Editor's Introduction.

Reliability Versus Accuracy: A Critical Distinction

Test reliability coefficients traditionally have been used to judge the quality of measurement. And reliability coefficients of .90 have often been considered adequate to assure the quality for standardized testing and large-scale assessment programs. However, a test reliability of .90 (or above) does not insure that individual test scores, such as national percentile ranks, are accurate. Consider, for example, a mathematics test with a reliability of .90 and imagine a student taking that test whose true score is at the 50th percentile; that is, we know that the student's actual capability is at that level. The probability is less than 1/3 (.309) that when the student takes the test, she will obtain a score within 5 percentile points of her true score, the 50th percentile (Rogosa 1999a,b).

The following informal example attempts to explain why high test reliability does not indicate good accuracy for an individual score, without the encumbrances of percentile rank scoring, complex measurement models, and other technical detail.

Shoe Shopping and the Reliability Coefficient

David Rogosa
Stanford University

Dedicated to Al Bundy,
A man who cares as much about good measurement as he does about his own children.

Try this on:

1. Customers are drawn from the combined population of male and female shoe-shoppers (female feet translated to the male shoe-size scale).

2. Mr. Bundy measures each shopper's shoe size as either too large or too small with equal probability.

- On a good day Mr. Bundy misses the correct shoe size by one-half size too big or one-half size too small.
- On other days Mr. Bundy misses the correct shoe size by a full size too big or a full size too small.

In each case the shoe size measurement error has mean 0 (overall and at each level of shoe size) and is uncorrelated with actual shoe size.

3. The accuracy of shoe fitting on the good day is poor, as most wearers would notice a half-size misfitting. On the other days the accuracy is totally unacceptable, as a full-size misfitting would presumably be unwearable.

4. The reliability coefficient for Al Bundy measurements on the good day is .944 (comparable to the very best reliabilities on nationally normed standardized tests, such as Stanford 9, Form T Total Reading in the lower grades), even though accuracy is poor. The reliability coefficient for Al Bundy making errors of a full shoe size is .808 (comparable to reliabilities on shorter standardized tests such as Stanford 9, Form T Science or Social Science in the upper grades or Spelling in some lower grades), even though accuracy is unacceptable.

Technical Notes

The notes below provide details on the construction of the empirical example and results for some alternative examples.

Error Process

Error variances are .25 for half-size errors and 1.0 for the full-size errors. A third (intermediate) error process can be constructed by pooling over good and bad Bundy days (or equivalently interleaving half-size and full-size errors), producing an error process with variance .625 (error distribution has mass 1/4 at values {-1, -1/2, 1/2, 1}). The resulting reliability for the empirical example is .871 (comparable to reliabilities for the total Mathematics score in the upper grades on the Stanford 9, Form T).

Empirical Shoe-size Distribution

Construction of the empirical shoe size example started with a histogram of male shoe sizes obtained from www.sizefinder.com. Information on constructing the female shoe size distribution and translating these into male sizes was obtained from www.genderweb.org and www.rei-outlet.com. That distribution, shown in Figure 1, has mean 7.7 and variance 4.21. An alternative empirical distribution is constructed from a mixture of male and female distributions taken from www.genderweb.org which give size in inches: male N(10.53, .53) female N(9.58, .51). Translating that mixture into discrete shoe sizes produces results similar to Figure 1, but with slightly larger variance of 4.47 and thus the reliabilities for Bundy measurements are slightly larger-- reliabilities .947, .817, and .877 for the half-size, full-size, and combined error processes.

Artificial examples for shoe size distributions

A previous version of this example used a discrete Uniform distribution: the distribution of sizes has support (5,15) with values at integer and half-integer sizes. This distribution produces a variance of true sizes 9.167 and thus much larger reliabilities of .973, .902, and .936 for the half-size, full-size, and combined error processes. Less unrealistic artificial examples would use a discrete triangular distribution: for example a triangular distribution with mass at at integer and half-integer sizes, endpoints at (2.5, 14.5) and mode at 6.5. This artificial distribution yields a mean size 7.83 and variance of 6.18 which implies reliabilities 0.961, 0.861, and 0.908 for the half-size, full-size, and combined error processes.

Acknowledgements

Thanks to Matt Finkelman for assistance in the construction of the empirical distribution example and the calculations. Thanks to the Google search engine for locating the web resources. The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B60002 and Award Number R305B960002-01 as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

References

Rogosa, D.R. (1999a). Accuracy of Individual Scores Expressed in Percentile Ranks: Classical Test Theory Calculations. CRESST Technical Report 509, September, 1999.

Rogosa, D.R. (1999b). How Accurate are the STAR National Percentile Rank Scores for Individual Students?--An Interpretive Guide. CRESST, August 1999.

Web Resources

"Average Bodies." http://www.genderweb.org/general/ave_bo.phtml

(17 October 2001).
"How rare are you?" http://www.sizefinder.com/new_page_2.htm (6 October 2001).
"How To Choose The Correct Shoe Size."
http://www.rei-outlet.com/reihtml/LEARN_SHARE/footwear/chcorrectsize.html
(21 October 2001).

note to editor:
Web references follow form from Columbia University Press online,
chapter 2.8 www.columbia.edu/cu/cup/cgos/idx_basic.html
the dates indicate the day that we viewed these sites.

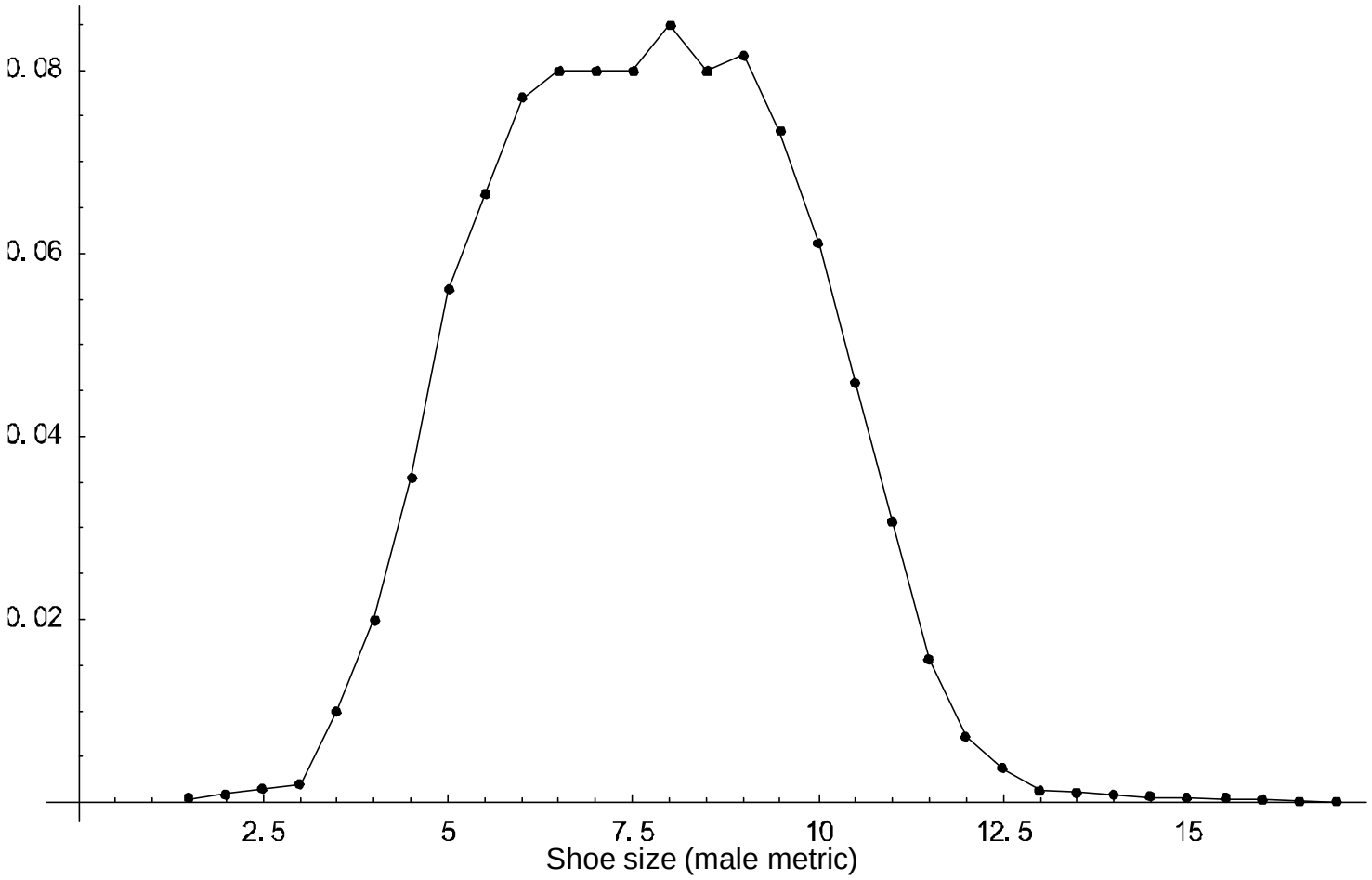

Figure 1. Empirical distribution of shoe size (male and female populations combined).