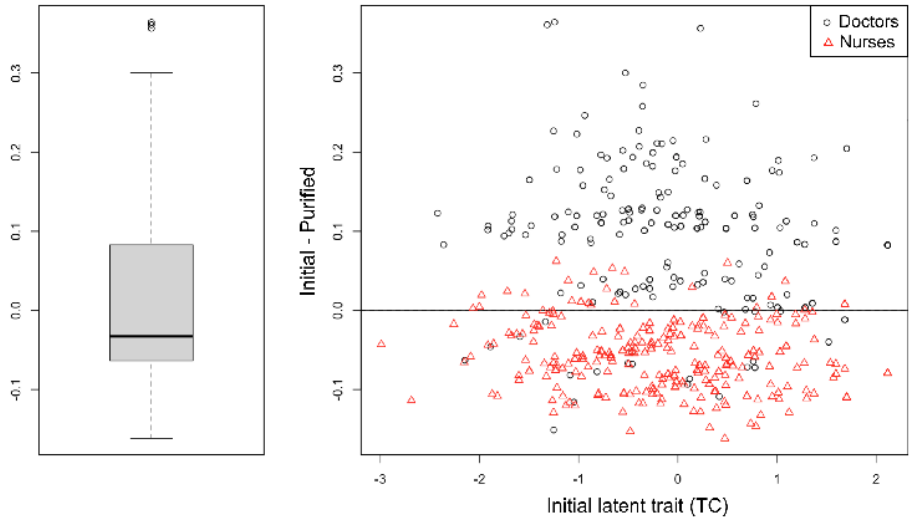

Figure 3 Difference between Initial and Purified TC Domain Scores for Each Respondent.
Note The y-axis is a purified score that accounts for DIF subtracted from the naïve score that ignores DIF.

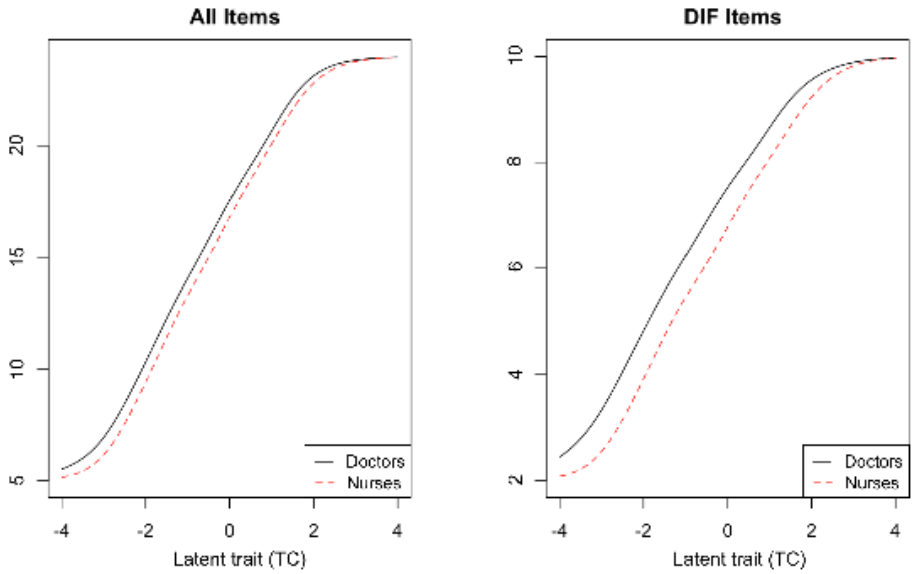

Figure 4 Test Characteristic Curves for All Items and DIF Items.

Note The y-axis is the sum of the item scores of the TC domain.

Discussion

This study aimed to test whether SAQ-K is a DIF-free instrument. To this end, we utilized the LR approach to handle the categorical response with a 5-point Likert scale of SAQ-K and revealed that 15 items had a statistically significant but practically minimal amount of DIF, thereby answering our research questions. Although providing detailed contextual implications of all DIF items is beyond the scope of this article, a few things about the SAQ-specific results are worth mentioning here.

First, TC1 ("The physicians and nurses here work together as a well-coordinated team") was tagged as a typical uniform DIF item that favored doctors across the entire TC trait continuum to a considerable degree (proportional β_1 change was bigger than 10%). Previous studies (for the hospital's internal use and, thus, not published) that did not account for DIF reported that the raw score from doctors was much higher than that of nurses. Some researchers have suggested that this phenomenon stems from the difference between doctors and nurses in how they define a "well-coordinated team." For instance, doctors may think of a good team as simply "doctors order and nurses follow them well" whereas nurses point out that their active participation in the decision-making process is essential for good teamwork [26]. Therefore, the detected DIF might have stemmed from the difference in how different groups interpret the item differently.

The other uniform DIF item, SC1 ("I would feel safe being treated here as a patient") also showed a higher score from doctors than nurses. The perception of the word "safe" might have differed between these groups. Doctors might perceive safe not as "free from medical errors," as SAQ originally intended, but as "clinical quality of treatment is reliable." Although this interpretation is just

our retrospective conjecture, it is worth conducting an in-depth investigation. Ultimately, an item's definition should be clearly presented in the survey instrument to prevent DIF that could have surely been avoided. We strongly recommend conducting a thorough pilot study for every instrument development process and resolving any discordance in item interpretation among different groups before rolling out the instrument widely.

Another issue to consider is the number of items to include on a survey questionnaire. For an instrument like the Patient-Reported Outcomes Measurement Information System (PROMIS), a relatively large number of items are included in a domain, at least in the item bank.[27] Yet, for many instruments designed for healthcare professionals, the number of items is reduced as much as possible considering their busy schedules. The SAQ-K is no exception; indeed, five out of six SAQ domains have only 4 to 6 items, and only the perception of management domain has 10 items. The problem is evident when multiple items show DIF; in other words, too few items are DIF free. Those non-DIF items serve as an anchor to calibrate DIF items across groups. Therefore, having enough DIF-free items in an instrument gives more stability. As it is practically not easy to increase the number of items in an instrument in a hospital setting, replacing DIF items with newly developed and tested non-DIF items might be a way to solve the issue.

In the statistical DIF detection process, type I and II errors (false positive and false negative, respectively) for DIF items can result from the impact of the other DIF items. DIF detection begins with measuring the latent trait; this trait level measurement itself is affected by DIF items.[10,28] To handle this issue, we applied Crane et al.'s approach, iterative detection and updated latent trait ability estimation.[2] In that approach, latent trait measurement is conducted with IRT and DIF is detected based upon the measurement. Then, the DIF items are purified. Using the new values, these steps are repeated until the same DIF items are flagged twice in a row.[20] The fundamental strength of this approach is that we retain all DIF items while checking whether the type I and II errors in DIF detection steps influenced the initial findings.[4,29] Considering the small number of items of SAQ domains, we think utilizing this approach while retaining all items is quite appropriate and, indeed, recommended.

The determination of a threshold to detect DIF has been actively discussed among researchers. Traditionally, we use a predetermined threshold, like alpha of 0.01, and compare the likelihood ratio χ^2 statistic to the threshold. A newly developed method is to utilize Monte Carlo simulation to set the empirical threshold.[20] To illustrate, instead of using a certain alpha value like 0.01, we can generate many simulated datasets from the original data and calculate the χ^2 statistic from each dataset. Then the 99th percentile of the statistics is the empirical threshold of DIF detection. For example, we run 1,000 cycles and the 10th smallest value is the threshold (empirical alpha value 0.01). If the value is 0.007, we can say we may expect to better control type I error with this threshold than with the nominal alpha, 0.01. The same logic can be applied to effect size measures too, but we do not recommend doing so because simulated R^2 is not meaningful in evaluating the magnitude of DIF.

In this particular study, we tried to detect DIF mainly using statistical significance, which naturally led us to address less the domain-or instrument-level impact of DIF items. Although not described in this article, we found that the favored group varies across items. To illustrate, for a certain uniform DIF item, doctors' scores can be higher; for another item, the opposite is true. In the domain level, these opposite DIF effects may be canceled out to some degree, and the total DIF effects get smaller, maybe even reaching a negligible level. For nonuniform DIF items, it is much more complicated. Therefore, caution should be exercised when evaluating the total amount of impact that DIF brings to a domain or instrument as a whole. At the population level, the proportion of respondents who answer a certain response category also influence the final effect. Let us assume that an item shows DIF mainly in a trait range corresponding to high response options (e.g., 4 or 5 on a 5-point Likert scale) and not much difference in the trait region of the lower response category (e.g., 1 or 2). If most respondents choose the lower response category, the population level effect of DIF would be minimized. Of course, in these examples of domain (instrument) level and population level, the direction of the DIF effect change can be reversed: The DIF impact would be more amplified instead of being canceled out or reduced.

This study is rather preliminary in that it uses a dataset from a single hospital, but we have shown that the methodology described here worked well in a healthcare setting. To build a broad and concrete knowledge base, further studies in different organizations are needed. Also, there are many different group variables in healthcare, such as work experience (duration), seniority, and full time versus part time. Even the size of a clinical area (number of healthcare workers) might cause DIF. Doctors and nurses are just one example. All these group variables are worth checking with DIF to maximize the effectiveness of an instrument, although this would be a huge undertaking. Obviously, it cannot be done in one night, but it should be done.

Conclusion

This is the fifth episode of the SAQ-K series. We are not sure whether it ends as a pentaptych or goes further. One thing is for sure: Safety culture plays a key role at some point in every accident causation [30]; thus, we have to shape the culture in an effort to improve patient safety. Thus, we developed and utilized instruments like SAQ. Thus far, the medical society has neglected DIF in safety culture measurements. Maybe we have been too busy solving impending problems threating patients, but that should not be an excuse any longer. We all know that, if we cannot measure safety culture, we cannot manage it. This study showed that we may not be fine with the measurement part; thus, a complete overhaul of the measurement instruments should be the first priority. We urge our colleague researchers to participate in this endeavor. Of course we will do so as well. We promise.

Acknowledgement

None.

Conflict of interest

None.

References

1. Glanz K, Rimer BK, Viswanath K. *Health Behavior and Health Education: Theory, Research, and Practice*. 4th edn. 2008;1–592.
2. Crane PK, Gibbons LE, Jolley L, et al. Differential item functioning analysis with ordinal logistic regression techniques: DIFdetect and difwithpar. *Med Care*. 2006;44(11 Suppl 3):S115–S123.
3. Crane PK, van Belle G, Larson EB. Test bias in a cognitive test: differential item functioning in the CASI. *Stat Med*. 2004;23(2):241–256.
4. Crane PK, Cetin K, Cook KF, et al. Differential item functioning impact in a modified version of the Roland–Morris Disability Questionnaire. *Qual Life Res*. 2007;16(6):981–990.

5. Gregorich SE. Do self–report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Med Care*. 2006;44(11 Suppl 3):S78–S94.
6. Jeong HJ, 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.
7. Jeong HJ, Jung SM, 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.
8. Jeong HJ, Minji Kim, 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.
9. Lee GS, Park MJ, Na HR, et al. A Strategy for Administration and Application of a Patient Safety Culture Survey. *Journal of Quality Improvement in Health Care*. 2015;21(1):80–95.
10. Millsap RE, Everson HT. Methodology review: Statistical approaches for assessing measurement bias. *Applied psychological measurement*. 1993;17(4):297–334.
11. 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.
12. Clauser BE, Mazor KM. Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*. 1998;17(1):31–44.
13. Camilli G, Shepard LA. Methods for identifying biased test items. *ERIC*. 1994;33(2):253–256.
14. Swaminathan H, Rogers HJ. Detecting differential item functioning using logistic regression procedures. *Journal of Educational measurement*. 1990;27(4):361–370.
15. Zumbo BD. *A handbook on the theory and methods of differential item functioning (DIF)*. Ottawa: National Defense Headquarters. 1999.
16. Rogers HJ, Swaminathan H. A comparison of logistic regression and Mantel–Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*. 1993;17(2):105–116.
17. Jodoin MG, Gierl MJ. Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*. 2001;14(4):329–349.
18. Agresti A. *Categorical data analysis*. John Wiley & Sons, USA. 1996;1–701.
19. Stata Corp. *Stata 14 Item Response Theory Reference Manual*. College Station, TX: Stata Press. 2015.
20. Choi SW, Gibbons LE, Crane PK. Lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *Journal of statistical software*. 2011;39(8):1–30.
21. 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):1–44.
22. Kvålseth TO. Cautionary note about R2. *The American Statistician*. 1985;39(4):279–285.
23. Allison PD. Measures of fit for logistic regression. *In Proceedings of the SAS Global Forum 2014 Conference*. 2014;1–13.
24. Zumbo B, Thomas D. *A measure of DIF effect size using logistic regression procedures*. National Board of Medical Examiners: Philadelphia, PA. 1996.
25. Crane PK, Gibbons LE, Ocepek–Welikson K, et al. A comparison of three sets of criteria for determining the presence of differential item functioning using ordinal logistic regression. *Qual Life Res*. 2007;16(Suppl 1):69–84.
26. http://www.ahrq.gov/professionals/quality–patient–safety/cusp/index.html.
27. Bevans M, Ross A, Cella D. Patient–Reported Outcomes Measurement Information System (PROMIS): efficient, standardized tools to measure self–reported health and quality of life. *Nurs Outlook*. 2014;62(5):339–345.
28. Holland PW, Wainer H. Differential item functioning. *Routledge*. 2012;1–470.
29. Reise SP, Widaman KF, Pugh RH. Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychol Bull*. 1993;114(3):552.
30. Jeong HJ, Julius C, Kim M, et al. Major cultural–compatibility complex: Considerations on cross–cultural dissemination of patient safety programmes. *BMJ Quality & Safety*. 2012;21(7):612–615.

Citation: Jeong HJ, Lee WC. Does differential item functioning occur across respondents' characteristics in safety attitudes questionnaire? *Biom Biostat Int J.* 2016;4(3):103–111. DOI: 10.15406/bbij.2016.04.00097