MedCrave
Step into the World of Research

Biometrics & Biostatistics International Journal

Research Article

Open Access

CrossMark

The pure and the overarching: an application of bifactor model to safety attitudes questionnaire

Abstract

Although this article is long, we believe that it is worth reading. It answers the most fundamental question that patient safety professionals have asked: How do we measure the overall safety attitudes score of each person? Broadly speaking, we showed how to structuralize various safety-related traits in one's mind, using the Safety Attitudes Questionnaire-Korean version as an example. We applied the bifactor model, which explicitly contains a general safety attitudes dimension that governs all SAQ-K items as well as six original SAQ-K domains. The major finding regarding the model structure is that several items might not fall under a specific SAQ domain, although they are still largely governed by safety attitudes in general; yet the stress recognition domain might not be part of the general construct—namely, safety attitudes. However, the more important information that we intended to share was that the bifactor model can effectively take control of the seemingly inadequate items or domains and calculate domain-specific scores and general safety attitudes score by providing different weights for each item. Thus, we can obtain much more purified domains scores from the data, compared to the traditional mean score approach. In addition, the item response theory-based approach used in this article gives more solid theoretical strength in handling the ordinal data and also offers the possibility of adding or dropping items or even a domain, while still allowing a longitudinal analysis—comparing scores from different versions of safety attitudes questionnaire. In addition to these theoretical strengths, the bifactor model provides exceptional computational efficiency compared to the other models we have tried, thereby allowing us to unlock an extremely large and complex dataset that could not be analyzed thus far. We hope the approach introduced in this article can help all the patient safety professionals achieve more precise and valid information from their already collected safety attitudes data as well as data collected in the future, ultimately saving more lives.

Volume 4 Issue 6 - 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: October 20, 2016 | **Published:** November 09, 2016

Disclaimer

The target audience of this article is those who actually use safety culture assessment tools in the healthcare setting, and we assumed they possess basic- to medium-level knowledge in statistics and psychometrics. Thus, we intentionally avoided showing complex formulae; instead, we tried to simplify, and sometimes oversimplify, the concepts introduced in this article to facilitate understanding. Terminology was also carefully chosen for those who did not major in statistics. In addition, we did not follow a typical research article style. We intentionally allocated many paragraphs usually reserved for the discussion section to the results section in order to facilitate understanding. Anyone curious about full mathematical descriptions are asked to please contact the authors.

Introduction

A dozen of our previous articles on the Safety Attitudes Questionnaire-Korean version (SAQ-K) introduced several new approaches to analyze collected data[1-5] and even novel methods to improve the instrument itself.[6-9] Yet we might not have answered a fundamental question: How are safety attitudes as a whole constructed in our minds? In other words, exactly how are SAQ domains interrelated, comprising more general higher level safety attitudes? To help understand this question, we would like to share what happened when we recently administered the survey in a hospital.

The anecdote actually begs a fundamental question as to how we have conceptualized the structure of SAQ domains thus far. For a smoother explanation, let us assume we are taking an English exam. The exam consists of grammar, reading, listening, writing, and speaking domains, each of which consists of a few questions. This is a very similar situation to SAQ. The difference is when we ask our English score, the teacher always provides both our overall English score and each domain score. We are not saying that the school system uses more advanced and elegant statistical models; rather, it is simply natural that we are curious about how we perform on a subject in general (overall score) in addition to the more specific scores within defined domains (e.g., grammar, reading, writing). Actually, more often than not, such a general score is treated as important as domain-specific scores.

A healthcare professional who had just completed the 34 items on the SAQ asked us, "What are you really measuring? Could you give me specific information, rather than just broadly saying safety culture or attitudes?" We answered that we were measuring six domains: teamwork climate (TC), safety climate (SC), job satisfaction (JS), stress recognition (SR), perception of management (PM), and working conditions (WC). She then stated, "I see. So you picked six domains comprising the concept of safety attitudes, meaning six scores will probably be calculated. But can you please show me how those domains form the safety culture or attitudes you said you want to measure?" It was a seemingly simple question so, with a gentle smile, we placed a piece of paper on the desk and drew the correlated factors (traits) model that we used for the confirmatory factor analysis during the instrument development phase (from now on, we depict only three of the six domains, TC, SC, and JS, to save space) (Figure 1).[10]

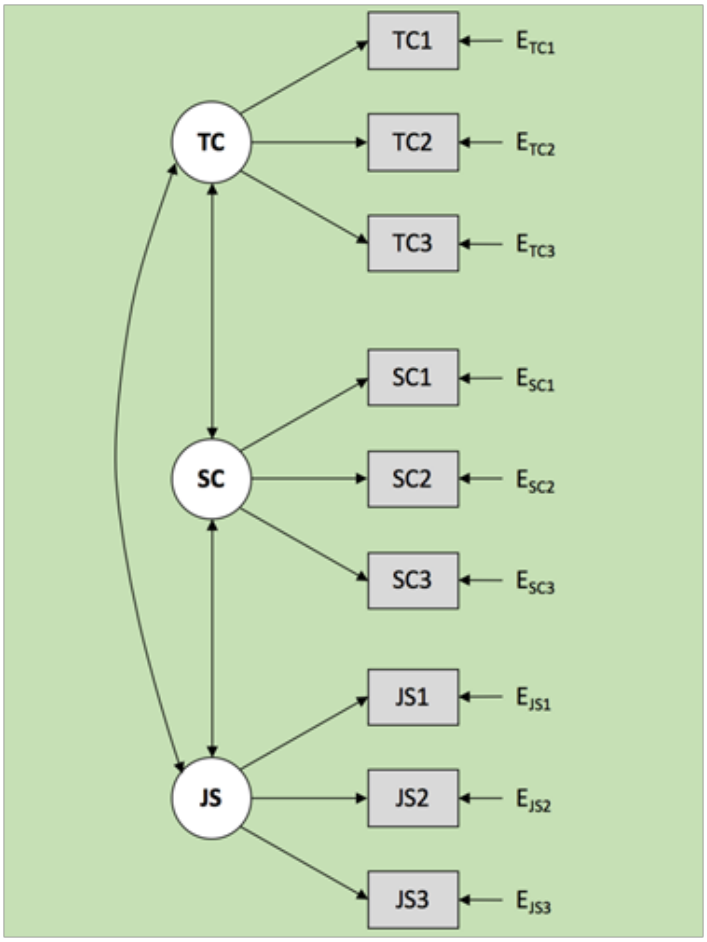

Figure 1 Correlated factors model for SAQ domains.

Note Rectangles are SAQ items; circles are SAQ domains; E means error term. These abbreviations will apply to all subsequent model presentations.

On the contrary, thus far, we have conventionally thought that SAQ consists of the six interrelated domains depicted in Figure 1, the correlated factors model. However, unfortunately, this model does not explicitly describe the higher level construct—namely, safety attitudes as a whole. Thus, we obtained only the six domain-specific scores and were unable to answer the healthcare professional's question about the overall score.

Despite this weakness, the correlated factors model has been a de facto standard or tradition that we have long followed without question. Yet history tells us "once the time comes, an old custom ought to be broken down"; we believe that time is now. We should answer the healthcare professional and ourselves. Through this article, we discuss the three most exhaustively tested models and their application to the SAQ data structure.

Finding the general score: second-order model

Let us begin this challenge by developing a model with the idea of a higher level overarching general construct. Such a model is often called the second-order model (Figure 2).[11]

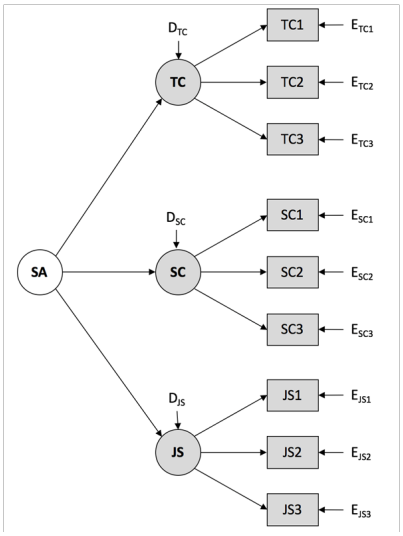

Figure 2 Second-order model.

Note D Disturbance for each domain.

Citation: Jeong HJ, Lee WC. The pure and the overarching: an application of bifactor model to safety attitudes questionnaire. *Biom Biostat Int J.* 2016;4(6):222–232. DOI: 10.15406/bbij.2016.04.00110

In the second-order model, the domains are dependent on only the general dimension (SA), and items depend on each of the specific domains. Therefore, SAQ domains are conditionally independent; in other words, the double-headed arrows among domains in the correlated factors model in Figure 1 are gone because the general SA dimension takes up their effects. The downside of this approach is that each domain might be treated as a disturbance of the general dimension, [12,13] which causes huge trouble for patient safety professionals, who mainly need domain-specific scores. In addition, the direct relationship between the overarching SA dimension and each item (rectangle) is not easily understood because it takes two steps for SA to reach each item. In sum, the price of the straightforward visual description of the model structure that takes into account the overarching SA dimension is not trivial: This model may not provide us with clear domain scores that most healthcare organizations want. Thus, we naturally look for another model that still embraces the overarching SA dimension while avoiding the weaknesses of the second-order model. Luckily, we have one.

Obtaining overarching safety attitudes score while estimating purified domain scores: bifactor model

In Figure 3 we present the bifactor model.[11]

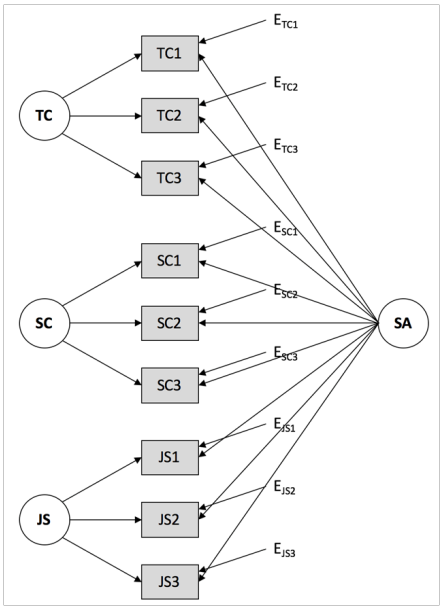

Figure 3 Bifactor model.

To clarify, we will review this model step by step. At first look, each and every item depends on both the general SA dimension and one of the specific SAQ domains (TC, SC, JS). In addition, no correlations are assumed (orthogonal) among specific domains because they are captured by the general SA dimension on which each item depends. This bifactor model might not be as intuitive as the second-order model for some readers, but the bottom line is that the bifactor model is actually a less restrictive model than the second-order model. In other words, the bifactor model is a generalization of the second-order model; thus, results that support the second-order model naturally support the bifactor model.[12] The take-away message is that the bifactor model covers both the general SA dimension and each domain simultaneously very clearly and explicitly, without the weakness of the second-order model.

For those who deal with SAQ data in the field, the advantage of applying the bifactor model to SAQ is huge. First, if we want to find a predictive relationship between SAQ domains and a specific clinical outcome (e.g., healthcare-associated infection rate), the domain scores from the bifactor model can be directly plugged in to a multiple regression model as independent variables.[13] Why is this so important? One of the fundamental assumptions of multiple regression is that independent variables should not be correlated with each other (multicollinearity). Thus, if we rely on the traditional correlated factors model (Figure 1), we are technically prohibited from building a predictive model with SAQ domains as covariates from the beginning because the model already connotes the correlations among the SAQ domains that we want to use as independent variables. Researchers frequently violate such a no- correlation assumption for independent variables in a multiple regression due to ignorance, negligence, or even groundlessly claimed practicality. However, we also know that a sandcastle eventually crumbles: Any analyses not founded on rock-solid methodology can never bring us precise results that lead to actual improvement in patient safety, especially in the long run. The bifactor model, specifically its orthogonality, enables us to get around such a multicollinearity issue and build a well-grounded and sound prediction model.

Then, how do we calculate the six SAQ domain scores and the overarching SA dimension score from the bifactor model? Some readers might think that simply calculating domain scores by averaging raw item scores (1 to 5) under the domain and one general SA score by averaging all 34 raw item scores are adequate. Figure 4 clearly depicts why such an approach is dangerous and how it could mislead us to a wrongful conclusion; in a word, simple mean domain scores and a mean general SA dimension score are invalid not only theoretically, but also practically.

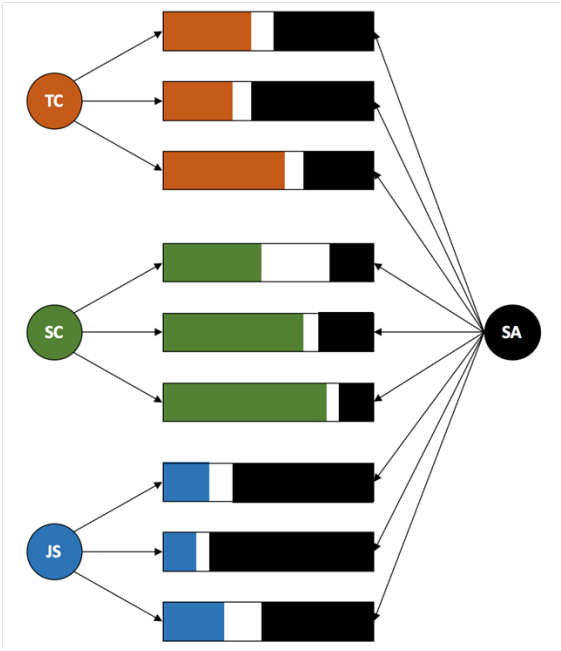

Figure 4 Variance bar graph for a bifactor model.

Note The colored portion on the left side of each rectangle is the variance of each item score explained by certain SAQ domains; the black part on the right side is the variance explained by the general SA dimension; the white portion in the middle is the measurement error.

Citation: Jeong HJ, Lee WC. The pure and the overarching: an application of bifactor model to safety attitudes questionnaire. *Biom Biostat Int J.* 2016;4(6):222–232. DOI: 10.15406/bbij.2016.04.00110

This figure did not come from actual data; we intentionally came up with the varying patterns over domains for illustration purposes. For the bottom three items (being shot by arrows from the JS domain), most of the variance of item scores is explained by the general SA dimension whereas the amount of effects from JS is minimal. Thus, if we calculate the JS score using a traditional simple mean score approach, what we get is a mixture of small JS domain-specific effects and the large bifactor effect. For the other domains, the same logic applies. On the other hand, if we calculate the overall SA score by the traditional mean of all items, then we again get a score that contains mixed effects of specific domains and the SA.

The good news is that the areas of the colored, white, and black rectangles as well as the factor loadings of all arrows in Figure 4 can be calculated. Thus, using the bifactor model, we can obtain purer scores of the six domains plus the overarching general SA dimension score simultaneously, which are much more precise than the traditional simple mean domain scores.[14]

Some readers might think that dividing rectangles means dividing item scores, such as if a person responds 5 (strongly agree), then 2 points might be given to the TC domain and 3 points to the general SA dimension. This is a huge misunderstanding. What we divide is the ownership of variance, which is reflected among domain scores throughout the model. An intentionally oversimplified explanation is that the length of the bars in the item rectangles stand for the weighting factors applied to item scores for calculating the weighted average score for each specific domain and the general SA dimension. (Theoretically, this is a skewed explanation, but it helps initial understanding.) Simply put, we will not get underestimated scores for each domain even when we apply the bifactor model; rather, we will obtain purer domain estimates after controlling for the general SA effect. The overarching general SA dimension can be understood in the same way.

End of the beginning: jumping to the pure and general world

If you accept the previously discussed concepts, the primary purpose of this article has already been accomplished. Yet we recognize too well that understanding this level of model from first glance is not easy at all. Thus, here we provide actual examples using our SAQ-K dataset containing responses from 1,142 participants collected from a tertiary hospital in Seoul, Korea, in 2013.[4] In this article, all scores are denoted as expected a posteriori (EAP; also called Bayes estimates) scores obtained using item response theory (IRT)—specifically, a graded response model (GRM). Thus, the scores will range from around -3 to 3 instead of the traditional 1 to 5 Likert scale score or its converted 0 to 100 score. You do not have to understand the algorithm; just be aware that IRT scores follow a normal distribution with a mean of 0 and a standard deviation of 1,[15] which can be transformed to a 0 to 100 scale at any time. To learn more about IRT and its application to SAQ, please refer to our previous articles in this series.[6,16]

One note that should be mentioned here is that the 5-point Likert score that SAQ uses is an ordinal scale.[17] For an ordinal scale, we can never know whether the difference between two adjacent values is the same as that of the other two adjacent values: How do we know the differences between strongly disagree and slightly disagree versus neutral and slightly agree are the same? We coded the scale from 1 to 5 for convenience's sake when entering the data into the computer. Thus, the numbers 1 to 5 from a Likert scale are just orders, not values, and their amount cannot be quantified. Theoretically, any calculations (e.g., adding, dividing, and averaging) are not allowed, albeit we traditionally do.[9] IRT can effectively handle this ordinal data structure, returning valid scores in various formats, of which we chose the most frequently used score type, EAP.

It might be hard to understand. However, rest assured: IRT and the bifactor model provide conversion tables from traditional summed scores (sum of raw item scores—1 to 5) to EAP scores for each domain, including the general dimension. (We discuss why we used summed scores instead of a mean score in a later section.) After entering the collected questionnaire responses into the computer, the EAP scores automatically appear on the screen. Of course, these scores are the purest possible domain scores for each participant and the score of overarching general safety attitudes dimension. Next we discuss purity and generality issues.

Methods

We developed two models, the correlated factors model (Figure 1) and the bifactor model (Figure 4), using the SAQ-K data. For both models, IRT GRM was used due to a 5-point ordinal scale of the SAQ response option structure. For the bifactor model, we used the usual Bock- Aitkin's algorithm (BAEM), but for the correlated factors model, we used the Metropolis- Hastings Robbins-Monro (MHRH) algorithm for calculation efficiency. Model fits were checked using limited-information fit statistics (M2 based) due to the exponential complexity stemming from the 34 items on the 5-point ordinal scale. Factor loadings from each model were calculated. The EAP score of each domain for each respondent were the estimated from both models, and their distributions were depicted as kernel density plot. Analyses were performed using statistical software packages flexMIRT 3.0 (Vector Psychometric Group, LLC, Chapel Hill, North Carolina) and Stata 14.2 (Stata Corp., College Station, Texas).

Results

Factor loadings matter

Table 1 shows the factor loadings for each item from both the correlated factors model and the bifactor model.

For the correlated factors model, factor loadings (lengths of grey bars) ranged from .54 (WC1) to .92 (JS3), showing relatively large values with stable distribution across all domains and items. Meanwhile, for the bifactor model, the factor loadings for items of each of the six SAQ domains (length of the bars in six different colors) hovered around .4 to .6, showing smaller values compared to the corresponding items from the correlated factors model; the colored bars are shorter than the grey bars. The clear exceptions were the SR domain items (purple bars), whose loadings were almost the same as the those from the correlated factors model (grey bars). In addition, some PM items (brown bars) showed exceptionally low loadings: PM1 (.12), PM2 (-.12), PM3 (.08), PM4 (.12), PM5 (.14), and PM7 (.19). It is noticeable that factor loadings for the SR items from the general SA dimension (black bars) were almost negligible, spanning from .05 to .09. Except for SR, the general SA dimension revealed similar factor loadings as those from the correlated factors model, with respect to both values and distribution.

Citation: Jeong HJ, Lee WC. The pure and the overarching: an application of bifactor model to safety attitudes questionnaire. *Biom Biostat Int J.* 2016;4(6):222–232. DOI: 10.15406/bbij.2016.04.00110

General and domain-specific scores from the models

Moving on to domain scores obtained through the model, Figures 5 and 6 show kernel density plots, with the EAP score on the x-axis and its density on the y-axis. You can think of these plots as simply a smoothed histogram of scores. Figure 5 is the kernel density plot of the general SA dimension from the bifactor model. Except for a peak around -.8, the plot shows a comparatively well-balanced bell-shaped distribution, where the majority of people scored between -2 and 2.

Table 1 Standardized factor loadings from correlated factors and bifactor models

ID	Items	Correlated Factor Model	Bifactor Model	
			SAQ Domains	General SA
Teamwork Climate				
TC1	Nurse input is well received in this clinical area	.73	.42	.62
TC2	Disagreements in this clinical area are resolved appropriately	.81	.44	.70
TC3	I have the support I need from other personnel to care for patients	.81	.46	.68
TC4	It is easy for personnel here to ask questions	.76	.46	.63
TC5	The physicians and nurses here work together as a well-coordinated team	.75	.42	.62
Safety Climate				
SC1	I would feel safe being treated here as a patient	.78	.31	.69
SC2	Medical errors are handled appropriately in this clinical area	.80	.42	.70
SC3	I know the proper channels to direct questions regarding patient safety	.72	.50	.60
SC4	I receive appropriate feedback about my performance	.80	.47	.69
SC5	I am encouraged by my colleagues to report any patient safety concerns	.74	.40	.64
SC6	The culture in this clinical area makes it easy to learn from the errors	.68	.35	.61
Job Satisfaction				
JS1	I like my job	.81	.64	.54
JS2	Working here is like being part of a family	.87	.60	.64
JS3	This is a good place to work	.92	.64	.67
JS4	I am proud to work in this clinical area	.89	.58	.69
JS5	Morale in this clinical area is high	.83	.48	.68
Stress Recognition				
SR1	When my workload becomes excessive, my performance is impaired	.68	.66	.05
SR2	I am less effective at work when fatigued	.72	.71	.09
SR3	I am more likely to make errors in tense or hostile situations	.78	.78	.07
SR4	Fatigue impairs my performance during emergency situations	.78	.78	.06
Perception of Management				
PM1	Unit management (UM) supports my daily efforts	.81	.12	.82
PM2	Hospital management (HM) supports my daily efforts	.63	-.12	.72
PM3	UM doesn't knowingly compromise patient safety	.86	.08	.87
PM4	HM doesn't knowingly compromise patient safety	.87	.12	.87
PM5	UM is doing a good job	.86	.14	.86
PM6	HM is doing a good job	.84	.40	.76
PM7	Problem personnel are dealt with constructively by our UM	.72	.19	.70
PM8	Problem personnel are dealt with constructively by our HM	.88	.57	.75
PM9	I get adequate, timely info about events that might affect my work from UM	.86	.60	.72
PM10	I get adequate, timely info about events that might affect my work from HM	.84	.53	.72
Working Condition				
WC1	The levels of staffing in this clinical area are sufficient	.54	.43	.37
WC2	This hospital does a good job of training new personnel	.80	.45	.66
WC3	All the necessary information for decisions is routinely available	.81	.52	.63
WC4	Trainees in my discipline are adequately supervised	.85	.54	.68

Note To save space, items were shortened as long as the original meaning was retained; the length of each bar indicates the factor loading of each item; although not shown, the standard errors (SE) ranged from 0.01 to 0.04 (mostly .02 to .03) for the items in the correlated factors model and from .02 to .07 (mostly 0.04 to 0.06) in the bifactor model.

Citation: Jeong HJ, Lee WC. The pure and the overarching: an application of bifactor model to safety attitudes questionnaire. *Biom Biostat Int J.* 2016;4(6):222–232. DOI: 10.15406/bbij.2016.04.00110

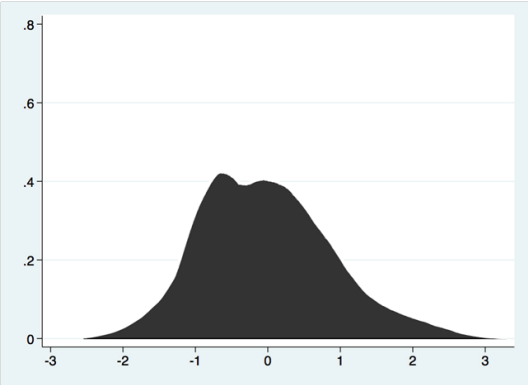

Figure 5 EAP Score distribution of the general SA dimension from bifactor model.

We also describe the logic used to obtain the scores in Figure 5, albeit in an oversimplified manner. SAQ-K has 34 items; thus, we have 34 responses from each respondent. Instead of calculating a simple average of the responses for each respondent, we applied the factor loadings of the general SA dimension from the bifactor model (the lengths of black bars in Table 1) as a kind of weighting value to get something like the weighted average of the 34 item responses. Therefore, among the 34 item responses, those from smaller loading items in the general SA dimension were naturally undervalued while those from higher loading items influenced the general SA dimension score more. As you might have already expected, SR items had only a minimal or even ignorable impact on the SA score because their factor loadings for the general SA dimension were very small—less than .1.

We then conducted comparisons for score distributions between the correlated factors model and the bifactor model for each domain (Figure 6). Within each domain, the factor loading- based weighting was applied to both models.

In general, score distributions from the correlated factors model (grey area) were spread wider and its density peaks were not as tall as those from the bifactor model (colored lines). As we mentioned earlier, for the bifactor approach, each item response is influenced by both the six SAQ domains and the general SA dimension. Thus, the scores from the bifactor model (colored lines in Figure 6) can be thought of as scores after filtering out the general SA dimension effect and, therefore, the purified domain specific scores. Furthermore, the bifactor model ensures no correlation among domains, and the score distributions in Figure 6 might be a much more precise reflection of true domain scores than those from the other models introduced in this article.

Note that the SR domain did not show much difference in score distribution between the correlated factors model and the bifactor model. By now, you may already understand why. For SR items, the factor loadings from the general SA domain were negligible, and the factor loadings from the correlated model and loadings from the bifactor model were almost the same. As a result, there was no substantive amount of effects from the general SA dimension to be filtered out, leaving the SR domain scores from the two models almost the same. The bottom line is that each item score is governed by its corresponding domain (and the general SA dimension for the bifactor model) while the factor loading is the manifestation of the amount of the governance in a particular model. The domain scores are a collective reflection of those factor loadings.

Although it would be a repetition, we graphically show this domain governance on items in Figure 7. We have no doubt you understand what is going on in Figure 7.

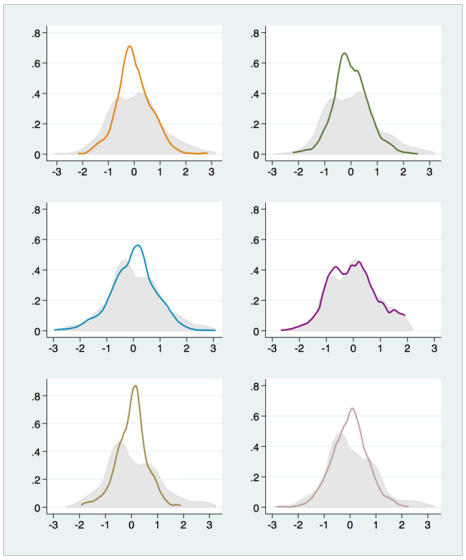

Figure 6 Score distribution comparisons for each SAQ domain between correlated factors and bifactor models.

Note Grey area: domain score distributions from the correlated factors model; lines: domain score distributions from the bifactor model.

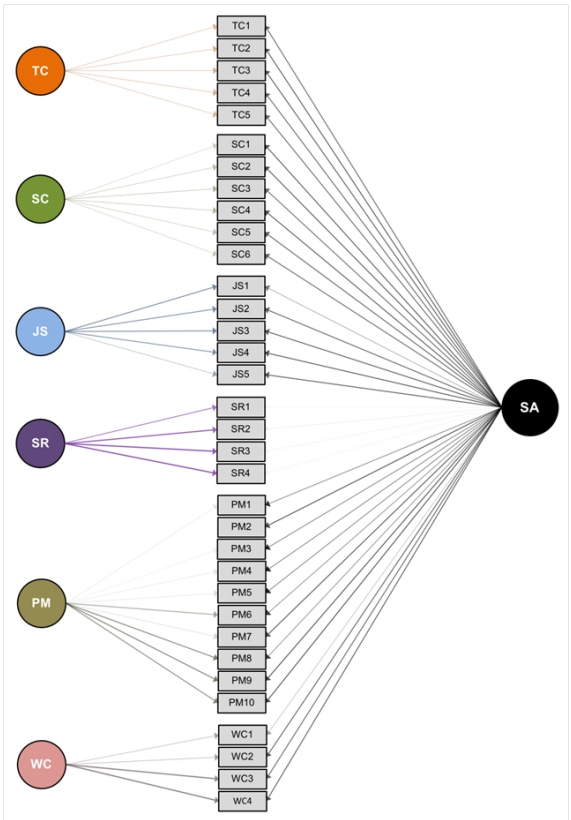

Figure 7 Factor loadings on SAQ domains and general SA dimension from bifactor models.

Note Color intensity of the arrows indicates the relative amount of factor loading, where darker is larger.

Citation: Jeong HJ, Lee WC. The pure and the overarching: an application of bifactor model to safety attitudes questionnaire. *Biom Biostat Int J.* 2016;4(6):222–232. DOI: 10.15406/bbij.2016.04.00110

The darkness of the arrows indicates the amount of contribution (factor loading) of each item score to each domain score and general SA dimension score. Again, focus on the SR domain and the four items under it. Almost inconspicuous arrows exist between the SR domain items and the general SA domain, but dark and deep arrows occur between items and the specific SR domain. Thus, the SR items are not under the influence of the general SA dimension; rather, the SR domain is almost a disconnected concept like Galapagos of SAQ. Simply put, SR is a different animal that should not be bred in the same cage as SAQ. This gives a full theoretical explanation as to why a few countries that adopted SAQ reported unreliable results from the SR domain; some of them even dropped the domain.[3,18-20] In addition, the arrows from the specific PM domain to PM1, 2, 3, 4, 5, and 7 are very hazy, depicting smaller factor loadings compared to the other items in the domain. To summarize, the SR items and five previously mentioned PM items were barely reflected in the estimations of the general SA dimension and the specific PM domain, respectively.

From summed score to expected a posteriori (EAP) scores

Some readers might wonder why we use EAP scores and how to use them without knowledge of IRT. This section addresses these concerns. Let us begin with the second part. Luckily, algorithms for converting between summed scores and EAP have already been developed,[21,22] although one hundred-percent accuracy is not guaranteed. Most software packages that support IRT produce such conversion tables for each IRT-based model developed. How does this apply to a real-world situation? If instrument validation is done by the developer, we may not need to run a complex model to obtain EAP scores ourselves; instead, the developers distribute the conversion table generated from the model in spreadsheet format with a built-in automatic matching function. As soon as we enter the collected responses into our computer, model-based factor loading-adjusted EAP scores are produced for each respondent. This is actually even easier and less cumbersome than calculating traditional scores in a domain-by- domain manner. In sum, you do not have to fully understand how IRT works to get pure scores within seconds.

To illustrate, Table 2a provides an excerpt of the conversion table for the general SA dimension. As we have 34 SAQ-K items rated on a scale of 1 to 5, the summed score spans from 34 to 170. Thus, the conversion is straightforward: Add all item scores and match the sum to the table to get the corresponding EAP score. For example, if a respondent's summed score is 102, then her EAP score is -0.85. However, the domain-specific score requires one more step, although it is still straightforward. Table 2b includes an excerpt of the conversion table for the TC domain. As the TC domain of SAQ-K consists of five items, the smallest possible summed score is 5 while the largest score is 25, meaning that 29 items occur in the remaining domains, and their summed scores range from 29 to 145. Thus, we calculate two summed scores—one for TC and the other for the other five domains—for each respondent, locate the score combination in the conversion table, and find the corresponding EAP score for the TC domain. For example, if a person's summed score is 17 for TC and the summed score from all other domains is 90, her EAP score is -0.15. For the other domains, the same logic applies.

Table 2a Excerpt of summed score to EAP score conversion table for general SA dimension

Summed score	EAP score
34	-4.286
.	.
.	.
98	-1.03
99	-0.99
100	-0.94
101	-0.89
102	-0.85
103	-0.08
104	-0.76
105	-0.71
106	-0.66
.	.
.	.
170	3.285

Table 2b Excerpt of summed score to EAP conversion table for TC domain

Summed score (Other domains)	Summed score TC	EAP score TC	EAP score general SA
29	5	-0.67	-4.29
.	.	.	.
.	.	.	.
90	5	-3.42	-0.80
90	6	-3.10	-0.74
90	7	-2.82	-0.68
90	8	-2.56	-0.63
90	9	-2.31	-0.58
90	10	-2.06	-0.53
90	11	-1.81	-0.47
90	12	-1.56	-0.42
90	13	-1.30	-0.37
90	14	-1.05	-0.32
90	15	-0.80	-0.26
90	16	-0.53	-0.21
90	17	-0.26	-0.15
90	18	0.02	-0.09
90	19	0.30	-0.03
90	20	0.59	0.03
90	21	0.88	0.09
90	22	1.18	0.16
90	23	1.50	0.23
90	24	1.84	0.31
90	25	2.21	0.39
.	.	.	.
.	.	.	.
145	25	0.33	3.29

One reason why we use a summed score instead of a simple mean score for this conversion is a matter of decimal. For domains with 4, 5, or 10 items, no problem arises. Summed scores for these domains can be divided by the number of items (i.e., 4, 5, or 10) clearly and easily. Yet for domains with 6 items, such as SC of SAQ-K, the mean score could have long or endless decimal numbers (e.g., 3.33333), which creates a huge problem. We typically use different software packages for generating basic tables and for complex analyses like the modeling introduced in this article. More often than not, the software packages have different algorithms for treating the floating point (i.e., dealing with decimals), such as rounding. Thus, if we copy and paste the calculated mean, the values with long decimal numbers might change, albeit in a small way to our eyes, which sometimes prohibits our software from finding and matching the mean score and EAP score in the table. The summed score frees us of this problem, and summing the item scores is actually easier than obtaining a mean score; thus, there is no reason to stick to the mean score.

Discussion

Use and misuse of models

This article aimed to answer the healthcare professional's question: "Where are the safety attitudes in general in your model?" We demonstrated the answer. We successfully obtained the general safety attitudes score; furthermore, we presented how to calculate purer and more precise scores for each of the six SAQ domains. We were happy; yet we soon began to ponder exactly what we did to answer the healthcare professional's question. Did we use different or any additional data to estimate the new scores we obtained? The answer was simply no. What made the difference was applying a different model to the very same data; through the model, we observed that the data told a completely different story. Then, how should we choose the most suitable model?

Actually, an infinite number of possible models could be used for the same dataset. For illustration purposes, let us assume an instrument with 12 items that depend on four domains. Obviously we could try correlated, second-order, or bifactor models, as introduced in this article. We could also develop a model with two correlated general dimensions as depicted in Figure 8—what we call a two-tier model.[23] Indeed, just drawing arrows among items and factors delivers different models; thus, the possibilities are virtually unlimited.

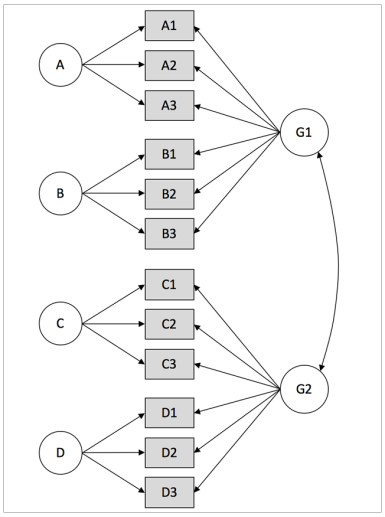

Figure 8 Two-tier model example.

Note A, B, C, and D are specific domains and G1 and G2 are general domains.

The good news is that our colleagues have tested various models and developed a decent list with a manageable number of model options for a given structure of the data and the purpose of the analysis. Thus, we should try those models and pick the one with the most substantive theoretical grounds, empirical fit, and parsimony.[24] This is exactly what we do when conducting all other analyses, such as linear regression, analysis of covariance, and even a *t*-test.

The bad news is that more often than not, we have very poor repertoire of models, sometimes acting like there is only one, the correlated factors model, especially for survey instrument validation. Along such a validation process, modeling occurs when checking the construct validity under the confirmatory factor analysis (CFA). We run a CFA, present a few model fit indices, and say that the model fit is good. Then we usually set aside the model with its vast amount of information. We then administer the "validated" instrument over the target population (if we have not already) and calculate what we call domain scores by simply averaging item responses. In this modus operandi, the step of choosing or developing a better model is not at all important because what we need is only a few acceptable model fit indices.

It feels smooth and sound, but that is only because we are so used to it. This seemingly natural process has serious flaws from a psychometrics perspective. First and foremost, recall there were varying levels of factor loadings among items and domains. As we showed in this article, each and every item has a different factor loading; thus, a simple mean domain score is not justified as long as there is no special reason. Consequently, we calculated factor scores that we expressed as a weighted domain score, where factor loadings were used as the weighting factor. (For statisticians, please see below regarding why we expressed them this way.) In sum, if we choose a model and use it to validate the instrument of our interest, we are naturally obliged to use the information from the model to calculate any scores from the instrument. If we calculate the simple average, it means that we regard all the factor loadings (arrows) from a domain (circle) to items (rectangles) as the same value. Have we ever seen the same factor loadings across all items in a model? Never. That is why we should use factor scores (weighted averages) yielded directly from the model itself. The premise is that we must develop a sound model that reflects the structure of the model components that we had in mind—namely, the structure based on theoretical backgrounds and empirical evidence.

In order to help you understand the magnitude of the difference in score distributions from the simple average approach and the model-based approach, we present Figure 9 using the TC domain as an example.

There is another reason why scoring through a model is important. Factor loadings are influenced by the relationships among domains. Thus, if we change the relationship structure in a model (two-headed arrow among circles: variance–covariance structure, statistically), the factor loadings naturally change accordingly, meaning we will obtain different domain scores. Simply put, modeling is the process of portraying how the various domains in our minds are related,[25] through which all the coefficients including factor loadings and scores align as they should. Recall that we could not answer the healthcare professional's simple question until we applied a different model: the bifactor model. The bottom line is that modeling is not a process confined to checking the construct validity; rather, it covers the whole process of research, from design through even scoring. The happiest news is that any decent software packages perform this whole process in one step, modeling, which means this approach is actually easier

Citation: Jeong HJ, Lee WC. The pure and the overarching: an application of bifactor model to safety attitudes questionnaire. *Biom Biostat Int J.* 2016;4(6):222–232. DOI: 10.15406/bbij.2016.04.00110

and takes much less time than our conventional approach introduced above.

At this point, we have to confess that factor scores are not technically a weighted average of item scores. Remember the direction of arrows in the models: from domains (circles) to items (rectangles). For weighted average, the direction should be reversed, as evident in the principal component analysis (PCA) approach. Strictly speaking, latent factors (domains) by definition govern the scores of each item through factor loadings and are not the other way around.[26] Yet we intentionally continued to misuse the term *weighted mean* or *average* because it facilitates readers' understanding, especially those who do not understand the latent variable concept. We also did so because such factor scores are automatically calculated by software packages and most of us do not need to understand the full behind-the-scenes logic.

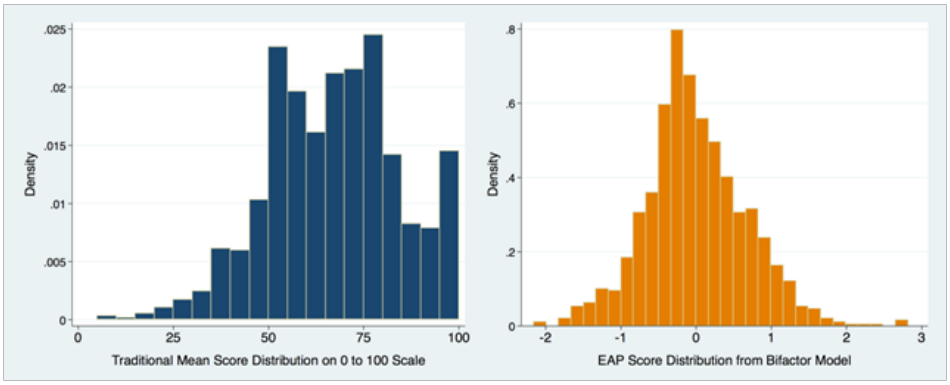

Figure 9 Contrast in the TC domain score distributions between ignoring and acknowledging models.

Note Left graph shows the distribution of the TC domain scores calculated by averaging raw item scores (ignoring model) on a 0 to 100 scale; right graph depicts the TC domain score distribution yielded directly from the bifactor model.

Strengths of item response theory and modeling combination

Let us add a comment on IRT. A very common misunderstanding of an IRT-based scale is about its 0 mean, 1 standard deviation (SD) scale. One might think that each sample group has its own mean and SD and, therefore, we cannot compare scores of groups. This idea came from classic test theory, which uses traditional and customary *z*- or *t*-scores. Luckily, this does not apply to IRT. IRT is sample independent: What IRT calculates first is each item's properties; once the parameters are calculated in one group, the same parameters can be applied to other groups.[21,27,28] Thus, people who have the same level of safety attitudes present the same EAP scores, regardless of the group in they are categorized. In some cases, the EAP scores of almost all respondents are below zero in one group and above zero in another group, sometimes highly skewed, proving that IRT and its EAP are not sample dependent.

The item parameter-focused characteristics of IRT allow us to conduct several analyses not possible with the classic test theory paradigm. For example, assume we had rolled out SAQ and, two years later, we found that the SR domain was not reliable and decided to drop it or we dropped some items from some of the PM domains. Unlike the classic test theory paradigm, IRT still can estimate valid scores from the remaining domains and also general SA scores. This allows us to conduct a solid longitudinal analysis. For a more comprehensive explanation on IRT and its application to SAQ, please refer to our previous articles.[6,16]

Interestingly, IRT and the bifactor model are inseparable in modeling SAQ data. The primary assumption of IRT is unidimensionality—namely, that items in a certain domain should be influenced by only one dimension.[29,30] To illustrate, items in the TC domain should be all about TC or at least TC should be the dominant dimension. This is a very strong assumption, but cannot easily be verified in many analyses. Actually, the traditional mean score approach relies more on this unidimensionality assumption; we do not even consider this issue when we calculate simple raw mean scores. However, we can quickly realize this assumption might not easily hold. Regardless of how high the Cronbach's alpha is, each item contains its own meanings that cannot be put on a single plate. In the bifactor structure, it is overtly portrayed. Each item is governed by two different dimensions: a specific domain and the general SA dimension. If the amount of influence from different domains is not negligible, as in the SAQ data, typical IRT with unidimensional assumption does not work; thus, we need to apply a multidimensional IRT model.[31] Thankfully, this issue is already addressed by the bifactor model, and we are relieved of the IRT's unidimensionality assumption.

Brief consideration of advanced application

In this article, we presented the orthogonality property among domains as one of the great advantages of the bifactor model: It allows for a multiple regression model with SAQ domains as covariates. Yet this does not mean the correlation coefficients among domain scores from the bifactor model are never statistically significant. Although the model clearly separates the domains, the possibility exists that the scores share the same pattern. However, the potential correlations from the bifactor model are fundamentally different from those of the correlated factors model, which assumes a complex correlation structure from the beginning. To make sure, we recommend conducting a multicollinearity checkup, such as by using the variance inflation factor (VIF) when developing a prediction model, even with scores from the bifactor model.

In closing this unexpectedly long article, we want to mention that the bifactor model is not the only model that should be applied to patient safety culture survey data, including SAQ. Another model

Citation: Jeong HJ, Lee WC. The pure and the overarching: an application of bifactor model to safety attitudes questionnaire. *Biom Biostat Int J.* 2016;4(6):222–232. DOI: 10.15406/bbij.2016.04.00110

might better answer a certain research question, also showing better model fit. Yet based on our years of experience, the bifactor model gives us what we needed most: the precise domain scores and the overall safety attitudes score. We hope that this article helps all patient safety personnel around the world open the door to the pure structure of healthcare professionals' safety attitudes and culture so we can truly improve safety and save more patients.

Conclusion

This article completes our long series of articles on SAQ-K. The journey began with a simple idea that there could be a cultural profile or pattern regarding patient safety for each organization or country, which we called the cultural-compatibility complex.[32] Through this series, we began by developing the SAQ-K,[4] then showed how to obtain a more precise work area and job type-specific scores as well as their interaction by applying an empirical Bayesian approach.[2,4] We suggested a solution to classify work areas by SAQ-K patterns to develop tailored safety programs.[3] We then presented a full description of IRT and its application to SAQ.[9,16] Of course, we did not miss the issue of how items are functioning differently across different groups.[6] Exploring different response options other than a 5-point Likert scale was a truly enjoyable experience.[7,8]

Now is the time to begin our new journey based on what we have done, and we certainly will. We will return to the Safety Genome Project, the terminology that we coined long ago. It will cover not only the safety culture profile, but also the impact on clinical indicators at the national and even international levels, which is way beyond the single organization level with which the current series has dealt. We sincerely thank all readers of this series, and we do hope that our work will help you help all your patients. Please enjoy saving lives, and stay tuned.

Acknowledgement

None.

Conflict of interest

None.

References

1. Jeong HJ, M Kim. A Practical Guide to Behavioral Theory–Driven Statistical Development of Quality and Safety Improvement Program in Health Care. *Biometrics & Biostatistics International Journal*. 2014;1(1):1–6.
2. Jeong HJ, 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):1–8.
3. 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):1–6.
4. 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):1–11.
5. Park Mi jin, Na Hae ran, Jeong HJ, 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.
6. Jeong HJ, WC Lee. Does Differential Item Functioning Occur Across Respondents' Characteristics in Safety Attitudes Questionnaire? *Biometics & Biostatistics International Journal*. 2016;4(3):1–9.
7. Jeong HJ, WC Lee. The level of collapse we are allowed: Comparison of different response scales in Safety Attitudes Questionnaire. *Biometrics & Biostatistics International Journal*. 2016;4(3):1–7.
8. Jeong HJ, Park MJ, Chul–Ho Kim, et al. Saving Lives by Saving Time: Association between Measurement Scale and Time to Complete Safety Attitudes Questionnaire. *Biometrics & Biostatistics International Journal*. 2016;4(5):1–4.
9. Jeong HJ, WC Lee. Ignorance or Negligence: Uncomfortable Truth Regarding isuse of Confirmatory Factor Analysis. *Journal of Biometrics & Biostatistics*. 2016;7(3):1–2.
10. De Mars CE. A tutorial on interpreting bifactor model scores. *International Journal of Testing*. 2013;13(4):354–378.
11. Reise SP, TM Moore, MG Haviland. Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *J Pers Assess*. 2010;92(6):544–559.
12. Yung YF, D Thissen, LD McLeod. On the relationship between the higher–order factor model and the hierarchical factor model. *Psychometrika*. 1999;64(2):113–128.
13. Chen FF, SG West, KH Sousa. A comparison of bifactor and second–order models of quality of life. *Multivariate Behavioral Research*. 2006;41(2):189–225.
14. Biderman M. Applications of Bifactor Models to Big Five Data. In The 28th Annual Conference of the Society for Industrial and Organizational Psychology Houston, TX, 2013;1–89.
15. Stata Corp, Stata 14 Item Response Theory Reference Manual, 2015. College Station, TX: Stata Press.
16. Jeong HJ, WC Lee. 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.
17. Agresti A. Categorical data analysis. New York: John Wiley & Sons, USA, 1996;996:1–721.
18. Gyeongsil Lee, Mi Jin Park, Hae–Ran Na, et al. Are Healthcare Workers Trained to be Impervious to Stress? *Biometrics & Biostatistics International Journal*. 2015;2(2):1–2.
19. 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*. 2010;10(1):234.
20. Nikoloz Gambashidze, Shoeb Ahmed Ilyas, Diana Pascu. Validation of the safety attitudes questionnaire (short form 2006) in Italian in hospitals in the northeast of Italy. *BMC Health Services Research*. 2015;15(1):284.
21. Thissen D, Pommerich M, Kathleen Billeaud, et al. Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement*. 1995;19(1):39–49.
22. Orlando M, CD Sherbourne, D Thissen. Summed–score linking using item response theory: Application to depression measurement. *Psychol Assess*. 2000;12(3):354–359.
23. Cai L. A two–tier full–information item factor analysis model with applications. *Psychometrika*. 2010;75(4):581–612.
24. Rijmen F. Formal Relations and an Empirical Comparison among the Bi–Factor, the Testlet, and a Second–Order Multidimensional IRT Model. *Journal of Educational Measurement*. 2010;47(3):361–372.

25. Rupp AJ, Templin, R Henson. Diagnostic measurement. Theory, methods and applications New York, NY: The Guilford Publication Inc. 2010.

26. Acock AC. Discovering structural equation modeling using Stata. College Station, Texas: Stata Press Books. 2013.

27. Stocking ML, FM Lord. Developing a common metric in item response theory. *Applied Psychological Measurement*. 1983;7(2):201–210.

28. Embretson SE, SP Reise. Item response theory. Psychology Press. 2012.

29. Mungas D, BR Reed. Application of item response theory for development of a global functioning measure of dementia with linear measurement properties. *Statistics in Medicine*. 2000;19(11–12):1631–1644.

30. Harvey RJ, AL Hammer. Item response theory. *The Counseling Psychologist*. 1999;27(3):353–383.

31. Fayers P. Item response theory for psychologists. *Quality of Life Research*. 2004;13(3):715–716.

32. Jeong HJ, Julius C Pham, Minji Kim, 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. The pure and the overarching: an application of bifactor model to safety attitudes questionnaire. *Biom Biostat Int J.* 2016;4(6):222–232. DOI: 10.15406/bbij.2016.04.00110