Determining an Appropriate Index of Reliability for the TAKS Mixed-Model Tests (Writing, Grade 9 Reading, and English-Language Arts)

Background

With the introduction of the Texas Assessment of Knowledge and Skills (TAKS) tests, the Texas Education Agency (TEA) seeks to determine an appropriate indicator of the reliability of tests that include a mixture of multiple-choice (dichotomous) items along with open-ended/essay (polytomous) items. Under the previous TAAS testing program TEA reported coefficient alpha based on just the multiple-choice component for writing tests. Since this indicator did not factor the essay response at all into the reported estimate of reliability it was desirable to look at other methods for estimating reliability that may provide more appropriate information for the TAKS. This document is intended to briefly record the procedures followed in determining an appropriate index of reliability for TAKS mixed-model assessments (i.e., those involving a mixture of multiple-choice and open-ended or essay questions).

Overview

As a starting point, the joint psychometric & quality assurance team from TEA and Pearson Educational Measurement (PEM) had a number of discussions regarding potential indicators of reliability and reviewed relevant literature on them. Under the direction of TEA, PEM psychometric staff used the Spring 2003 TAKS administration data to provide estimates of reliability for the indicators under consideration to TEA for review and discussion. Among those reliability indicators was the traditional coefficient alpha (including the polytomous items as appropriate), a stratified coefficient alpha, the Angoff-Feldt coefficient, and (for the ELA tests only) the Feldt-Gilmer coefficient. Each of these measures is only briefly described below but references are provided for those interested in examining more detailed background on each of these indicators.

Coefficient Alpha (Cronbach, is probably the most frequently referenced measure of internal-consistency reliability. For purposes of computation herein the SPSS reliability procedure was used to compute this index. For the ELA tests the weighted essay score was used instead of the unweighted essay score.

The stratified coefficient alpha (Cronbach, Schonemann, & McKie, 1965) is determined as if each content component area (multiple-choice, open-ended, and/or essay) is a subtest, or category as follows:

$$Strat\ \alpha\rho_{XX'} = 1 - \frac{\sum \sigma_{X_j}{}^2 \left(1 - \alpha\rho_{X_J X'_J}\right)}{\sigma^2_{Total}}$$

An internal-consistency measure of reliability for each component is needed for this computation. For the multiple-choice component the standard coefficient alpha was used. For essay prompts, the inter-rater correlation (of the first two raters) was used as the estimate of reliability for that component. For the open-ended (short-answer) item component a standard coefficient alpha was used. Once again it should be noted that the weighted essay score was used for the ELA tests. Also, for the ELA tests the stratified

alpha was derived over 4 parts, where the component parts were: (1) Reading M-C items, (2) Writing M-C items, (3) Open-Ended (short answer) items, (4) Essay prompt. The correlation between rater1 and rater2 as the estimate of coefficient alpha for the Essay, actual coefficient alpha were obtained for all other component parts.

The Angoff-Feldt coefficient (Angoff, 1953; Feldt, 1975; Feldt & Brennan, 1989) is useful for tests with two component parts. In the instance of TAKS the two parts were (1) the multiple-choice component, and (2) the polytomous component (which may be essay only, open-ended only, or a combination of open-ended with weighted essay). The coefficient, discussed in the recent Feldt and Charter (2003) article is determined as follows:

$$r_{xx'} = \frac{4(r_{12})(S_{x_1})(S_{x_2})}{S^2_{x_{tot}} - \left(\frac{S^2_{x_1} - S^2_{x_2}}{S_{x_{tot}}} \right)^2}$$

Finally, for the TAKS ELA tests it may be informative to determine reliability from a 3-part measure (multiple-choice, open-ended, and essay components separately). This was done using the Kristof (1974) coefficient which is purportedly equivalent to the Feldt-Gilmer coefficient (Gilmer & Feldt, 1983) that was requested by TEA. For a three-part test (in this instance multiple-choice, open-ended, and weighted essay) the Kristof coefficient is determined from the component parts covariance matrix as follows:

$$r_{xx'} = \frac{\left(S_{X_1X_2} S_{X_1X_3} + S_{X_1X_2} S_{X_2X_3} + S_{X_1X_3} S_{X_2X_3} \right)^2}{\left(S_{X_1X_2} S_{X_1X_3} S_{X_2X_3} S^2_X \right)}$$

Application of Indices to TAKS Spring 2003 Test Data

The following table provides estimates of reliability for the TAKS mixed-model tests based on the Spring 2003 test administration.

Grade/Test	Angoff-Feldt	Kristof/ Feldt-Gilmer	Alpha (SPSS)	Stratified Alpha
11 ELA	0.7684	0.7533	0.8666	0.9363
10 ELA	0.7847	0.7573	0.8521	0.9247
9 Reading	0.8077		0.8745	0.8821
7 Writing	0.8809		0.8804	0.8897
4(E) Writing	0.8242		0.7979	0.8263
4(S) Writing	0.7358		0.7826	0.7970

Simulation Studies

Since multiple measures of reliability were able to be determined, the question became one of determining which indicator was most appropriate for use in the TAKS program. To address this question Dr. Miller from TEA designed and conducted a small-scale simulation study which used the Spearman-Brown corrected split-half reliability coefficient as the "gold standard" against which all others would be compared. Results from this simulation study identified the Stratified Coefficient α as the coefficient which was consistently the closest in value to the Spearman-Brown corrected split-half reliability coefficient.

Dr. Miller from TEA contacted Dr. Leonard Feldt at the University of Iowa, widely regarded as an expert in the area of reliability estimation, regarding this situation, sharing the results of his small-scale simulation. Following is a portion of Dr. Feldt's e-mailed response to Dr. Miller:

> Given the way in which the data were produced, your MC and OE scores did not satisfy the conditionof tau equivalence, Rather, they were what Joreskog (and those of us who followed his lead) called congeneric. This means that the Angoff-Feldt coefficient, the usual alpha coefficient based on item scores, and alpha based on the MC and OE scores all have their assumptions violated. Only the split-half and stratified alpha coefficients have their assumptions well met. As theory would lead us to expect, the coefficients with violated assumptions exhibit various biases. Only the stratified alpha comes close to the split half coefficient, which is appropriately regarded as the standard, as the data were generated.

Dr. Miller then conducted another simulation study in which the MC and OE scores did satisfy tau-equivalence. In this study, the reliability estimates obtained using the Angoff-Feldt, the Angoff-Gilbert, and the Stratified α all were consistently close to that obtained using the (Spearman-Brown corrected) split-half reliability coefficient.

The above indicates a problem that may exist with the use of many reliability coefficients on real datasets, however: reliability estimators such as the Cronbach's α, the Angoff-Feldt coefficient, and the Angoff-Gilbert coefficient are more susceptible to violation of assumptions (i.e., tau equivalence) than a congeneric reliability estimators such as the Stratified α or Joreskog's congeneric reliability estimate (1985). As Raykov (1977) points out, the congeneric model is the least restrictive, most general model of use for reliability estimation: the congeneric model assumes that each individual item measures the same latent variable but with possibly different scales, possibly different degrees of accuracy, and possibly different amounts of error.

Previous Investigation of Reliability Indicators with TAAS

Dr. Miller from TEA had also done some previous examination of the stratified coefficient alpha indicator as it might have been applied to the TAAS writing assessment. This section (below) includes his work examining the comparison of three measures of coefficient alpha as applied to the spring 1998 exit level TAAS Writing test.

Method

Using the student response data from the spring 1998 TAAS exit level writing test, Coefficient α, Standardized Coefficient α, and Thissen's Stratified α were all computed to estimate the reliability of that test. This test had 40 MC items and one four-point essay item that was weighted such that the essay was worth 40 points. Thus, the entire test was worth 80 points.

Applying Cronbach's α to the 41 items as scored above yielded an estimated reliability of 0.583.

Applying Standardized Cronbach's α to the 41 items as scored above yielded an estimated reliability of 0.882.

To compute Thissen's Stratified α statistic, Dr. Miller utilized estimates of (1) the correlation between the MC total score and the essay score and (2) the parallel forms reliability of the essay. For the spring 1998 exit writing TAAS, the estimated (Pearson) correlation between the MC total score and the essay score is 0.57. Not having field test results from spring 1998, he was unable to compute an estimate for the parallel forms reliability of the essay. Dr. Miller, therefore, used the estimated parallel forms reliability of the SAT II Writing Test essay as an estimate, which is approx. 0.60.

Applying Thissen's Stratified α to the spring 1998 exit level writing TAAS yielded an estimated reliability of 0.831.

Taking a closer look at Cronbach's α and Standardized Cronbach's α for use with mixed MC/CR tests, Dr. Miller argues that the two cannot be recommended for such cases.

Consider Cronbach's α. It is known to be a lower bound for reliability unless all items are τ-equivalent. For the exit TAAS writing test, the τ-equivalence assumption is blown away since, in addition to the score range for item 41 being different than the score ranges for the other items, item 41's weight (w_{41}) toward the total score is 10 compared to every other item's weight (w_1, …, w_{40}) being 1. Thus, the low reliability estimate of 0.583. As a matter of fact, one may note that as $w_{41}/w_i \to \infty$, where i=1,…,40, Cronbach's $\alpha \to 0$. Thus, Cronbach's α cannot be recommended for tests which have some items weighted nor for tests which contain items with different score ranges.

Consider Standardized Cronbach's α. Standardizing all the variables on the exit writing TAAS essentially forces the the essay to count one point rather than the 40 points it actually counts toward the total score (i.e., the reliability is calculated as if the essay counts 1/41 of the total test score rather than 1/2 of the total test score). This is unreasonable. Cronbach's alpha and Standardized Cronbach's alpha do not even yield the same estimate even when no weighting is performed and only dichotomous items are used. For example, using only the 40 MC items on the spring 1998 exit writing TAAS, the following was obtained:

Cronbach's α =0.868
Standardized Cronbach's α=0.875.

For this special case, standardizing made very little difference in the estimates obtained -- even so, it can still be seen that Cronbach's α and Standardized Cronbach's α theoretically estimate something different. In fact, Standardized Cronbach's α does not theoretically estimate the reliability of ANY test unless the test is scored using the z-score values of each item response rather than the item responses themselves. There may be cases where using the z-scores may be preferable; for example, when all items are scored on the same scale, no items are weighted, and there is reason to believe that all item variances are equal regardless of the values of the estimated item variances.

Conclusion

Given the internal discussion between TEA and PEM psychometric staff, results of Dr. Miller's work in this area, and the advice from Dr. Feldt, it was recommended by the PQA team that the stratified coefficient alpha be used as the indicator of reliability for the TAKS mixed-model assessments (grades 4 and 7 Writing, grade 9 Reading, and grades 10 & 11 English Language Arts).

This documentation was developed in order to record the process through which such a recommendation came to be made.

References

Angoff, W. H. (1953). Test reliability and effective test length. *Psychometrika*, *18*, 1-14.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334.

Feldt, L. S. (1975). Estimation of the reliability of a test divided into two parts of unequal length. *Psychometrika*, *40*, 557-561.

Feldt, L. S., and Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.) Educational Measurement (3rd Edition, pp. 105-146). New York: American Council on Education and Macmillan.

Feldt, L. S., and Charter, R. A. (2003). Estimating the reliability of a test split into two parts of equal or unequal length. *Psychological Methods*, *8*, 102-109.

Gilmer, J. S., and Feldt, L. S. (1983). Reliability estimation for a test with parts of unknown lengths. *Psychometrika*, *48*, 99-111.

Joreskog, K.G. and Sorbom, D. (1985). LISREL VI: Analysis of linear structural relationships by the method of maximum likelihood. User's Guide. University of Uppsala, Uppsala, Sweden.

Kristof, W. (1974). Estimation of reliability and true score variance from a split of a test into three arbitrary parts. *Psychometrika*, *39*, 491-499.

Raykov, T. (1977). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, 21, 173-184.