How Much Is Lost in Using Single Items?

Ron D. Hays, PhD[1], Steven Reise, PhD[2], and José Luis Calderón, MD[3]

[1]Department of Medicine, University of California Los Angeles, Los Angeles, CA, USA; [2]Department of Psycholgy, University of California Los Angeles, Los Angeles, CA, USA; [3]Charles Drew University, Los Angeles, CA, USA.

J Gen Intern Med 27(11):1402–3
DOI: 10.1007/s11606-012-2182-6

In this issue of *JGIM*, West et al.[1] demonstrate with multiple data sets that single items perform similarly to the Maslach Burnout Inventory (MBI) long-form measures of emotional exhaustion and depersonalization in terms of associations with suicidality, serious thoughts of dropping out of medical school, endorsing dishonest behavior, disagreeing with an altruistic attitude, and perceived major medical error. When differences were found between the single items and the full-length scales, the single items "tended to slightly underestimate the magnitude of association." The underestimates for single items are consistent with lower reliability of measurement.

Multiple items are used to sample a range of content when one item is not a sufficient indicator of a construct. Multiple-item scales yield scores that are generally more reliable (and potentially valid) than those produced by single items. Reliability of scores from one item can be estimated using the intraclass correlation. Reliability of scores from multiple-item scales can be estimated using a special case of the Spearman-Brown prophecy formula,[2] where K is the number of items and I is the intraclass correlation: (K*I)/(1+(K−1)*I).

What may be surprising to readers is how well the single items performed relative to the full scales. But this phenomenon has been demonstrated before. For example, Robins, Hendin, and Trzesniewski[3] showed that a single-item measure of self-esteem performed similarly to the ten-item Rosenberg self-esteem scale. So how much is lost when using single rather than multiple items? If a concept is substantively complex, a single item will not represent the construct as well as multiple items. If a concept is conceptually narrow, it is possible for a single item to represent it about as well as multiple items.

Similar results for scores estimated from full length versus subsets of items are expected if the two are *very highly* correlated. But even if the longer and shorter variants of a measure are correlated at r=0.70, they can have substantially different correlations with a criterion. Consider a single item (S1), a multi-item scale (S2), and a criterion (C). The correlation between S1 and C can be estimated using a formula from Cohen, Cohen, West, and Aiken:[4*]

$$r_{S1.C} = r_{S1.S2}r_{S2.C} \pm \sqrt{\left(1 - r^2_{S1.S2}\right)\left(1 - r^2_{S2.C}\right)}$$

If the correlation between S1 and S2 is 0.70 and the correlation between S2 and the criterion C is 0.50, the correlation between S1 and C is constrained between 0.35± 0.62. That is, the S1 and C correlation is free to take on values anywhere between −0.27 and 0.97, so there is no guarantee that it is even positive. Thus, even if the correlation between a single item and the multiple item scale is large, it is possible that there could be different patterns of correlations with other variables.

The single items examined by West et al.[1] were those found in previous analyses to have the largest factor loadings on emotional exhaustion ("I feel burned out from my work") and depersonalization ("I have become callous toward people since I took this job"). The Spearman correlations of these items with the full length emotional exhaustion and depersonalization scales ranged from 0.76 to 0.83 and 0.61–0.72, respectively.[5] The item with the largest factor loading is most related to the scale score, but each item in a unidimensional scale is positively related with the scale score and may have similar associations with other variables. It would have been informative if West et al.[1] had reported the performance of all the items in the emotional exhaustion and depersonalization scales. In addition, they could have evaluated whether associations of individual items with the criterion variables vary by characteristics such as gender and age.

The demonstration by West et al.[1] of the similarity of a single item subset to the sum of items measuring a single concept is the underpinning of short-form measures. For example, short-form measures of patient, physician, and other hospital employee perceptions of hospital care were derived by selecting subsets of items that accounted for at least 90 % of the variance in the long-form scale.[6,7] A similar approach was used in selecting the items for the Medical Outcomes Study 36-Item Short-Form Health Survey.[8]

Published online August 10, 2012

*Typos in the published formula were corrected by Dan Ozer, Ph.D., UC Riverside.

Item response theory (IRT) is grounded in knowing that any subset from a pool of unidimensional items can be used to represent the underlying concept.[9,10] IRT has well-known advantages over the simple-summated scoring used in the MBI example provided by West et al.[1] In two-parameter IRT models, item difficulty (how likely in general people are to response high versus low to an item) and discrimination (how strongly related the item is to the underlying score) are estimated for each item in the scale. These parameters indicate the degree to which different items yield information for each person assessed. The best item is the one with the highest discrimination (analogous to the highest factor loading) that is closest in difficulty to where the person is on the underlying continuum ("ability").

Within the limits of the reliability of the full item set, a subset of items can be selected that yields whatever target level of reliability is desired for a particular application. Different subsets of items can be used for different respondents to estimate scale scores as efficiently as possible. This tailored item selection process is the essence of computer-adaptive testing.[11] Because all items are scored on the same metric, the estimate of the underlying score (theta) is more accurate and appropriate than assuming equal item parameters like West et al.[1] did when they arbitrarily equated one item to the total score by multiplying by the number of items in the scale. Reliability adequate for individual-level purposes (i.e., >0.90) has been achieved with about five items in the Patient-Reported Outcomes Measurement Information System (PROMIS) project: http://www.nihpromis.org/measures/selectinginstrument.

In summary, all items in unidimensional scales are positively associated with one another and the total scale score. For scales that represent narrow concepts, the items will tend to be more highly correlated, and the total score may be estimated accurately with fewer items. In addition, it is possible for a multi-item scale to have a single item that is very highly associated with the scale total (i.e., has a high factor loading or IRT discrimination parameter) and to perform like the full-length scale. The lower response burden makes parsimonious measures desirable when they retain the psychometric strengths of longer measures. However, equivalence of scores produced from single versus multiple items needs to be demonstrated and tradeoffs carefully considered on a case-by-case basis.

Corresponding Author: *Ron D. Hays, PhD; Department of Medicine, University of California Los Angeles, 911 Broxton Avenue, Los Angeles, CA 90024, USA (e-mail: drhays@ucla.edu).*

REFERENCES

1. **West CP, Dyrbye LN, Satele DV, Sloan JA, Shanafelt TD.** Concurrent validity of single item measures of emotional exhaustion and depersonalization in burnout assessment. J Gen Internal Med. 2012; doi:10.1007/s11606-012-2015-7
2. **Clark EL.** Spearman-Brown formula applied to ratings of personality traits. J Ed Psych. 1935;26:552–555.
3. **Robins RW, Hendin HM, Trzesniewski KH.** Measuring global self-esteem: construct validation of a single-item measure and the Rosenberg self-esteem scale. Pers Soc Psych Bull. 2001;27:151–161.
4. **Cohen J, Cohen P, West SG, Aiken LS.** Applied multiple regression/correlation analysis for the behavioral sciences. 3rd ed. Mahwah: Erlbaum; 2003.
5. **West CP, Dyrbye LN, Sloan JA, Shanafelt TD.** Single item measures of emotional exhaustion and depersonalization are useful for assessing burnout in medical professionals. J Gen Internal Med. 2009;24:1318–1321.
6. **Hays RD, Larson C, Nelson EC, Batalden PB.** Hospital quality trends: a short-form patient-based measure. Med Care. 1991;29:661–668.
7. **Nelson EC, Larson CO, Hays RD, Nelson SA, Ward D, Batalden PB.** The physician and employee judgment system: reliability and validity of a hospital quality measurement method. Qual Rev Bull. 1992;18:284–292.
8. **Ware JE Jr, Sherbourne CD.** The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care. 1992;30:473–483.
9. **Hambleton RK, Swaminathan H.** Item response theory: principles and applications. Boston: Kluwer-Nijhoff; 1985.
10. **Hays RD, Morales LS, Reise SP.** Item response theory and health outcomes measurement in the 21st Century. Med Care. 2000;38(Suppl): II28–42.
11. **Choi SW, Reise SP, Pilkonis PA, Hays RD, Cella D.** Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. Qual Life Res. 2010;19:125–136.