Educational Measurement
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
Summer 2015, Vol. 34, No. 2, pp. 6–7

Comments on "A Note on Subscores" by Samuel A. Livingston

Sandip Sinharay, *Pacific Metrics Corporation,* and Shelby J. Haberman, *Educational Testing Service*

According to the Standards 1.14 and 2.3 of the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014), empirical evidence should justify reporting of any scores including subscores. Following those guidelines, the use of real data provides the basis of assessment of the value of subscores in our publications such as Haberman (2008), Haberman, Sinharay, and Puhan (2009), Puhan, Sinharay, Haberman, and Larkin (2010), and Sinharay (2010, 2014) where we have evaluated subscores for individuals and for institutions and have examined both estimation of true subscores and estimation of true residuals from regression of subscores on total scores. Numerous specific tests have been examined by the use of data from actual test administrations. Sinharay, Rijmen, Choi, and Dorans (2014) also emphasized the use of real data to assess the value of subscores.

In contrast, Livingston (2015) provided, instead of real data, some hypothetical scenarios related to gain scores. As a consequence, it is not possible to know the relevance of his comments to real life. Nonetheless, if real rather than hypothetical data are used, then some general comments can be made concerning appropriate practice.

To begin, if changes in student performance are to be evaluated, then the following steps should be followed in order to apply the approach of Haberman (2008) to individuals or the approach of Puhan, Sinharay, Haberman, and Larkin (2010) to groups:

- Compute changes in each subscore for each student between the beginning and end of the course. That is, compute Subscore 1 at the end of course minus Subscore 1 at the beginning of course, Subscore 2 at the end of course minus Subscore 2 at the beginning of course and so on, for each student.[1]
- Apply the method of Haberman (2008) to the above changes in subscores instead of to the subscores themselves. The computations would involve the reliabilities of the changes in the subscores and the covariances between the changes. If institutions are studied, then proceed as in Haberman, Sinharay, and Puhan (2009).

Consider application of the recommended procedure to the hypothetical examples in Livingston (2015). In the first hypothetical example, changes in the three true subscores are perfectly correlated with the change in the total score, so that the subscores provide no useful information concerning individuals or even groups that is not provided by the total score. For any individual or institution, all inferences should be based on the total score. Presumably reporting of summary information would indicate that, relative to the change in true scores for Subscore 3, the change in true scores is always 10 greater for Subscore 2 and 20 greater for Subscore 1, but this fact has no value for any specific individual or group, because the result applies to everyone. Similar comments appear to apply to the second hypothetical example of Livingston concerning proportional changes in observed scores, although three hypothetical observations hardly constitute a substantive example.

In real applications involving evaluation of performance change between two time periods, reliability can be a major problem even for total scores, as noted by Lord and Novick (1968, p. 76), who described gain scores as "notoriously unreliable." For instance, Allison (1990, p. 95) included an example where scores at Time 1 and Time 2 have the same variance and each has reliability of .7 (as in the first table of Livingston, 2015) and are correlated .6; the reliability of the change score is .25. Therefore, any defense of subscores based on an argument concerning changes over time is unlikely to satisfy professional standards. Nonetheless, we would welcome a real example of longitudinal data on individuals or institutions which did indeed demonstrate the value of observed or augmented changes in subscores.

In addition to their hypothetical nature, the examples appear to have little relationship to reality. Research has shown that subscores in educational testing are highly correlated (e.g., Sinharay, 2010; Wainer & Feinberg, 2015). Therefore, it is extremely unlikely that students would gain five times as much on Subscore 1 as on Subscore 2 (second table of Livingston, 2015) or gain two standard deviations on average on Subscore 1 and zero on average on Subscore 3 (first table).

At this point, the challenge is for Livingston (2015), or for someone else, to provide a real example of subscores from an educational assessment that are useful in the context of gain scores, but are not useful when considered at a single time point. Until such an example is provided, the issue raised by Livingston and cited as a concern of his as early as 2007 in Puhan et al. (2010, p. 283) is of no practical interest.

Note

[1] Note that one has to ensure that the subscores at the beginning and end of the course are on the same scale. If not, one has to apply an equating procedure before computing the differences. Puhan and Liang (2011) discussed equating of subscores.

Sandip Sinharay, Pacific Metrics Corporation, Monterey. CA 93940; ssinharay@pacificmetrics.com. Shelby J. Haberman, Educational Testing Service, Princeton, NJ 08541; shaberman@ets.org.

References

Allison, P. D. (1990). Change scores as independent variables in regression analysis. *Sociological Methodology*, *20*, 93–114.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*, 204–229.

Haberman, S. J., Sinharay, S., & Puhan, G. (2009). Reporting subscores for institutions. *British Journal of Mathematical and Statistical Psychology*, *62*, 79–95.

Livingston, S. A. (2015). A note on subscores. *Educational Measurement: Issues and Practice*, *34*(2), 5.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Puhan, G., & Liang, L. (2011). Equating subscores under the nonequivalent anchor test (NEAT) design. *Educational Measurement: Issues and Practice*, *30*(1), 23–35.

Puhan, G., Sinharay, S., Haberman, S. J., & Larkin, K. (2010). The utility of augmented subscores in a licensure exam: An evaluation of methods using empirical data. *Applied Measurement in Education*, *23*, 266–285.

Sinharay, S. (2010). When can subscores be expected to have added value? Results from operational and simulated data. *Journal of Educational Measurement*, *47*, 150–174.

Sinharay, S. (2014). Analysis of added value of subscores with respect to classification. *Journal of Educational Measurement*, *51*, 212–222.

Sinharay, S., Rijmen, F., Choi, S., & Dorans, N. J. (2014). The revised standards and its role in research on educational measurement. *Educational Measurement: Issues and Practice*, *33*(4), 36–38.

Wainer, H., & Feinberg, R. (2015). For want of a nail: Why unnecessarily long tests may be impeding the progress of Western civilisation. *Significance*, *12*, 16–21.