Standards for Interpreting Reliability Coefficients

Randall E. Schumacker, Ph.D.

The Standards for Educational and Psychological Testing provides direction for test score reporting and usage in the credentialing of persons in many occupations and professions (1999). Tests used in credentialing by professional organizations are designed to determine whether the essential knowledge and skills of a specific domain have been mastered by the examinee. Credentialing typically covers a number of related but distinct areas and is an ongoing process with several tests given on a regularly scheduled basis, so revisions and/or newer versions are often necessary. Important testing information including reliability coefficients are useful in comparing scores from these different tests, but interpretation allowances must be made for the variability of scores from different samples of examinees, administration techniques from which the reliability coefficients were obtained, the source(s) of error indicated by the reliability coefficient, the number of items on the test, and the length of time allowed for testing. Nunnally and Bernstein (1994) provided guidance in the interpretation of the reliability coefficient by stating that a value of .70 is sufficient for early stages of research, but that basic research should require test scores to have a reliability coefficient of .80 or higher. When important decisions are to be made with test scores, a reliability coefficient of .90 is the minimum with .95 or higher a desirable standard.

Item Covariance Effects on Internal Consistency Reliability Coefficient

The following inter-item correlation matrix indicates that one item (Item 11) doesn't correlate positively with the other items. The consequence of a low or negative inter-item correlation (even with just one item) is a reduction in the value of the reliability coefficient. Removal of the one bad item (Item 11) will result in an internal consistency reliability coefficient of .91 (10 item test) compared to .77 (11 item test). Given the valid ten item test, a higher reliability of scores can be achieved by dropping the one bad item; resulting in 91% of the obtained score being true score with only 9 % measurement error.

```
Inter-Item Correlation Matrix (N of Cases = 137)
Item1 1.000
Item2 .626 1.000
Item3 .551 .644 1.000
Item4 .654 .634 .624 1.000
Item5 .503 .626 .708 .623 1.000
Item6 .542 .496 .604 .601 .656 1.000
Item7 .683 .611 .625 .652 .591 .749 1.000
Item8 .394 .184 .237 .297 .267 .183 .323 1.000
Item9 .344 .536 .354 .456 .481 .396 .471 .397 1.000
Item10 .373 .369 .472 .437 .430 .408 .442 .548 .658 1.000
Item11 -.060 -.010 -.041 .006 .000 -.060 .105 -.066 .025 .019 1.000
```

SOLUTIONS

Classical test analysis is useful when constructing tests, especially when examining traditional coefficients of reliability. Validity of test scores is the more important concept, but validity is limited by the reliability of a set of scores. Professional associations should seek test scores that have a high degree of validity as well as a high degree of reliability. Professional associations should consider computing reliability using Rasch measurement models (Schumacker, 2004). We are well trained in providing classical, generalizability theory, and latent trait theory reliability estimates for test data.

REFERENCES

Nunnally, J.C. & Bernstein, I.H. (1994). Psychometric Theory (3rd Edition). McGraw-Hill Series in Psychology, McGraw-Hill, Inc., New York: NY, 264-265.

Schumacker, R.E. (2004). Rasch Measurement: The Dichotomous Model. In Introduction to Rasch Measurement, Smith, R. and Smith, E. (Eds.), Chapter 10, JAM Press, Maple Grove, MN.

Standards for Educational and Psychological Testing (1999). Joint Committee of the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, AERA: Washington: DC.