

Running Head: A GENERALIZATION OF STRATIFIED α

A Generalization of Stratified α that Allows for Correlated
Measurement Errors between Subtests

Leslie Keng
Pearson

G. Edward Miller
Private Consultant

Kimberly O'Malley
Pearson

Ahmet Turhan
Pearson

Author Note:

Leslie Keng, Pearson, Austin, Texas.

Requests for reprints should be sent to Leslie Keng, Pearson, 400 Center Ridge Drive,

Austin, TX 78753, E-mail: Leslie.Keng@Pearson.com

ABSTRACT

This paper presents a generalization of Stratified α that allows for correlated measurement errors between some subtest scores that make up a composite score. The Generalized Stratified α offers simplicity of calculations and can be applied to subtests that are unidimensional or multidimensional. The Generalized Stratified α was used to estimate the reliability of the spring 2006 grade 3 Texas English Language Proficiency Assessment System (TELPAS) composite score, a composite score of student performance on listening, speaking, writing, and reading subtests. The listening, speaking, and writing subtest scores for each student are ratings typically given by the student's English language teacher. As such, these subtest scores are anticipated to have correlated measurement errors.

A Generalization of Stratified α that Allows for Correlated Measurement Errors between Subtests

Examinees are often administered a set of related subtests and then evaluated based on a composite of the subtests' scores. Well-known examples of composite scores include ACT scores (which are composites of English, Math, Reading, and Science subtest scores) and total SAT scores (which are composites of Reading, Mathematics, and Writing subtest scores). The use of composite scores has become more widespread with federal testing requirements under Title III of No Child Left Behind now calling for states to assess students with limited English proficiency (LEP) annually from kindergarten through 12th grade in the four language domains of listening, speaking, reading and writing. A composite of the student's performance on each of these domains is calculated to represent the student's overall English language proficiency. The composite scores of LEP students in a campus or school district are then combined for federal accountability reporting. To calculate a composite score, some states give equal weighting to each of the language domain subtest scores; while other states apply different weights to each subtest. A state's decision about subtest weighting is usually based on program-specific policies or requirements.

Several methods exist for estimating the reliability of a composite score. These methods include the Coefficient α (Cronbach, 1951), the Stratified α (Cronbach, Schonemann, & McKie, 1965; Feldt & Brennan, 1989), the Angoff-Feldt coefficient (Angoff, 1953; Feldt & Brennan, 1989), the Kristof/Feldt-Gilmer coefficient (Kristof, 1974; Gilmer & Feldt, 1983), Maximal Reliability (Li, Rosenthal, & Rubin, 1996), McDonald's ω (McDonald, 1970, 1999), Multidimensional ω (Kamata, Turhan, & Darandari, 2003), and the structural equation modeling

(SEM) approach of Raykov (1997, 2002). If the subtest scores that make up the composite do not assess the same underlying dimension, several of these reliability coefficients have been found to be biased (Zimmerman, Zumbo, & Lalonde, 1993; Komaroff, 1997; Murphy & DeShon, 2000; Osbourn, 2000; Raykov, 1998, 2001; Raykov & Shrout, 2002; Kamata et al., 2003); exceptions to this are the Stratified α , Maximal Reliability, Multidimensional ω , and the SEM approach of Raykov, which shares the same formal basis as Multidimensional ω . Kamata et al. (2003) found that Stratified α , when directly compared with Maximal Reliability and Multidimensional ω under five different multidimensional factor-structure, generally performed the best. As such, the Stratified α served as the starting point for this research.

An inherent assumption in all but one (the Raykov SEM approach) of the reliability coefficients listed above is uncorrelated measurement errors of the test components or subtests. In practice, this assumption cannot always be expected to hold. For example, in an English language proficiency assessment for LEP students who are assessed in listening, speaking, writing, and reading, each student's scores on the speaking, listening, and writing subtests can be ratings given by the same person (e.g. the student's English language teacher). Thus, the measurement errors in the ratings for these three subtests cannot be assumed to be uncorrelated. In cases such as this, a method for computing composite score reliability given correlated measurement errors should be utilized. The obvious choice would seem to be Raykov's SEM approach. However, Raykov's SEM approach to reliability estimation does have its limitations. First, a large number of subjects are required because it uses a SEM method based on asymptotic theory. Second, quoting Raykov and Shrout (2002), "the approach requires for its application a tenable overall model. This necessitates that the researcher correctly specifies the number of underlying constructs and the particular way the employed measures load on them. If this

specification is incorrect or done in an insufficiently informed way, misleading results can as well follow” (pp. 207-208). Third, this approach, at least as presented by the authors, does not make use of any known subtest reliabilities but instead estimates them from the covariance matrix of the subtests. This can result in incorrect model-estimated subtest reliabilities.

This study serves two purposes. First, it presents a generalization of Stratified α that allows for correlated measurement errors between subtests that is both simple and does not have the inherent limitations of the Raykov SEM approach to reliability estimation. Second, the Generalized Stratified α is applied to operational testing data to estimate reliability of a composite score that measures English language learner (ELL) performance on subtests for the listening, speaking, writing, and reading domains.

A GENERALIZATION OF STRATIFIED α

Let X_i , $i=1, \dots, k$, represent the score on test i of k tests (or subtests) which cannot be presumed to be parallel, essentially tau-equivalent, nor even congeneric. In other words, unidimensionality of the k subtests is not assumed. In addition, assume that the measurement error associated with each test score may be correlated with the measurement errors of some or all of the other k tests. Define composite score Z as

$$Z = \sum_{i=1}^k w_i X_i \quad (1)$$

where $\sum_{i=1}^k w_i = 1$. If uncorrelated measurement errors between the k subtests could be assumed, the reliability of Z could be estimated by Stratified α , given by,

$$\alpha_{\text{Strat}} = 1 - \frac{\sum_{i=1}^k w_i^2 \sigma_{X_i}^2 (1 - \rho_{X_i X_i'})}{\sigma_Z^2} \quad (2)$$

Note that the numerator of Equation 2 is simply the error variance of Z given uncorrelated measurement errors among tests. By replacing that error variance with the error variance of Z , given possible correlated measurement errors between subtests, the revised formula results in a Generalized Stratified α . The generalized formula (see Appendix 1 for the derivation) is:

$$\alpha_{\text{Generalized Strat}} = 1 - \frac{\sum_{i=1}^k w_i^2 \sigma_{X_i}^2 (1 - \rho_{X_i X_i'}) + \sum_{i \neq j}^k w_i w_j \rho_{e_{X_i} e_{X_j}} \sigma_{X_i} \sigma_{X_j} \sqrt{(1 - \rho_{X_i X_i'}) (1 - \rho_{X_j X_j'})}}{\sigma_Z^2} \quad (3)$$

$$= \alpha_{\text{Strat}} - \frac{\sum_{i \neq j}^k w_i w_j \rho_{e_{X_i} e_{X_j}} \sigma_{X_i} \sigma_{X_j} \sqrt{(1 - \rho_{X_i X_i'}) (1 - \rho_{X_j X_j'})}}{\sigma_Z^2} .$$

EXAMPLE

The Generalized Stratified α was used to estimate the reliability of the spring 2006 grade 3 Texas English Language Proficiency Assessment System (TELPAS) composite score, a composite score of student performance on listening, speaking, writing, and reading subtests. The listening, speaking, and writing subtest scores for each student were ratings (1, 2, 3, or 4) given by the student's English language teacher; hence, these three scores likely have correlated measurement errors. The reading subtest was a multiple-choice reading test for which the student received two scores: a scale score and a categorized score of 1-4 based on scale score cutoffs; the categorized reading subtest score was the one used in the computation of the TELPAS composite score. The measurement errors of the reading subtest scores can be assumed to be uncorrelated with the measurement errors of the listening, speaking, and writing subtests given the different assessment method for reading compared with the method for the other three domains.

Descriptive statistics for the subtest scores used in the computation of the TELPAS composite score are presented in Tables 1 and 2.

Insert Tables 1 and 2

The TELPAS composite score is a weighted combination of the listening, speaking, writing, and reading subtest scores with the formula:

$$\text{TELPAS composite score} = 0.05 X_L + 0.05 X_S + 0.15 X_W + 0.75 X_R \quad (4)$$

where X_L denotes the listening subtest score, X_S denotes the speaking subtest score, X_W denotes the writing subtest score, and X_R denotes the reading subtest score. A much higher weight was assigned to the reading subtest score because it is assumed to be the most reliable score.

Estimation of the reliability of the TELPAS composite score was performed using the Generalized Stratified α given by Equation 3. This required estimates of the variances of the scores for all four subtests, estimates of the reliabilities for all four subtests, an estimate of the variance of the composite score, and estimates of the measurement error correlations between the four subtests. For the TELPAS subtests, only some of these were readily available or calculable from the spring 2006 student-level test results. Estimates of the variances of the scores of all four subtests and the composite score were calculable from the student scores. The three measurement error correlations between the reading subtest and each of the listening, reading, and writing subtests were assumed to be zero, as stated previously. However, the measurement error correlations between the listening, speaking, and writing subtests were unknown. Likewise, the reliabilities of the listening, speaking, and writing tests were unknown – no internal consistency estimate of reliability is possible for these single-rating subtests and no students had been re-tested on any of the subtests in order to compute test-retest reliability

estimates. An internal consistency estimate of the reliability of the reading subtest *scale score* was available and its value was 0.94. The reading subtest scale score for this grade ranges from 316 to 888 and it is transformed into a *categorized rating score* ranging from 1 to 4 before it is used in the computation of the TELPAS composite score (in Equation 4). The reliability of the categorized rating score for reading was unknown. Therefore, ad-hoc methods were needed to estimate the four subtest score reliabilities and the measurement error correlations between the speaking, listening, and writing subtest scores.

First, the reliability of the *categorized rating score* for the reading subtest was estimated.

The procedure for doing this was as follows:

1. From the known distribution of observed reading subtest *scale scores* (denoted $X_{RScale,1}, \dots, X_{RScale,n}$), generate random scale scores $W_{RScale,1}, \dots, W_{RScale,n}$.

2. Compute

$$Y_{RScale,i} = \left(1 - \rho_{X_{RScale}, X'_{RScale}} - \sqrt{\rho_{X_{RScale}, X'_{RScale}}^2}\right) \mu_{X_{RScale}} + \rho_{X_{RScale}, X'_{RScale}} X_{RScale,i} + \sqrt{1 - \rho_{X_{RScale}, X'_{RScale}}^2} W_{RScale,i}$$

for $i = 1, \dots, n$. The correlation coefficient between the X_{RScale} 's and Y_{RScale} 's is approximately $\rho_{X_{RScale}, X'_{RScale}}$, the reliability of the uncategorized reading subtest scale scores. This was estimated to be 0.94 based on the internal consistency estimate of the reliability for TELPAS reading scores.

3. Transform both the X_{RScale} 's and Y_{RScale} 's into the categorized rating scores (i.e. 1 to 4) using the known category boundaries of the reading subtest scale scores. Denote the categorized rating scores for X_{RScale} and Y_{RScale} as X_R and Y_R , respectively.

4. Compute the Pearson correlation coefficient between X_R and Y_R . This is the estimated reliability of the categorized rating scores for the TELPAS reading subtest.

Applying the procedure above to the spring 2006 grade 3 TELPAS reading scale scores yielded an estimated reliability of 0.88 for the categorized rating scores of the TELPAS reading subtest.

Next, estimates of the reliabilities of the listening, speaking, and writing subtest scores were computed. As stated earlier, no estimate of internal consistency reliability or test-retest reliability could be computed for any of these subtests because of the format of these subtests. A SEM analysis on the TELPAS data was, thus, performed to obtain estimates of the *minimum* reliabilities for the listening, reading, and writing subtest scores. A pictorial representation of the final structural equation model is given in Figure 1.

Insert Figure 1

The SEM analysis was performed using the software package EQS (Bentler, 1995). A copy of the EQS code used is given in Appendix 2. This analysis yielded the following estimated minimum reliabilities for the listening, speaking, and writing subtest scores, respectively,

$$r_{X_L X'_L} = 0.62, r_{X_S X'_S} = 0.66, \text{ and } r_{X_W X'_W} = 0.76.$$

The same SEM analysis was also used to estimate the measurement error correlations between the listening, speaking, and writing subtests. Preliminary analyses had revealed that the measurement error correlations between the writing subtest and the other two subtests were near zero, so these two measurement error correlations were set to zero for the final analysis. The estimated measurement error correlation between the listening and speaking subtest obtained from the final analysis was: $\hat{\rho}_{e_{X_L} e_{X_S}} = 0.624$.

With these estimates computed, Equation 3 was then used to estimate the TELPAS composite score reliability, yielding Generalized Stratified α value of 0.909.

DISCUSSION

This study presents a generalization of Stratified α that allows for correlated measurement errors between some of the subtest scores that make up a composite score. The Generalized Stratified α is computationally simple, being easily computed with a calculator or a spreadsheet. As with the Stratified α , unidimensionality of the subtest scores is not assumed with the Generalized Stratified α . Application of the Generalized Stratified α is demonstrated by using it to estimate the composite score reliability for one of the grades in a statewide ELL assessment program, the Texas English Language Proficiency Assessment System (TELPAS). The TELPAS composite score is a weighted sum of each student's performance on four English language domain subtests in listening, speaking, writing, and reading.

The TELPAS dataset itself presented challenges in that (1) only the reading subtest had a known reliability (and even that reliability was not for the scale that was actually used in the composite score); and (2) the measurement error correlations between the listening, speaking, and writing subtests were unknown. Procedures for dealing with both of these challenges were also presented in the paper. These challenges are by no means unique to the TELPAS dataset. For example, any testing program that uses composites made up of subtest scores that are single ratings (such as ratings on a portfolio or writing selection) likely face the issue where at least some of the subtest reliabilities are unknown.

The derivation of the method and procedures in this paper originated from the need to address operational issues in a statewide ELL assessment program (TELPAS) to satisfy federal testing requirements. As such, the scenarios presented in this paper are realistic and the method and procedures presented in this paper provide a simple and flexible way of estimating composite score reliability that can prove useful for many statewide testing programs.

REFERENCES

- Angoff, W. H. (1953). Test reliability and effective test length. *Psychometrika*, *18*, 1-14.
- Bentler, P.M. (1995). *EQS structural equations program manual*. Encino CA: Multivariate Software.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.
- Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334.
- Cronbach, L.J., Schönemann, P., & McKie, D. (1965). Alpha coefficient for stratified-parallel tests. *Educational and Psychological Measurement*, *25*, 291-312.
- Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 105-146). New York: Macmillan.
- Gilmer, J. S., and Feldt, L. S. (1983). Reliability estimation for a test with parts of unknown lengths. *Psychometrika*, *48*, 99-111.
- Kamata, A., Turhan, A., & Darandari, E. (2003). *Estimating reliability for multidimensional composite score scale scores*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.
- Komaroff, E. (1997). Effect of simultaneous violations of essential tau-equivalence and uncorrelated error on coefficient alpha. *Applied Psychological Measurement*, *21*, 337-348.
- Kristof, W. (1974). Estimation of reliability and true score variance from a split of a test into three arbitrary parts. *Psychometrika*, *39*, 491-499.
- Li, H., Rosenthal, R. & Rubin, D.B. (1996). Reliability of measurement in psychology: From Spearman-Brown to maximal reliability. *Psychological Methods*, *1*, 98-107.

- McDonald, R.P. (1970). Theoretical foundations of principal factor analysis and alpha factor analysis. *British Journal of Mathematical and Statistical Psychology*, 23, 1-21.
- McDonald, R.P. (1999). *Test Theory: Unified Treatment*. Lawrence Erlbaum Associates.
- Murphy, K.R. & DeShon, R.P. (2000). Inter-rater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology*, 53, 873-900.
- Osbourn, H.G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods*, 5, 343-355.
- Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, 21, 173-184.
- Raykov, T. (1998). Cronbach's alpha and reliability of composite with interrelated nonhomogeneous items. *Applied Psychological Measurement*, 22, 375-385.
- Raykov, T. (2001). Bias of coefficient alpha for congeneric measures with correlated errors. *Applied Psychological Measurement*, 25, 69-76.
- Raykov, T. & Shrout, P. (2002). Reliability of scales with general structure: Point and interval using a structural equation modeling approach. *Structural Equation Modeling*, 9(2), 195-212.
- Zimmerman, D.W., Zumbo, B.D., & Lalonde, C. (1993). Coefficient alpha as an estimate of test reliability under violation of two assumptions. *Educational and Psychological Measurement*, 53, 33-49.

APPENDIX 1

Derivation of Equation 3

Let $X_i, i=1, \dots, k$, represent the score on test i of k tests (or subtests) for which the following usual Classical True-Score Theory assumptions hold

1. $X_i = T_i + e_{X_i}$
2. $E(X_i) = T_i$
3. $\rho_{e_{X_i}T_i} = 0$
4. $\rho_{e_{X_i}T_j} = 0$ for $i \neq j$

but for which the following one usual Classical True-Score assumption does not hold:

5. $\rho_{e_{X_i}e_{X_j}} = 0$ for $i \neq j$.

In addition, assume the k tests are not parallel nor essentially tau-equivalent. (The fact that Assumption 5 does not hold precludes the tests from being parallel or essentially tau-equivalent anyway.)

Define composite score Z as

$$Z = \sum_{i=1}^k w_i X_i .$$

The composite score reliability of Z is defined to be

$$\rho_c = 1 - \frac{\sigma_{e_z}^2}{\sigma_Z^2} .$$

Now,

$$\sigma_{e_z}^2 = \text{Var}(e_Z)$$

$$\begin{aligned}
 &= \text{Var}\left(e_{\sum_{i=1}^k w_i X_i}\right) \\
 &= \text{Var}\left(\sum_{i=1}^k w_i e_{X_i}\right) \\
 &= \sum_{i=1}^k \text{Var}(w_i e_{X_i}) + \sum_{i \neq j}^k \text{Cov}(w_i e_{X_i}, w_j e_{X_j}) \\
 &= \sum_{i=1}^k w_i^2 \sigma_{e_{X_i}}^2 + \sum_{i \neq j}^k w_i w_j \rho_{e_{X_i} e_{X_j}} \sigma_{e_{X_i}} \sigma_{e_{X_j}} \\
 &= \sum_{i=1}^k w_i^2 \sigma_{X_i}^2 (1 - \rho_{X_i X_i'}) + \sum_{i \neq j}^k w_i w_j \rho_{e_{X_i} e_{X_j}} \sigma_{X_i} \sigma_{X_j} \sqrt{(1 - \rho_{X_i X_i'})(1 - \rho_{X_j X_j'})}
 \end{aligned}$$

since $\sigma_{e_{X_i}} = \sigma_{X_i} \sqrt{1 - \rho_{X_i X_i'}}$ = standard error of measurement for X_i .

Hence,

$$\rho_c = 1 - \frac{\sum_{i=1}^k w_i^2 \sigma_{X_i}^2 (1 - \rho_{X_i X_i'}) + \sum_{i \neq j}^k w_i w_j \rho_{e_{X_i} e_{X_j}} \sigma_{X_i} \sigma_{X_j} \sqrt{(1 - \rho_{X_i X_i'})(1 - \rho_{X_j X_j'})}}{\sigma_Z^2}$$

APPENDIX 2

EQS Code for Structural Equations Model Analysis

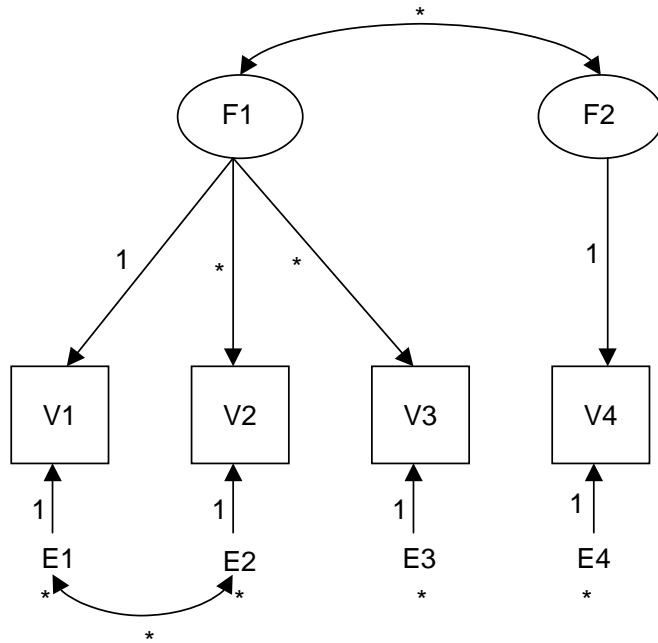
```

/TITLE
  Spr 2006 Grade 3 TELPAS SEM Analysis
/SPECIFICATIONS
  VARIABLES=4; CASES=81448;
  MATRIX=COV;
  METHOD=ML;
/EQUATIONS
  v1=f1+e1;
  v2=*f1+e2;
  v3=*f1+e3;
  v4=f2+e4;
/VARIANCES
  e1,e2,e3,e4=*; f1=*; f2=*;
/COVARIANCES
  e1,e2=*; f1,f2=*;
/CONSTRAINTS
  .12(f2,f2)-.88(e4,e4)=0;
/PRINT
  CORRELATIONS=YES
/MATRIX
  0.888 0.789 0.607 0.578
  0.789 0.942 0.651 0.593
  0.607 0.651 0.894 0.631
  0.578 0.593 0.631 1.136
/END

```

Notation: v1 denotes TELPAS Listening subtest score
v2 denotes TELPAS Speaking subtest score
v3 denotes TELPAS Writing subtest score
v4 denotes TELPAS Reading subtest score

Figure 1
Structural Equation Model for TELPAS Subtests



Model Constraint: $0.12 \text{ Var}(F2) = 0.88 \text{ Var}(E4)$

Variable Notation:

- V1 = TELPAS Listening Score
- V2 = TELPAS Speaking Score
- V3 = TELPAS Writing Score
- V4 = TELPAS Reading Score

Table 1: Descriptive Statistics for TELPAS subtests – Grade 3 (Spring 2006)

Statistic\Subtest	<i>Listening</i>	<i>Speaking</i>	<i>Writing</i>	<i>Reading</i>
<i>N</i>	81,448	81,448	81,448	81,448
<i>Mean</i>	2.78	2.61	2.28	2.95
<i>Standard Deviation</i>	0.94	0.97	0.95	1.07

Table 2: Correlations between TELPAS subtests – Grade 3 (Spring 2006)

Subtest	<i>Listening</i>	<i>Speaking</i>	<i>Writing</i>	<i>Reading</i>
<i>Listening</i>	1.00	0.86	0.68	0.57
<i>Speaking</i>		1.00	0.71	0.57
<i>Writing</i>			1.00	0.63
<i>Reading</i>				1.00