MedCrave
Step into the World of Research

Biometrics & Biostatistics International Journal

Research Article

Open Access

CrossMark

A practical approach to develop a parsimonious survey questionnaire—taiwanese patient safety culture survey as an example

Abstract

This study aims to suggest a method to choose what items can be removed from a survey instrument, thereby lessening survey participants' burden. We used the Taiwanese Patient Safety Culture (TPSC) survey as an example—in particular, the emotional exhaustion (EE) and work–life balance (WLB) domains. A traditional factor analysis approach, whereby a Likert scale is treated as a linear variable, was applied, and an item response theory (IRT) graded response model (GRM), where a response is treated as an ordinal scale, was used. From both methods, five out of nine EE items and three of seven WLB items were filtered out for removal; these items provided significantly smaller information for domain score estimation than other items. To check whether the downsized versions showed sufficient validity, we tested the factor structure of the remaining items and compared fit indices between the models with all original items and the new models with the remaining items. The new models in both EE and WLB domains showed better model fit than the models with all original items. In this study, we showed how to select more important items while ensuring that the selected items as a whole can provide a better fit compared to the original instrument. This interesting phenomenon may be due to the translation step and different cultural soils where surveys are administered. We strongly recommend checking the possibility that items can be removed when adopting and validating a survey instrument—not only to shorten the overall length, but also to increase validity.

Volume 6 Issue 3 - 2017

Hsun-Hsiang Liao,[1] Wei-Chiang Lee,[2] Yuh-Ling You,[3] Heng-Lien Lo,[3] Cheng-Fan Wen,[3] Heon-Jae Jeong,[4] Pa-Chun Wang[5]

[1]Deputy CEO, Joint Commission Taiwan
[2]Department of Medical Affairs and Planning, Taipei Veterans General Hospital & National Yang-Ming University School of Medicine, Taipei, Taiwan
[3]Division of Quality Improvement, Joint Commission of Taiwan, Taipei, Taiwan
[4]Advisor, Joint Commission Taiwan, Taipei, Taiwan
[5]CEO, Joint Commission Taiwan

Correspondence: Heon-Jae Jeong, Advisor, Joint Commission Taiwan, 5F, NO.31, Sec. 2, Sanmin Rd., Banqiao Dist., New Taipei City 22069, Taiwan, Tel 886-2-8964-3000, Fax 886-2-2963-4292, Email hj9571@gmail.com

Received: August 23, 2017 | **Published:** September 05, 2017

Introduction

Healthcare professionals tend to work within extremely tight schedules. Yet those in the field of improving the quality and safety of care almost always want to administer survey questionnaires to these professionals to collect information on various aspects of healthcare. Such information is certainly crucial, but it is also certain that the surveys would put much burden on healthcare professionals. In this regard, it is our obligation to control the burden expected when administering a survey. From a more practical standpoint, fatigue from frequent surveys may result in a low response rate that endangers the representativeness of the survey itself,[1] which could lead to less useful survey results.

In addition to being frequently asked to participate in surveys, the burden of completing a survey also resides in the characteristics of each questionnaire, such as how many questions are asked, how long each item is, and even what measurement scale is used. Indeed, Jeong and Lee (2016) showed that the number of response options causes significant differences in the time to complete surveys; for example, a 3-point Likert scale is much faster to complete than a 5-point Likert scale.[2] In addition to reducing the time required, they also showed that a 3-point and a 5-point Likert scale yielded very similar results in a population level.[3] Although their studies used a specific instrument, the Safety Attitudes Questionnaire-Korean version (SAQ-K), the results suggest that we can modify and streamline survey questionnaires without undermining their validity.

Among the above-mentioned various characteristics of questionnaire, the number of items play an important role in deciding the amount of a survey's burden; thus, this is one of the aspects that should be addressed first. As such, we tried to establish or at least suggest a clear and structured approach to reducing survey items without compromising validity for the target readers, who are quality and safety personnel in the field but may not have experience in advanced statistics.

To this end, we used the Taiwanese Patient Safety Culture Survey Questionnaire (TPSC) as an example. TPSC has been administered for the past decade to hospitals in Taiwan by the Joint Commission of Taiwan (JCT). It consists of 46 main items in addition to 14 items asking about demographics and a few issues specific to the JCT's needs. Each of the 46 main items came from three different instruments measuring the corresponding constructs: 30 items from safety attitude questionnaire (SAQ), one of the most widely used patient safety culture measurement instruments,[4-9] nine items in the emotional exhaustion (EE) domain from the Maslach Burnout Inventory,[10] and seven work–life balance (WLB) domain items from the study "Accelerating the coping process".[11]

In this study, we applied two different approaches to choose appropriate items that can be removed: a traditional factor analysis and item response theory (IRT). Although they share some commonalities, each is rooted in different foundations. From the practical perspective, IRT has its unique strength in showing the amount of information of

each item.[12] On the other hand, the factor analysis approach is still the most popular method for survey development,[13] and TPSC was also validated by such approach before it was rolled out. By using them in tandem and checking whether the results were in concordance, we obtained much more stable results. Let us describe the procedures and results in detail.

Methods

Selecting target domains to reduce

The SAQ section of TPSC is the overall safety culture construct consisting of six domains, each of which contains an average five items. These already compact item numbers in each domain did not provide us much opportunity to drop items. Thus, we focused on the EE and WLB domains to reduce the number of items. As Table 1 shows, each of these domains have relatively more items compared to SAQ domains, suggesting there is room for reduction.

Table 1 List of items of EE and WLB domains

ID	Items
EE1	I feel like I am at the end of my rope
EE2	I feel burned out from my work
EE3	I feel frustrated by my job
EE4	I feel I am working too hard on my job
EE5	I feel emotionally drained from my work
EE6	I feel used up at the end of the workday
EE7	I feel fatigued when I get up in the morning and have to face another day on the job
EE8	Working with people all day is really a stain for me.
EE9	Working with people directly puts too much stress on me
WLB1	Skipped a meal
WLB2	Ate a poorly balanced meal
WLB3	Worked through a day/shift without any breaks
WLB4	Changed personal/family plans because of work
WLB5	Had difficulty sleeping
WLB6	Slept less than 5 hours in a night
WLB7	Arrived home late from work

Nine EE items were measured using a 5-point likert scale (1 = agree strongly, 2 = agree slightly, 3 = neutral, 4 = disagree slightly, 5 = disagree strongly), and seven WLB items were measured using a 4-point scale (1 =always, 2 =occasionally, 3 = a few times occasionally, 4 = rarely or none). For this study, we used 2016 TPSC data that JCT collected from various hospitals across Taiwan, which is the most recent and largest TPSC dataset as of this writing

Selection of the Items to be retained and their confirmation

First, we utilized the most commonly used form of factor analysis—namely, principal component factor analysis (PCFA)—and calculated the factor loading of each item. We also applied the IRT graded response model (GRM) considering the responses on 5- and 4-point Likert scales for EE and WLB domains, respectively, to obtain the item information function (IIF) describing the amount of information each item provides. By examining PCFA and IIF results together, we found items that convey a smaller amount of information than the others. We then confirmed our decision on the instrument validity perspective by conducting a confirmatory factor analysis (CFA) and comparing model fit indices between an all-item model and reduced-item model. For all analyses introduced in this article, Stata 14.2 (StataCorp, College Station, Texas) was used.

Results

Characteristics of participating hospitals and respondents

As described in Table 2, the 2016 TPSC dataset contains a total of 83,807 respondents from 98 hospitals, which are categorized into four different levels according to the Taiwanese healthcare system. The majority of both the number of organizations and the number of respondents were regional hospitals: 53 (54.1%) and 41,583 (49.6%), respectively. Furthermore, 34,443 (41.1%) respondents were from only 16 (16.3%) medical centers, which is understandable considering that medical centers are the biggest facilities. The dataset included 28 district hospitals (28.6%) and one psychiatric hospital (1.0%), from which 7,279 (8.7%) and 502 (0.6%) respondents, respectively, completed the survey.

Table 2 Characteristics of participating hospitals

Hospital level	Number of hospitals		Number of respondents	
	N	%	N	%
Medical center	16	16.3%	34,443	41.1%
Regional hospital	53	54.1%	41,583	49.6%
District hospital	28	28.6%	7,279	8.7%
Psychiatric hospital	1	1.0%	502	0.6%
Total	98	100.0%	83,807	100.0%

Table 3 shows respondents' job types. According to the data, 44,000 (52.5%) respondents were nurses, 13,273 (15.8%) were administration staff, and 9,135 (10.9%) were technologists. Physicians accounted for 7,837 (9.4%) of respondents, followed by pharmacists and rehabilitation staff at 3,085 (3.7%) and 1,400 (1.7%), respectively.

Table 3 Characteristics of respondents

Job type	N	%
Physician	7,837	9.4
Nurse	44,000	52.5
Technologist	9,135	10.9
Pharmacist	3,085	3.7
Administration	13,273	15.8
Rehabilitation	1,400	1.7
Others	5,077	6.1
Total	83,807	100

Exploring items to drop

Once each of the EE and WLB domains was independent construct in TPSC, we analyzed them separately. Thus, we describe the results by each domain.

Citation: Liao HH, Lee WC, You YL, et al. A practical approach to develop a parsimonious survey questionnaire—taiwanese patient safety culture survey as an example. *Biom Biostat Int J.* 2017;6(3):350–355. DOI: 10.15406/bbij.2017.06.00169

We begin with the EE domain and its nine items. First, we conducted a PCFA. From this exploratory factor analysis, we obtained factor loadings and factor scoring coefficients; we then chose four items showing clearly smaller coefficients and scores compared to the others—namely, EE 1, 4, 8, and 9—as well as one item that was somewhat equivocal in value (EE 3). Next, we conducted an IRT GRM on the EE domain and drew IIF curves. We found the same four items and possibly EE3 were underperforming. The PCFA and IRT GRM revealed only ignorable differences, justifying our filtering process.

We include only the IIF curves from IRT GRM due to space limits (Figure 1). Interpreting the IIF curves was straightforward: Higher curves provide more precise information for calculating the EE domain score; in other words, items with lower curves are less effective or rather redundant. In IIF curves, EE 1, 4, 8, and 9 obviously provided ignorable information, so we removed those four items. As in the PCFA, EE 3 was hard to determine decisively; thus, we decided to keep it in this step.

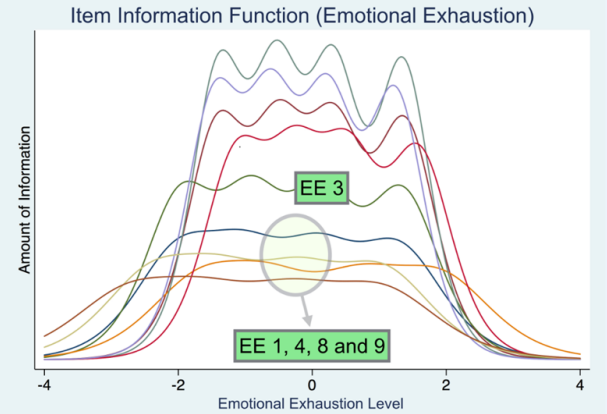

Figure 1 Item information function for emotional exhaustion domain.

Note X-axis is the level of respondents' EE, scaled to zero mean and one standard deviation (SD)

For the WLB domain, we conducted the same analyses. Figure 2 shows the results for the IIF curve format. Clearly, EE 5, 6, and 7 provided significantly less information; thus, we dropped these items. Although not described here, the PCFA showed the same results.

Confirmation of retained items

To ensure that the validity was not compromised by dropping items, we conducted CFAs and compared model fit indices of the full model with all the original items and the reduced or nested model with only the items we chose. For the EE domain, we developed three models: model 1 with all the original nine items; model 2 with items 2, 3, 5, 6, and 7 (please recall that item 3 was equivocal and thus could be dropped or retained); and model 3 with items 2, 5, 6, and 7, the most parsimonious model. In the following table, grey cells stand for the values that were within a satisfactory range, according to generally accepted model fit guidelines. The cutoff values for good model fit were alpha > 0.70, RMSEA < 0.08, CFI > 0.95, TLI > 0.95, and SRMR < 0.08, as summarized by Acock (2013), which reflect recent trends in using more stringent standards.[14]

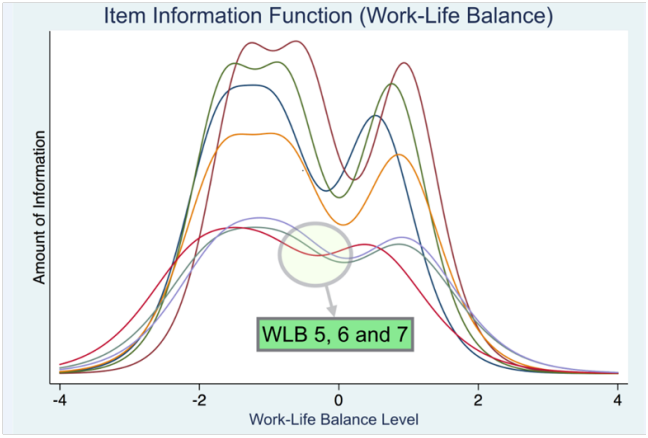

Figure 2 Item information function for work–life balance domain.

Note X-axis is the level of respondents' EE, scaled to zero mean and one standard deviation (SD)

As depicted in Table 4, model 3 (the most parsimonious model, with EE 2, 5, 6, and 7) showed the best values for all model fit indices—even better than the full model—while satisfying the general guidelines introduced earlier.

Table 4 Model fit indices of EE domain item combinations

	Model 1	Model 2	Model 3
Included items	All 9 items	2, 3, 5, 6 and 7	2, 5, 6 and 7
Alpha	0.920	0.908	0.900
RMSEA	0.198	0.176	0.018
CFI	0.831	0.953	1.000
TLI	0.774	0.905	0.999
SRMR	0.075	0.038	0.002

Note Grey cells mean good fit for each index

Finally, we compared the EE domain scores from the three models using the simple average scores of included items that we normally use. The correlation coefficients were 0.980 between models 1 and 2, 0.992 between models 2 and 3, and 0.961 between models 1 and 3, suggesting that the reduced-item model(s) provided EE domain scores without much difference from the full model with all the original items. As a result, utilizing a model with only four items (EE 2, 5, 6, and 7) was a statistically reasonable choice.

Unlike EE, the WLB domain did not include any equivocal items to address. Therefore, we developed and compared only two models: model 1 with all seven original WLB items and model 2 with WLB 1, 2, 3, and 4. As demonstrated in Table 5, whereas values in the grey cells satisfy external model fit index criteria, model 2 showed significantly better fit than model 1 with all the original items. RMSEA was the only index that model 2 failed to meet, but considering that its cut-off value was 0.08, the current results strongly suggest that

model 2 is the optimal item set for the WLB domain. The correlation between scores from the two models was 0.951, suggesting that the four-item questionnaire derived from model 2 can serve as a stable surrogate for the full seven-item questionnaire; indeed, it might be a better instrument.

Table 5 Model fit indices of WLB domain item combinations

	Model 1	Model 2
Included items	All 7 items	1, 2, 3 and 4
Alpha	0.894	0.868
RMSEA	0.149	0.095
CFI	0.916	0.991
TLI	0.873	0.972
SRMR	0.053	0.017

Note Grey cells mean good fit for each index

Discussion

When diagnosing a patient, we order many tests in the hopes that more tests would provide a more precise diagnosis. There is no problem in doing so, as long as conducting the tests is supported by strong clinical evidence. Yet there might be problem in practicality; too many tests probably place an avoidable burden not only on patients, but also on the hospital system itself. As a result, system-wide, the benefit from the extensive number of tests could be less than we expected. We can expand on this idea using the development of a statistical model as an analogy: When we develop a multiple regression model, it is recommended to ensure model parsimony as a model with too many covariates is generally penalized.[15,16] In diagnosis, those covariates correspond to the ordered tests while the dependent variable might be the disease to be diagnosed. Clearly, the best-case scenario would be predicting the dependent variable using the smallest possible number of covariates, if no decrease in precision is guaranteed.

Administering a survey questionnaire to improve quality and safety in healthcare organizations is similar: More information could be collected with more items in a questionnaire, but there is no doubt that already overworked healthcare professionals would suffer from completing such a survey. More importantly, the burden from a long questionnaire increases the survey refusal and drop-out rate as well as the number of missing values. As this missing structure due to the length of the survey is neither completely random nor random, modern imputation methods are not easily applicable by definition.[17] List-wise deletion cannot solve the problem either because of the same issue of missing structure, in addition to the loss of too many responses.[18-20] All in all, we should control our greed and keep the length of the entire survey questionnaire within a manageable range for healthcare professionals' busy schedule. Of course, the validity of the optimized version must be ensured.

This study distinguishes itself from other similar studies by using IRT and factor analysis approach in tandem. It is important to describe why we chose to do so. Those two approaches are, by definition, built on two different paradigms. Traditional factor analysis assumes the measurement scale is linear[21]—basically, a ratio scale or at least an interval scale where we can freely obtain central tendency and variation.[3] However, in many survey instruments, including the TPSC we analyzed, a response is measured on an ordinal scale, usually Likert, which is not linear at all. Indeed, we can never say the difference between "disagree slightly" and "neutral" or between "neutral" and "agree slightly" is the same.[22] Therefore, applying a traditional factor analysis approach to the TPSC is, technically speaking, inappropriate.[23]

On the contrary, IRT, especially GRM, treats responses as an ordinal scale, not a ratio or interval, as it should.[24,25] This is well illustrated in Figures 1 and 2 presented earlier, where the IIF curves of items were depicted. Here, the point is the curves, not straight lines, and humps caused by response category changes. The format of IIF curves gives us an intuitive grasp of the relative importance of each item.[12] Such information naturally begs the question: Why did we use a traditional factor analysis in conjunction with IRT instead of just sticking to IRT? The answer lies in the dimension of history. The development of original instruments was based on a factor analysis paradigm; thus, simultaneously applying the traditional factor approach and relatively new IRT paradigm is more useful than we might think. Each methodology can serve as safeguard to the other. If the results from the two methods had been different, that would have signaled the need for a deep, exhaustive examination of each item. In this study, the results from the two converged, indicating that our findings were reliable.

In the first step, an exploratory factor analysis by PCFA would need to be explained in more detail. We used PCFA because of its popularity,[14] but had its communality assumption not held, we prepared other factor analysis methods as fallbacks,[14,21] although they turned out not to be necessary. Thus, for an exploratory factor analysis, there is huge room for applying different methodology chosen for each specific situation. So why conduct such a seemingly unstable analysis? Conducting an exploratory factor analysis has more meaning than just screening less important items. In this study, the results showed that each of the current EE and WLB domain items was aligned well under its corresponding single latent factor, regardless of the existence of less informative items. This result directly supports the empirical evidence that IRT's unidimensional assumption was not violated,[26] which enabled us to proceed to IRT GRM and IIF calculation. Although there are more sophisticated approaches for checking and handling such an assumption,[27,28] we chose not to go further due to one of the goals of this study—namely, providing guidelines for non-statistician practitioners.

In the last step, the CFA of candidate models had a significant implication, much beyond just choosing the best item combination. For both the EE and WLB domains, reduced models showed better model fit than the full model with all the original items. Several scenarios can explain this, such as the difference in language and cultural soil between where the instrument was developed and where it is being adapted and administered.[29] This suggests that we may not need to be obsessive in translating and using all the items in the original version, as we habitually do. Instead, checking the potential to downsize is preferable considering the burden to participants and even better model fit. It cannot be overstated that more items do not always guarantee higher validity.

Readers may wonder to what number of items we can reduce an instrument. Conventionally, each domain in a CFA with multiple domains should include at least two items.[30] Yet we strongly recommend at least four items, especially when an individual domain-specific score is the important information the survey aims to measure; two is too small to guarantee the centrality and spread of a domain score. Furthermore, testing differential item functioning (DIF) is an issue to consider. Although seldom tested and adjusted,

DIF is a major source of bias to be considered. To analyze DIF with responses in an ordinal scale like a 5-point Likert scales, at least four items are needed to obtain stable results, even with the Makov chain Monte Carlo method.[31, 32]

Although briefly mentioned earlier, we would like to reiterate that every step shown in this article could be done solely by IRT. Even exploratory and confirmatory factor analyses can be done under the name of item factor analysis.[33,34] However, few survey instruments in healthcare have been developed and validated with IRT. Thus, until all or at least the majority of instruments are developed or updated using IRT, our two-pronged approach fills the gap, sufficing both IRT-oriented and traditional factor approach-oriented personnel.

Conclusion

It is hard to resist the temptation to include more items in a survey questionnaire, especially when we have the "more is better" mindset. Yet it may be time to reconsider that groundless axiom. Indeed, we showed that more may not always be better; rather, it could be worse in terms of both instrument validity and length. If you get feedback from your colleagues that they are too busy to complete a long survey, it is an indisputable signal for you to examine your instrument and consider item reduction. In the healthcare setting, we should never forget that completing a survey takes up participants' precious time that could be used for direct patient care.

We close this article by quoting Mark Twain: "I did not have enough time to write short." Yes, it is an oxymoron, but it may be the very thing that we do when we develop a survey questionnaire. It is time to stop Mark Twain's joke. We have too many lives to save.

Goodness of fit indices used in this study

- RMSEA: The Root Mean Square Error of Approximation
- CFI: Comparative Fit Index
- TLI: Tucker-Lewis Index (also known as non-normed fit index (NNFI))
- SRMR: Standardized Root Mean Square Residual

This study was approved by the Institutional Review Board of the National Cheng Kung University in Taiwan: B-ER-106-194-T. (Safety culture questionnaire survey for hospital employees seldom requires IRB approval. Yet, for future study we received one)

We do hope the methodology introduced in this article helps your endeavor. Please do not hesitate to contact the corresponding author for all Stata commands for the analyses in this article.

Acknowledgements

None.

Conflicts of interest

None.

References

1. Fowler FJ. Improving survey questions: Design and evaluation. *Social Research Methods*. 1995;38:200.
2. Heon–Jae Jeong, Mijin Park, 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):00105.
3. 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):00100.
4. Sexton JB, Helmreich RL, Neilands TB, et al. The Safety Attitudes Questionnaire: psychometric properties, benchmarking data, and emerging research. *BMC Health Serv Res*. 2006;6(1):44.
5. Lee WC, Wung HY, Liao HH, et al. Hospital safety culture in Taiwan: a nationwide survey using Chinese version safety attitude questionnaire. *BMC Health Serv Res*. 2010;10:234.
6. Etchegaray JM, EJ Thomas. Comparing two safety culture surveys: safety attitudes questionnaire and hospital survey on patient safety. *BMJ Quality & Safety*. 2012;21(6):490–498.
7. Carvalho REFLd, SHDB Cassiani. Cross–cultural adaptation of the Safety Attitudes Questionnaire–Short Form 2006 for Brazil. *Revista Latino–Americana de Enfermagem*. 2012;20(3):575–582.
8. Colla JB, Bracken AC, Kinney LM, et al. Measuring patient safety climate: a review of surveys. *Qual Saf Health Care*. 2005;14(5):364–366.
9. Relihan E, Glynn S, Daly D, et al. Measuring and benchmarking safety culture: application of the safety attitudes questionnaire to an acute medical admissions unit. *Ir J Med Sci*. 2009;178(4):433–439.
10. Maslach C, SE Jackson, MP Leiter. Maslach Burnout Inventory. Consulting Psychologists Press, Palo Alto, CA, USA. 1986.
11. Pennebaker JW, M Colder, LK Sharp. Accelerating the coping process. *J Pers Soc Psychol*. 1990;58(3):528–537.
12. 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):00079.
13. Reise SP, KF Widaman, RH Pugh. Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychol Bull*. 1993;114(3):552.
14. Acock AC. Discovering structural equation modeling using Stata. Stata Press Books, USA, 2013;pp. 306.
15. Stolzenberg RM. Multiple regression analysis. In: Melissa Hardy & Alan Bryman (Eds.), Handbook of Data Analysis. Publisher, City, USA, 2004;pp. 165–208.
16. Daniel WW, WD Wayne. Biostatistics: a foundation for analysis in the health sciences, (10th end) Publisher, City, USA. 1995.
17. Rubin DB. Multiple imputation for nonresponse in surveys. John Wiley & Sons, USA. 2004.
18. Roth PL. Missing data: A conceptual review for applied psychologists. *Personnel Psychology*. 1994;47(3):537–560.
19. Gary King, James Honaker, Anne Joseph, et al. List–wise deletion is evil: what to do about missing data in political science. Annual Meeting of the American Political Science Association, Boston, MA, USA. 1998.
20. Mazza GL, CK Enders, LS Ruehlman. Addressing Item–Level Missing Data: A Comparison of Proration and Full Information Maximum Likelihood Estimation. *Multivariate Behavioral Research*. 2015;50(5):504–519.
21. Thompson B. Exploratory and confirmatory factor analysis: Understanding concepts and applications. American Psychological Association, USA. 195. 2004.

22. Agresti A, Kateri M. Categorical data analysis. Springer, USA, 2011;pp. 206–208

23. Jeong HJ, WC Lee. Ignorance or Negligence: Uncomfortable Truth Regarding isuse of Confirmatory Factor Analysis. *Journal of Biometrics & Biostatistics*. 2016;7(3):298.

24. Samejima F. Estimation of latent ability using a response pattern of graded scores. *Wiley Online Library*. 1969;1968(1):169.

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

26. Drasgow F, CL Hulin. Item response theory. In: Dunnette MD & Hough LM (Eds.), Handbook of Industrial and Organizational Psychology, Consulting Psychologists Press, CA, USA. 1990;1:577–636.

27. Nandakumar R. Assessing Dimensionality of a Set of Item Responses–Comparison of Different Approaches. *Journal of Educational Measurement*. 1994;31(1):17–35.

28. Harrison DA. Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics*. 1986;11(2):91–115.

29. Beaton DE, Bombardier C, Guillemin F, Ferraz MB et al. Guidelines for the process of cross–cultural adaptation of self–report measures. *Spine* (Phila Pa 1976). 2000;25(24):3186–3191.

30. Cole DA. Utility of confirmatory factor analysis in test validation research. *J Consult Clin Psychol*. 1987;55(4):584.

31. Jeong HJ, WC Lee. Does Differential Item Functioning Occur Across Respondents' Characteristics in Safety Attitudes Questionnaire? *Biometics & Biostatistics International Journal*. 2016;4(3):00097.

32. Choi SW, LE Gibbons, PK Crane. Lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *J Stat Softw*. 2011;39(8):1–30.

33. Jeong HJ, WC Lee. The Pure and the Overarching: An Application of Bifactor Model to Safety Attitudes Questionnaire. *Biometrics & Biostatistics International Journal*. 2016;4(6):110.

34. Maydeu–Olivares A, H Joe. Assessing approximate fit in categorical data analysis. *Multivariate Behav Res*. 2014;49(4):305–328.

Citation: Liao HH, Lee WC, You YL, et al. A practical approach to develop a parsimonious survey questionnaire—taiwanese patient safety culture survey as an example. *Biom Biostat Int J.* 2017;6(3):350–355. DOI: 10.15406/bbij.2017.06.00169