Survey research: be careful where you step ...

J M Etchegaray, W Fischer

Common misconceptions about the reliability and validity of different constructs

Two interesting articles recently appeared in *QSHC*. Kho *et al*¹ published their findings about the test-retest reliability and internal consistency of different forms of the Safety Climate Survey,² and this was followed by some guidelines by Pronovost and Sexton³ about the measurement of culture. Given that researchers are focusing more on using surveys to measure different issues related to culture, the focus of the present discussion is to examine common misconceptions about the reliability and validity of different constructs. Advances in other fields are also mentioned to shed light on some of the perplexing issues faced by health services researchers attempting to study culture. While Pronovost and Sexton³ are correct in stating that “we must understand these sources of variation in order to target” different improvement opportunities and that “failure to understand these important issues can cause managers to make incorrect inferences regarding scores” (page 231) on any measure, we believe that several incorrect statements in their editorial need to be examined so that culture research (and survey research in general) can take the correct step in the journey to better measurement and assessment of constructs.

RELIABILITY AND VALIDITY

Reliability focuses on our inference concerning the level of consistency demonstrated by scales or items that comprise a variable.⁴ As psychometric experts⁵ explained, a variable (for example, teamwork) is measurable through survey questions while a construct is a latent expression of the variable and is therefore not directly measurable (that is, observable). There are many different ways to compute reliability depending on the specific inference one wants to make about a measure. Test-retest reliability allows researchers to examine reliability over time. While this is used sometimes by researchers,¹ the more popular choice for estimating reliability is internal consistency. Internal consistency refers to the

“average correlation among items within a test” (page 251).⁵

Importantly, internal consistency is an indication of the *reliability* of a measure, not *construct validity* as Pronovost and Sexton³ asserted. While Pronovost and Sexton praised Kho *et al*¹ for “their rigorous evaluation of construct validity and reliability”, their study should be viewed as important work that examined three scales in an ICU, collected field data, and added knowledge about only the *reliability* of the measures. Unfortunately, their study does not enhance the validity—construct or otherwise—of the Safety Climate Survey. Reliability determines the upper limit of validity (see page 219, Formula 7 in Guion⁴), but the two concepts are distinct. Further, as noted by Guion,⁴ “internal consistency is necessary but not sufficient for construct validity” (page 90).

The second misinterpretation concerning internal consistency is that high levels of internal consistency can serve as an indication of unidimensionality of a measure. Pronovost and Sexton³ made this assertion when noting that “the high score on internal consistency suggests that” the Safety Climate Survey reflects “a single construct—safety culture” (page 231). As the psychometric literature⁴ indicates, lengthening a scale (by adding additional items) can result in a more reliable scale (as evidenced by internal consistency). Given this fact, it is possible to develop a scale that measures multiple dimensions and has high internal consistency. For example, a scale that has five items about attitudes to teamwork (where these items are strongly correlated with each other) and five items about attitudes to culture (with strong correlations also existing between these items) will probably have high internal consistency, yet measure multiple dimensions. While many researchers use a minimum level of reliability to determine whether to include a scale in their analysis,¹ reliability can serve as a precursor to informing whether they want to examine the dimensionality of a scale but it *does not ensure* unidimensionality of a scale.

Unidimensionality requires a different type of statistical test altogether. Principal axis factor analysis, which is one type of exploratory factor analysis,⁶ is used to test for the unidimensionality of factors. Using this technique, unidimensionality is established if the first eigenvalue for each factor is greater than one and accounts for more than 20% of the variance.⁶ The presence of small values for the second eigenvalues and their respective percentage of variance accounted for also contributes to the establishment of unidimensionality.

There are many different types of validity that researchers seek to find for their scales, including internal, statistical, construct, content, criterion, and external. Basically, validity is defined as the extent to which an item or scale measures the construct it was designed to measure. A scale that measures a construct consistently is viewed as reliable; a scale that measures the intended construct is viewed as valid. There is much literature^{4,5,7} that discusses how to assess the validity of measures; suffice it to say, the assessment of validity is drastically different from the assessment of reliability.

CONCLUSION

Health services researchers need to assess the validity and reliability of their measures correctly. It is hoped that this commentary will dispel some misconceptions offered by previous researchers about these important issues. If researchers are not careful about how they conceptualize reliability and validity related to any survey instrument, they might encourage the scientific community and healthcare organizations to go in the wrong direction. Pronovost and Sexton³ started their article with Latsus's famous quotation: “*The journey of a thousand miles begins with one step*”. We agree with that quotation, but hope that health services researchers take the *correct* first step in inferring information about the reliability and validity of their instrument.

Qual Saf Health Care 2006;15:154–155.
doi: 10.1136/qshc.2005.016659

Authors' affiliations

J M Etchegaray, W Fischer, The University of Texas M D Anderson Cancer Center, Houston, TX 77030, USA

Correspondence to: Dr J M Etchegaray, The University of Texas M D Anderson Cancer Center, 1515 Holcombe Blvd, Unit 141, Houston, TX 77030, USA; jmetcheg@mdanderson.org

Funding: none.

Competing interests: none declared.

REFERENCES

- 1 **Kho ME**, Carbone JM, Lucas J, *et al*. Safety Climate Survey: reliability of results from a multicenter ICU survey. *Qual Saf Health Care* 2005;**14**:273–8.
- 2 **Sexton JB**, Thomas EJ, Helmreich RL. Error, stress, and teamwork in medicine and aviation: cross sectional surveys. *BMJ* 2000;**320**:745–9.
- 3 **Pronovost P**, Sexton B. Assessing safety culture: guidelines and recommendations. *Qual Saf Health Care* 2005;**14**:231–3.
- 4 **Guion RM**. *Assessment, measurement, and prediction for personnel decisions*. Mahwah, NJ: Lawrence Erlbaum Associates, 1998.
- 5 **Nunnally JC**, Bernstein IH. *Psychometric theory*. New York: McGraw Hill, 1994.
- 6 **Reckase MD**. Unifactor latent trait models applied to multi-factor tests: results and implications. *J Educ Stat* 1979;**4**:207–30.
- 7 **Messick S**. Validity of psychological assessment: validation of inferences from persons' responses and performance as scientific inquiry into scoring meaning. *Am Psychologist* 1995;**9**:741–9.