A Personal Reflection on Measurement from Three Decades Away

I have been involved in the nursing profession since 1982. As I look back and reflect upon my own participation in the evolution and maturity of measurement within nursing, I feel simultaneously excited and sheepish. My excitement stems from seeing a continued proliferation of high-quality measurement articles, such as the one in this issue of *Research in Gerontological Nursing*, “Diagnostic Accuracy of the Cognitive State Test in the Detection of Dementia Among Iranian Older Adults” (Lotfi, Tagharrobi, Sharifi, & Abolhasani, 2015). My sheepishness arises from my early grasp of measurement as a naïve “paint by numbers” approach involving a handful of statistical approaches: Cronbach's alpha for internal consistency, check; factor analysis to confirm the subscales, check; item-scale correlation, check; and finally, bivariate correlations or Kappa agreement for interrater reliability, check. Perhaps my experiences were a reflection of how measurement in nursing education and research has matured over the past 30 years.

To provide historical context, there was heated disagreement about the minimum entry level into nursing that progressed throughout the 1980s (Van Winkle, 1987). My home state of Pennsylvania was awash with diploma and associate degree programs, all graduating competent nurses (Pennsylvania Nurses Association, 1980). The ability to conduct research was apparently one of the things that differentiated us from diploma and associate degree nurses. As a result, those of us in bachelor programs were told that we were going to be the leaders and change agents in nursing. I could barely change a tracheostomy dressing, but...I digress. Part of my preparation to become this change agent was plodding through a nursing research class senior year. The final assignment involved developing and administering a survey, which my group and I did with haphazard gusto and minimal comprehension about validity and reliability assessments. None of us involved in undergraduate education would conceive such a project today because we have the benefit of three decades of nursing education research that informs us otherwise. Even with doctoral students, we are careful about having them engage in instrument development because we know it is a lengthy and systematic process.

The work by Lotfi et al. (2015) illustrates such a systematic process. The original instrument, the Turkish Cognitive State Test, was developed by a neurologist in Turkey whose practice comprised large numbers of older adults with minimal to no formal education (Babacan-Yildiz et al., 2013). Babacan-Yildiz et al. (2013) noted that the most common screening tests for dementia were developed in literate populations and the items were literate-dependent—and thus invalid—for illiterate patients. The instrument contains 38 items to assess 11 cognitive domains. Scores range from 0 to 30, with higher scores reflecting better cognitive states. Lotfi et al. (2015) adapted the Turkish Cognitive State Test to produce the Persian Cognitive State Test (P-COST). They also conducted validity and reliability assessments, which have been published elsewhere. Their current study represents another important and systematic step in their trajectory of instrument modification and evaluation: diagnostic accuracy.

The study will satisfy clinicians and researchers because of the methodical way in which the authors provide support for the cutoff value of 25.5 to discriminate between individuals with and without dementia. The first strength is the use of the known groups method, the gold standard in this case being diagnoses received as a result of radiological and clinical examinations. The authors offer support for their selection of the cutoff value via the receiver operating characteristic (ROC) curve and Youden Index. Both statistics involve comparisons between sensitivity and specificity values. The ROC curve compares the true positives to...
false positives for each score (Polit & Yang, 2015). The area under the curve should be as close to 1 as possible; for this instrument, the area under the curve is 0.95. The Youden Index, which is calculated by subtracting the sums of the sensitivity and specificity values from 1 \([(\text{sensitivity} + \text{specificity}) - 1]\), can range from 0 to 1. A score of 0 indicates that the instrument has no discriminatory value, whereas a score of 1 indicates perfect discriminatory value (Polit & Yang, 2015). Lotfi et al. (2015) provide a table showing sensitivity, specificity, and the Youden Index values for several proposed cutoff scores. The remaining diagnostic parameters of the P-COST are favorable.

The efforts of Babacan-Yildiz et al. (2013) and Lotfi et al. (2015) also demonstrate a growing sophistication among clinicians in the area of instrument development, refinement, and evaluation. Those involved with instrument development and adaptation are becoming more selective with their choices for reliability and validity assessments, which often depend on the theoretical basis for instrument development (e.g., classical test theory, item-response theory) (Polit & Yang, 2015). There was a time when a measure of internal consistency, such as Cronbach’s alpha, was \textit{de rigueur} for any instrument—which may have resulted in perfectly good and clinically valid instruments being rejected, because the individual items contributed to a formative index, not a reflective scale (Polit & Yang, 2015). This distinction is important. The items that comprise a reflective scale are considered the result of the underlying trait and are conceptually viewed as effect indicators (Polit & Yang, 2015). Examples of reflective scales are instruments that measure depression. The items reflect the extent of the depression. If all items capture some aspect of the construct, depression, the responses to all items will be related to the amount of depression experienced by the individual. Hence, the likelihood that a positive response to a hypothetical item (e.g., “I feel blue”) will correlate to a positive response to a hypothetical item (e.g., “I feel sad”) is great. Cronbach’s alpha is an appropriate assessment for internal consistency when evaluating reflective scales.

On the other hand, items that comprise a formative index characterize the construct. Examples of formative indices are behavioral checklists. One example is the Resistiveness to Care Scale and subsequent modified versions (Jablonski-Jaudon et al., in press; Volicer, Hurley, & Mahoney, 2015). Care-resistant behavior is not an internal trait; it is an observable behavior that can be quantified using a validated and reliable instrument. The items do not correlate; for example, an individual who exhibits the behavior “turn away” is not more or less likely to “grab object.” These behaviors do not necessarily correlate. Cronbach’s alpha for an assessment of internal consistency is an inappropriate choice. Another aspect that adds to the confusion is the use of the words scale and index. Scale should refer to reflective instruments and index should refer to formative instruments (Polit & Yang, 2015). This convention is not consistently followed.

In conclusion, occasional reflection can help remind us of where we have been and the progress we are making—not just in the areas of gerontological nursing and research, but in measurement as well. I predict that we will see continued progress and changes within the science of measurement. Progressive advances in statistics and computing will open new vistas in reliability and validity assessment techniques.

In an ironic twist of fate, I now teach a measurement class to doctoral students: full circle.

REFERENCES

\textit{Rita A. Jablonski-Jaudon, PhD, CRNP, FAAN}
\textit{Associate Professor}
\textit{School of Nursing}
\textit{The University of Alabama at Birmingham}
\textit{Birmingham, Alabama}

\textit{The author has disclosed no potential conflicts of interest, financial or otherwise.}

doi:10.3928/19404921-20151104-01