Improving Validity and Reliability in STAT 2000 Assessments

by

Gregory Gerard Jansen

(Under the Direction of Jennifer J. Kaplan)

Abstract

STAT 2000 is the introductory statistics course given at the University of Georgia and is taken by 1300 students each fall and spring semester. This thesis analyzes the reliability and validity of the third and fourth tests given in the spring 2012 and fall 2012 semesters. The reliability coefficients for the fall tests, after the introduction of a new validity check that included a more direct analysis of the questions as well as the tests themselves, were compared to the reliability of the spring tests. The new validity check increased reliability for both tests, and the reliability for test four was significantly higher in the fall than for the spring. While further improvements can be made, this formal validity check can increase the reliability, improve the assessments, and the instructors' understanding of the students' understanding.

Index words: Validity, Reliability, Introductory Statistics, Educational Measurement, Statistics Education

Improving Validity and Reliability in STAT 2000 Assessments

by

Gregory Gerard Jansen

B.S., Clemson University, 2011

A Thesis Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Master of Science

Athens, Georgia

2013

Improving Validity and Reliability in STAT 2000 Assessments

by

Gregory Gerard Jansen

Approved:

Major Professors: Jennifer J. Kaplan

Committee: Stacey Neuharth-Pritchett
Jack Morse

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2013

Acknowledgments

I wish to thank my advisor, Dr. Jennifer Kaplan, and my committee members Mr. Jack Morse and Dr. Stacey Neuharth-Pritchett for their continual assistance throughout this entire process. I'd also like to thank Elizabeth Amick, Tracy Kimethu, and Jack Morse for serving as the STAT 2000 experts for my validity determinations.

Lastly, I'd like to thank my family, all of whom have believed in me from the start.

Table of Contents

List of Tables

Chapter 1

Introduction

1.1 STAT 2000

STAT 2000 is the introductory statistics course offered at the University of Georgia (UGA). It satisfies the quantitative reasoning general education requirement for undergraduate students at the UGA, and educates approximately 2800 students per academic year. From the UGA undergraduate bulletin for the academic year 2012-2013, the course includes

> the collection of data, descriptive statistics, probability, and inference. Topics include sampling methods, experiments, numerical and graphical descriptive methods, correlation and regression, contingency tables, probability concepts and distributions, confidence intervals, and hypothesis testing for means and proportions (University of Georgia, 2013).

Each fall and spring semester approximately 1300 students take the course, separated into seven lecture classes taught by three or four different lecturers. The summer sections are taught by two or three lectures, who together teach approximately 200 students. The students also attend a weekly lab led by one of nine teaching assistants from the Department

of Statistics graduate program. Due to the large number of students and teaching assistants used each semester, the course is run by a STAT 2000 coordinator to regulate the course and material taught. The coordinator creates labs and tests that are used by all of the students. In the labs, students answer questions through the online testing software WebAssign. It is also in these lab classes that students complete four examinations via WebAssign, with an optional cumulative fifth assessment available during finals week. The fifth optional exam takes the place of the final and is available for students who wish to have the opportunity to improve their grade. For each assessment all students receive the same test questions, but the tests are changed slightly for each student by having the numbers within the question and order of the questions randomized. This ensures the students receive the same test while also preventing against possible dishonest test-taking practices that are hard to track with computer-based testing. The students are given 35 to 45 minutes to complete each lab, and 50 minutes to complete each test. Each test consists of multiple-choice questions (approximately 60% of the test questions) and fill-in-the-blank calculation questions (40%).

1.2 Research Goals

The goals of this investigation are to improve the reliability and validity of the assessments given in STAT 2000. Reliability is the amount of assurance that a student's results from an exam can be replicated if the student were tested multiple times under similar conditions (Crocker and Algina, 1986). Validity analyzes the assessment for completeness and appropriateness (Miller, Linn, and Gronlund, 2013). For example, consider a math test that aims to test a student's knowledge of the order of operations. It would be incomplete if it did not include questions regarding exponents and inappropriate if it asked the student to calculate a logarithm.

The possible outcomes of analyzing this information include an investigation into the current state of validity and reliability, an exploration of possible improvements to be made, how to make these improvements (if necessary), and an improved understanding of the students' knowledge.

The last two assessments before the final exam (test 3 and test 4) given in the spring and fall 2012 semesters will be analyzed. Test 3 tests knowledge and understanding of statistical inference (confidence intervals and hypothesis testing) for one sample, while test 4 assesses understanding of inference for two samples as well as Chi-Square. The spring assessments serve as a baseline and consider the test creation process used by the STAT 2000 coordinator in previous years. The fall assessments will undergo a validity process by a committee of four individuals familiar with STAT 2000 during the test creation. The reliabilities for each test will then be compared after the assessments have been given.

The tests and the test questions will not be provided in this document. Since the questions on the tests are continually refined from year to year, placing them in this document could lead to a possible information leak. The results will still discuss different validity issues encountered and how they were overcome without stating the specific questions.

The theory behind both reliability and validity will be explored in chapter 2. A more formal explanation of the methodology and procedure of determining reliability and validity can be found in chapter 3. The results follow in chapter 4, with a summary of these results in chapter 5. The document ends with my final thoughts in chapter 6.

Chapter 2

Theory

2.1 Validity

Validity "refers to the degree to which all of the accumulated evidence supports the intended interpretation of test scores for the proposed purpose" (Hubley and Zumbo, 2011, p. 220) and is considered on a continuous scale. Furthermore, the degree of validity is not measured directly, but rather is inferred from the assessment and the goals of the test provider. Therefore, whether an assessment has high or low validity is a judgement by those familiar with the content and goals of the assessment (Miller et al., 2013).

Assessment validity can be separated into three categories: content, construct, and criterion. Content validity ensures that the questions on an assessment are representative of the objectives of the assessment. Construct validity focuses on the alignment of the assessment intent and assessment content, including underrepresentation and irrelevance of specific content areas. Criterion validiation is studied when the test user wishes to use the result to infer real-world performance in the area (Crocker and Algina, 1986). For the purposes of STAT 2000, criterion validity is not applicable; therefore, this research will not consider this aspect of validity.

2.1.1 Content Validity

Content validation requires an explicit definition of measurable objectives and a judgement (from experts) of the assessment and how well the items represent these objectives (Crocker and Algina, 1986). This can be achieved through a myriad of ways, though one popular method is by creating a table of specifications. A table of specifications organizes each question by the objective or topic (a column down the left side) and by the level of the question (across the top row).

Measurable Objectives

A measurable objective consists of a specific content item that the students are expected to know, with a description of the conditions under which this item should be known (Mager, 1975). For example, one objective from test 3 states "Given an interval, students will be able to determine the point estimate." The content item in this case is the determination of the point estimate, and the condition is when they are given an interval. A second objective from this same test is "Given data or a summary of data, students will calculate the point estimate." For the second objective, the content item (calculating a point estimate) is the same, but the condition under which the point estimate is calculated is "given data or a summary of data" rather than "given an interval." As a result there are two different objectives to account for the calculation of the point estimate. The topics and the objectives generated from these topics can be found in appendix C.

When determining content validity, a test creator must consider the following questions:

1. Should objectives be weighted based on their importance?

2. How should the experts judge the assessment?

The test creator has the option to weigh objectives equally, or he or she can choose to ask more questions about certain objectives. This decision must be made by the test creator

based on his or her belief of the importance of each objective. He or she must also take into account the goals of the assessment in making this decision.

Experts judging the assessment can do so either by matching the questions to objectives, rating the questions dichotomously (0 if it does not meet any listed objectives, 1 if it does), or by using a Likert scale, where 1 indicates a poor fit and 5 indicates an excellent fit. If a Likert scale is used, any number of options can be used depending on the test creator's needs (Crocker and Algina, 1986).

Table of Specifications

The level of understanding required by the students for each question can be shown using a table of specifications. In each of the cells of the table, a placeholder, either a tick mark or the question number, is placed to indicate a question of a given topic at a given level of understanding. There are several arrangements of levels, with Bloom's taxonomy being the best known. Bloom uses six levels of thinking in his taxonomy to categorize cognitive levels. In order from simplest to most complex, they are knowledge, comprehension, application, analysis, synthesis, and evaluation (Crocker and Algina, 1986). Explanations of these levels can be found in appendix D. In general, tables of specifications can also be used to ensure proper weighting of the objectives by providing the test creator with a visualization of which topics are covered by which questions. For examples of these tables, refer to section 4.1.1.

2.1.2 Construct Validity

Constructs are hypothetical concepts that aim to explain our understanding of latent variables. They underlie what is directly measured, and what they are truly attempting to measure must be inferred from what can be measured. A test developer uses previously developed items, research, and/or other tools to determine how measurable information can be related in support of a construct. The observer then develops an instrument to measure

their construct. This instrument can be a checklist, a rating system, or a series of questions. These are all directly tangible behaviors or performances for the observer to use in his or her judgement of the construct. In classes such as STAT 2000, or most other classes throughout all levels of education, this instrument is an achievement test (Crocker and Algina, 1986).

Validity then becomes the assurance that the test is measuring the intended construct. The items on a given test should strive to assess all facets of the construct, and only those facets of the construct. Underrepresentation of content is a construct violation because it does not encompass all possible dimensions of the construct. Furthermore, irrelevant information also reduces construct validity. This extraneous information can distract an examinee from the important areas of the question, thus affecting his or her performance on the assessment (Crocker and Algina, 1986).

2.2 Reliability

Reliability is a prerequisite for validity, but the presence of reliability is not enough to ensure validity. Unlike validity, reliability is measured directly from the numerical assessment outcomes. Reliability measures the amount of error due to measurement and determines the consistency of the test if it were given to the same student many times. Therefore, students who take assessments with high reliability can expect similar results each time. Reliability is viewed as a correlation between two tests or two test forms, and is numerically represented as a number between 0 and 1 just as correlation is. The value of this reliability is called a reliability coefficient (Miller et al., 2013).

There is no hard and fast determination for "good" reliability, but higher reliability values are preferred. Teacher-made tests generally have reliability coefficients of about 0.70 (Miller et al., 2013; Neukrug and Fawcett, 2006). These tests will usually have lower reliability coefficients than nationwide standardized tests, which tend to have reliability coefficients

above 0.90. This is partially due to the amount of debate and consideration test questions receive for inclusion in nationwide tests, but also a result of the length of these assessments (Neukrug and Fawcett, 2006).

One of the main types of reliability is Internal-Consistency Reliability, used to look for consistency of items within one assessment. Methods to measure this reliability include the Split-Half (or Spearman-Brown) formula, the Kuder-Richardson 20 formula, and Cronbach's α. There is also Alternate Form Reliability which measures the consistency among multiple forms of a test, and Stability Reliability which employs a test-retest method (Miller et al., 2013). For the purposes of this research only Internal-Consistency Reliability will be considered. Alternate Form Reliability is not applicable for this research because the goal is not to compare one test form to another. Due to the setup of STAT 2000 as well as time constraints, there are no pre- or post-tests given that would allow for Stability Reliability analysis and therefore Stability Reliability is omitted from this investigation as well.

2.2.1 The True Score Model

British psychologist Charles Spearman modeled test scores as the sum of two components, true score and error, with the equation

$$X = T + E \tag{2.1}$$

where X represents the observed test score, T represents the true score, and E symbolizes random error. The true score is considered to be a measure of the actual knowledge of a student. It is the mean (or expected value) of all scores the student receives on an infinite number of the same test. Therefore,

$$E(X_i) = T_i \tag{2.2}$$

for a student i. For example, assume there is a test that a student will take. If we have the ability to test the student an infinite number of times on a selected topic and we graph the number of questions answered correctly, the expectation is that the student's true score will occur more often than any other score. However, due to time, test fatigue, and a myriad of other reasons we do not have the ability to test students on the same assessment repeatedly. Because we have only one observation of the student's test score, the true score is a number that is unobtainable (Crocker and Algina, 1986).

Error is a consequence of many possibilities; everything from selection factors (guessing, incorrect markings) to outside circumstances (the student's health on that given day, the testing environment, etc.) contributes to an observed score different from the true score. By rearranging the True Score Model Equation, error is defined as

$$E = X - T \tag{2.3}$$

This model has three assumptions

1. the expected value for the errors is zero
2. the true scores and error scores for a population of students are independent
3. the error of one installment of the test is independent of any other installments

(Crocker and Algina, 1986).

2.2.2 Covariance and Reliability

Parallel Forms

If there are two parallel (equivalent) forms of a test (X and X') given to the same student, we can determine the covariance between the two tests. By the definition of covariance,

$\rho_{XX'} = \dfrac{\sigma(X, X')}{\sigma(X)\sigma(X')}$.

$$\begin{aligned} \rho_{XX'} &= \frac{\sigma(X, X')}{\sigma(X)\sigma(X')} && (2.4) \\ &= \frac{\sigma(T + E, T' + E')}{\sigma(X)\sigma(X')} && (2.5) \\ &= \frac{\sigma(T, T')}{\sigma(X)\sigma(X')} && (2.6) \end{aligned}$$

Because $T \equiv T'$ and $X \equiv X'$ due to the parallelism of the tests, then

$$\begin{aligned} \rho_{XX'} &= \frac{\sigma^2(T)}{\sigma(X)\sigma(X)} && (2.7) \\ &= \frac{\sigma_T^2}{\sigma_X^2} && (2.8) \end{aligned}$$

(Lord and Novick, 1968)

Mathematically, the reliability of a test is the squared correlation between the true score (T) and the observed score (X). Define this as ρ_{XT}^2. By the properties of correlation,

$$\rho_{XT}^2 = (\rho_{XT})^2 = \left(\frac{\sigma_{XT}}{\sigma_X \sigma_T}\right)^2$$

σ_{XT} is the covariance of X and T. By the derivation in Appendix A.1, we know that $\sigma_{XT} = \sigma_T^2$, and hence

$$\rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_X^2} \qquad (2.9)$$

(Lord and Novick, 1968)

By combining equations 2.8 and 2.9, we can determine

(a) $\rho_{XX'} = \rho_{XT}^2$, and

(b) because ρ_{XT}^2 is a reliability coefficient, $\rho_{XX'}$ is also a reliability coefficient

ρ^2_{XT} cannot be calculated directly due to the σ^2_T in its numerator. As stated previously, the true score cannot be calculated directly, and thus the variance of the true score (σ^2_T) also cannot be calculated directly. However, because $\rho_{XX'}$ is equivalent to ρ^2_{XT}, and $\rho_{XX'}$ is a reliability coefficient, we can calculate a reliability coefficient using information from both test forms.

Single Form Tests

The use of $\rho_{XX'}$ requires students to take two forms of a test which is not feasible for many classes, but because tests are made up of individual scores summed together, we can think of one test score as the sum of individual question scores. Thus, the observed score is the sum of the observed score of each question. In general, if we have equally-weighted questions, we can mathematically represent X, T, and E as

$$X = \sum_{i=1}^{n} Y_i \tag{2.10}$$

$$T = \sum_{i=1}^{n} T_i \tag{2.11}$$

$$E = \sum_{i=1}^{n} E_i \tag{2.12}$$

where Y is the score of a question of the test, the subscript i represents the specific questions, and n refers to the number of questions that make up the assessment (Lord and Novick, 1968).

2.2.3 The Spearman-Brown Formula and the Split-Half Methods

If individual measurements are parallel (i.e. $Y_1 = Y_2 = \cdots$), then the variances and covariances of each of these measurements are equal. Thus,

$$\sigma_X^2 = \sigma^2\left(\sum_{i=1}^{n} Y_i\right) \tag{2.13}$$

$$= \sum_{i=1}^{n}\sigma^2(Y_i) + \sum_{\substack{i=1 \\ i\neq j}}^{n}\sum_{j=1}^{n}\sigma(Y_i, Y_j) \tag{2.14}$$

$$= n\sigma_Y^2 + n(n-1)\sigma(Y_i, Y_j) \tag{2.15}$$

$$= n\sigma_Y^2 + n(n-1)\sigma_Y^2\rho_{YY'} \tag{2.16}$$

$$= n\sigma_Y^2(1 + (n-1)\rho_{YY'}) \tag{2.17}$$

and

$$\sigma_T^2 = \sigma^2\left(\sum_{i=1}^{n} T_i\right) \tag{2.18}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n}\sigma(T_i, T_j) \tag{2.19}$$

Because individual measurements are parallel, $T_i = T_j = T_1$.

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma(T_i, T_j) = n\sum_{i=1}^{n}\sigma^2(T_1) \tag{2.20}$$

$$= n(n\sigma^2(T_1)) \tag{2.21}$$

$$= n^2\sigma^2(T_1) \tag{2.22}$$

Using equation 2.8, substituting the values for σ_X^2 and σ_T^2 gives us

$$\rho_{XT} = \frac{n\rho_{YY'}}{1 + (n-1)\rho_{YY'}} \tag{2.23}$$

This result is the Spearman-Brown formula, and says that if we have parallel components (Y and Y'), we can use the correlation of scores between these two components of the assessment to find the reliability of the overall assessment (Lord and Novick, 1968).

The formula does not specify how to create two components of a test. Thus, split-half methods are typically used to separate the test into two halves. The goal, to satisfy the assumption used to find the Spearman-Brown formula, is to create two halves that are as parallel as possible. Possible methods are random assignment, even-odd separation, and "matched" sub-tests, in which items with similar content are separated into each of the two sub-tests. Each individual's score is calculated for each half, and the correlation between all individuals' sub-test 1 vs. sub-test 2 is calculated. This correlation is the value for $\rho_{YY'}$ in equation 2.23 (Crocker and Algina, 1986).

2.2.4 The Kuder-Richardson 20 Formula

The split-half methods have one glaring pitfall to their calculation: there is no way to obtain a unique value. Because there are multiple methods to creating sub-tests, each method will yield a different value for reliability. Furthermore, there is no way to ensure two halves that are exactly parallel (Kuder and Richardson, 1937)

Kuder and Richardson (1937) found that if the assessment has n items, then

$$\sigma_X^2 = \left(\sum_{i=1}^{n}\sqrt{p_i q_i}\right)^2 \rho_{YY'} - \sum_{i=1}^{n} p_i q_i \rho_{YY'} + \sum_{i=1}^{n} p_i q_i$$

This formula rearranged yields $\rho_{YY'} = \frac{\sigma_X^2 - \sum_{i=1}^{n} p_i q_i}{(n-1)\sum_{i=1}^{n} p_i q_i}$. Substituting this into the Spearman-Brown Formula (equation 2.23) allows us to write ρ_{XT} as

$$\rho_{XT} = \frac{n}{n-1}\left(1 - \frac{\sum_{i=1}^{n} p_i q_i}{\sigma_X^2}\right) \tag{2.24}$$

the derivation of which can be found in Appendix A.2. Kuder and Richardson show that this formula is minimally computationally intensive, but has nearly exact reliability measurements (within 0.001) as compared to the more rigorous equations that they derived earlier in their paper (Kuder and Richardson, 1937). This formula is named the Kuder-Richardson (or KR) 20 formula because it is the 20^{th} equation in their article (Crocker and Algina, 1986). Henceforth, it will be referred to as r_{20}.

2.2.5 Cronbach's α

As with the Spearman-Brown Formula, the Kuder-Richardson 20 has a major shortfall in that it only works for items scored dichotomously. In his paper describing his calculation of α, Cronbach rewrote r_{20} almost immediately as

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_{i=1}^{n} \sigma^2(Y_i)}{\sigma_X^2}\right) \tag{2.25}$$

(Cronbach, 1951). In Cronbach's formula, $\sum_{i=1}^{n} p_i q_i$ is changed to $\sum_{i=1}^{n} Y_i$. In the r_{20} formula, $p_i q_i$ is the variance of the scores for question i (Kuder and Richardson, 1937). Thus, $\sum_{i=1}^{n} p_i q_i$ is the sum of the variances for each question. Cronbach determined that each question has

variance whether or not it is scored dichotomously, and therefore generalized the KR 20 into one that will work across a range of possible scores for the question. He notes that the formula reduces to the KR 20 when the items are scored dichotomously (Cronbach, 1951).

2.2.6 Standard Error of Measurement

The standard error of measurement (SEM) is the average standard deviation of all individual students' error distributions. Each individual error is indeterminable due to the single observation, but we are able to determine the group SEM using the True Score Model and the corresponding assumptions.

$$X = T + E \tag{2.26}$$

$$\sigma^2(X) = \sigma^2(T + E) \tag{2.27}$$

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2 + 2\sigma_{TE} \tag{2.28}$$

Due to the independence of test scores and error scores (the second assumption in section 2.2.1) $\sigma_{TE} = 0$, and $\sigma_X^2 = \sigma_T^2 + \sigma_E^2$. Taking this one step further,

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2 \tag{2.29}$$

$$1 = \frac{\sigma_T^2}{\sigma_X^2} + \frac{\sigma_E^2}{\sigma_X^2} \tag{2.30}$$

Using the derivation from section 2.2.2 that $\frac{\sigma_T^2}{\sigma_X^2} = \rho_{XX'}$, we can simplify equation 2.30 and reorganize the result into our final form

$$\sigma_E = \sigma_X \sqrt{1 - \rho_{XX'}} \tag{2.31}$$

(Crocker and Algina, 1986; Lord and Novick, 1968)

The calculated SEM is the value of the standard error used for all student scores in the population, and is the error used when an instructor is concerned with the difference between the true score and observed score. This is the standard error used when creating a confidence interval to estimate a student's true score (Crocker and Algina, 1986).

2.3 Generating a Hypothesis Test

Feldt (1969) set out to create a hypothesis test that would allow for the comparison of two reliability coefficients. In an earlier paper, he found that the sampling distribution of r_{20} follows an F-distribution with degrees of freedom $(n-1)$ and $(n-1)(k-1)$, where n is the number of observations (or in our case, students), and k is the number of items on the assessment (Feldt, 1965). The test statistic follows the form

$$F_{(n-1),(n-1)(k-1)} = \frac{1 - \rho_{20}}{1 - r_{20}} \tag{2.32}$$

where ρ_{20} is the hypothesized population value for the Kuder-Richardson 20 formula.
Looking to test the hypothesis $H_0 : \rho_1 = \rho_2$, Feldt expanded on his previous work, and denoted a test statistic W to represent the sampling distribution required to test the hypothesis, where

$$W = F_1 \cdot \frac{1}{F_2} \tag{2.33}$$

$$= \left[\frac{1-\rho_1}{1-r_1}\right]\left[\frac{1-r_2}{1-\rho_2}\right] \tag{2.34}$$

Henceforth, for simplicity of notation, $F_2' = \frac{1}{F_2}$. Because the null hypothesis denotes that $\rho_1 = \rho_2$, we can simplify W into

$$W = \frac{1-r_2}{1-r_1} \tag{2.35}$$

W follows an F-distribution with degrees of freedom ν_1 and ν_2. To calculate these degrees of freedom, note that F_1 has degrees of freedom $df_1 = (n_1 - 1)$ and $df_2 = (n_1 - 1)(k_1 - 1)$ and F_2' has degrees of freedom $df_3 = (n_2 - 1)(k_2 - 1)$ and $df_4 = (n_2 - 1)$. Feldt used moments of these F-distributions to find

$$A = \frac{df_4}{df_4 - 2} \cdot \frac{df_2}{df_2 - 2} \tag{2.36}$$

$$= \frac{\nu_2}{\nu_2 - 2} \tag{2.37}$$

$$B = \frac{(df_1 + 2)df_4^2}{(df_4 - 2)(df_4 - 4)df_1} \cdot \frac{(df_3 + 2)df_2^2}{(df_2 - 2)(df_2 - 4)df_3} \tag{2.38}$$

$$= \frac{(\nu_1 + 2)\nu_2^2}{(\nu_2 - 2)(\nu_2 - 4)\nu_1} \tag{2.39}$$

Solving this system for ν_1 and ν_2, the resultant degrees of freedom for W are

$$\nu_1 = \frac{2A^2}{2B - AB - A^2} \tag{2.40}$$

$$\nu_2 = \frac{2A}{A - 1} \tag{2.41}$$

(Feldt, 1969).

Chapter 3

Methodology

There are four tests given to students in STAT 2000 over the course of the semester, with an optional fifth exam given during finals week. The focus of this research is on the third and fourth tests given in the spring of 2012 and the fall of 2012.

3.1 Differences between test creation

3.1.1 Spring 2012

In the spring of 2012 tests 3 and 4 consisted of questions chosen from previous tests (approximately 80%), and new questions from textbooks and other sources (20%). Using a list of topics for each test, a question is selected for inclusion that covers the topic with the goal of having one question per topic. These questions did not go through a formal validity check, but the STAT 2000 lecturers during this semester did review the questions to determine if any changes needed to be made based on previous student responses, difficulty, clarity, and importance to the course.

3.1.2 Fall 2012

The test questions for the fall 2012 assessments were based on the measurable objectives. Questions similar to previous test questions as well as examples from the notes were matched to an objective and used in these tests. The measurable objectives were created out of the previously-used topics, and were checked and approved by the current coordinator after revisions. Twenty-five questions were then created, the coordinator's intended length of the assessment, based on these objectives.

3.2 Checking Validity

3.2.1 Content Validity

From earlier in this document, the questions that must be addressed by a test writer are

1. Should objectives be weighted based on their importance?

2. How should the experts judge the assessment?

For these tests, the objectives were weighted equally for the reason that the more important topics had more objectives related to them. Therefore, by creating one question per objective, the test was already weighted more toward the important topics.

The validity committee for the spring 2012 tests consisted of the STAT 2000 Coordinator, two graduate students, one of whom has four semesters of TA experience and the other a former teacher familiar with introductory statistics, and the author, who has three semesters of experience as a teaching assistant and two semesters of experience as an instructor for one of the sections. Because the author created the tests for fall 2012 he did not analyze the tests for validity, but received feedback from the other three coders mentioned previously. The questions asked to the committee members were

- Is the content being assessed by the question clear?
- Is the question important to assess in an introductory statistics class?

The validity sheets for the fall included an additional question of

- Is the question assessing just one objective?

because questions were created based on objectives for those tests, and not for the spring tests.

The committee of experts was asked to rate each of these questions for each of the test items using a Likert scale of 1 (complete disagreement) to 5 (complete agreement). For the overall analysis the mean of the values is reported. The five-point Likert scale was used because it allowed for individual interpretation of the scores on a scale, rather than a dichotomous "yes" or "no." Using a dichotomous code would not allow for questions that are only slightly unclear due to word choice or syntax, but still have content that is appropriate for the assessment. The use of a five-point Likert scale allows for degrees of validity, as mentioned in chapter 2. It is also a scale very familiar to the coders and would not require much explanation to understand other than an explanation of the meaning of each endpoint.

The full spreadsheets that summarize the responses to these questions are available in section 4.1.2.

3.2.2 Construct Validity

The construct used for tests 3 and 4 was statistical inference. Statistical inference refers to the process of drawing conclusions about populations from sample data subject to variation. In many introductory statistics courses, students begin their exploration of inference by estimating a population value from a sample of data by creating a confidence interval. They follow with conducting hypothesis tests to test a claim about the population value.

These two ideas comprise test 3 in STAT 2000. Test 4 continues with the remaining introductory statistics inference topics: confidence intervals and hypothesis tests to compare two populations, and categorical data analysis using Chi-Square.

To determine construct validity, two open-ended questions were asked about the assessments as a whole:

- Does the assessment cover all intended topics?
- Does the assessment weigh one topic too much or too little?

and one yes-or-no question:

- Do any questions contain any irrelevant or unfamiliar information that may detract students?

The first two questions ensure that there is no underrepresentation or overrepresentation of topics under the construct, and that each of the topics is relevant to the construct. The third question ensures that all of the information stated in the question is useful and there is no information that will distract students from answering the question given. While extraneous information may be purposefully used to separate those less versed in the construct, this is not one of the goals for STAT 2000 assessments.

Again, the full spreadsheets that discuss the responses to these questions are available in section 4.1.2.

3.3 Selecting a Reliability Coefficient

Because there is no partial credit on these tests, the value of r_{20} is equal to the value of Cronbach's α, and thus either could be used to calculate and report reliability. Reporting both would be redundant, and therefore I will report Cronbach's α without any loss of generality or interpretation.

Chapter 4

Results

4.1 Validity Results

4.1.1 Tables of Specifications

Because the goal was to create one question per objective, the topics, rather than the objectives, are listed along the left-most column. For the levels of thinking, the six levels from Bloom's taxonomy were used. For the spring tests, the tables are described based on the tests taken by the students. The tables for the fall tests are described before revisions, and do not reflect the tests the students took.

Below are the tables of specifications for each of the four tests. The number-letter combination in the table refers to the question number and part that fits the given topic and level of understanding required by the student. A description of the acronyms used can be found in Appendix B.

A number in parentheses indicates a question that required the use of technology to answer correctly. Because the use of technology shows up as separate objectives, the question number is repeated in the row that indicates the context of the question.

Table 4.1: Table of Specifications, Spring 2012 Test 3

Topic	Knowledge	Comprehension	Application	Analysis	Synthesis	Evaluation
Creating CIs	(2b), 5a, 7b			2c		
Finding PE, MOE, Confidence Letter	3a	2a, 3b, (8a)				
Using Technology	2b, 8a					
Assumptions	3c, 9b					
Sample Size Calculations	4a			1a, 1b, 4b		
Errors	10b					
P-Values, Conclusions		6a, 5b	13a			
Hypotheses	11a					
Test Statistic	9a, 12a		12b	10a		
Hypothesis Test/CI Relations			7a		12c	
Response Variables						

Table 4.2: Table of Specifications, Fall 2012 Test 3

Topic	Knowledge	Comprehension	Application	Analysis	Synthesis	Evaluation
Creating CIs	5a, (7a)		5b			
Finding PE, MOE, Confidence Letter	4a, 6a	(4b)	8a			
Using Technology	4b, 7a					
Assumptions	3b	13a				
Sample Size Calculations	9a			9b		
Errors	11b, 12b					
P-Values, Conclusions		2b, 3a, 4c, 11a	12c			12a
Hypotheses	10a					
Test Statistic			10b			
Hypothesis Test/CI Relations			13c			
Response Variables	2a	1a				

Table 4.3: Table of Specifications, Spring 2012 Test 4

Topic	Knowledge	Comprehension	Application	Analysis	Synthesis	Evaluation
Variables						
Experiments and Studies	11b	10b	9a			
Creating CIs						
Finding PE, MOE, Confidence Letter	2a, 15b					
Using Technology	15a					
P-Values, Conclusions		1a, 2b, 6b, 7a, (15a)	1b	4a	8a	12a
Hypotheses	11a, 13a					
Test Statistic	3a					
Hypothesis Test/CI Relations						
Chi-Square Expected Counts	5b, 6a					
Degrees of Freedom	1c	5a				
Assumptions						
Errors	14a					

Table 4.4: Table of Specifications, Fall 2012 Test 4

Topic	Knowledge	Comprehension	Application	Analysis	Synthesis	Evaluation
Variables	3a, 4a, 5a					
Experiments and Studies	1a, 3b	5b				
Creating CIs	(13a)					
Finding PE, MOE, Confidence Letter	2a	6a		6b		
Using Technology	13a					
P-Values, Conclusions		2b, 5c, 9b, 11c, 14a	8a		4b	
Hypotheses	7a					
Test Statistic	7b, 11b			12a		
Hypothesis Test/CI Relations		2c				
Chi-Square Expected Counts		9a				
Degrees of Freedom	11a					
Assumptions	10a					
Errors						

4.1.2 Committee Results

For the spring tests the validity check was done retroactively. Therefore the validity information is described based on the actual tests the students took and the results did not inform the editing of the tests. For the fall tests there was the benefit of time between test creation and administration. This time was used to review the validity responses with the coordinator to fix any issues that arose. The results described for the tests in the fall refer to the committee members' responses before the test was administered. In addition, the coordinator performed a second round of validity check before administering the test to the students.

Spring 2012, Test 3

This test had the weakest content validity of any of the four, but no questions received an average score lower than 4.5 for clarity or for importance. Three questions (1a, 1b, and 4b) did not get a perfect score for clarity, while eight questions were potentially unimportant to assess in an introductory statistics setting. In the future, the coordinator should have a discussion with the other instructors over whether the content covered by these eight questions is important to assess in the course.

This test also had the most issues regarding the amount of each topic covered. Mentions from the committee were made that confidence intervals, sample size calculations, and test statistic calculations were overrepresented, while p-values and errors were underrepresented in this test.

Further exploring construct validity, question 4 on this test was considered potentially confusing due to the terminology ("tax assessor") used in the question. Students may not be aware of what a tax assessor is, and their focus in trying to determine what a tax assessor does could detract from the sample size calculation in the question. This idea could also be applied to question 10 with the use of the phrase "BCS system" in the question.

Table 4.5: STAT 2000 Test 3 Validity Form Summary, Spring 2012

Rate on a scale of 1 (fully incomplete) to 5 (fully complete)

Question	Part	Is the content being assessed by the question clear?	Is the question important to assess in an introductory statistics class?
1	a	4.5	4.5
	b	4.5	4.5
2	a	5	5
	b	5	4.75
	c	5	5
3	a	5	5
	b	5	5
	c	5	5
4	a	5	4.5
	b	4.5	4.5
5	a	5	5
	b	5	5
6	a	5	5
7	a	5	4.75
	b	5	5
8	a	5	4.75
9	a	5	5
	b	5	5
10	a	5	5
	b	5	5
11	a	5	5
12	a	5	5
	b	5	4.5
	c	5	5
13	a	5	5

Topics Overrepresented: Confidence Intervals; Sample Size; Test Statistics
Topics Underrepresented: P-values; Error vs. Significance Level

Spring 2012, Test 4

For this test, only one question received a clarity score lower than a 5 (4.5), and four questions may not be applicable for an introductory statistics course. Once again, no average score for any of the headings was less than a 4.5.

Test 4 overrepresented errors, p-values, and conclusions (because errors are covered in the third test, they are not assessed in the fourth test). The test underrepresents assumptions (for two-sample hypothesis tests as well as chi-square tests), determining experimental units, hypothesis test and confidence interval relations, and creating confidence intervals.

The setup of the first question used the phrase "arthroscopic surgery" which may not be a term known to all students. In addition, the mean scores given in this question did not have units, which could cause confusion in interpreting the numbers. Numbers 3 and 7 both contained irrelevant information. Question 15a contained a grammatical error. The question reads "...in samples such as this *one*." (emphasis added), but the question itself asks about a two-sample t-test, which by definition must have two samples.

Table 4.6: STAT 2000 Test 4 Validity Form Summary, Spring 2012

Rate on a scale of 1 (fully incomplete) to 5 (fully complete)

Question	Part	Is the content being assessed by the question clear?	Is the question important to assess in an introductory statistics class?
1	a	5	5
	b	5	4.75
	c	5	5
2	a	5	5
	b	5	5
3	a	5	4.5
	b	5	5
4	a	5	5
5	a	4.5	5
	b	5	5
6	a	5	5
	b	5	5
7	a	5	5
8	a	5	5
9	a	5	5
	b	5	5
10	a	5	5
	b	5	5
11	a	5	5
	b	5	5
12	a	5	5
13	a	5	5
14	a	5	4.5
15	a	5	5
	b	5	4.5

Topics Overrepresented: Errors (Test 3 Material); P-values and Conclusions
Topics Underrepresented: Assumptions; Experimental Units; Hypothesis Test and CI Relations; Creating CIs

Fall 2012, Test 3

One question on this test received a score lower than a 5 for clarity (4.33), and three questions highlighted an issue of measuring more than one objective within the same question. Questions 4a, 5a, and 9a highlighted the need to adjust the wording of the question to ensure it was only testing one objective. For question 11a, the preamble of the problem was reworded to more specifically dictate the goal of the question.

All questions were considered to be important to assess in STAT 2000. In addition, it was uniformly agreed that all topics and objectives were covered with correct weight.

The third question made reference to an American president's approval rating, but did not mention that it was taken by poll. Especially for students who did not grow up in the United States, the word "poll" could distract them needlessly if they do not know approval ratings are determined by polling Americans. The terminology in question 6 implies the data given is from the population, and not from the sample, which affects the students' approach to the problem. Number 12 contains much more information than is necessary to assess a student's understanding of the objective.

Table 4.7: STAT 2000 Test 3 Validity Form Summary, Fall 2012

Rate on a scale of 1 (fully incomplete) to 5 (fully complete)

Question	Part	Is the content being assessed by the question clear?	Is the question important to assess in an introductory statistics class?	Is the question assessing just one objective?
1	a	5	5	5
2	a	5	5	5
	b	5	5	5
3	a	5	5	5
	b	5	5	5
4	a	5	5	4
	b	5	5	5
	c	5	5	5
5	a	5	5	4.33
	b	5	5	5
6	a	5	5	5
7	a	5	5	5
8	a	5	5	5
9	a	5	5	5
	b	5	5	4.33
10	a	5	5	5
	b	5	5	5
11	a	4.33	5	5
	b	5	5	5
12	a	5	5	5
	b	5	5	5
	c	5	5	5
13	a	5	5	5
	b	5	5	5
	c	5	5	5

Topics Overrepresented: *(None)*
Topics Underrepresented: *(None)*

Fall 2012, Test 4

All questions were considered to be clear for this test. One question assessed multiple objectives, and two questions should be reconsidered for use in the test. 7b and 10a were deleted completely because the question asked about a topic that is traditionally not tested in STAT 2000. They were replaced with questions from the STAT 2000 bank that covered the respective objectives. For 9a, a discussion with the coordinator determined that the issue with the question was misplaced, and should have been with construct validity (described below), not content validity. Just as with test 3, all parties agreed that the topics and objectives were covered with correct weight.

Questions 3, 7, and 9 all contained irrelevant information not needed to answer the question. Question 6 compares two locations to their relative distance from the ocean, but does not specify which one is closer to the ocean, which may be misleading to students. The setup to the questions contained in numbers 10 and 11 could be reworded, although the questions themselves were clear.

Table 4.8: STAT 2000 Test 4 Validity Form Summary, Fall 2012

Rate on a scale of 1 (fully incomplete) to 5 (fully complete)

Question	Part	Is the content being assessed by the question clear?	Is the question important to assess in an introductory statistics class?	Is the question assessing just one objective?
1	a	5	5	5
2	a	5	5	5
	b	5	5	5
	c	5	5	5
	d	5	5	5
3	a	5	5	5
	b	5	5	5
4	a	5	5	5
5	a	5	5	5
	b	5	5	5
	c	5	5	5
6	a	5	5	5
	b	5	5	5
7	a	5	5	5
	b	5	3.67	5
	c	5	5	5
8	a	5	5	5
9	a	5	5	4.67
	b	5	5	5
10	a	5	4.33	5
11	a	5	5	5
	b	5	5	5
12	a	5	5	5
13	a	5	5	5
	b	5	5	5

Topics Overrepresented: *(None)*
Topics Underrepresented: *(None)*

4.2 Quantitative Results

The data gathered from each of the four tests consisted of each student's score for each question part, as well as their overall score on the test.

4.2.1 Data Summary

Simple summary statistics for the overall scores are provided in the table below.

Table 4.9: Summary Statistics for Final Scores

		Mean	Standard Deviation	Median	n
Spring 2012	Test 3	74.89	14.97	76	1151
	Test 4	80.89	13.79	84	1147
Fall 2012	Test 3	74.92	15.65	76	1278
	Test 4	80.47	14.51	84	1268

The mean and median for both tests were approximately equal between the two semesters. The standard deviation, however, was higher for both tests in the fall than in the spring. There were also more students enrolled in the class in the fall than in the spring. All four tests had 25 items, each worth 4 points.

4.2.2 Reliability Coefficients

A table of the reliability coefficients are below. The reliability coefficients for the fall reflect the assessment given to the students after two rounds of validity check.

Table 4.10: Cronbach's α Values

	Test 3	Test 4
Spring 2012	0.737	0.741
Fall 2012	0.754	0.771

4.2.3 Standard Error of Measurement

The standard error of measurement for each assessment is given in the table below:

Table 4.11: Standard Error of Measurement

	Test 3	Test 4
Spring 2012	7.683	7.018
Fall 2012	7.754	6.940

4.2.4 Testing the Hypotheses

The hypothesis test will have hypotheses

$H_0 : \rho_1 = \rho_2$

$H_A : \rho_1 < \rho_2$

discussed in section 2.3. For clarity moving forward, $\rho_1 = \rho_{SP}$ and $\rho_2 = \rho_{FA}$ represent the spring 2012 and fall 2012 tests, respectively. The hypothesis test will be done twice, once each for test 3 and test 4. Accordingly, the test statistic is $W = \frac{1-r_{FA}}{1-r_{SP}}$.

Test 3

For test 3, the individual degrees of freedom are shown in the table below:

Table 4.12: Degrees of Freedom for Test 3

df_1	df_2	df_3	df_4
1150	27,600	30,648	1277

Plugging these values into the formulas for A and B (equations 2.37 and 2.39) yield

$$A = 1.0016 \tag{4.1}$$

$$B = 1.0067 \tag{4.2}$$

Plugging these values into equations 2.40 and 2.41 yield the degrees of freedom of the F-distribution followed by the test statistic, W.

$$\nu_1 = 1109 \tag{4.3}$$

$$\nu_2 = 1221 \tag{4.4}$$

The test statistic has the value

$$W_3 = \frac{1 - r_{FA}}{1 - r_{SP}} \tag{4.5}$$

$$= \frac{1 - 0.754}{1 - 0.737} \tag{4.6}$$

$$= 0.935 \tag{4.7}$$

The p-value of this test statistic is 0.126.

Test 4

Following the same method as in the previous section,

Table 4.13: Degrees of Freedom for Test 4

df_1	df_2	df_3	df_4
1146	27,504	30,408	1267

$$A = 1.0017 \tag{4.8}$$

$$B = 1.0068 \tag{4.9}$$

$$\nu_1 = 1104 \tag{4.10}$$

$$\nu_2 = 1211 \tag{4.11}$$

$$W_4 = \frac{1 - r_{FA}}{1 - r_{SP}} \tag{4.12}$$

$$= \frac{1 - 0.771}{1 - 0.741} \tag{4.13}$$

$$= 0.884 \tag{4.14}$$

The p-value of this test statistic is 0.018.

Chapter 5

Results Summary

5.1 Validity

Test 3 in the spring had the most concerns regarding clarity of the questions written, and tests 3 and 4 in the spring had the most concerns regarding relevance in introductory statistics. These two tests also had the most disparty when determining over or underrepresentation of the subjects tested. This means the construct (statistical inference) is not fully represented by the tests given. All four tests had concerns with construct validity on the individual question level due to the terminology used, irrelevant information, and grammatical errors, all of which could misrepresent the items that make up the construct.

The validation procedure was much more formal in the fall than in the spring with the use of ratings to analyze the tests before they were given, rather than the formal analysis done post facto. But given that the spring test questions have been informally studied and altered for many years, the content validity of the spring questions is similar to the content validity of the fall after both rounds of validity check. However, in terms of the underrepresentation of topics construct validity was not satisfied in the spring as it was in the fall.

Analyzing both content and construct validity formally before the time of assessment can ensure validation of the content areas and construct of statistical inference. As stated previously, there were many topics in the spring that were either over or underrepresented, and with a more formal validity check (including the use of a table of sepcifications) this issue can be rectified fairly easily. Furthermore, questions with irrelevance of information or unfamiliar terminology can be amended before administering the test.

5.2 Reliability

As stated previously, teacher-made tests typically have a reliability coefficient of approximately 0.70. All four assessments were above this value, with both of the assessments in the fall of 2012 above 0.75. The fall 2012 reliability increased by 1.7 percentage points for test 3 and three percentage points for test 4 over the spring 2012 reliability. From the spring to the fall, the standard error of measurement increased a small amount for test 3 and decreased a small amount for test 4. The hypothesis tests indicate that the reliability increase for test 3 was not very significant, but the p-value for the test 4 increase intimates strong evidence against the null hypothesis.

Chapter 6

Final Thoughts

The STAT 2000 coordinator and instructors are already doing a good job of ensuring validity and reliability on their assessments. The reliabilities of the two exams investigated in the spring are 0.737 (test 3) and 0.741 (test 4), which are both adequate coefficient values for reliability for the purposes of their assessments. Their own validity process received high comments for the examination questions. Work still needs to be done in the construct validity areas of underrepresentation, unfamiliar terminology, and irrelevant information.

Checking validity formally slightly improved reliability, probably because it was not much more intensive than what is already done. Nevertheless, because reliability did improve and the validity check is not much more intensive, I would recommend performing a validity check in this manner. It not only rates the questions (and thus puts its clarity and importance on a scale), but also considers the possibility of underrepresentation of certain topics in the test. Underrepresentation can also be avoided by using a table of specifications, which gives the added benefit of analyzing each question for the level of thinking it requires from the student. Including questions from all levels of thinking allows for the instructors to have a better understanding of the students' understanding.

Creating questions off of the measurable objectives rather than topics can also improve the assessments. By testing specific ideas rather than general topics, it will help the STAT 2000 instructors zero in on the information that students are having the most trouble understanding. Including a table of specifications can add a subsequent degree of the instructors' understanding by analyzing what level of thought the students aren't grasping.

The objectives students are not grasping can also be determined by performing discrimination analysis. Discrimination analysis analyzes multiple-choice questions by creating a high-scoring and a low-scoring group based on overall test performance and calculating a discrimination value for each possible answer choice. The discrimination values for incorrect answers allow the researcher to determine which answer choices might be distractors for students who do not completely understand the material. For correct answers it determines which questions discriminate the most between the highest scorers and the lowest scorers. This information can be used to determine if the choices within the multiple-choice question need to be changed, or if the question is discriminating between the two groups too much or not enough.

The importance of assuring valid and reliable testing for a class such as STAT 2000 cannot be overstated. With the large number of students that take the course every semester, it is necessary to have clear and relevant questions to avoid widespread confusion during the testing period about the goal of each question. In addition, the assessments must cover all of the topics within the intended construct to ensure they fully understand the construct. By formalizing their validity checks and considering all aspects of each type of validity, the STAT 2000 coordinator and instructors can make certain that all of their tests valid and reliable.

Appendix A

Derivations

A.1 Derivation of $\sigma_{XT} = \sigma_T^2$

Here is the derivation of $\sigma_{XT} = \sigma_T^2$ referenced in Section 2.2.2. This comes from page 57 of *Statistical Theories of Mental Test Scores* (Lord and Novick, 1968). The assumptions mentioned below are stated at the end of section 2.2.1.

$$\begin{aligned}
\sigma_{XT} &= E(XT) - E(X)E(T) && \text{(A.1)}\\
&= E[(T+E)T] - E(T+E)E(T) && \text{(A.2)}\\
&= E(T^2) + E(ET) - (E(T))^2 - E(E)E(T) && \text{(A.3)}\\
&= \sigma_T^2 + E(ET) - E(E)E(T) && \text{(A.4)}\\
&= \sigma_T^2 && \text{(A.5)}
\end{aligned}$$

since $E(ET) = E(E)E(T)$ by independence of true scores and error scores (assumption 2) and $E(E) = 0$ (assumption 1).

A.2 Derivation of the Kuder-Richardson 20 Formula

The derivation of the Kuder-Richardson 20 Formula (Equation 2.24 in section 2.2.4) is found on pages 156 to 158 of the *Psychometrika* article *The Theory of the Estimation of Test Reliability* (Kuder and Richardson, 1937).

As stated in that section, Kuder and Richardson discovered a formula for σ_X^2 that could be rearranged to be this equation,

$$\rho_{YY'} = \frac{\sigma_X^2 - \sum_{i=1}^{n} p_i q_i}{(n-1)\sum_{i=1}^{n} p_i q_i}$$

which is used in the first step of the derivation of the KR 20.

$$\rho_{XT} = \frac{n\rho_{YY'}}{1+(n-1)\rho_{YY'}} \tag{A.6}$$

$$= \frac{n\left[\frac{\sigma_X^2 - \sum_{i=1}^{n} p_i q_i}{(n-1)\sum_{i=1}^{n} p_i q_i}\right]}{1+(n-1)\left[\frac{\sigma_X^2 - \sum_{i=1}^{n} p_i q_i}{(n-1)\sum_{i=1}^{n} p_i q_i}\right]} \tag{A.7}$$

$$= \frac{\frac{n}{n-1}\left[\frac{\sigma_X^2 - \sum_{i=1}^{n} p_i q_i}{\sum_{i=1}^{n} p_i q_i}\right]}{1+\frac{\sigma_X^2 - \sum_{i=1}^{n} p_i q_i}{\sum_{i=1}^{n} p_i q_i}} \tag{A.8}$$

$$= \frac{n}{n-1} \cdot \frac{\frac{\sigma_X^2 - \sum_{i=1}^{n} p_i q_i}{\sum_{i=1}^{n} p_i q_i}}{\frac{\sum_{i=1}^{n} p_i q_i + \sigma_X^2 - \sum_{i=1}^{n} p_i q_i}{\sum_{i=1}^{n} p_i q_i}} \tag{A.9}$$

$$= \frac{n}{n-1} \left(1 - \frac{\sum_{i=1}^{n} p_i q_i}{\sigma_X^2} \right) \tag{A.10}$$

Appendix B

Terms used in the Tables of Specifications

Below is a list of some of the terms used in the Tables of Specifications in section 4.1.1.

CI - Confidence Interval

PE - Point Estimate

MOE - Margin of Error

Confidence Letter - a z- or t-score calculated as part of the margin of error for a confidence interval

Assumptions - Requirements for a valid confidence interval or hypothesis test

Appendix C

Topics and Objectives

C.1 Topics

Below are the topics used to determine questions for the spring 2012 tests.

C.1.1 Test 3

CI General

What does 95% confidence mean?

What happens to margin of error and width of the interval if the level of confidence is changed?

What happens to the margin of error and width of the interval if the sample size is changed (only for p)?

What is the center of the interval?

Given point estimate and margin of error or standard error, determine the interval CI for p

Find the appropriate z-value

Calculate the margin of error

Calculate limit(s) using StatCrunch

Check conditions

CI for mu

Find the appropriate t-value

Calculate the margin of error

Calculate limit(s) using StatCrunch

- Given summary information
- Given actual data

Check conditions

N for p

With prior estimate

Without prior estimate

N for mu

HT general

Type I, Type II error

Indicate the relationship between a Type I error and the significance level

Understand what the p-value means and how it is used to make a decision

Set up hypotheses given a scenario

Indicate the connection between a CI and a 2-sided test

State a conclusion in context

Indicate the same general format for test statistics and confidence intervals

Determine a p-value for a t or z test statistic

Indicate the p-value for a two sided test is twice the p-value for a one-sided test.

HT for mu

Identify assumptions/conditions

Identify the test statistic

Carry out a test given actual or summary data

C.1.2 Test 4

Experiments

Identify response and explanatory variables

Identify treatments, experimental units

Indicate that treatments are values of the explanatory variable

Indicate type of experiment: completely randomized design, randomized block design, matched pairs design

Indicate whether an experiment is single-blind or double-blind

Identify a scenario as an experiment or observational study

Indicate the types of conclusions that can be drawn from experiments and observational studies.

CI General

What does 95% confidence mean?

What happens to margin of error and width of the interval if the level of confidence is changed?

Find point estimate given an interval

Given point estimate and margin of error or standard error and df, determine the interval

HT general

Type I, Type II error

Understand what the p-value means and how it is used to make a decision

Set up hypotheses given a scenario

Indicate the connection between a CI and a 2-sided test

State a conclusion in context

Indicate the same general format for test statistics and confidence intervals

Determine a p-value for a t or z test statistic

Indicate the p-value for a two sided test is twice the p-value for a one-sided test.

Paired difference test

Indicate whether data represents independent or dependent samples

Identify assumptions/conditions

Identify the test statistic

Determine xbar and s for the differences, given data

Carry out a test given actual data

Carry out a test given summary data?

HT p1-p2

Identify assumptions/conditions

Determine test statistic given necessary information

Determine confidence interval given adequate information

HT & CI mu1-mu2

Identify assumptions/conditions

Determine test statistic given necessary information

Determine confidence interval given adequate information

HT Chi-Square Goodness of Fit and Test for Independence

State hypotheses and conclusions

Identify conditions/assumptions

Determine Expected counts for cells

Determine test statistic

Use Chi-Square calculator with appropriate df

Carry out a goodness of fit test using StatCrunch

C.2 Measurable Objectives

Below are the measurable objectives created based off of these topics, used to create the questions for the fall 2012 tests

C.2.1 Test 3

- Students will categorize a given response variable as quantitative or qualitative
- Given the response variable type, students will choose the appropriate inference test
- Students will be able to identify a short run interpretation of a confidence interval
- Students will be able to identify a long run interpretation of a confidence interval
- Given a change in confidence level, students will be able to determine the effect on the confidence interval
- Students will calculate an interval given the point estimate, and margin of error or standard error and the level of confidence
- Given an interval, students will be able to determine the point estimate
- Given data or a summary of data, students will calculate the point estimate
- Given an interval, students will be able to determine the margin of error
- Given a level of confidence, students will use StatCrunch to find the appropriate t- or z-score
- Students will calculate margin of error
- Students will use StatCrunch to calculate the limits
- Students will identify conditions required for a valid confidence interval
- Given required parameters, students will be able to calculate a sample size to achieve the desired margin of error for an interval

- Students will be able to identify Type I and Type II errors
- Students will calculate a p-value
- Given a p-value, students will correctly identify the result of the significance test
- Students will be able to set up hypotheses
- Students will use the null hypothesis and a confidence interval to determine significance
- Students will state a conclusion of a hypothesis test, in context of the problem
- Students will identify assumptions required for a valid hypothesis test
- Students will calculate a test statistic
- For a sample size computation, given a change in confidence level, margin of error, or standard deviation, students will determine the change in sample size required
- Students will correctly identify the relationship between alpha and p-value given a conclusion about a hypothesis test
- Students will correctly identify the relationship between the significance level of a hypothesis test and the confidence level of a confidence interval

C.2.2 Test 4

- Given a study, students will indicate the type of study
- Students will identify a scenario as experimental or observational
- Students will identify response and explanatory variables
- Students will identify treatments of the explanatory variable
- Given a scenario, students will correctly choose the type of experiment
- Students will indicate whether an experiment is single- or double-blind

- Given the type of experiment, students will conclude whether causality is possible
- Students will indicate whether two samples were selected independently or dependently
- Students will categorize a given response variable as quantitative or qualitative
- Given the response variable type, students will choose the appropriate inference test
- Students will calculate an interval given the point estimate, and margin of error or standard error and the level of confidence
- Given an interval, students will be able to determine the point estimate
- Given a level of confidence, students will use StatCrunch to find the appropriate t- or z-score
- Students will use StatCrunch to determine limits of a confidence interval
- Students will calculate margin of error
- Students will interpret a confidence interval, in context of the problem
- Given a p-value, students will correctly identify the result of the significance test
- Students will be able to set up hypotheses for two populations
- Students will interpret the p-value
- Students will use the null hypothesis and a confidence interval to determine significance
- Students will calculate a test statistic for two populations
- Students will be able to select hypotheses for a Chi-Square test
- Students will identify assumptions required for a Chi-square test
- Students will determine expected counts for a Chi Square test
- Students will use StatCrunch to determine the Chi-Square statistic for a Goodness of Fit test

- Given a chi-square problem, students will correctly calculate degrees of freedom
- Students will be able to use the Chi-Square calculator using degrees of freedom
- Students will find degrees of freedom given a two-population scenario
- Students will calculate the Chi-Square statistic for a single category
- Students will determine the effect a change in the observed count has on the Chi-Square statistic
- Students will determine the correct conclusion for a Chi-Square scenario
- Given a Chi-Square test statistic, students will find the p-value

Appendix D

Explanation of Bloom's Levels of Thinking

These explanations of the levels of Bloom's taxonomy are derived from the definitions and examples on page 73 of *Introduction to Classical and Modern Test Theory* (Crocker and Algina, 1986).

Knowledge - Recall of factual material, usually by memorization

Comprehension - Interpreting previous information

Application - Manipulating or applying previous knowledge

Analysis - Identifying patterns and recognizing trends

Synthesis - Creating new ideas, inferring, or predicting

Evaluation - Creating and developing judgements of an outcome or idea

Bibliography

Crocker, L. M. and Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart, and Winston, New York.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–333.

Feldt, L. S. (1965). The approximate sampling distribution of kuder-richardson reliability coefficient twenty. *Psychometrika*, 30(3):357–370.

Feldt, L. S. (1969). A test of the hypothesis that cronbach's alpha or kuder-richardson coefficient twenty is the same for two tests. *Psychometrika*, 34(3):363–373.

Hubley, A. M. and Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research*, 103(2):219–230.

Kuder, G. F. and Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, 2(3):151–160.

Lord, F. M. and Novick, M. R. (1968). *Statistical theories of mental test scores.* Addison-Wesley, Reading, Mass.

Mager, R. F. (1975). *Preparing Instructional Objectives*. Fearon Publishers, Belmont, Calif., 2nd edition.

Miller, M. D., Linn, R. L., and Gronlund, N. E. (2013). *Measurement and assessment in teaching*. Pearson, Boston, 11th edition.

Neukrug, E. S. and Fawcett, R. C. (2006). *Essentials of testing and assessment: a practical guide for counselors, social workers, and psychologists*. Thomson Brooks/Cole, Canada.

University of Georgia (2013). Bulletin - Courses Home. Retrieved Feb. 9, 2013. http://bulletin.uga.edu/CoursesHome.aspx.