Procedures for Estimating Internal Consistency Reliability

Prepared by the
Iowa Technical Adequacy Project (ITAP)

July 22, 2003

Table of Contents:

❶ Introduction

This guide has been prepared to assist educators in estimating the reliability of a set of scores resulting from the local administration of an assessment instrument. The procedures outlined within this protocol can be generalized to almost any context—regardless of the number and type(s) of items/tasks on the assessment. After the presentation of the mathematical formula for coefficient alpha (Part 2), a detailed description of the steps involved in calculating coefficient alpha is presented (in Part 3), along with an example of how the steps could be followed "by hand." Then, after a brief discussion of interpretation issues (in Part 4), another illustration is provided that shows how the steps can be completed either by hand or by using the *Excel* computer program.

What do we mean by "reliability?"

Reliability is a characteristic of a set of test scores. It is information that tells us how accurate the scores are—how much they might be contaminated by errors that often cause scores to be higher or lower than they really ought to be. Often we think of reliability as being synonymous with "consistency." If we could repeat the assessment, would we get the same answer? Would students get about the same scores?

What kinds of things could happen to cause students to obtain scores that have errors in them? Why don't individuals who are tested receive the exact score they really ought to get? Let's review the categories of errors that can impact students' scores on an assessment before moving on:

- **Random variation within an individual** can cause a student to perform differently from day to day because of health problems, low motivation, distractions that affect concentration, and just not being able to recall some things that actually have been learned. Sometimes guessing is a factor—a student is much luckier or more unlucky than usual. To the extent that all of these factors were to affect a student's test score (in a negative or positive way) on a certain day (a very uncommon occurrence), the student would not obtain a score that is very representative of his/her actual achievement.

- **Situational (environmental) factors** are things that happen in the assessment environment that could interfere with getting a true indication of students' achievements. Many of us have been distracted by someone with the sniffles or the cracking of gum during an otherwise quiet testing period. A room that is too hot and stuffy, or one that is cold enough to bring on the shakes, can be a distraction that disturbs concentration and interferes with trying to remember or apply learned ideas.

- **Instrumentation variables** mainly refer to the specific group of questions that appear on a test that is designed to measure achievement in a particular area. These questions represent a sample from the pool of all of the possible questions that could have been asked. If, in fact, a different but similar sample of questions would have been used, some test takers might have

received higher scores while others received lower scores. Such errors, which are due to "content sampling," can be quite influential in the scores from achievement assessments.

- **Rater idiosyncrasies and subjectivity** are factors that can arise when scoring is done for constructed-response, essay, and performance assessments. When individuals obtain scores that depend somewhat on **who** did the scoring, these kinds of errors can creep into the scores. When some scorers are too harsh or too lenient, or when scorers get tired over time or become unduly influenced by the kind of responses they most recently read, the scores assigned can become contaminated by these factors. The scores represent actual achievement less well than they would if scorers were well trained and consistent in applying the scoring rubric.

How can the presence of errors be detected?

An important step in trying to find out how much measurement errors might be affecting a set of scores is to identify the kinds of errors that might most likely occur in a given assessment circumstance. For example, with scores from a multiple-choice test, rater idiosyncrasies and subjectivity would not be a factor. Machines often do the scoring. So in this situation, random errors associated with health, fatigue, and guessing could occur; conditions in the room during the assessment could present problems; and content sampling errors might be a factor. Once we have an idea of the kinds of errors that could be most intrusive in our circumstance, we can consider a method for gathering information that could help estimate the extent to which those kinds of errors are creating "noise" in the scores. Although there are basically four different types of errors—within examinee, situational, instrumentation, rater/subjectivity—this protocol focuses on estimating **internal consistency** (the magnitude of errors associated with instrumentation, i.e., content sampling) because this type of error is quite common and information about its magnitude is practically useful for all achievement assessment situations.

There are several methods associated with estimating internal consistency. These sets of methods differ in that they require each student to either: a) take two versions of the test (on the same or different days), b) take the same version of the test twice (on different days), or c) take one version of the test one time. Clearly, giving only one version of the test on a single occasion is more practical than having students taking the same test (or different versions of the test) on two different occasions. Some of the names used in discussing this type of approach include "Coefficient alpha" and "K-R20." We can use these methods to determine how extensively content sampling errors might be affecting the scores, but they don't allow us to detect the presence of errors due to temporary factors within test takers (e.g., ill health) or within the assessment environment (e.g., too much noise). Nonetheless, reliability estimates from this approach are helpful because most educators believe that content sampling errors are the most frequently occurring and most damaging types of errors for most students' scores. If items/tasks from the assessment are scored subjectively (e.g., using a rubric or scoring guide), then in addition to internal consistency you will also be interested in detecting errors attributable to the scoring process (e.g., percent of exact rater agreement).

❷ Description of Coefficient Alpha

There are different methods available for estimating internal consistency reliability based on a single administration of a given assessment because assessments differ in composition. For example, KR-20 was developed for use with assessments where each item/task is scored right/wrong (right = 1 point, wrong = 0 points). Many assessments, however, consist of items/tasks that are scored in different ways—some items are scored right/wrong and others might be scored using a scoring guide and worth two or more points each. Coefficient alpha is a method of estimating reliability for a test that is composed of any combination of item types. Thus, coefficient alpha is considered to be most useful for applications to a wide variety of contexts. (This is the reason why we are using it in this guide.). Before moving on to the steps involved with calculating coefficient alpha, it is important to consider the mathematical definition, as provided below.

$$Coefficient\ \alpha = \frac{k}{k-1}\left[1 - \frac{\sum s_i^2}{s_t^2}\right]$$

where:

k = number of separately scored test items/tasks
Σ = the operation symbol meaning "the sum of"
s_i^2 = variance of students' scores on a particular test item/task
$\sum s_i^2$ = sum of the item variances for all test items/tasks
s_t^2 = variance of the total test scores

Variance:

$$s^2 = \frac{\sum\left(X - \overline{X}\right)^2}{n}$$

If calculating s_i^2 (i.e., variance of the scores for a given item/task), then

Σ = the operation symbol meaning "the sum of"
X = score received by a particular student on a particular test item/task
$\overline{X}$ = mean score for a particular test item/task
n = number of students (i.e., number of scores for the test item/task)

If calculating s_t^2 (i.e., variance of the total test scores), then

X = total test score received by a particular student
$\overline{X}$ = mean total test score
n = number of students (i.e., number of total test scores)

❸ Steps in Estimating Coefficient Alpha

1. Prepare data set

To estimate coefficient alpha you must first have access to each score obtained by the students—for each of the test item/task. For example, if the test consists of 10 questions you need to know the score (i.e., number of points) each student received on each of the 10 questions. This information then needs to be presented in a two-dimensional table, with each row representing a given student and each column representing a given item/task. The table below provides an example of the scores for eight students on a 4-item test where each item was scored right/wrong (i.e., correct = 1 point & incorrect = 0 points). Notice that there are no blanks. If a student left a question blank, the student was awarded no points for this question. Thus, the student would have received a "0" for the question. If an item/task was worth more than 1 point, the number of points each student received on the item/task should be recorded. For example, for a 3-point item the possible points a student could receive might range from 0 to 2 or 1 to 3 (depending on the rubric).

Example:

Student	Item 1	Item 2	Item 3	Item 4
Kara	1	0	0	0
Adam	1	1	0	1
Claire	0	1	0	0
Jacob	1	0	0	1
Graham	1	1	0	1
Shea	1	0	1	1
Patricia	1	1	1	1
Alex	1	1	1	1

2. Calculate the total test score on the test for each student

Most often the "total test score" is determined by adding together the points received on each of the separate items/tasks on the assessment. This information needs to be added to the two-dimensional table of student scores. Thus, the sample data presented above would be modified as follows.

Example:

Student	Item 1	Item 2	Item 3	Item 4	**Total Score**
Kara	1	0	0	0	**1**
Adam	1	1	0	1	**3**
Claire	0	1	0	0	**1**
Jacob	1	0	0	1	**2**
Graham	1	1	0	1	**3**
Shea	1	0	1	1	**3**
Patricia	1	1	1	1	**4**
Alex	1	1	1	1	**4**

3. Calculate the mean for each item and for the total test score

Calculating the mean for each item/task and for the total test score is actually an intermediate step because it is the *variances* that are used in the formula—NOT the means. However, as you can see on page 3, when calculating the variance "by hand" you need to know the mean. If you are going to be using a computer program, like *Excel*, to calculate the variances you do not need to independently calculate the means. However, the mean for each item/task and for the total test score is information that is beneficial for describing student performance and is very useful information to have on file. The means for the sample data are as follows.

Example:

Student	Item 1	Item 2	Item 3	Item 4	Total Score
Kara	1	0	0	0	1
Adam	1	1	0	1	3
Claire	0	1	0	0	1
Jacob	1	0	0	1	2
Graham	1	1	0	1	3
Shea	1	0	1	1	3
Patricia	1	1	1	1	4
Alex	1	1	1	1	4
Means	**0.875**	**0.625**	**0.375**	**0.750**	**2.625**

4. Calculate the variances for each item and for the total test score

$$s^2 = \frac{\sum\left(X - \overline{X}\right)^2}{n}$$

If calculating s_i^2 (i.e., variance of the scores for a given item/task), then

Σ = the operation symbol meaning "the sum of"
X = score received by a particular student on a particular test item/task
$\overline{X}$ = mean score for a particular test item/task
n = number of students (i.e., number of scores for the test item/task)

If calculating s_t^2 (i.e., variance of the total test scores), then

X = total test score received by a particular student
$\overline{X}$ = mean total test score
n = number of students (i.e., number of total test scores)

Example: *For Item 2, with mean = 0.625*

$$s^2 = \frac{\sum\left(X - \overline{X}\right)^2}{n}$$

$$= \frac{(0-.625)^2 + (1-.625)^2 + (1-.625)^2 + (0-.625)^2 + (1-.625)^2 + (0-.625)^2 + (1-.625)^2 + (1-.625)^2}{8}$$

$$= \frac{(.3906 + .1406 + .1406 + .3906 + .1406 + .3906 + .1406 + .1406)}{8}$$

$$= \frac{1.875}{8} = 0.234$$

The variance for each of the four items and for the total scores has been added to the table below.

Student	Item 1	Item 2	Item 3	Item 4	Total Score
Kara	1	0	0	0	1
Adam	1	1	0	1	3
Claire	0	1	0	0	1
Jacob	1	0	0	1	2
Graham	1	1	0	1	3
Shea	1	0	1	1	3
Patricia	1	1	1	1	4
Alex	1	1	1	1	4
Means	0.875	0.625	0.375	0.750	2.625
Variances	**0.109**	**0.234**	**0.234**	**0.188**	**1.234**

5. Calculate coefficient alpha

$$Coefficient\ \alpha = \frac{k}{k-1}\left[1 - \frac{\sum s_i^2}{s_t^2}\right]$$

k = number of separately scored test items/tasks

Σ = the operation symbol meaning "the sum of"

s_i^2 = variance of students' scores on a particular test item/task

$\sum s_i^2$ = sum of the item variances for all test items/tasks

s_t^2 = variance of the total test scores

Example:

The variances for the sample set of data have been reproduced in the table below. These values have then been "plugged" into the formula for coefficient alpha.

	Item 1	Item 2	Item 3	Item 4	Total Score
Variances	0.109	0.234	0.234	0.188	1.234

$$\begin{aligned} Coefficient\ \alpha &= \frac{k}{k-1}\left[1 - \frac{\sum s_i^2}{s_t^2}\right] \\ &= \frac{4}{4-1}\left[1 - \frac{(0.109 + 0.234 + 0.234 + 0.188)}{1.234}\right] \\ &= \frac{4}{3}\left[1 - \frac{0.765}{1.234}\right] \\ &= 1.333(1 - 0.620) \\ &= 1.333(0.380) \\ \alpha &= 0.507 \end{aligned}$$

6. Summarizing the information

Prior to using the obtained reliability estimate, to make a determination regarding the consistency of the scores, it is useful to have a thorough understanding of the context of the assessment and the set of scores used to estimate the reliability. Thus, the following information should be summarized and available to those individuals who make decisions regarding the use of the scores from the given assessment.

Description of Assessment	
Name:	*What is the name of the assessment?*
Grade Level:	*For which grade level(s) was the assessment designed?*
Time of Year Administered:	*In which month was the assessment administered?*
Number and Type(s) of Tasks:	*What type(s) of test questions (e.g., multiple choice or constructed response) are on the assessment and how many of each type of question are there?*
Type of Scoring:	*What type(s) of scoring are required—objective, subjective, or a combination? If a rubric is used, how are score points assigned?*
Purpose:	*How are the scores from the assessment used?*
Reliability Estimate	
Type of reliability estimate:	Internal consistency based on coefficient alpha.
Reason why this type of estimate is important:	*Why is internal consistency information important and applicable to the scores from this assessment? If subjective scoring is used, what was the extent of rater consistency?*
Type of Score(s):	*What type(s) of scores were used to estimate the reliability (e.g., raw score)?*
Grade Tested & Administration Date:	*What grade level was tested and when was the assessment administered?*
Number of Students:	*What was the number of students whose scores were used for the analysis?*
Estimate:	*What was the numerical value of the reliability estimate?*

❹ Interpretation Issues: *What does the reliability estimate mean?*

Reliability estimates can be interpreted much like a correlation coefficient. They have values that can range from 0.00 up to +1.00 (on rare occasions it could be negative), and the higher the value is, the less the scores are affected by measurement errors. But reliability estimates do not necessarily have to be very high in order for us to find them acceptable. How we should interpret a reliability coefficient depends a great deal on how the scores will be used. Let's look at some examples of "internal consistency" reliability estimates.

1. **Standardized math assessment used to estimate annual student growth or select students for a program (like TAG or Title I).** Reliabilities in these situations should be in the .80-.90 (or higher) range because the scores tend to be used alone to make important instructional decisions about individuals.

2. **Classroom science assessments or textbook tests that are used to determine quarter grades.** Reliabilities for these sets of scores often are in the .40-.50 range. But ordinarily these scores are combined with other grading information, so the reliability of the scores from the single assessment doesn't need to be terribly high. No decisions are generally made on the basis of these scores alone. It's the *total* score for the grading period that determines the grade.

3. **Districtwide reading assessment used to determine the percent of students in a grade who are proficient.** These reliabilities tend to be around .70-.90 (or higher). Because the scores are being used to make decisions about a group (e.g., all eighth-grade students), the reliability need not be quite as high as when scores would be used to make decisions about individual students.

Surely, the higher the better is a good way to think about how high reliability estimates ought to be. But certain situations permit us to tolerate lower values than others. If scores from two or more assessments are to be combined into a total score, which in turn is used to make decisions, a lower level of reliability can be acceptable for the component scores that form the totals. If decisions are to be made about groups rather than individual students, lower values might be more acceptable. The more a decision rests on a single score (which rarely should be the case), the higher score reliability ought to be. That's why scores used to make college admission decisions, employment screening decisions, or decisions about professional licensure need to have particularly high reliabilities, preferably in the 0.90s.

❺ An Illustration

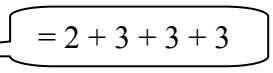
This illustration is based on a set of scores from 12 students who took a four-item test. Each item was a constructed-response task, scored by raters using a 4-point rubric (0, 1, 2, 3). Because responses to these tasks were scored subjectively, rater consistency information would be needed in addition to internal consistency information. For the purpose of this illustration, however, only internal consistency is being estimated.

Step 1: Prepare Data Set: Matrix of Student Scores on Each Item

Student	Item 1	Item 2	Item 3	Item 4
001	2	3	3	3
002	2	1	2	3
003	2	3	3	3
004	2	3	2	1
005	3	3	1	1
006	2	3	1	1
007	2	3	1	1
008	1	1	0	0
009	2	3	1	3
010	2	3	3	3
011	3	3	2	3
012	1	0	1	1

Step 2: Calculate the Total Test Score for each Student

Student	Item 1	Item 2	Item 3	Item 4	**Total Test Score**
001	2	3	3	3	**11**
002	2	1	2	3	**8**
003	2	3	3	3	**11**
004	2	3	2	1	**8**
005	3	3	1	1	**8**
006	2	3	1	1	**7**
007	2	3	1	1	**7**
008	1	1	0	0	**2**
009	2	3	1	3	**9**
010	2	3	3	3	**11**
011	3	3	2	3	**11**
012	1	0	1	1	**3**



Step 3: Calculate the Item and Total Test Score Means

Student	Item 1	Item 2	Item 3	Item 4	Total Test Score
001	3	3	3	3	11
002	2	1	2	3	8
003	3	3	3	3	11
004	2	3	2	1	8
005	3	3	1	1	8
006	2	3	1	1	7
007	2	3	1	1	7
008	1	1	0	0	2
009	2	3	1	3	9
010	2	3	3	3	11
011	3	3	2	3	11
012	1	0	1	1	3
Means	**2.00**	**2.417**	**1.667**	**1.917**	**8.00**

If using Excel:

◇	A	B	C	D	E	F
1	Student	Item 1	Item 2	Item 3	Item 4	Total Test Score
2	001	2	3	3	3	11
3	002	2	1	2	3	8
4	003	2	3	3	3	11
5	004	2	3	2	1	8
6	005	3	3	1	1	8
7	006	2	3	1	1	7
8	007	2	3	1	1	7
9	008	1	1	0	0	2
10	009	2	3	1	3	9
11	010	2	3	3	3	11
12	011	3	3	2	3	11
13	012	1	0	1	1	3
14	Means	*=average(b2:b13)*	*=average(c2:c13)*	*=average(d2:d13)*	*=average(e2:e13)*	*=average(f2:f13)*

Type the formulas into each cell corresponding to a given item and to the total test score.

Step 4: Calculate the Item and Total Test Score Variances

Sample calculation for Item 1, with mean = 2.0

Student	Item 1	$X - \overline{X}$	$(X - \overline{X})^2$
001	2	2 – 2 = 0	$0^2 = 0$
002	2	2 – 2 = 0	$0^2 = 0$
003	2	2 – 2 = 0	$0^2 = 0$
004	2	2 – 2 = 0	$0^2 = 0$
005	3	3 – 2 = 1	$1^2 = 1$
006	2	2 – 2 = 0	$0^2 = 0$
007	2	2 – 2 = 0	$0^2 = 0$
008	1	1 – 2 = -1	$(-1)^2 = 1$
009	2	2 – 2 = 0	$0^2 = 0$
010	2	2 – 2 = 0	$0^2 = 0$
011	3	3 – 2 = 1	$1^2 = 1$
012	1	1 – 2 = -1	$(-1)^2 = 1$
Σ	24	0	4

$$s^2 = \frac{\sum (X - \overline{X})^2}{n} = \frac{4}{12} = 0.333$$ = variance for Item 1

If using Excel:

◇	A	B	C	D	E	F
1	Student	Item 1	Item 2	Item 3	Item 4	Total Test Score
2	001	2	3	3	3	11
3	002	2	1	2	3	8
4	003	2	3	3	3	11
5	004	2	3	2	1	8
6	005	3	3	1	1	8
7	006	2	3	1	1	7
8	007	2	3	1	1	7
9	008	1	1	0	0	2
10	009	2	3	1	3	9
11	010	2	3	3	3	11
12	011	3	3	2	3	11
13	012	1	0	1	1	3
14	Means	2.000	2.417	1.667	1.917	8.000
15	Variances	*=varp(b2:b13)*	*=varp(c2:c13)*	*=varp(d2:d13)*	*=varp(e2:e13)*	*=varp(f2:f13)*

Type the formulas into each cell corresponding to a given item and to the total test score.

Results:

	Item 1	Item 2	Item 3	Item 4	Total Score
Means	2.000	2.417	1.667	1.917	8.000
Variances	0.333	1.076	0.889	1.243	8.333

Step 5: Calculate Coefficient Alpha

$$\alpha = \frac{k}{k-1}\left[1-\frac{\sum s_i^2}{s_t^2}\right] = \frac{4}{4-1}\left[1-\frac{(0.333+1.076+0.889+1.243)}{8.333}\right]$$

$$= \frac{4}{3}\left[1-\frac{3.541}{8.333}\right] = \frac{4}{3}[1-0.4249] = 1.333(0.5751)$$

$$\alpha = 0.767$$

If using Excel:

◇	A	B	C	D	E	F
1	Student	Item 1	Item 2	Item 3	Item 4	Total Test Score
2	001	2	3	3	3	11
3	002	2	1	2	3	8
4	003	2	3	3	3	11
5	004	2	3	2	1	8
6	005	3	3	1	1	8
7	006	2	3	1	1	7
8	007	2	3	1	1	7
9	008	1	1	0	0	2
10	009	2	3	1	3	9
11	010	2	3	3	3	11
12	011	3	3	2	3	11
13	012	1	0	1	1	3
14	Means	2.000	2.417	1.667	1.917	8.000
15	Variances	0.333	1.076	0.889	1.243	8.333
16	Coef-α	*type in formula*				

$$= \frac{k}{k-1}\left[1-\frac{\sum s_i^2}{s_t^2}\right]$$

= (4/(4-1))*(1-(sum(b15:e15)/f15))

$k = 4$ items

Step 6: Summarizing the Information

Date:	May 30, 2003
School District:	Merida Community Schools
Contact Person:	Kris Waltman

Description of Assessment

Name:	Math Assessment: Problem Solving
Grade Level:	Grade 5
Time of Year Administered:	April
Number and Type(s) of Tasks:	4 constructed-response tasks
Type of Scoring:	Responses to each task are scored subjectively; 4-point rubric (0-3)
Purpose:	Scores are used to monitor progress towards local goals; scores are reported to the community on an annual basis. Scores are <u>not</u> used to determine annual student growth, determine individual student grades, or to make promotion/retention decisions.

Reliability Estimate

Type:	Internal consistency using coefficient alpha
Reason:	Helps to understand the magnitude of measurement error associated with this sample of 4 tasks. *It does NOT help to understand errors that are due to subjective scoring—rater consistency information is needed for this.*
Type of Score:	Raw scores
Scores from:	5^{th}-grade students tested in 2003
Number of Students:	12
Estimate:	0.767

Judgment: The two primary sources of error associated with the scores from this assessment are related to content sampling and rater subjectivity associated with scoring the constructed-response tasks. Estimates of the magnitude of both of these sources of errors need to be considered when determining how confident we can be in using these scores. It is recommended that rater consistency information (e.g., percent exact agreement) be considered first—prior to making a judgment regarding errors related to content sampling (e.g., coefficient alpha). In this illustration, if the errors due to rater subjectivity were quite low we could then look at the obtained reliability estimate to determine the magnitude of errors associated with content sampling. Given that the scores from this assessment are used to monitor progress towards local goals, and are <u>not</u> used to determine annual student growth or to make promotion/retention decisions, the reliability estimate of 0.767 indicates that the errors due to content sampling are small enough that we can confidently use scores from this assessment to monitor group-level achievement towards our local goals.