Barriers to reporting medication errors: a measurement equivalence perspective

Jason M Etchegaray,[1] Terry Throckmorton[2]

[1]The University of Texas Medical School at Houston, The University of Texas-Houston Memorial Hermann Center for Healthcare Quality and Safety, Houston, Texas, USA
[2]The Methodist Hospital, Houston, Texas, USA

Correspondence to
Dr Jason M Etchegaray, The University of Texas Medical School at Houston, The University of Texas-Houston Memorial Hermann Center for Healthcare Quality and Safety, 6410 Fannin Street, UTPB 11.08, Houston, TX 77030, USA; jason.etchegaray@uth.tmc.edu

Accepted 9 May 2009
Published Online First 4 August 2010

ABSTRACT

Objectives To demonstrate a statistical analysis for testing the measurement equivalence of a patient safety survey instrument. The survey instrument examined in the present study is the Medication Administration Error Reporting Survey.

Methods Surveys were posted to a random sample of registered nurses in the State of Texas, with 435 nurses completing the survey. The surveys contained questions about various error reporting issues, including the 16-item, Medication Administration Error Reporting scale. Nurses were divided into one of two samples—calibration and holdout—to ensure replicability of the results. Within each sample, two groups were created based on nurse tenure on the job.

Results Multiple Group Confirmatory Factor Analysis was conducted across nurses with varying levels of experience for the calibration and holdout samples. For each sample, a baseline model was estimated, where model parameters were allowed to vary across the nursing groups, and compared with more restrictive models. The results provided support for the factor structure of the Medication Administration Error Reporting System but yielded mixed results concerning the equivalence of the measure across nursing groups.

Conclusions The present study provides an explanation of how to examine the measurement equivalence of survey instruments and demonstrated that the Medication Administration Error Reporting scale might not be equivalent across nurses who differ with respect to experience levels.

The increase in patient safety research during the last few years has led researchers to focus on different aspects of medication errors, including error reporting.[1–4] Research has indicated that while direct observation of medication administration yields significantly better detection rates of errors than that found from chart reviews and incident report reviews, medication errors are severely underdetected and under-reported in practice.[4] To better understand why medication errors are not reported, Wakefield *et al*[5] created the Medication Administration Error Reporting (MAER) survey. The present study extends previous research on the MAER by examining whether it demonstrates measurement equivalence, an important psychometric property of surveys that health services researchers need to examine prior to making comparisons between groups.[6 7]

Researchers have identified four measurement factors for the MAER, with each factor demonstrating an acceptable level of internal consistency.[5 8] The factors for the MAER, which are the focus of the present study, are (1) 'Disagree with definition,' which focuses on the ways nurses conceptualise the definition of errors (four items), (2) 'Fear,' which measures blame and negative outcomes for nurses when they report errors (five items), (3) 'Administrative response,' which examines administration's handling of the reporting of errors (four items), and (4) 'Reporting effort,' which examines time and effort needed to report an error (two items). While these four factors seem reasonable, given the analysis and interpretation from these authors,[5 8] a limitation the authors noted is that their studies have been focused on nurses in only one state. Further, they have not examined measurement equivalence, which refers to the finding that the dimensionality of a measurement scale is conceptualised similarly by different participants and that the items used to measure the scale relate to the construct being measured in a similar manner for different raters.[9] A lack of measurement equivalence means that differences between groups responding to a test (eg, measuring intelligence[10]) or survey (MAER, in the present study) are not reflective of true differences on the construct being measured (eg, intelligence and error reporting) but rather due to bias in some aspect of the test/survey process (eg, biased questions). Often, researchers want to compare responses from groups who differ on job types,[11] geography,[12] gender[13] or across time[10] to determine whether between-group differences exist and if interventions are needed to address such differences. Measurement equivalence is necessary to have confidence that the between-group differences are reflective of real differences on the construct (ie, intelligence) being measured, and not an artefact of the measurement process.

Measurement equivalence has been examined for numerous tests/surveys across different fields. Facteau and Craig[11] showed that performance ratings of managers from multiple raters (ie, self, peer, supervisor) were equivalent, and Etchegaray[13] demonstrated equivalence of ratings of executive performance when the raters differed on gender. Conversely, Wicherts *et al*[10] demonstrated a lack of equivalence for some intelligence tests over time, and Tucker *et al*[12] found support for the lack of equivalence of a life satisfaction scale between Russians and North Americans respondents. Mark and Wan[6] have also examined the measurement equivalence of a patient satisfaction survey across participants differing with respect to gender, race and time, while Hurtado *et al*[7] focused on equivalence across Spanish and English versions of an inpatient care quality survey.

One test that allows for an examination of measurement equivalence is multigroup confirmatory factor analysis (MGCFA). Researchers using

the measurement equivalence framework compare a baseline model, where factor loadings are allowed to be free across the two (or more) groups for which equivalence is being tested, to a number of more restrictive models. To the extent that one of the more restrictive models does not fit the data significantly worse than the baseline model, measurement equivalence is established. Previous literature[4 6–13] provides additional insight about the methodology used to assess measurement equivalence.

For the present study, it was decided to create groups for the measurement equivalence analysis based on nurse experience because previous research[3] has shown that RNs differing in experience report a significantly different number of errors, with more experienced nurses reporting fewer errors. It is reasonable to expect that perceptions of errors are influenced by nursing experience, given that the nursing literature[14 15] has indicated that differences between experts and novices occur in different areas of work. Therefore, it is possible that nurse experience plays a role in reasons why nurses report errors. To the extent that future researchers want to compare reasons nurses report errors based on experience, it is important to establish the measurement equivalence for the MAER. The present study has two primary objectives: (1) demonstrate how the measurement equivalence approach can be used on an important issue in patient safety (ie, why medication errors are not reported) and (2) examine whether nurses who differ with respect to nursing experience respond to the MAER similarly.

METHODS

Setting and participants

A list of registered nurses in the State of Texas was obtained from the Board of Nurse Examiners. IRB approval and informed consent were obtained. From the list, a random sample of registered nurses was selected for inclusion in the present study. This sample was mailed a survey along with a self-addressed stamped envelope for survey return. Four thousand two hundred and fifty nurses were mailed surveys, with 435 returning a completed survey. Three hundred and ninety-six women and 39 men completed the survey. The ethnicity of participants was 82% Caucasian, 5% African–American or Black, 5% Hispanic, 5% Asian, and 3% other. The average number of years licensed was 20.3 (SD=11.4), and the average tenure with their current organisation was 12.1 (SD=17.5).

Measures

The 16-item MAER survey, focused on reasons that nurses do not report errors, was used in the present study.

Procedure

The surveys contained a code number that linked the individual nurse's name with the survey; the code number was used only to determine whether a second survey should be sent to a nurse and was held by a third-party vendor throughout the project. Second surveys were mailed out only if the participant did not complete the first survey. All surveys were anonymous once entered into the database.

Data analysis

A median split was performed on the nurses based on their years since licensure to test the hypothesis for the present study. Group A consisted of nurses with less than 20 years of experience, and group B consisted of nurses with 20 or more years of experience. To increase confidence in the findings, nurses were divided into a calibration and hold-out sample, resulting in 192 nurses (ie, 96 nurses in each of the two groups) in the calibration sample and 192 nurses in the hold-out sample.

AMOS 17 was used to conduct the MGCFA on this four-factor measure. Given that a four-factor solution for this measure has been found to fit the data best,[5 8] the four-factor model was examined in the present study with all factors allowed to covary with each other. MGCFA was used to estimate the fit of a baseline model for both groups; to the extent that the baseline model provides adequate fit, measurement equivalence may be examined by comparing the baseline model to a second model, referred to as an equal factor loadings model. The equal factor loadings model specifies that the factor loadings for each item are equal for nurses in groups A and B. If the equal factor loadings model (ie, measurement equivalence) provides a similar fit to the baseline model, one may conclude that the ratings are equivalent.

Assessment of model fit for CFA

Model fit may be assessed from statistical and practical perspectives.[16] The χ^2/df ratio, with ratios less than three indicating good model fit, was examined. Researchers[17] have also suggested examining additional fit indices to assess model fit. Three fit indices were used in the present study: Non-Normed Fit Index (NNFI[17]), Comparative Fit Index (CFI[18]) and the Root Mean Square Error of Approximation (RMSEA[19]). Researchers have suggested that values for the NNFI and CFI should be greater than 0.90, and RMSEA values should be 0.05 or less.[20 21]

RESULTS

Although Wakefield *et al*[5 8] initially proposed 16 items for the MAER, one of the items did not load on any of the factors and is therefore not examined in the present study. The MGCFA results are contained in table 1 for the calibration and holdout samples. Estimation of the baseline model for the calibration sample indicated acceptable fit, with χ^2 (df=168)=248, $p<0.05$, χ^2/df=1.48, NNFI=0.90, CFI=0.92, RMSEA=0.05. Factor loadings for all of the items in the a priori baseline model were significant for all nurses, regardless of their level of experience. Given that the baseline model for the calibration sample had acceptable fit, the equal factor loadings model was estimated by constraining the factor loadings for the items to be equal across both groups of nurses. The equal factor loadings model for the calibration sample provided a significantly worse fit than the baseline model, with χ^2 (df=179)=269.99, $p<0.05$, NNFI=0.89, CFI=0.90, RMSEA=0.05. The χ^2 difference test, which examines the decrement in fit between the revised baseline and equal factor loadings models, indicated a significant decrement change in the value of χ^2:$\Delta\chi^2$ (df=11)=21.99, $p<0.05$.

Given that the equal factor loadings model provided a worse fit than the baseline model for the calibration sample, a partial measurement invariance framework[22] approach was employed, with each factor loading individually allowed to covary (with all other factor loadings remaining constant). This approach indicated that the equal factor loadings model could be significantly improved, while not indicating a significantly worse fit than the baseline model, if the factor loadings for all items in the 'Disagree with Definition' factor were not forced to be equal across the two groups. The modified model with equal factor loadings yielded χ^2 (df=176)=259.64, $p<0.05$, NNFI=0.90, CFI=0.91, RMSEA=0.05. The χ^2 difference test, which compared the modified equal factor loadings model with the baseline model, did not indicate a significant decrement in model fit: $\Delta\chi^2$ (df=8)=11.64, $p>0.05$.

The baseline model examined for the holdout sample was the same model examined for the calibration sample. Results

Table 1 Confirmatory factor analysis for calibration sample

Sample	Model	χ^2	df	χ^2/df ratio	Significance of $\Delta\chi^2$	Tucker-Lewis Index (TLI)	Comparative fit index	Root mean square error of approximation
Calibration	Baseline	248.00	168	1.476	—	0.90	0.92	0.05
	Equal factor loadings	269.99	179	1.508	0.02*	0.89	0.90	0.05
	Modified equal factor loadings model	259.64	176	1.475	0.17†	0.90	0.91	0.05
Hold-out	Baseline	249.54	168	1.485	—	0.90	0.92	0.05
	Equal factor loadings	260.82	176	1.482	0.19	0.90	0.91	0.05
	Equal intercepts	273.18	191	1.430	0.65	0.98	0.99	0.05
	Equal uniqueness	290.52	206	1.410	0.30	0.99	0.99	0.05
	Equal factor variance	298.59	210	1.422	0.08	0.99	0.99	0.05

*Represents a significant decrement in fit from the less restricted model.
†Represents a test for decrement in fit compared with the baseline model.

indicated χ^2 (df=168)=249.54, p<0.05, χ^2/df=1.49, NNFI=0.90, CFI=0.92, RMSEA=0.05. The equal factor loadings model examined for the holdout sample was the modified model examined for the calibration sample. For the holdout sample, the equal factor loadings model, along with the more restrictive models examining measurement equivalence, indicated acceptable fit, as shown in table 1.

DISCUSSION

The present study examined the extent to which perceptions of barriers to error reporting, collected via the MAER, were perceived similarly for nurses differing on experience. The results indicate that with the exception of the 'Disagree with Definition' factor, where lack of measurement equivalence was identified, the MAER appears to measure the factors equivalently across groups. The reasons for the lack of equivalence for this factor are unclear, but further research is required to (1) replicate the results of the present study and, if replication occurs, (2) examine alternative items asking about the same content. Given the results of the present study, it appears that the nurses in the two groups conceptualised the items for this scale differently. As such, comparisons between nurses in these two experience groups for this factor should not be made, as such comparisons would not indicate actual differences between groups.

Despite the lack of measurement equivalence for the calibration sample, the results of the baseline model for both samples indicated that the four-factor model provided a good fit to the data. The replication of the four-factor structure via MGCFA, combined with previous research around the MAER, indicates that these dimensions appear to measure distinct dimensions of barriers to error reporting.

The present study had three significant limitations. First, the sample size was relatively small. Examining a larger sample size will strengthen confidence in the findings by allowing for multiple statistical approaches (MGCFA and IRT) and the use of calibration and hold-out samples. Next, the sample was limited to only nurses in the State of Texas. While this sample provides a difference from previous research on this measure,[5 8] the inclusion of nurses from multiple states will allow for a more generalisable study. Third, the nurses were divided into two groups based on a median split. Although the median split approach allowed for a demonstration of MGCFA, it is doubtful that nurses at the upper end of Group A (19 years) would be considered to be that different (based on experience) from those at the lower end of Group B (20 years). Further, there are alternative ways to measure nursing experience, such as tenure since being licensed. A larger sample size and alternative conceptualisations of experience would allow for a more thorough examination of the measurement equivalence of the MAER based on experience.

CONCLUSION

There are two practical findings from the present study. First, prior to assuming that measures are equivalent across multiple groups, researchers need to empirically demonstrate such equivalence. The findings from the present study support the notion that nurses who differ on experience conceptualise items related to barriers to error reporting differently. Second, there are many surveys currently used to examine perceptions about differing concepts, for example, patient satisfaction, safety culture and safety climate, that need to examine their measures in a similar manner to ensure that equivalence exists prior to making between-group comparisons. Hopefully, researchers will build on studies[6–13] using the measurement equivalence framework to develop a large body of knowledge on which to rely so that future scientific and applied research might yield meaningful results.

Acknowledgements The authors would like to thank E Thomas, for his guidance during manuscript development, and QSHC reviewers/editor, for comments on a previous version of this manuscript.

Funding Funding for the first author provided by The University of Texas-Houston Memorial Hermann Center for Healthcare Quality and Safety, The University of Texas Center of Excellence for Patient Safety Research and Practice (Agency for Healthcare Research and Quality Grant no 1 PO1 HS11544-01), and a K02 award from the Agency for Healthcare Research and Quality (Grant no 1 K02 HS017145-02). Funding for the second author provided by Texas Nurses' Association, District 9.

Competing interests None.

Ethics approval Ethics approval was provided by the The University of Texas MD Anderson Cancer Center.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. *Advances in patient safety: from research to implementation*. **Vols 1—4**, AHRQ Publication Nos. 050021 (1—4). Rockville (MD): Agency for Healthcare Research and Quality, 2005. http://www.ahrq.gov/qual/advances/.
2. **Institute of Medicine.** *To err is human*. Washington: National Academy Press, 1999.
3. **Walters JA.** Nurses' perceptions of reportable medication errors and factors that contribute to their occurrence. *Appl Nurs Res* 1992;**5**:86—8.
4. **Flynn EA,** Barker KN, Pepper GA, *et al*. Comparison of methods for detecting medication errors in 36 hospitals and skilled-nursing facilities. *Am J Health Syst Pharm* 2002;**59**:436—46.
5. **Wakefield DS,** Wakefield BJ, Uden-Holman T, *et al*. Understanding why medication administration errors may not be reported. *Am J Med Qual* 1999;**14**:81—8.
6. **Mark BA,** Wan TTH. Testing measurement equivalence in a patient satisfaction instrument. *West J Nurs Res* 2005;**27**:772—87.
7. **Hurtado MP,** Angeles J, Blahut SA, *et al*. Assessment of the equivalence of the Spanish and English versions of the CAHPS Hospital Survey on the quality of inpatient care. *Health Serv Res* 2005;**40**:2140—61.
8. **Wakefield BJ,** Uden-Holman T, Wakefield DS. *Development and validation of the medication administration error reporting survey. advances in patient safety: from research to implementation*. **Vol 4**, AHRQ Publication Nos. 050021 (1—4). Rockville: Agency for Healthcare Research and Quality, 2005. http://www.ahrq.gov/qual/advances/.
9. **Vandenberg RJ,** Lance CE. A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organizational Research Methods* 2000;**3**:4—70.

10. **Wicherts JM,** Dolan CV, Hessen DJ, *et al*. Are intelligence tests measurement invariant over time? investigating the nature of the flynn effect. *Intelligence* 2004;**32**:509—37.
11. **Facteau JD,** Craig SB. Are performance appraisal ratings from different rating sources comparable? *J Appl Psychol* 2001;**86**:215—27.
12. **Tucker KI,** Ozer DJ, Lyubomirsky S, *et al*. Testing for measurement invariance in the satisfaction with life scale: A comparison of Russians and North Americans. *Soc Indic Res* 2006;**78**:341—60.
13. **Etchegaray JM.** Measurement equivalence of executives' performance ratings for same- and opposite-gender dyads. *Journal of Leadership Studies* 2007;**1**:21—32.
14. **Benner P.** *From novice to expert: Excellence and power in clinical nursing practice*. Menlo Park: Addision-Wesley Publishing Company, 1984.
15. **Bennett DS,** Dune L. Everyday thoughts: Harnessing the thought process toward a practical framework for increasing critical thinking and reducing error. *Crit Care Nurs Clin North Am* 2002;**14**:385—90, viii—ix.
16. **Reise SP,** Widaman KF, Pugh RH. Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychol Bull* 1993;**114**:552—66.
17. **Bentler PM,** Bonnet DG. Significance tests and goodness of fit in the analysis of covariance structures. *Psychol Bull* 1980;**88**:588—606.
18. **Bentler PM.** Comparative fit indexes in structural models. *Psychol Bull* 1990;**107**:238—46.
19. **Steiger JH,** Lind J. Statistically based tests for the number of common factors. *Paper presented at the annual meeting of the Psychometric Society*, Iowa City, IA, (1980, May).
20. **Bollen KA.** *Structural equations with latent variables*. New York: Wiley, 1989.
21. **Kline RB.** *Principles and practice of structural equation modeling*. New York: The Guilford Press, 2005.
22. **Byrne BM,** Shavelson RJ, Muthen B. Testing for the equivalence of factorial covariance and mean structures: the issue of partial measurement invariance. *Psychol Bull* 1989;**105**:456—66.