RELIABILITY OF COMPOSITE TEST SCORES: AN EVALUATION OF TEACHERS' WEIGHTING METHODS

By

BONITA MARIE STEELE
B.A., B.Ed., Laurentian University, Canada, 1992
Diploma in Special Education, The University of British Columbia, Canada, 1997

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF ARTS

In

THE FACULTY OF GRADUATE STUDIES

DEPARTMENT OF EDUCATIONAL AND COUNSELING PSYCHOLOGY, AND SPECIAL EDUCATION

With Specialization in

MEASUREMENT, EVALUATION, AND RESEARCH METHODOLOGY

We accept this thesis as conforming to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

April, 2000

© Bonita Marie Steele, 2000
In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Educational Counseling Psychology, + Special Educat

The University of British Columbia
Vancouver, Canada

Date April 18/2000
ABSTRACT

Teachers often combine component scores from varied item formats to create composite scores. These composite scores are supposed to reflect students' ability or achievement level in a specific domain such as mathematics or science. Sometimes these same teachers weight the component scores before combining them in an effort to increase the validity of the inferences made from the resulting composite scores (Chen, Hanson, & Harris, 1998). These resulting composite scores have validity-related intuitive merit for teachers who use them, but no research-demonstrated value. The present study examined some consequences of such teacher-designed ad-hoc weighting methods so that the resulting psychometric properties of these methods could be evaluated.

The item formats of the two components examined in this study were constructed-response and multiple-choice. Because of the unequal variance of the two item format component scores, and the dissimilar metric used for each of the two item formats, the resulting composite score is considered a congeneric measure.

What this present study attempted to do was to discover which of the typically-used teacher-designed ad-hoc weighting methods used for the weighting of the two component scores is best according to a particular criterion. The criterion adopted to evaluate the worth of a weighting method was the bias in the reliability estimates of the weighted composite scores.

For the evaluation of bias, the deviations in reliability estimates between un-weighted composite scores and corresponding weighted component scores were examined. Reliability estimates of un-weighted composite scores provide the true
estimations calculated from the actual test data, while the reliability estimates of the weighted composite scores reflect the post-hoc manipulations imposed by some teachers.

If a teacher decided to use an ad-hoc weighting method in an effort to increase the validity of a composite score formed by combining multiple-choice and constructed-response items, then he/she should do it without causing much of a change, or bias, in the reliability of the resulting composite test score. Such manipulations as component-score weighting can artificially inflate or deflate the reliability estimate of the resulting composite score. This in turn could create a composite score that is of an undesirable psychometric deviation from its original form, which could in turn result in a score-dictated misrepresentation of an examinee's ability or achievement level.

In all cases of test design and item combining, the validity of the resulting scores are of paramount concern. Many teachers have an intuitive sense of this, but lack the psychometric expertise to critically examine test scores from this perspective. In an intuitive-based fashion, teachers have often made attempts at ensuring the validity of scores by imposing some ad-hoc weighting methods on the multiple-choice and constructed-response components involved.

Four of the weighting methods examined in this research were chosen because of their common usage in educational research and public education, and one method was derived from measurement theory. The component score weighting methods are: score points as weights; number of items as weights; number of minutes (time) as weights; equal weights; and z-scores as weights.

With the minimal amount of measurement training that most classroom teachers have before they receive their teaching certification, they become accustomed to relying
on intuition, rather than measurement theory, for their test design methodology. Teacher-designed ad-hoc weighting methods are examples of intuition-based logic that has not yet been examined in a psychometric fashion. The research question is as follows: What commonly-used teacher-designed ad-hoc method of weighting and combining multiple-choice and constructed-response component scores is the best? The selection of the best method was done according to the criteria of minimal artifactual bias in the resulting composite score reliability estimates.

With the increased use of assessments that include mixed item types, it is certain that for some school administrators and parents, a composite test score would be desired. Encompassed in a composite score is the dilemma of how to combine the component scores in a manner that would maximize the validity and reliability of the composite score, without artificially inflating or deflating it and violating the principles of measurement. Crocker and Algina (1986) stated that to interpret composite scores in a meaningful way, in other words, to make valid inferences from the scores, it is important to understand how the statistical properties of the scores on each subtest or component influence those of the composite score.

When attempting to combine component test scores to produce a single composite score, a problem does arise if the component scores are not calculated using the same item format. When multiple-choice and constructed-response are the two item formats of the respective components, strict measurement principles of parallel measures will reveal that test homogeneity does not exist. Because of this, the scalings of the measurements of the two components differ by more than just an additive constant, and hence, a composite score is illogical (Lord & Novick, 1968).
Multiple-choice items are typically scored in a dichotomous manner, whereas constructed-response items are typically scored in a polytomous manner. Dichotomously scored items are scored without a magnitude of quantity, just a simple binary code that reveals if the item was correct or not. Polytomously scored items are scored in a continuous manner, where the content of such item responses is rated on an arbitrary scale in an attempt to measure the magnitude of the attribute in question (Russell, 1897).

Like many things in the realm of educational psychology, pure measurement principles provide good guidelines; but in application, they can also provide unrealistic restrictions. Current educational/psychological assessment research in North America is riddled with violations of measurement principles, but that doesn't mean that psychological constructs are not being measured in a relatively efficient and revealing manner. In small-scale assessments across North America, component scores of unlike item types are often combined to form composite scores, regardless of the fact that these scores may be deemed illogical by researchers such as Michell (1990) and Lord and Novick (1968).

It is likely that many future assessments will be made up of a combination of multiple-choice and constructed-response item types (Wainer & Thissen, 1993). Often, weighting methods are imposed on such item components with minimal psychometric foundation. For the sake of fair measurement of students' abilities and achievement levels, research regarding this psychometric query is necessary. The research findings of this present study will provide the basis of forethought and understanding in terms of the selection of an ad-hoc weighting method. This forethought and understanding will enable teachers to be more certain about their precision of reported composite scores.
The recent evolution of educational assessment as described above, has resulted in a shift in focus in educational research. An example of this is Chen, Hanson, and Harris' (1998) findings of classroom teachers' typical ad-hoc methods of weighting multiple-choice and constructed-response item scores in an effort to produce a composite score. The unknown/questionable nature of the reliability of the composite scores derived using these ad-hoc weighting methods is essentially the problem that was investigated in this thesis.

The main research task posed by the present study was to find out which teacher-designed ad-hoc component-score weighting method, designed for the combining of two component scores (one made of multiple-choice items and the other of constructed-response items), was the best. The evaluative criteria of composite score reliability bias was examined across variations in class size and item ratios (multiple-choice: constructed-response) to ensure that conclusions were reflective of real-life classroom sizes and test-item ratios. By examining a variety of aspects of three factors (class size, test mixture, and weighting method), an evaluative look at the general tendency for each weighing method to cause artificially deflated/inflated component-score reliability estimates was examined. It was found that the component-score weighting method that treated the components equally, called the equal weights (EQUAL) weighing method in the present study, lead to the least amount of deflation/inflation in the composite score reliability estimates. Please remember the limitations of the study: the results were calculated using test scores from the academic domain of science only, and these test scores were only gathered from Canadian grade eight students.
The Third International Mathematics and Science Study (TIMSS) data was used for this research. The TIMSS data set was selected because it included both multiple-choice and constructed-response item formats.
TABLE OF CONTENTS

Abstract........................................................................................................................................... ii
List of Tables....................................................................................................................................... x
List of Figures..................................................................................................................................... xi
Dedication.......................................................................................................................................... xii

Chapter One: Introduction
  1.1 Introduction................................................................................................................................. 1
  1.2 Weighing Methods....................................................................................................................... 2
  1.3 Historical Events Leading to the Act of Combining Multiple-Choice and Constructed Response Items.............................................................................................................................. 3
  1.4 Comparison of Multiple-Choice and Constructed-Response Items............................................. 6
  1.5 Item Scoring.................................................................................................................................... 9
  1.6 The Problem................................................................................................................................... 10
  1.7 Purpose of the Study and Research Question.............................................................................. 11
  1.8 Contribution.................................................................................................................................. 12
  1.9 The Present Study....................................................................................................................... 12
  1.10 A Preview of Chapter Two.......................................................................................................... 12

Chapter Two: Literature Review
  2.1 Introduction.................................................................................................................................. 14
  2.2 Historical Perspective.................................................................................................................... 14
      2.2.1 Measuring Psychological Variables.................................................................................. 14
      2.2.2 Similarities, Differences, and Limitations of Constructed-Response and Multiple-Choice Items.................................................................................................................................................. 18
      2.2.3 Related Issues of IRT and Unidimensionality........................................................................ 28
  2.3 Recent Literature.......................................................................................................................... 29
      2.3.1 The Respective Composite, and a Comparison of Constructed-Response and Multiple-Choice Items................................................................................................................................. 29
      2.3.2 Related Issues of IRT............................................................................................................ 40
  2.4 Methodology of forming Composite Scores.................................................................................. 42
      2.4.1 Weighting Mixed Item-Type Composite Scores................................................................. 42
      2.4.2 Issues Related to the Reliability of Mixed Item-Type Composite Scores.......................... 44
      2.4.3 Reliability Estimates of Congeneric Measures................................................................... 49
  2.5 Summary...................................................................................................................................... 52

Chapter Three: Methodology
  3.1 Data Source................................................................................................................................... 54
      Third International Mathematics and Science Study (TIMSS)....................................................... 54
  3.2 Design of the Present Study.......................................................................................................... 57
      3.2.1 Independent Variables.......................................................................................................... 58
      3.2.2 Dependent Variables........................................................................................................... 64
LIST OF TABLES

Table 1: A Tabular Representation of the Independent Variables..............................58
Table 2: Multiple-Choice and Constructed-Response Ratios Sorted by Item Ratio.........60
Table 3: Multiple-Choice and Constructed-Response Number of Items......................61
Table 4: Reliability Estimates of Component Scores...............................................68
Table 5: Mean Reliability Estimates for Raw Scores..............................................69
Table 6: One-Way Repeated-Measures ANOVA Results for Methods Effect.............70
Table 7: Pairwise Comparisons of Mean Reliability Estimates..................................71
Table 8: Correlation of Raw Composite Scores for each Booklet............................72
Table 9: Deviation of Mean Reliability Estimates for Each Composite Method.........74
Table 10: Descriptive Statistics of Effect Sizes....................................................75
LIST OF FIGURES

Figure 1: Effect Sizes Corresponding to the Weighting Methods.......................... 75
DEDICATION

To Paul Davidson, to whom I would like to thank for his un-dying support and emotional encouragement. The honesty of his laughter, the strength of his convictions, and the depth of his love; are things very rare and wonderful. I wanted to mention them here in an effort to pay homage and express my gratitude for knowing him.

To Dr. Nand Kishor, to whom I would like to thank for sharing his wisdom, his patience, his kindness, and most of all, his intellect. The greatest lesson that I learned from Dr. Kishor is the true meaning of integrity. By knowing him, I am forever changed.
CHAPTER ONE

1.1 Introduction

Teachers often combine component scores from varied item formats to create composite scores. These composite scores are supposed to reflect students' ability or achievement level in a specific domain such as mathematics or science. Sometimes these same teachers weight the component scores before combining them in an effort to increase the validity of the inferences made from the resulting composite scores (Chen, Hanson, & Harris, 1998). These resulting composite scores have validity-related intuitive merit for teachers who use them, but no research-demonstrated value. The present study examined some consequences of such teacher-designed ad-hoc weighting methods so that the resulting psychometric properties of these methods could be evaluated.

With regards to the notion of component score weighting, Burt (1950) suggested that "a priori or subjective weighting may be necessary when questions of value are concerned" (p. 122). Also, Wang and Stanley (1970) stated that "the weights may be chosen so as to maximize certain internal criteria such as the reliability of the composite measure" (p. 663). Thus, component score weighting has at times been advocated by those well-versed in psychometric theory. Admittedly, the above advocations for component score weighting were not mentioned here to conclude that it should be done; rather, they were provided here to show that a teacher's notion of component score weighting does have some psychometric foundation.

The item formats of the two components examined in this study were constructed-response and multiple-choice. Because of the unequal variance of the two item format component scores, and the dissimilar metric used for each of the two item formats, the resulting composite score is considered, by definition, a congeneric measure.
What this present study attempted to do was to discover which of the typically-used teacher-designed ad-hoc weighting methods used for the weighting of the two component scores is best according to a particular criterion. The criterion adopted to evaluate the worth of a weighting method was the bias in the reliability estimates of the weighted composite scores. Because of the congeneric nature of the composite score, a corresponding congeneric technique of estimating the reliability of composite scores had to be used.

For the evaluation of bias, the deviations in reliability estimates between un-weighted composite scores and corresponding weighted component scores were examined. Reliability estimates of un-weighted composite scores provided the true estimations calculated from the actual test data, while the reliability estimates of the weighted composite scores reflect the post-hoc manipulations imposed by some teachers.

If a teacher decided to use an ad-hoc weighting method in an effort to increase the validity of a composite score formed by combining multiple-choice and constructed-response items, then he/she should do it without causing much of a change, or bias, in the reliability of the resulting composite test score. Such manipulations as component-score weighting can artificially inflate or deflate the reliability estimate of the resulting composite score. This in turn could create a composite score that is of an undesirable psychometric deviation from its original form, which could in turn result in a score-dictated misrepresentation of an examinee's ability or achievement level.

1.2 Weighting Methods

Weighting component scores before combining them can be done using various methods. Four of the weighting methods examined in this research were chosen because
of their common usage in educational research and public education, and one method was
derived from measurement theory. The component score weighting methods are: score
points as weights; number of items as weights; number of minutes (time) as weights;
equal weights; and z-scores as weights.

With the minimal amount of measurement training that most classroom teachers
have before they receive their teaching certification, they become accustomed to relying
on intuition, rather than measurement theory, for their test design methodology. Teacher-
designed ad-hoc weighting methods are examples of intuition-based logic that has not yet
been examined in a psychometric fashion. Consequently, the research question of the
present study asked: What commonly-used teacher-designed ad-hoc method of weighting
and combining multiple-choice and constructed-response component scores is the best?
The selection of the best method was done according to the criteria of minimal artifactual
bias in the resulting composite score reliability estimates.

1.3 Historical Events Leading to the Act of Combining Multiple-Choice and
Constructed-Response Items

The usefulness of objectively scored standardized tests (using multiple-choice
items), was demonstrated with the use of the Army Alpha test in World War I. The
Army Alpha test was used as a mass-testing tool for the classification of military
personnel. Using the multiple-choice-item model of the Army Alpha test, the first
Scholastic Aptitude Test (SAT) was designed in 1926 to classify college candidates. By
1937, the American College Board’s essay tests were replaced with multiple-choice
achievement tests (Frederiksen, 1984). The continued use of multiple-choice tests by the
military was seen through World War II. These tests became more useful and efficient
for large-scale assessments with the development of a primitive IBM test-scoring machine. More efficient test-scoring machines were developed over time; by the 1980s, the Educational Testing Service had created machines that could score 10,000 multiple-choice item answer sheets an hour (Frederiksen).

By the 1980s, the multiple-choice test, along with the efficient test-scoring machines, made very large-scale testing "feasible" (p. 193); something that would not have been possible with constructed-response tests (Frederiksen, 1984). By 1982, 1.5 million students wrote the SAT. The efficient and economical test-scoring machines had "driven out other types of examinations, at least for large-scale testing programs" (Frederiksen, p. 194). The increasing role that certain large-scale multiple-choice tests were having in "holding schools accountable for the achievement of their students" (p. 194), presented a potential bias in the educational process (Frederiksen). Both Frederiksen and Popham (1993) explored this bias.

Popham (1983) explained that large-scale assessments played a small role in the direction that educational instruction took in the 1950s and early 1960's. During these years, the lack of media coverage of test scores resulted in the scores being used for ability grouping and for identifying students who needed help (Popham). Then in 1965, the Elementary and Secondary Education Act (ESEA) was passed in the United States; it mandated that programs funded by this act must be evaluated to ensure further funding (Popham). "Tests were being employed to make keep-or-kill decisions about educational programs" (Popham, p. 23). In the 1970s, some American states passed laws that required the use of competency tests for making decisions such as the granting of a high school diploma (Popham). In 1979, an American federal judge ruled that unless states
provided students with appropriate preparation for the competency tests, these tests violated the Constitution of the United States (Frederiksen, 1984).

Over time, with pressure to provide sufficient preparation for students, the mean scores on these statewide competency tests saw an increase. Frederiksen (1984) stated that although the “improvement in basic skills is of course much to be desired, and the use of tests to achieve that outcome is not to be condemned” (p. 195), a “reliance on objective tests to provide evidence of improvement may have contributed to a bias in education that decreases effort to teach other important abilities that are difficult to measure with multiple-choice tests” (p. 195).

This bias was widely recognized in 1982 when the American National Assessment of Educational Progress (NAEP) committee reported that for the previous ten years, on state mandated competency tests, performance on items measuring basic skills had either increased or stayed the same; but performance on items measuring more complex cognitive skills had decreased (Frederiksen, 1984). Frederiksen acknowledged that there could be many causes of the decrease in performance of the complex cognitive skill testing items, “but the possibility must be considered that the mandated use of minimum competency tests, which use the multiple-choice format almost exclusively, may have discouraged the teaching of abilities that cannot easily be measured with multiple-choice items” (p. 195).

The preceding description of historical events described the trend in large-scale assessment design formats during the last century. Early in the 1900's, multiple-choice items were used exclusively. By the 1970's, both multiple-choice and constructed-response items were used. Also by the 1970's, item-types used in large-scale assessments
began to have a substantial influence on small-scale testing. By this time, small-scale test
 designers such as teachers, started to follow the trend set by large-scale assessments.
 They began to include both multiple-choice and constructed-response item formats in
 their tests. This trend was adopted by many teachers who also took steps to increase the
 validity of the resulting composite scores by weighting the component scores of each
 item type before combining them. The weighting methods used were typically derived
 from intuition, and hence, needed to be psychometrically examined.

1.4 Comparison of Multiple-Choice and Constructed-Response Items

Bennett, Rock, and Wang (1991) stated that there are many unanswered questions
 about the differences between the multiple-choice and constructed-response item formats.
 Some questions are about the difference in measurement characteristics, others are about
 the difference in constructs being measured, the predictive validity of scores, and the
 reliability of scores (Bennett et al., 1991).

Multiple-choice test items are typically portrayed as “assessing simple factual
 recognition” (p. 77), and constructed-response test items are portrayed as “evaluating
 higher order thinking” (p. 77) (Bennett et al., 1991). Bennett et al. (1991) stated that such
 differences should be seriously examined because “they imply a mismatch between the
 highly valued thinking skills schools are lately attempting to impart, and the methods
 used for determining if those goals are being achieved” (p. 77).

The multiple-choice format is noted as being objective and efficient; while at the
 same time is criticized for the following: not resembling criterion behaviors, its limited
 use for instructional diagnosis, and its limited capability to measure certain cognitive
 processes (Bennett, Rock, Braun, Frye, Spohrer, & Soloway, 1990). Bennett et al. (1990)
stated that suggestions for greater reliance on constructed-response items are often given
to address the limitations of the multiple-choice item test format. This is done because
constructed-response items seem to have the ability to depict real-life educational and
work-setting tasks (Sebrechts, Bennett, & Rock, 1991).

Constructed-response items can be understood to be “any task for which the space
of examinee responses is not limited to a small set of presented options” (Bennett, 1993,
p. 100). Bennett stated that for the constructed-response format, the “examinee is forced
to formulate, rather than recognize, an answer” (p. 100). Birenbaum and Tatsuoka (1987)
stated that constructed-response items could also reveal information about the examinee's
of constructed-response items are that “(1) they provide a more direct representation of
content specifications . . . (2) they provide more diagnostic information about
examinees’ learning difficulties from their responses, (3) examinees prefer them to
multiple-choice items, and (4) the test formats may stimulate the teaching of important
skills, such as problem solving and essay writing” (p. 51).

Along with the advantages of the constructed-response item format are the
disadvantages. This format is expensive to score and it is time consuming for both
examinees and scorers. Because of the time needed to complete a constructed-response
item, relatively few can be answered by examinees, compared to multiple-choice items.
This lower number of items will often result in relatively low reliability estimates of the
scores, because score reliability estimates are a function of the number of items. Wainer
and Thissen (1993) found that it took 4 to 40 times longer to administer, and hundreds to
thousands of times more expensive to score a constructed-response test that would reveal scores with similar reliability to that of a corresponding multiple-choice test.

The merits of multiple-choice items are many. For example, they are favorable in terms of time, money, and objective and reliable scoring. The potential drawbacks of these item types are that they may give credit for rote-memory recall rather than critical thinking, they allow for guessing on behalf of the examinee, and they typically do not give credit for divergent critical analysis.

The merits of constructed-response items are noteworthy as well. This item type provides the examinee with tasks that appear more "real-life", as well as a chance to express relatively high levels of cognitive functioning. The potential drawbacks of constructed-response item types are that they are time consuming, expensive, and typically reveal scores that are estimated to have low reliability coefficients. Further, constructed-response items tend to access a relatively limited domain sample, and they also produce results that are confounded with the individual examinee's competency of writing skills.

The combining of these two item types is done in an attempt to reap the benefits of both, such as the "real-life"/valid nature of the scores from the constructed-response items and the reliable nature of scores from the multiple-choice items. In creating a composite score, the shortcomings of each item type remain, but they may now be examined in an effort to produce the most valid and reliable composite score. Hence, with the examination of the benefits and shortcomings of the two item types, a researcher may combine the component scores in a way that may better reveal magnitudes of the attribute being measured than one item type could do on its own.
In an early study in this area, Wainer and Thissen (1993) concluded that the multiple-choice item component yielded more precision in less testing time than the constructed-response item component. They suggested that "maybe measuring something that is not quite right *accurately*, may yield far better measurement than measuring the right thing *poorly*" (pp. 115). Contrary to this, Messick (1995) pointed out that methodological conclusions based on reliability analyses alone are weak as they fail to include aspects of validity. Both Samuel Messick (1995) and Lee Cronbach (1971) have stated that it is the valid inference of test scores that is paramount in designing a test. It has been suggested that the inclusion of real-life constructed-response items in a test that is composed of multiple-choice items, will provide an increase of the validity of the inferences of the resulting composite scores (Messick).

In all cases of test design and item combining, the validity of the resulting scores are of paramount concern. Many teachers have an intuitive sense of this, but lack the psychometric expertise to critically examine test scores from this perspective. In an intuitive-based fashion, teachers have often made attempts at ensuring the validity of scores by imposing some ad-hoc weighting methods on the multiple-choice and constructed-response components involved.

### 1.5 Item Scoring

Multiple-choice items are typically scored in a dichotomous manner, whereas constructed-response items are typically scored in a polytomous manner. Dichotomously scored items are scored without a magnitude of quantity, just a simple binary code that reveals if the item was correct or not. Polytomously scored items are scored in a
continuous manner, where the content of such item responses is rated on an arbitrary scale in an attempt to measure the magnitude of the attribute in question (Russell, 1897).

### 1.6 The Problem

With the increased use of assessments that include mixed item types, it is certain that for some school administrators and parents, a composite test score would be desired. Encompassed in a composite score is the dilemma of how to combine the component scores in a manner that would maximize the validity and reliability of the composite score, without artificially inflating or deflating it and violating the principles of measurement. Crocker and Algina (1986) stated that to interpret composite scores in a meaningful way, in other words, to make valid inferences from the scores, it is important to understand how the statistical properties of the scores on each subtest or component influence those of the composite score.

When attempting to combine component test scores to produce a single composite score, a problem does arise if the component scores are not calculated using the same item format. When *multiple-choice* and *constructed-response* are the two item formats of the respective components, strict measurement principles of parallel measures will reveal that test homogeneity does not exist. Because of this, the scalings of the measurements of the two components differ by more than just an additive constant, and hence, a composite score is illogical (Lord & Novick, 1968).

Like many things in the realm of educational psychology, pure measurement principles provide good guidelines; but in application, they can also provide unrealistic restrictions. Current educational/psychological assessment research in North America is riddled with violations of measurement principles, but that doesn’t mean that
psychological constructs are not being measured in a relatively efficient and revealing manner. In small-scale assessments across North America, component scores of unlike item types are often combined to form composite scores, regardless of the fact that these scores may be deemed illogical by researchers such as Michell (1997) and Lord and Novick (1968).

The recent evolution of educational assessment as described above, has resulted in a shift in focus in educational research. An example of this is Chen, Hanson, and Harris’ (1998) findings of classroom teachers’ typical ad-hoc methods of weighting multiple-choice and constructed-response item scores in an effort to produce a composite score. The unknown/questionable nature of the reliability of the composite scores derived using these ad-hoc weighting methods is essentially the problem that was investigated in this thesis.

1.7 Purpose of the Study and Research Question

The purpose of this thesis was to discover which of the teacher-designed ad-hoc methods of weighting and combining multiple-choice and constructed-response component scores, as presented by Chen et al. (1998), was the best. The evaluative criteria of composite score reliability bias was examined across variations in class size and item ratios (multiple-choice: constructed-response) to ensure that conclusions were reflective of real-life classroom sizes and test-item ratios. The research question follows as such: Which of the teacher-designed ad-hoc methods of weighting and combining multiple-choice and constructed-response component scores, is the best?
1.8 Contribution

It is likely that many future assessments will be made up of a combination of multiple-choice and constructed-response item types (Wainer & Thissen, 1993). Often, weighting methods are imposed on such item components with minimal psychometric foundation. For the sake of fair measurement of students' abilities and achievement levels, research regarding this psychometric query is necessary. The research findings of this present study will provide the basis of forethought and understanding in terms of the selection of an ad-hoc weighting method. This forethought and understanding will enable teachers to be more certain about their precision of reported composite scores.

1.9 The Present Study

The Third International Mathematics and Science Study (TIMSS) data was used for this research. The TIMSS data set was selected because it included both multiple-choice and constructed-response item formats. For each of the five component weighting methods, various sample sizes and various test mixtures (dictated by the item-type ratios built into the quasi-parallel booklets used by TIMSS) of the science questions were examined in the light of deviation in reliability estimates of the resulting weighted and un-weighted composite scores. The effect sizes of the deviations provided insight for making appropriate decisions about methods of weighting component scores when dealing with both multiple-choice and constructed-response test items.

1.10 A Preview of Chapter Two

The primary issues addressed in the literature review of this thesis are related to composite score reliability. Secondary issues addressed are related to the reliability and validity of test scores; the comparison and equivalence of constructed-response and
multiple-choice items; how each method of creating a composite score from components of unlike scoring has its limitations and strengths; how measurement theory reveals complications of combining two test component scores whose scaling differ by more than just an additive constant; and how the appropriateness of using the congeneric model of parallel measures applies.
CHAPTER TWO

Literature Review

2.1 Introduction

The scope of this literature review will include four sections. The first section will give a historical review of the concepts and questions to be explored in this study. This will provide the reader with the insight as to how prominent this query has been in educational research during the past century. The second section will present recent and current research literature with regards to this study. This will provide the reader with an up to date perspective of the research literature pertaining to this topic. The third section of this chapter will present literature that will support the method of analysis used in this study. The forth and final section of this chapter will present a discussion of the research question that is raised by the literature review, as well as a suggested research hypothesis.

2.2 Historical Perspective

2.2.1 Measuring Psychological Variables

This sub-section is intended to provide a historical perspective of the issues related to the measurement of psychological variables (in particular, the combining of mixed data sources to produce a measurement) in a general context at first, then within the context of educational measurement, and then finally, in the context of the assessment of science achievement. Relevant literature will be presented in chronological order, and will be discussed in light of the quest to gather a historical perspective on issues related to measuring educationally-related psychological variables. In this sub-section it is revealed how the trend of educational testing evolved from fixed-choice-item tests to
performance-assessment tests; and how logistics, politics, and economics led to the need for combining the two.

Nafe (1942) stated that when all is not known about the attribute at hand, quantification could not be made possible by one simple measure alone. Nafe added that if the attribute is not fully understood, an attempt at quantification could only be realized through logical processes that take into consideration all of the pertinent and experimentally observed facts. It seemed that Nafe endorsed the use of both qualitative and quantitative observations in an effort to reveal information about an attribute. The key point of this article is that it gave direction to the ideas of measuring a variable that is not fully understood. Nafe's ideas led researchers to consider using diverse data collection procedures to measure attributes that are not fully understood.

In the educational domain, there are many attributes that are not fully understood. But it was not until the 1980’s that the need to do a better job of assessing thinking skills in certain contexts became widely recognized. In 1989, Nickerson published an article in which controversial issues with regards to educational testing were explored. Nickerson opened with what he thought to be an issue of great importance: the need for the development of policies and procedures that do a better job of evaluating educational progress, and especially the extent to which students were learning to perform tasks that require high-level cognitive abilities.

From contemporary literature on the topic, Nickerson (1989) deduced that “a major criticism of both standardized tests and tests constructed by teachers for purposes of determining students’ grades, is that they tend to emphasize recall of declarative or procedural knowledge, and provide little indication either of the level at which students
understand subject matter or of the quality of their thinking” (p. 3). He pointed out that because higher-order cognitive functioning is a major goal of education, inadequate tools for measuring this means that educators are at a loss to judge the success of the education enterprise as a whole (Nickerson).

Nickerson (1989) stated that in the past, for tests to be widely used in the education enterprise, it was considered important that they can be easily administered, and easily and unambiguously scored. These constraints usually meant that tests were actualized with the use of multiple-choice or other highly structured items that allowed for single-right-answers. A major problem with these highly structured items is their failure to give evidence of the nature of the reasoning process by which choices were made (Nickerson). Nickerson argued that a “systematically valid evaluation of higher-order cognitive skills will probably depend on the use of tests that call for the performance of extended cognitive tasks and require considerable judgement in scoring” (Nickerson, p. 5). The key point of this article is that Nickerson openly argued that the inclusion of constructed-response items in educational assessments is necessary if the education community wants accurate evaluation of important higher-order cognitive skills.

By 1991, Collis and Romberg had explored advances in the understanding of how children learn, specifically mathematics. The rationale that these authors gave for pursuing this effort was based on the notion that during the 1980’s, there was a consistent call for new procedures for assessing mathematical performance in the United States (Collis & Romberg). They found that current assessment procedures were based on an old view of mathematics and learning, and hence, provided inadequate information about
examinees' performance (Collis & Romberg). One criticism that Collis and Romberg (1991) had of current assessment measures was the fragmented view of mathematics that these tests reflected, rather than a view of mathematics as an integrated whole. A second criticism was that current mathematics tests predominantly consisted of items requiring factual knowledge or procedural skills rather than items that assess students' cognitive abilities (Collis & Romberg).

Collis and Romberg (1991) stated that the emphasis of a mathematics test should not only be on the solutions, but also on processes and strategies. Tests should include constructed-response items for which the examinee has to create a response rather than select a response from a list (Collis & Romberg). The author of the present study thought that the logic of mathematics testing described above by Collis and Romberg, applies to science testing as well.

Collis and Romberg (1991) warned researchers of practicality-related concerns that may discourage the implementation of the constructed-response test items. One concern was that of the complexity of constructed-response items and the time needed by students to construct a response. Because of this, few items can be used at once, and this will lead to problems of sample selection and test score reliability (Collis and Romberg). Another concern they warned of was that knowledgeable humans, rather than machines, must do the scoring; this will lead to greater expenses and problems with interjudge agreement (Collis & Romberg). Eventually, it would be suggested that the combining of the multiple-choice and constructed-response items offered a remedy to the problems discussed by Collis and Romberg.
2.2.2 Similarities, Differences, and Limitations of Constructed-Response and Multiple-Choice Items

This sub-section is intended to provide a historical perspective on the issues related to the similarities, differences, and limitations of both constructed-response and multiple-choice item response formats. In some of the following studies, constructed-response and multiple-choice item formats were explicitly compared, while in other studies, only one of the two formats were examined.

In Frederiksen’s (1984) research, he warned that the abilities which are most easily and economically tested become the ones that are most often taught. He admitted that the representation of basic skills is certainly desired, but Frederiksen expressed a concern that the reliance on objective tests to provide evidence of improvement may contribute to a bias in education that decreases the effort to teach other important abilities that are difficult to measure with multiple-choice tests. He stated that if educational tests “fail to represent the spectrum of knowledge and skills that ought to be taught, they may introduce bias against teaching important skills that are not measured” (Frederiksen, p. 193).

Frederiksen (1984) cited a 1982 study by the National Assessment of Educational Progress (NAEP) that found that in the previous 10 years, performance on items measuring basic skills had not declined, but there was a decrease of items that reflected more complex cognitive skills. Of the many causes of this decline, Frederiksen considered the possibility that the mandated use of minimum competency tests, which used the multiple-choice format exclusively, “may have discouraged the teaching of abilities that cannot easily be measured with multiple-choice items” (p. 195).
Frederiksen (1984) reported the results of a study that categorized the multiple-choice items of the GRE Advanced Psychology Test in terms of cognitive difficulty. It was reported that 70% of the items measured memory, 15% measured comprehension, 12% required analytic thinking, and only 3% required evaluation. Hence, this multiple-choice item format test is primarily a measure of factual knowledge. Even though advocates of multiple-choice items claim that the multiple-choice item type can probe the higher levels of cognition, the difficulty of devising multiple-choice items that measure higher order cognitive skills "tend to result in tests that elicit factual knowledge rather than more complex processes" (p. 196).

Frederiksen (1984) found that many studies which sought a relationship between constructed-response and multiple-choice test items found very high correlations, often in the .90's. In the past, these correlations were generally interpreted to mean that the choice of the above-mentioned item formats does not influence what a test measures. Frederiksen stated that this interpretation was erroneously based on evidence in which the constructed-response items used for the correlation studies, were simple adaptations of the pre-existing multiple-choice items. This being the case, the constrained nature of multiple-choice items was reproduced for the corresponding constructed-response items. Correlating the constrained multiple-choice items with the corresponding "adapted" constructed-response items, would not surprisingly yield a strong relationship, but this is more about the unrealistic limitations put on the constructed-response items with regards to their purpose of being unconstrained (Frederiksen).

In support of his criticisms, Frederiksen (1984) reported the findings of a study that used the test called "Formulating Hypotheses" to find how highly correlated the
formats would be if each format provided items that required higher level cognition. It was reported that the correlations between the corresponding constructed-response and multiple-choice items were very low, "for example, for scores reflecting the mean quality of the ideas, the correlation between the formats was .18, and for scores based on number of ideas, the correlation was .19" (p. 198). Frederiksen stated that these correlations were much lower than the score reliabilities (as correlations always are), and this provided evidence that the two formats did not measure the same construct. Frederiksen concluded with the following call to educators with regard to test formats and the educational process:

The "real test bias"... has to do with the influence of tests on teaching and learning. Efficient tests tend to drive out less efficient tests, leaving many important abilities untested—and untaught. An important task for educators and psychologists is to develop instruments that will better reflect the whole domain of educational goals and to find ways to use them in improving the educational process. (p. 201)

Guthrie (1984) followed Frederiksen's (1984) line of thinking. He expressed a belief that "learning that is the essence of education consists of interpretations, problem solving, and applications of learned principles" (p. 190). Like Frederiksen, he contended that multiple-choice items typically measured low level cognitive abilities, and that higher level cognitive operations demanded a different response format, the constructed-response. Guthrie pleaded that although constructed-response items pose serious challenges of time, logistics, and reliability, "the need for exploring them is directly proportional to the pervasiveness of testing as a dimension of schooling" (p. 190).
Thissen, Steinberg, and Fitzpatrick (1989) explored the use of IRT for item analysis to describe the performance of the answer choices for multiple-choice items, using a sample of 1000 examinees. For this analysis, trace lines were statistically inferred from the data because the data set was not sufficient to provide them directly (Thissen et al.) due to a relatively small sample size. In a similar study that used a much larger sample size of 800,000, the researchers used empirical trace lines. Because of the weak findings that resulted from the smaller (and insufficient) sample size, Thissen et al. advised that IRT procedures be limited for use with large-scale testing programs only; they should not to be used for examination of classroom tests. The key point relevant to the present study is that although IRT may show promise for item analysis, and may reveal the strengths of each item with both multiple-choice or constructed-response items, it does this with only very large samples. Typical class size samples are implicitly (and explicitly by Thissen et al.) excluded.

Birenbaum and Tatsuoka (1987) made comparisons between constructed-response and multiple-choice items in which procedural tasks of arithmetic operations were examined. Birenbaum and Tatsuoka administered tests that were intended to measure ability with regards to fraction-addition arithmetic. Approximately half of the participants were given tests with constructed-response items, while the other half were given tests with the same number of corresponding multiple-choice items. The results of the study indicated considerable differences between the item formats, with more "favorable results" (p. 391) for the constructed-response format (Birenbaum & Tatsuoka). Birenbaum and Tatsuoka found that for constructed-response items, students who have not mastered certain procedural tasks, tend to be more consistent in applying their rules
of operation for solving procedural tasks than for multiple-choice items. For the constructed-response item format, a factor-analytic examination of the underlying structure revealed two distinct clusters, one for items that required the addition/subtraction of fractions with like denominators and one for questions that required the addition/subtraction of fractions with unlike denominators. The formation of these clusters accurately mirrors real-life breakdown in mathematics understanding. In contrast, for the analysis of the multiple-choice item format, no such clusters were identified; there was no distinct separation between the like and unlike denominator items.

Birenbaum and Tatsuoka (1987) stated that although multiple-choice items are considerably easier to score, they might not provide the appropriate information for identifying students’ misconceptions with respect to the given subject matter; the constructed-response items seem more appropriate for that purpose.

There are two key limitations of Birenbaum and Tatsuoka’s (1987) study. First, the subjects were all pulled from a “mathematical laboratory” of a high school. This “laboratory” accommodated students that were not mainstreamed for reasons ranging from poor mathematical ability to scheduling conflicts. This sample was not representative of the school’s population; hence, limited inferences can be made from the findings of Birenbaum and Tatsuoka (1987). Second, both the multiple-choice and the constructed-response test subjects were given the same number of items to solve, 19. Constructed-response items typically take much more time to complete than multiple-choice items, so it can be understood that the two groups’ testing time varied dramatically.
Having mentioned the key limitations of the study, certain key issues were raised that need to be addressed as well. One very interesting consequence of having both cohorts answer the same number of items, is that an analysis of the reliability of the scores for each item format can be accurately estimated in an effort to compare the two formats with respect to internal consistency. It has been typically reported that constructed-response item format tests yield less reliable scores than multiple-choice item format tests, even though internal consistency is a factor of number-of-items, and multiple-choice items typically outnumber the constructed-response items in an item count. So, in this rare situation where the number of items are the same, the reliability coefficients for the scores from the two item format tests, were found to be very similar. The Cronbach alpha reliability coefficient was .98 for the constructed-response item test and .97 for the multiple-choice item test. The author of the present study recognizes that because of the dichotomous nature of multiple-choice test items, Cronbach alpha is the inappropriate coefficient for the reliability of the multiple-choice scores; It should have been KR-20.

In another study that examined the differences between constructed-response and multiple-choice test-response items, Bennett, Rock, Braun, Frye, Spohrer, and Soloway (1990) explained that multiple-choice items have been the heart of standardized testing in the United States for the past century. They claimed that the use of the multiple-choice item format is justified by its objectivity and efficiency, and more recently by the development of such statistical models as IRT for its analysis (Bennett et al., 1990). Bennett et al. (1990) explained that multiple-choice items have been criticized because they don’t always “directly resemble criterion behaviors, they are of limited utility for
instructional diagnosis, and they might not be capable of measuring certain cognitive processes or skills” (p. 151). In response to these criticisms, Bennett et al. (1990) suggested that there should be a heavier reliance on constructed-response items because they can present tasks to the testees that are similar to those encountered in real educational and work-settings. Bennett et al. (1990) expanded their response by stating that constructed-response item responses can offer information on problem-solving processes, something that multiple-choice items often fail to do (Bennett et al., 1990).

Bennett et al. (1990) described the main disadvantages of constructed-response items for major testing programs, as the subjectivity and high cost associated with the scoring. They suggested that if a machine-scorable constructed-response item type could be developed, “problems associated with scoring cost and reliability might be substantially reduced” (Bennett et al., 1990, p. 151). The purpose of the study by Bennett et al. (1990) was to assess the relationship between machine-scorable computer program results of both multiple-choice and constructed-response items from the College Board’s Advanced Placement Computer Science (APCS) Examination. Although it was not the primary objective of the study, what appeared significant to Bennett et al. (1990) was the strong relationship that existed between the multiple-choice and the constructed-response items.

Bennett et al. (1990) stated that even though multiple-choice and constructed-response items appeared to measure the same essential construct, there were good reasons to maintain and increase the role of constructed-response items in educational assessment. The most compelling reason given by Bennett et al. (1990) was that successful completion of constructed-response items was central to future curriculum
Bennett et al. (1990) offered more reasons, stating that the inclusion of constructed-response items in large scale tests would emphasize to students and teachers the need to focus on the development of this skill, while the multiple-choice format may measure and encourage the development of irrelevant skills (Bennett et al., 1990). Their final words on the subject suggested that they were strongly in favor of the addition of constructed-response items to multiple-choice items for large scale assessments:

"Because of their perceived relevance, the inclusion of constructed-response items should help respond to these concerns, thereby increasing the credibility of the measures taken" (Bennett et al., 1990, p. 162).

Bennett, Rock, and Wang (1991) proceeded to investigate/replicate some of the findings of the above study by Bennett et al. (1990), by specifically examining the equivalence of multiple-choice and constructed-response items. In the study by Bennett et al. (1990), little support was found for the existence of construct differences between the item formats; Bennett et al. (1991) specifically examined whether the two formats assessed the same construct.

Bennett et al. (1991) claimed that this topic was of particular interest because the multiple-choice and constructed-response item formats are often portrayed in educational research literature "as measuring disparate cognitive constructs" (p. 77) as well as measuring constructs of different value (Bennett et al., 1991). Bennett et al. (1991) found that from a recent review of literature on the topic, the two formats did appear to measure different abilities, but the nature of these differences was unclear. They argued that the reason these differences were unclear was due to the nature of the construction of constructed-response items under investigation. These constructed-response items
appeared to differ from pre-existing multiple-choice items only in response format, and in such cases, "the constructed-responses will measure the same limited skills as the multiple-choice items" (Bennett et al., 1991, p. 78).

Bennett et al. (1991) used the same ACPS examination as was used in the Bennett et al. 1990 study, and claimed that the constructed-response items of this examination appeared to be more than simple adaptations of the multiple-choice items (Bennett et al., 1991). They speculated that this factor would ensure that the APCS examination would provide a reasonable opportunity for any real differences in the underlying constructs measured by these formats to be revealed (Bennett et al., 1991).

In a factor-analytic examination of the APCS data, Bennett et al. (1991) found that a single factor provided the most parsimonious fit, meaning that the evidence of the study offered little support for the notion that multiple-choice and constructed-response formats measure substantially different constructs. (Bennett et al., 1991).

Sebrechts, Bennett and Rock (1991) entered this research domain of item-type analysis by stating that constructed-response items were often argued to be more effective than multiple-choice items for educational assessment. They claimed that this argument was based on the notion that constructed-response items better reflect the tasks that examinees faced in real academic and work domains (Sebrechts et al.). Hence, constructed-response items "engender better measurement of higher order skills, permit responses to be evaluated diagnostically according to both the processes used to arrive at a solution and the degree of correctness, and communicate to teachers and students the importance of practicing these real-world tasks" (Sebrechts et al., p. 856).
Sebrechts et al. (1991) noted a key limitation of actualizing the inclusion of constructed-response items in educational assessment: the substantial costs associated with the scoring by knowledgeable human readers. In an effort to reduce the needed funding for constructed-response items for large scale testing programs, Sebrechts et al. devised a computer program that could potentially replace the human readers. This program was evaluated by examining the agreement between the computer-generated scores and human-generated scores.

Sebrechts et al. (1991) reported high correlations between the two methods of score generation, but two limitations of this study cause doubt about the worth of these findings. One limitation was the sample size of 30 for a correlational study. The other limitation is the method of development for the constructed-response items that ensured a constrained constructed-response format; The constructed-response items were simply “adapted from standard, five-option, multiple-choice word problems” (Sebrechts, p. 857). An ideal constructed-response item would be free of constraints; hence, the constructed-response items used in this study do not represent ideal constructed-response items.

Further expanding the idea of computer programs designed to score constructed-response type items, Kaplan (1992) attempted to create an automated system that could analyze “natural language” responses to constructed-response items. In an experiment similar to that of Sebrechts et al. (1991), Kaplan found that when responses were short and “the training set relatively representative” (p. 18), the scoring program matched the scoring of a human rater without much disagreement. But for longer item responses, “the training and scoring processes was more complex because of the complexity of the responses, with the result of poorer performance of the scoring program” (p. 18). Kaplan
suggested that the incorporation of a thesaurus into the scoring program may remedy the poor performance of the computerized scoring program. Kaplan’s suggestion is good in that it is a possible solution to the problem of his program. The capability of computers seems boundless with knowledgeable programming, so even though Kaplan’s program did not succeed on a variety of complex responses, it did succeed with some. I think that the search for the effective and accurate programming for constructed-response item scoring will lead to the eventual alleviation of the negative aspects of constructed-response items, such as time consumption, manpower, and cost.

2.2.3 Related Issues of IRT and Unidimensionality

This sub-section is intended to provide a historical perspective of the issues related to IRT and its limiting assumption of unidimensionality. In an effort to explain the basic ideas of item response theory (IRT), Hambleton, Swaminathan, and Rogers (1991) described the two postulates that the theory rests upon. The first postulate is that “the performance of an examinee on a test can be predicted (or explained) by a set of factors called traits, latent traits, or abilities” (p. 7); the second postulate is that “the relationship between examinees’ item performance and the set of traits underlying item performance can be described by a monotonically increasing function called an item characteristic function or item characteristic curve (ICC)” (Hambleton et al., p. 7).

Hambleton et al. highlighted two benefits of a properly fit IRT model: examinee ability estimates are not test dependant, and item indices are not group-dependant.

There are two main limitations of IRT models with respect to the thesis at hand. One limitation is caused by the mathematical model’s assumption of unidimensionality, and the other is the large number of subjects needed to make IRT models fit test data.
Hambleton et al. (1991) stated that for the assumption of unidimensionality to be met adequately by a set of test data, there must be a dominant component that influences test performance. For composite tests that have both multiple-choice and constructed-response item types, this assumption of unidimensionality is questionable. Frederiksen (1984) claimed that this composite does not meet the assumption of unidimensionality, while Wainer and Thissen (1993) claimed that it does. This debate will not be addressed explicitly in the present study, rather just recognized for the sake of awareness. The limitation that has more relevance to the study at hand is the large number of subjects needed to make IRT models fit test data. This limitation prevents IRT modeling from being used with typical classroom-sized samples (Hambleton et al.).

2.3 Recent Literature

2.3.1 The Respective Composite and a Comparison of Constructed-Response and Multiple-Choice Items

Part of this subsection is intended to provide a current perspective on the issues related to the strengths, weaknesses, and limitations of both constructed-response and multiple-choice items. Another part of this subsection is intended to provide a current perspective on the respective composite in which constructed-response and multiple-choice items are put together to form one test that reveals one score. The final part of this subsection is intended to provide a current perspective on the notion of congeneric measures and how they are related to the above composite dilemma.

Bennett (1993) wrote of concerns about the conception of a new test theory that would integrate constructed-response testing, artificial intelligence, and model-based measurement. Bennett described constructed-response items as ones for which “the
space of examinee responses is not limited to a small set of presented options” (p. 100). He explained that in this new test theory, the examinee is forced to formulate rather than recognize a response to the given constructed-response item. Bennett recognized that this definition implies a range in the complexity of the responses, but declared that the focus of his study was on complex constructed-response items that couldn’t be scored “immediately and unambiguously using mechanical application of a limited set of explicit criteria” (Bennett, 1993, p. 100).

Regardless of the difficulties in the scoring of constructed-response items, Bennett (1993) suggested that such research into the understanding of efficient and accurate methods is worthy of attention because constructed-response items “represent real-world tasks” (p. 100). Bennett suggested that constructed-response items “should more readily engage many of the higher-order cognitive processes required in academic and work settings” (p. 100).

The key contribution of Bennett’s (1993) findings, is his support of the constructed-response item format. In conclusion, Bennett argued that important constructs that are presumably not measured by multiple-choice items, are more likely to be measured via the constructed-response items.

Wainer and Thissen (1993) presented the most influential piece of literature on the articulation of the problem posed by this thesis. The need for educational research to be done in the field of mixed-format composite tests was acknowledged and expanded by Wainer and Thissen. They gained much attention from other researchers in this area, and hence, their paper has become a seminal article and has provided a base for research in the field of mixed-format composite tests. Furthermore, their research brought forth to
the research community the urgency and importance of the problem at hand: How do we best combine mixed-format component scores to produce an accurate composite score?

Wainer and Thissen (1993) examined cases in which the two components of the composite were made of multiple-choice and constructed-response items respectively. They identified the theoretical measurement problems of combining the component scores of each, and suggested various methods to work around these problems.

With regards to the mixed-format composite test described above, Wainer and Thissen (1993) identified and critiqued various methods of forming a composite score. Such methods were reliability weighting, Item Response Theory weighting and predictive validity weighting. They found that it did not appear practical to equalize the reliability of the different mixed-item components, or to examine predictive validity weighting in many cases, and hence, suggested that consideration must be given to IRT weighting (Wainer & Thissen). Realizing that they had not found the perfect theoretically correct solution, Wainer and Thissen openly invited other researchers to provide counter-evidence to what they proposed.

One fault was the fact that a practical limitation of IRT weighting was not identified by Wainer and Thissen (1993); that is of the restrictiveness of the very large sample size for which IRT becomes effective. Although Wainer and Thissen suggested IRT weighting, they did not pursue such an investigation; consequently, no further comments on the issue of IRT can be made here.

Another fault in their research is that from their limited reliability analyses, Wainer and Thissen made methodological inferences about the usefulness of constructed-response test items. Rather than pursuing an IRT analysis, Wainer and Thissen (1993)
devoted a major part of their article to the examination of their reliability data. They questioned the need for constructed-response items based on their findings that multiple-choice items are much “more precise” (p. 113); they actually referred to the lower reliability of the constructed-response component of a composite test as “unfortunate” (p.112). They considered the multiple-choice item component score to be more precise than the constructed-response item component score simply because it typically yielded scores that were more reliable. They made three errors in their logic. One error was made by considering reliability (and not validity) as the only factor in rating the precision/accuracy of test scores. A second error was made by under-emphasizing the fact that reliability is a function of the number of items, and in a two-component composite test as described earlier, multiple-choice components typically have far more number of items than a constructed-response component. This alone could account for the difference in the computed reliability coefficients. A third error was made by stating methodological inferences about the precision of each item type based on reliability analysis alone. Inferences made by Wainer and Thissen (1993) which suggested that constructed-response items are not as precise/accurate/good as multiple-choice items, were made on the basis of a reliability analysis alone. An example of what they reported is provided here: “One reviewer (of like mind to us) suggested that one ought not use constructed-response items. We almost agree” (Wainer & Thissen, 1993, p. 114). As Moss (1994) noted earlier in her paper, recent researchers following the work of Messick and Cronbach “have stressed the importance of balancing concerns about reliability, replicability, or generalizability with additional criteria such as ‘authenticity’, ‘directness’, or ‘cognitive complexity’” (p. 27).
Finally, Wainer and Thissen (1993) made a failing attempt to demonstrate an unbiased query of the usefulness of the constructed-response items. Wainer and Thissen (1993) cited Wittrock and Baker (1991) as an example of the researchers who purported that constructed-response items assess better deep understandings and allow broader analyses than do multiple-choice items. Then, without presenting an unbiased critique of Wittrock and Baker's findings, Wainer and Thissen contradicted the findings of the above research by claiming that in comparisons of the efficacy of the two formats, they had only found evidence that demonstrates that the multiple-choice item format is superior to the constructed-response item format. The attempt failed when Wainer and Thissen gave no citations of these comparisons.

In response to the above findings by Wainer and Thissen, it should be noted that for the development of constructed-response items, no guidelines set by a governing body such as AERA or APA have been developed, published, and enforced. Because of this, the results of research that compares these two item types vary a great deal. The most commonly used taxonomy in formulating levels of questions is Bloom's Taxonomy (Bloom, Englehart, Furst, Hill & Krathwhol, 1956). In this cognitive taxonomy, there are six levels: (1) knowledge; (2) comprehension; (3) application; (4) analysis; (5) synthesis; (6) evaluation. One of the purported benefits of constructed-response items is that they can access the higher levels of an examinee's cognition. So, it would seem that a well-designed constructed-response item would require and enable the examinee to access and use higher levels of cognitive functioning, such as synthesis and evaluation, to produce a successful answer.
Some constructed-response items are developed by simply extending multiple-choice items into questions that require written answers to low-level cognitive tasks such as knowledge and comprehension. These constructed-response items would not necessarily reveal any better, or deeper understandings than the multiple-choice items. Well-designed constructed-response items are developed in a manner that does enable the items to assess better and deeper understandings than multiple-choice items; they allow for broader analyses of the construct in question as Wittrock and Baker (1991) have claimed.

If the constructed-response items were simply an extension of the multiple-choice items, I would expect the correlation between the scores from the two item formats to be strong and positive. This high correlation should not be interpreted as demonstrative that both item formats measure the same construct, because the correlation of the item formats is dependant on the method of development, and level of cognitive complexity of each item. Hence, the correlations may vary as the depth of cognition sought by each multiple-choice/constructed-response item varies.

Lukhele, Thissen, and Wainer (1994) examined the relative value of multiple-choice and constructed-response items for large-scale achievement tests. They stated that the primary motivation for the use of constructed-response items is the idea that they can measure traits that cannot be measured by multiple-choice items, such as the assessment of an examinee’s dynamic cognitive process. Also, Lukhele et al. stated that constructed-response items “are thought to replicate more faithfully the tasks examinees face in academic and work settings” (p. 235). One limitation of constructed-response items
noted by Lukhele et al., is that because of limits on testing time, topic coverage using constructed-response items may be more limiting than multiple-choice items.

Lukhele et al. (1994) reported implications about the value of the constructed-response item. From their data, they concluded that constructed-response items “provide less information in more time at greater cost than do multiple-choice items” (Lukhele et al., p. 245). And although Lukhele et al. expressed sympathy toward the arguments in favor of constructed-response items, they stated that they had yet to see convincing psychometric evidence supporting their use. Lukhele et al. (1994) recognized the key limitations of their study such as the constructed-response items of the tests used may not have been state of the art, that the scorers were not as well trained as they might have been, and that the scoring rubrics are deficient in some unknown, but important, way.

Wilson and Wang (1995) examined the issues that arise when different modes, or item types, of assessment are combined. They described the main advantages of performance-based items, such as constructed-response, as “they provide more direct representation of content specifications, . . . they provide more diagnostic information about examinees’ learning difficulties from their responses, . . . the examinees prefer them to multiple-choice items, and the test formats may stimulate the teaching of important skills such as problem solving and essay writing” (Wilson & Wang, p. 51). They also described the advantages of the multiple-choice item format in that they are more economical to score “and have well-established patterns of reliability” (Wilson & Wang, p. 51).

Although the two item formats are very different from each other, Wilson and Wang (1995) stated that there exists the suggestion of “combining the scores from the
different assessment modes in order to take advantage of the positive aspects of each and
to attempt to avoid the negative aspects” (p. 51). Wilson and Wang recognized that in
order for this suggestion to be justified, each assessment mode must be found to be
sufficiently similar for measuring some construct or ability. They explained this notion
with their claim that “in order to combine scores from the different assessment modes, an
assumption of common latent variable is necessary” (Wilson & Wang, p. 52).

Wilson and Wang (1995) found that a typical open-ended (constructed-response)
item provided approximately 4.5 times more information than a typical multiple-choice
item. In addition, they found that constructed-response items provided more information
for high-ability examinees, but the multiple-choice items provided more information for
average examinees. Wilson and Wang found no conclusive evidence for the support of
the theory that both item formats measured one dimension.

Following the same line of research as Wilson and Wang (1995), Wang and
Wilson (1996) recognized the complementing advantages and disadvantages of multiple-
choice and constructed-response items. With regards to the reliability and validity of
scores for each type, test constructors have begun to incorporate these items together in
one test. In an effort to better understand the limitations and commonalties of the
composite test, Wang and Wilson attempted to illustrate how to “(a) link different test
forms through common items and common raters, (b) detect rater severity, and (c)
compare item information” (p.170) between multiple-choice items and constructed-
response items via item response modeling.

From their findings, Wang and Wilson (1996) stated that regardless of the
information gain of constructed-response items, its practical use is not justified. They
emphasized that the objectivity of raters is the most criticized issue of constructed-response items, and that any procedure to score constructed-response format tests or tests with both multiple-choice and constructed-response items will be misleading if rater severity is not taken into account. The key point of this study is Wang and Wilson’s warning of the impact of raters’ scoring objectivity on the reliability and validity of the test scores.

Along with Wang and Wilson (1996), Ercikan et al. (1998) recognized a “heightened interest” (p. 137) in tests that contain both multiple-choice and constructed-response items. They identified a critical question: can the two item types be combined to create a common scale? And if so, Ercikan et al. stated that the combining of the two item types “raises the issue of the relative weight of each item when producing a single score” (p. 137). Ercikan et al. described some commonly used methods of combining the scores from the two item types, as well as their view of the limitations of each method.

Weighting methods such as equal weighting, weighting according to testing time required for each item type, and weighting according to the number of raw score points, are stated by Ercikan et al. (1998) to be commonly used methods for combining the scores from multiple-choice and constructed-response formats. Having been demonstrated by Wainer and Thissen (1993), Ercikan et al. warned that for these methods, “anomalous results can arise” (p. 138) because these three weighting methods are likely to “produce a composite reliability that rarely exceeds the reliability of the more reliable part” (p. 138). This should be expected, as it is what is touted as part of the rationale for combining the scores: compromising the high reliability of scores from the
multiple-choice items for the high validity of the scores of the constructed-response
items.

Earlier in this chapter, related research findings by Birenbaum and Tatsuoka
(1987) were presented that seemed to provide a contrast to the above warnings by
Ercikan et al. (1998). Birenbaum and Tatsuoka's research findings demonstrated that the
lower internal-consistency based estimates of reliability of scores from a constructed-
response item section of a test may simply result from the smaller number of items in
comparison to the number of multiple-choice items in a composite test.

In response to the above warnings, Ercikan et al. (1998) proposed using an "IRT
model based approach to combining scores from the two item types" (p. 138). Of the
issues pertaining to the appropriateness of the use of IRT models for the present study
and its small sample sizes, the one that limits the inference of Ercikan et al.'s study, is the
large required sample size. It was discussed and demonstrated by Hambleton et al.
(1991) that IRT models do not fit data from typical classroom samples.

Qualls (1995) examined the issue of combing the scores of the two item types
from yet a different perspective. She stated that if scores are to be combined, another
issue that must be examined is the method of estimating the reliability of a test that
contains multiple item formats (Qualls). Qualls found that the two most common internal
consistency estimation techniques, the split-halves and the Cronbach’s alpha coefficient,
“would generally be inappropriate for multidimensional instruments” (p. 111). She found
that a test comprised of multiple item formats would most likely result in “parts with
differing functional lengths” (Qualls, p. 111); by definition, such an instrument could
only be represented though the adoption of the congeneric model. The term "functional
length" can be loosely interpreted as "degree of importance". From this, Qualls concluded that an appropriate reliability estimate for this type of instrument could be derived using the techniques of Raju (1977) and Feldt (1989) that will be explored in greater detail in a section on the methodologies of combining scores.

Qualls (1995) wrote of the three degrees of parallelism in part tests: the classically parallel, tau-equivalently parallel, and congenerically parallel. She stated that the difference among the three degrees of parallelism is based on the distributions of true score, error scores, and observed scores of the parts or items. She clarified that in all three degrees of parallelism, “the underlying skills are presumed to be identical, but units of measurement, the metric of measurement, may or may not be the same” (Qualls, p. 112). Thus in the case of congenerically parallel situations, the score distributions may have different part-test parameters “that in part result from variations in functional length of the parts” (Qualls, p. 112).

In a congenerically-parallel measurement situation, Qualls (1995) stated that although two item types may be viewed as measures of the same variable, the difference in the item format requires examinees to process information in different ways. She explained that when a researcher must consider weighting of the part scores, regardless of the rationale, they result in different units of measurement for the parts, hence, the functional length (degree of importance) of the parts will vary (Qualls). Qualls claimed that “the dissimilarity in functional length that characterizes the questions in a multi-item format test can only be modeled through the adoption of a congeneric model for part scores” (p. 113). Further elaboration of Qualls suggestions are presented in the methodology section of this literature review.
2.3.2 Related Issues of IRT

This subsection is intended to provide a current perspective of the issues related to IRT with regards to calculating a composite score from two test components made of different item formats.

Mislevy (1993) examined how contemporary test theory falls short in its scope by not accounting for the different "caricatures of ability" (p. 19) that have been found to be fundamentally important in the understanding of the domains of cognitive and educational psychology. Such caricatures are the students' internal representation of knowledge, problem-solving strategies, and the reorganization of knowledge as they learn.

Mislevy (1993) contended that the IRT model is never correct, and that "a single variable that accounts for all nonrandomness in examinees' responses is not a serious representation of cognition" (p. 25). Mislevy pointed out that like the Classical Test Theory (CTT), IRT captures examinees' tendencies to give correct responses, it does not capture all the systematic information in the data. Mislevy used the example that if a high-scoring examinee missed items that are generally easy, an examination of these unexpected responses could "reveal misconceptions or atypical patterns of learning" (p. 26). Mislevy continued this argument by stating that, to understand the data, one must "look beyond the simple universe of the IRT model to the content of the items, the structure of the learning area, the pedagogy of the discipline, and the psychology of the problem-solving tasks the items demand" (p. 26). Mislevy stated that if educational decisions depend on the patterns of an examinee's cognition, as well as, or instead of, overall proficiency, neither IRT nor CTT are the right tool.
Mislevy (1993) pointed out another application of IRT models that may fail: measuring learning. He stated that IRT characterization is complete for only a highly constrained type of change, for example, "an examinee's chance of success on all items must increase or decrease by exactly the same amount (in an appropriate metric)" (p. 27). Because of this, Mislevy contended that the patterns of learning that could be the crux of instruction can not be revealed via IRT.

Thissen, Wainer and Wang (1994) examined the assumption of unidimensionality for IRT when both multiple-choice and constructed-response items are combined in a test. They stated that to use IRT to score such a test, the test must be essentially unidimensional. The composite test will be expected to be unidimensional when the multiple-choice and constructed-response items are "effectively counterparts of one another with different formats" (p. 114), and multidimensional if the two item formats tap different skills (Thissen et al.).

For the tests that Thissen et al. (1994) examined, they found that the knowledge and skills required to answer the constructed-response items were very similar to those required for the multiple-choice items. This meant that they would be examining format differences when a comparison between the constructed-response and multiple-choice components was made (Thissen et al); not knowledge and skill differences.

Thissen et al. (1994) found that for the tests they examined, the constructed-response items intended to measure the same skill/ability/construct as the multiple-choice items. The authors were keen to recognize a key limitation to their findings though. They explained that for a test that used both item formats to measure two different aspects of one construct, they would expect results that indicated multidimensionality.
They even suggested that if that was the case, the component scores should be reported separately, and not as a composite score (Thissen et al.).

2.4 Methodology of Forming Composite Scores

2.4.1 Weighting Mixed Item-Type Component Scores to Form a Reliable and Valid Composite Score

This subsection is intended to provide an understanding of the pervasive issues related to the weighting of mixed item-type component scores in an effort to form composite scores that are maximally reliable and valid. Wang and Stanley (1970) examined various methods of differentially weighting component scores when these scores are to be combined to form a composite score. In discussing the weighting methods, Wang and Stanley stated that the "weights may be chosen so as to maximize certain internal criteria such as the reliability of the composite measure.... All methods weight most heavily those component measures which are "best" according to the particular criterion adopted, and they weight least, perhaps negatively, those which are "worst" (p. 663). The author of the present study believed that the above logic applies to validity as well as reliability.

Wang and Stanley (1970) conducted an in-depth review of empirical studies that examined the component-score weighting issue. They found that for many empirical studies, using/involving a wide variety of methods, weighted and unweighted test scores correlated nearly perfectly, with correlation coefficients in the mid to high nineties. This evidence led Wang and Stanley to conclude "that weighting was not worth the trouble" (p. 689).
The main limitation of Wang and Stanley’s (1970) conclusion is that it was made by examining studies in which the weighting was done on components of the same item format, hence the findings do not transfer to the weighting of dissimilar item formats. On the same note, the key point that I inferred from the findings of Wang and Stanley was that for components of like items, weighting is not very useful; but for components of unlike items, no such limiting evidence exists. So, in the present study, the weighting of different item type components will be examined in an effort to produce more reliable composite scores.

Chen, Hanson, and Harris (1998) also examined the weighting of different components to form a composite, but they did so with mixed item type composites. The specific composite models that they examined were combinations of a multiple-choice item component and a constructed-response item component that were found to be typical of classroom teachers. Chen et al. stated that the intent of their study was to “provide some information regarding the effects of using different methods of combining scores to form a composite, and to explain the issue of forming composites scores from the initial weighting issue through properties of the reported scores” (p. 1).

Chen et al. (1998) operationalized commonly used weighting methods of forming composite scores from multiple-choice and constructed-response format components. Using factors such as number of score points, number of items, time allocated for the items, and equal importance weighting, Chen et al. calculated the nominal weights of each component. Because Chen et al. presented their findings in an incomplete fashion at an academic conference, a complete set of results is not given. Admittedly, the incomplete nature of Chen et al.’s study does limit possible usefulness to the present
study. Regardless, the methodology presented by Chen et al. is judged as 'sound' by a panel of Canadian teachers (Misch, Kavenah, Milne, Andrews, & Davidson, 1999, personal correspondence). Hence, their methods of combining component scores to produce composite scores mirror what is being done in Canadian classrooms today.

One weighting method that was not mentioned by Chen et al. (1998) is that of z-score (standard score) weighting. I found some merit in examining this method along with the other four methods proposed by Chen et al. A z-score is a type of standard score that is frequently used in educational research. A standard score is a form of derived score that uses standard deviation units to express an individual's performance relative to the group's performance.

2.4.2 Issues Related to the Reliability of Mixed Item-Type Composite Scores

This sub-section is intended to provide an understanding of past and present issues related to reliability, and more specifically, the reliability of mixed item-type composite scores. To understand the concept of reliability, it is useful to explore the work of Kuder and Richardson (1937) who examined several approximations of the theoretical formula designed to estimate the reliability coefficient of test scores. The approximations were derived “with reference to several degrees of completeness of information about the test and to special assumptions” (p. 151). Kuder and Richardson explicitly stated that “no matter how computed, the reliability coefficient is only an estimate of the percentage of total variance that may be described as true variance, i.e., not due to error” (p. 151).

Because reliability coefficients provide the basis on which the weighting methods used in combining multiple-choice and constructed-response item components will be
evaluated in this thesis, a key point made by Kuder and Richardson (1937) will be presented here in an effort to further enhance an understanding of reliability. Kuder and Richardson stated that there is no unique value of the reliability coefficient; the coefficient is method dependent.

Further examining the issue of the reliability of test scores, Moiser (1943) focused on the special case of a weighted-composite score. He stated that for a composite score, "in absence of a measurable external criterion, ... the maximum reliability of the composite is certainly as legitimate a basis for assigning weights as is the intuitive assignment of small whole numbers" (Moiser, p. 161). Another point made by Moiser was that research that was intended to investigate the reliability of weighted composites, should be supported because "the investigation of the effect of the weighting of the components on the reliability of the composite is a problem of considerable, even if secondary, importance" (p.161).

Moiser (1943) concluded his examination of weighted-composite scores with an equation, from which it can be stated that for maximum reliability of the composite, "the weight of the component is directly proportional to the weighted sum of its intercorrelations with the remaining components, and inversely proportional to its error variance" (p. 168). With this, he presented the basic formula "by which the reliability of the weighted composite may be estimated from a knowledge of the dispersions, the weights, the reliabilities, and the intercorrelations of the components" (p. 162).

The following five steps replicate Moiser's prelude to the actual formula, concluding with his formula in step 6:
1) A composite variable $y$ consists of the weighted sum of $n$ component scores,

$$x_1, x_2, \ldots, x_j, \ldots, x_n.$$

2) For any examinee, $i$, 

$$y_i = W_1x_{i1} + W_2x_{i2} + \ldots + W_jx_{ij} + \ldots + W_nx_{in}, \quad (1)$$

where both $y_i$ and $x_i$ are deviation scores, and $W_i$ is the weighing coefficient.

3) The variance due to error of the $j$th component (equal to the square of its standard error) is 

$$\sigma^2_j = \sigma_j^2(1 - r_{jj}), \quad (2)$$

where $r_{jj}$ is the reliability of component $x_j$, and $\sigma_j^2$ is the variance of the $j$th component.

4) Since the error variance of each component is multiplied by its weight to give the error variance of the composite, and the error variances of the components are uncorrelated, the error variance of the composite variable may be written as such:

$$\sigma^2_y = \sum_{j=1}^{n} W_j^2 \sigma_j^2 = \sum_{j=1}^{n} W_j^2 \sigma_j^2 (1 - r_{jj}) = \sum_{j=1}^{n} W_j^2 \sigma_j^2 - \sum_{j=1}^{n} W_j^2 \sigma_j^2 r_{jj}. \quad (3)$$

5) The total variance of the composite is given by

$$\sigma^2_y = \sum_{j=1}^{n} W_j^2 \sigma_j^2 + 2 \sum_{j=1}^{n} \sum_{k=1}^{n} W_j W_k \sigma_j \sigma_k r_{jk} \quad (j<k). \quad (4)$$

6) Hence, the reliability of the composite can be expressed as a function of the ratio of the error variance to the total variance:

$$r_{yy} = 1 - \frac{\sigma^2_y}{\sigma^2_y} = 1 - \frac{\sum_{j=1}^{n} W_j^2 \sigma_j^2 - \sum_{j=1}^{n} W_j^2 \sigma_j^2 r_{jj}}{\sum_{j=1}^{n} W_j^2 \sigma_j^2 + 2 \sum_{j=1}^{n} \sum_{k=1}^{n} W_j W_k \sigma_j \sigma_k r_{jk}} \quad (j<k). \quad (5)$$

In contrast to the conclusions of Moiser (1943) with regards to the calculation of the reliability estimate for composite test scores, Rozeboom (1987) stated that the
"traditional formulas for estimating the reliability of a composite test from its internal item statistics are inappropriate to judge the reliability . . . of weighted composites of subtests that are appreciably nonequivalent" (p. 277). Rozeboom’s quest was focused on answering the question of how the reliability of a composite score can be discovered by only using the existing component score information. The importance of this quest was revealed when Rozeboom stated that in the case when the composite score is an “observable surrogate” (p. 279) for some hard-to-measure construct, the square root of the reliability estimate of the composite is a useful upper bound on the validity of the test’s scores. Rozeboom derived and reported an equation that answered/ended his quest. The following steps replicate Rozeboom’s prelude to his equation, concluding with his actual formula:

1) For a linear composite score of nonequivalent subtests, $$x^* = w_1x_1 + \ldots + w_mx_m$$ when the component scores are $$x_1, \ldots, x_m$$, and $$w_1, \ldots, w_m$$ are weights whose component reliability estimates are $$r_{x_1}, \ldots, r_{x_m}$$.

2) Because subtests $$x_1, \ldots, x_m$$ are not presumed equivalent, the reliability of the composite can not be estimated using the traditional internal consistency formulas. However, the reliability of the composite can be determined by the subtests’ reliabilities, their variances/covariances, and their corresponding weights, when one assumes that measurement errors are uncorrelated between subtests.

3) This is Rozeboom’s resulting formula:
The main contribution of Rozeboom (1987) was his equation, and its special features that could have potentially met the needs of the present study. He stressed that currently in educational measurement, the question of how internal-consistency estimates of reliability “are inflated by correlated errors is still an empirical unknown” (Rozeboom, p. 283). But for his equation, the issue of correlated errors is removed. Rozeboom stated that for his equation’s rationale, the test components are not just sub-groups of one test that had one item format, administered at one point in time. Rather, each component score is derived from different kinds of measures, obtained in different ways, and possibly, in different settings. Rozeboom claimed that with his equation, the “presumption of uncorrelated errors between subtests should be as true to fact as this classic ideal ever approaches in practice” (p. 283). A key limitation of the equation by Rozeboom, which is also a limitation of many other reliability estimates, is its assumption of non-correlated error. Although it can be gathered from his argument that mixed-item composites would have a better chance at having non-correlated errors than the like-item composites, there is no assurance that this will always be the case.

For a more general method of estimating reliability, Allen and Yen (1979) stated that when a test is divided into $k$ components, the variances of scores on the components
and the variance of the total test scores, can be used to estimate the reliability of the test's scores. They stated that if the components are essentially τ-equivalent, then coefficient α will produce the reliability of the scores; but if the components are not essentially τ-equivalent, then coefficient α will underestimate the reliability of the scores. The formula for internal-consistency reliability as spoken of by Allen and Yen is:

\[ \alpha_{xx'} \geq \alpha = \frac{k}{k-1} \left[ \frac{\sigma^2 x - \sum_{i=1}^{k} \sigma^2 y_i}{\sigma^2 x} \right] \]  

(8)

where \( x \) is the observed score for a test formed from combining \( k \) components, such that \( x = \sum_{i=1}^{k} y_i \); \( \sigma^2 x = \) the population variance of \( x \); \( \sigma^2 y_i = \) the population variance of the \( i \)th component, \( y_i \); and \( k = \) the number of components that are combined to form \( x \).

### 2.4.3 Reliability Estimates of Congeneric Measures

This sub-section is intended to provide a fuller understanding of the concept of congeneric measures as well as the corresponding congeneric method of estimating reliability. Earlier in this literature review, Qualls (1995) research was introduced with the purpose of providing a description of the notion of congeneric measures. Her work will be re-introduced here in an effort to provide a rationale for selecting the congeneric method for estimating the reliability of test scores when the test is composed of multiple-item formats.
Qualls (1995) examined the problem of reliability estimates for tests composed of multiple item formats, and found that for this type of test, the only appropriate model of part-test parallelism is the congenerically parallel model because it carries "more flexible implications for the part-score distributions" (Qualls, p. 113). For the congeneric model, "each coneneric part will contribute a unique amount to the total test true score" (Qualls, p. 113). Because the coneneric parts "vary in their contribution to total test true score variance" (p. 113), neither Spearman-Brown nor Coefficient Alpha methods of reliability estimates would be appropriate considering the embedded assumptions of each. Both methods would yield negatively biased estimates of the reliability coefficient if the test were made of coneneric parts (Qualls).

For a test composed of coneneric parts, each reliability estimation method described above could be used, but only the coneneric method, introduced by Raju (1977) and expanded upon by Feldt and Brennan (1989), would not underestimate it.

Raju's reliability coefficient is useful and appropriate when a researcher is working with coneneric measures of which all components are scored in a dichotomous manner. In the case where the test used all dichotomously scored discrete item types that are equally weighted, "\( \lambda_j \) can often be defined by the number of items, \( (n) \), in the part in relation to the total number of items, that is \( \lambda_j = n_j / (n_1 + ... + n_k) \)" (Qualls, p. 114). Raju presented a formula to estimate the reliability of such a composite score:

\[
R_{pxx'} = \frac{\sigma^2_x - \sum \sigma^2_y}{(1 - \sum \lambda^2_j)\sigma^2_x}.
\]  

(9)

when \( x \) is the observed composite score, and \( y_1, y_2, ..., y_k \) are the component scores of \( k \) separately scored parts; consequently, \( x = y_1 + y_2 + ... + y_k \).

\[
\]  

(10)
The Raju formula (9) does not appear as straightforward when some scores are dichotomous and others are polytomous.

For the classically congeneric model, in which the functional lengths may vary, Qualls (1995) stated that it is possible to derive an appropriate reliability estimate with Raju’s formula. In this model, the functional lengths, \( \lambda_1, \ldots, \lambda_k \) when \( \sum \lambda_j = 1 \), must be estimated from part test data (Qualls). The solution for \( \lambda_j \) can be found as follows:

\[
\lambda_j = \frac{\sigma^2 \gamma_j + \sigma \gamma_j \gamma_1 + \sigma^2 \gamma_2 + \sigma \gamma_2 \gamma_1 + \ldots + \sigma \gamma_j \gamma_k}{\sigma^2 x} = \sigma \gamma_j \gamma_1 \sigma^2 x,
\]

which is getting quite complicated.

Feldt and Brennan (1989) expanded on Raju’s findings and developed a variation of Raju’s formula by making the assumption that the components of the composite test contribute to the total test parameters “in direct proportion to their functional length” (p. 115). With this assumption, Feldt and Brennan (1989) recognized that the true score variance for each component will equal \( \lambda_j^2 \sigma^2 T_x \)

\[
(12)
\]

and the error score variance will be \( \lambda_j \sigma^2 \bar{e} \)

\[
(13)
\]

when the true score component for part \( j \) is \( T_j = \lambda_j T_x + C_j \).

\[
(14)
\]

For this variation of Raju’s formula, called the Feldt-Raju (F-R) equation, the \( \lambda_j \) substitutes the \( \lambda_j \) from equation 9. Qualls (1995) stated that “in essence, the sum of the elements in any row, \( j \), of the variance/covariance matrix of classically congeneric parts or the covariance of scores on part \( j \) with the total score equals \( \lambda_j \sigma^2 x \) and \( \lambda_j \) can be easily estimated from the sample matrix, \( \hat{\lambda}_j = (\text{sumrow}_j) / \sigma^2 x \)” (p. 116). The simplicity of the \( \hat{\lambda}_j \) calculation as opposed to the more ardent \( \lambda_j \) calculation is the reason that it was
chosen for the present study. The assumption that is made when the \( \hat{\lambda}_j \) calculation is used rather than the \( \lambda_j \) (that the components of the composite test contribute to the total test parameters in direct proportion to their functional length). With assumptions stated, simply substituting \( \hat{\lambda}_j \) for \( \lambda_j \) in Raju’s reliability formula (9) will yield the Feldt-Raju formula for estimation of the reliability of congeneric measures:

\[
F - R_{\text{pxp}} = \frac{\sigma_x^2 - \sum \sigma_j^2 y_j}{(1 - \sum \hat{\lambda}_j^2)\sigma_x^2}.
\]  

(15)

2.5 Summary

The present study attempted to answer the research question, “Which of the teacher-designed ad-hoc methods of weighting and combining multiple-choice and constructed-response component scores, is the best?” In an effort to present a strong case for the relevance and importance of this study, I presented an extensive list of the research findings and opinions of researchers in the field of educational measurement. With varying opinions of the appropriateness and usefulness of incorporating constructed-response items into the traditional fixed-response tests, the mixed item format did actualize and is very common today for both small and large-scale assessments. For this reason, the focus of the present study was not be on the argument of which item type is best, but rather the effects upon composite reliability estimates of different weightings of multiple-choice and constructed-response component scores.

On the combining of the scores to produce a composite score, there exists the debate of the usefulness and appropriateness of the use of IRT for the production of composite scores, but the IRT debate will not be the focus of the present study because
IRT models do not work for all sample sizes. IRT models may very well solve the mystery of the composite score for large-scale assessment, but, in its present form, it does not do the same for typical classes of 15 to 45 students.

Classroom teachers have their own methods of weighting and combining the scores, but these methods may not be grounded in strict measurement theory. Until this study, the effect of component-score weighting upon the reliability of composite scores of these methods has not being widely examined by educational researchers. With regards to the ranking of the reliability of the various weighting methods, it is worth noting that Wang and Stanley stated that the "weights may be chosen so as to maximize certain internal criteria such as the reliability of the composite measure.... All methods weight most heavily those component measures which are "best" according to the particular criterion adopted, and they weight least, perhaps negatively, those which are "worst" (p. 663).

What the present study did was evaluate each of four commonly-used teacher methods, plus one based on measurement theory, in terms of the reliability of the composite score, while also considering some other factors that may affect the reliability such as class size and test mixture.
CHAPTER THREE

Methodology

This chapter provides a detailed description of the methodology of the present study. The TIMSS study and the selection of TIMSS data for use in the present study are discussed, followed by a description of the design and procedures.

3.1 Data Source

Third International Mathematics and Science Study (TIMSS)

Operating under the patronage of the International Association for the Evaluation of Educational Achievement (IEA), TIMSS was a comparative study of student achievement in mathematics and science in 45 countries. TIMSS was a relatively recent, professionally organized, large-scale assessment of mathematics and science achievement that used a combination of multiple-choice and constructed-response items. The data set provided a current, relevant, reputable, and applicable data set.

The data collection took place in 1994 and 1995. Each country from the TIMSS divided its examinees into three “populations” based on age at the time of testing. Population 1 consisted of students at the earliest point at which most children are considered old enough to respond to written test questions; Population 2 consisted of students who had finished primary education and were beginning secondary education; and Population 3 consisted of students nearing the end of their secondary education.

TIMSS sought to obtain reliable and valid scores by combining “traditional objective multiple-choice items with innovative and challenging free-response questions” (Martin, 1996, p. 1-16). Martin stated that as a consequence of the emphasis on “authentic and innovative assessment” (p. 1-17), much of the testing time was allocated
for answering the free-response (constructed-response) items. For the actual test booklets used in the TIMSS, there was an elaborate test-development process performed by, what TIMSS administrators referred to as, "the leading experts" (Martin, 1996, p.1-15) from 50 countries. First, there was the development of an item pool, then pilot-tests of these items, then augmentation of the item pool, and finally, the creation of eight quasi-parallel test booklets.

For the present study, science-achievement data collected from Population 2 was selected for two reasons. First, there was a larger incidence of science-achievement questions in the form of polytomously scored constructed-response items in science than there were in mathematics; second, Population 2 was compulsory for all 45 countries. Population 2 encompassed “all students enrolled in the two adjacent grades containing the largest proportion of students of age 13 at the time of testing” (Martin, 1996, p. 1-8).

Some constraints of the TIMSS test design were listed by TIMSS administrators and are presented here to enhance the understanding of the data:

a) The total testing time for individual students in Population 2 was not to exceed 90 minutes;

b) There were to be a maximum of eight different test booklets for Population 2, and were to be distributed randomly (rotated) in each sampled classroom;

c) There was to be a small subset of items common to all test booklets, to allow the verification of within-school booklet rotation and testing of the equivalence of the student samples assigned each booklet;
d) The booklets were to be approximately parallel in content coverage, although it was acceptable for some booklets to have more mathematics than science content and vice versa;

e) The booklets were to be approximately parallel in difficulty;

f) Each booklet was to contain some mathematical and some science items;

g) Content coverage in each subject (mathematics and science) was to be as broad as possible. In Population 2, items from 12 mathematics content areas and 10 science areas were to be included;

h) It was anticipated that resources would be insufficient to provide precise estimates of item-level statistics for all items in the TIMSS item pool. Therefore, it was decided that such estimates would be required only for a subset of test items. Those items should, therefore, appear in as many booklets as possible. Others would appear in only one or two booklets, and would have less precise item-level statistics; and

i) For a subset of the items, all of the item intercorrelations were required, whereas for the remaining items only a sample of item intercorrelations was estimated. (Adams & Gonzalez, 1996, p. 3-2)

Adams and Gonzalez (1996) described the Population 2 test design as follows:
a) All test items were grouped into 26 mutually exclusive item clusters;

b) There is a core cluster that appears in the second position in every booklet;

c) Each cluster contains items in mathematics, science, or both;

d) Each booklet comprises up to 7 item clusters;
e) The booklets are divided into two parts and administered in two consecutive testing sessions;

f) One cluster appears in all booklets and some clusters appear in three, two or only one booklet;

g) Of the 26 clusters, 8 take 12 minutes, 10 take 22 minutes, and 8 take 10 minutes; and

h) Thus, the design provides 396 unique testing minutes, 198 of which are for mathematics.

Both the raw TIMSS data and some preliminary findings are now available to the public for further research. Although the TIMSS had data from 45 countries, only the Canadian Population 2 data were used in the present study.

3.2 Design of the Present Study

In the present study, there were three independent variables and two dependent variables. The three independent variables were sample size, test mixture, and weighting method. Sample size and test mixture were used to examine the weighting methods' performance under varying typical assessment conditions, but weighting method was the independent variable of most interest. One dependent variable was the reliability estimate of the unweighted and weighted composite scores. The other dependent variable was the deviations between the weighted composite score reliability estimates, and the corresponding un-weighted composite reliability estimates, calculated as effect sizes, Cohen's $d$. An examination of the effect sizes calculated from each of the five weighting methods, over eight test mixtures and three sample sizes, revealed the "best" weighting method. The "best" weighting method was revealed by having an effect size
mean closest to zero, and by having the smallest range. See Table 1 for a tabular representation of the variables.

Table 1

A Tabular Representation of the Independent Variables

<table>
<thead>
<tr>
<th>TM</th>
<th>Sample Size</th>
<th>Weighting Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>15</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td></td>
<td>45</td>
<td>1 2 3 4 5</td>
</tr>
</tbody>
</table>

Note: TM represents test mixtures. In the final analysis, each of the 120 cells contain the reliability estimates of the corresponding weighted composite scores.

3.2.1 Independent Variables

Sample size (class size) was an independent variable because it represented typical class sizes in Canada. Depending on various demographic factors, class sizes of 15, 30, and 45 are typical. Test mixture was an independent variable because it represented typical combinations of multiple-choice and constructed-response items. The TIMSS was designed in such a way that a researcher using the data set may explore the effects of different test mixtures. For the present study, the test mixture variable was
explored by examining the different test mixtures as set by the TIMSS. Each test mixture corresponded to a booklet; there were eight booklets for population 2, therefore there were eight test mixtures to be explored. Finally, weighting method was the principle independent variable explored because the purpose of this study was to find the best weighting method.

Sample Size

In this study, the use of the term sample size maybe confusing, so a thorough explicit description will be given here. In the introduction of this study, the term sample size was used to describe the typical class sizes that were simulated in an effort to mimic real-life class sizes in Canada. This will remain the exclusive use of the term.

Following this, there is the number of participants involved in each of the eight TIMSS booklets, approximately 2000. The scores from the participants of the eight booklets are never combined in this study; the sample size of each of the eight sub-populations will be referred to as the number of participants corresponding to a specific booklet.

From each of the eight sub-populations, random samples of 15, 30, and 45 were selected and their impact examined on the reliability of composite scores to reveal a trend in the findings. The effect that sample size had on reliability estimates in this context was examined to see if sample size was an influential factor.

Test Mixture

With each sample size, for each weighting method, various test mixtures were examined. These test mixtures were dictated by the item ratio within each of the eight
TIMSS booklets. For example, in booklet eight, the item ratio of multiple-choice to constructed-response (mc:cr) is 2.2:1 as shown in Table 2.

Table 2

Multiple-Choice and Constructed-Response Ratios Sorted by Item Ratio

<table>
<thead>
<tr>
<th>Booklet</th>
<th>Score Ratio*</th>
<th>Item Ratio*</th>
<th>Time Ratio*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.7:1</td>
<td>33:1</td>
<td>2.75:1</td>
</tr>
<tr>
<td>3</td>
<td>0.84:1</td>
<td>16:1</td>
<td>2.46:1</td>
</tr>
<tr>
<td>5</td>
<td>0.78:1</td>
<td>8.7:1</td>
<td>1.37:1</td>
</tr>
<tr>
<td>4</td>
<td>0.24:1</td>
<td>5:1</td>
<td>2.14:1</td>
</tr>
<tr>
<td>7</td>
<td>0.21:1</td>
<td>4:1</td>
<td>1.65:1</td>
</tr>
<tr>
<td>2</td>
<td>0.21:1</td>
<td>3.9:1</td>
<td>2.21:1</td>
</tr>
<tr>
<td>6</td>
<td>0.18:1</td>
<td>3.4:1</td>
<td>2.21:1</td>
</tr>
<tr>
<td>8</td>
<td>0.10:1</td>
<td>2.2:1</td>
<td>0.96:1</td>
</tr>
</tbody>
</table>

Note: '*' indicates the ratio of multiple-choice to constructed-response.

Weighting Method

For the weighting method variable, four weighting methods designed to calculate composite scores were revealed by classroom teachers (Chen et al., 1998). These four methods, plus one based on measurement theory, are described below:

a) composite score using score points as weights;
b) composite score using number of items as weights;
c) composite score using time-per-item as weights;
d) composite score using equal weights;
e) composite score using equal variances as weights.
To actualize the weighting methods, various aspects of the test mixtures need to be revealed. In Table 3, there is a complete breakdown of all aspects of the eight test mixtures. For example, for booklet eight, there are 22 multiple-choice items and 10 constructed-response item; there are 22 multiple-choice score points and 211 constructed-response score points; and there are 22 multiple-choice minutes and 23 constructed-response minutes.

Table 3

Multiple-Choice and Constructed-Response Number of Items, Score Points and Time

<table>
<thead>
<tr>
<th>Booklet</th>
<th># Items</th>
<th># Score Points</th>
<th># Minutes (Time)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mc</td>
<td>cr</td>
<td>mc</td>
</tr>
<tr>
<td>1</td>
<td>33</td>
<td>1</td>
<td>33</td>
</tr>
<tr>
<td>3</td>
<td>32</td>
<td>2</td>
<td>32</td>
</tr>
<tr>
<td>5</td>
<td>26</td>
<td>3</td>
<td>26</td>
</tr>
<tr>
<td>4</td>
<td>30</td>
<td>6</td>
<td>30</td>
</tr>
<tr>
<td>7</td>
<td>28</td>
<td>7</td>
<td>28</td>
</tr>
<tr>
<td>2</td>
<td>31</td>
<td>8</td>
<td>31</td>
</tr>
<tr>
<td>6</td>
<td>31</td>
<td>9</td>
<td>31</td>
</tr>
<tr>
<td>8</td>
<td>22</td>
<td>10</td>
<td>22</td>
</tr>
</tbody>
</table>

Note: 'mc' represents 'multiple-choice' and 'cr' represents 'constructed-response'.

The five weighting methods along with the rationale for the component/composite score calculations will be expanded upon here using booklet eight as an example. In each case, the scores were standardized so that the composite score totaled 100 points.

Method One: Score Points as Weight (SCPT)

The first method of forming a composite score was score points as weight. The maximum total raw score for each component was used to dictate the weights. For
example, in the multiple-choice component of booklet eight, there are 22 dichotomously-scored items and therefore a maximum of 22 score points. For the constructed-response component of the same booklet, there are 10 polytomously-scored items, with a maximum of 211 score points in total. For both components, the total number of raw score points for this booklet is $22 + 211 = 233$. It can be calculated that for booklet eight, $\frac{22}{233} = 9\%$ of the composite score came from the multiple-choice component score, and $\frac{211}{233} = 91\%$ came from the constructed-response component score. Thus,

$$\frac{mc}{22} = X_1 \div 9 \quad \frac{cr}{211} = X_2 \div 91$$

$$X_1 = mc \left( \frac{9}{22} \right) \quad X_2 = cr \left( \frac{91}{211} \right)$$

$$X_1 = mc \left( .41 \right) \quad X_2 = cr \left( .43 \right)$$

Composite Score ($Y$) = $X_1 + X_2$

$$Y = mc(.41) + cr(.43)$$

**Method 2: Number of Items as Weight (ITEM)**

The second method of forming a composite score was number of items as weight. For booklet eight, 22 multiple-choice items and 10 constructed-response items make up the booklet, hence a total of 32 items. Consequently, for booklet eight, $\frac{22}{32} = 69\%$ of the composite score came from the multiple-choice component score, and $\frac{10}{32} = 31\%$ came from the constructed-response component score. Thus,

$$\frac{mc}{22} = X_1 \div 69 \quad \frac{cr}{211} = X_2 \div 31$$

$$X_1 = mc \left( \frac{69}{22} \right) \quad X_2 = cr \left( \frac{31}{211} \right)$$

$$X_1 = mc \left( 3.1 \right) \quad X_2 = cr \left( .15 \right)$$

Composite Score ($Y$) = $X_1 + X_2$

$$Y = mc(3.1) + cr(.15)$$
Method 3: Allocated Time Per-Component as Weight (TIME)

The third method of forming a composite score was time per component as weight. For booklet eight, the time allotted the participants for the entire booklet is reported as being 90 minutes. Because approximately one half of the test is made of science items, I assumed that 45 minutes were allotted for the science items. TIMSS test designers assumed that one minute would be spent on each multiple-choice item (Adams & Gonzalez, 1996). Using the one-minute-per-multiple-choice-item guideline, it can be calculated that for booklet eight, 22 of the 45 science-related minutes, about 49%, were spent on multiple-choice items; and 45 - 22 = 23 minutes, about 51%, were spent on the constructed-response items. Consequently, 49% of the composite score came from the multiple-choice component score, and 51% came from the constructed-response component score. Thus,

\[ \frac{mc}{22} = X_1 / 49 \quad \frac{cr}{211} = X_2 / 51 \]

\[ X_1 = mc (49/22) \quad X_2 = cr (51/211) \]

\[ X_1 = mc (2.2) \quad X_2 = cr (.24) \]

Composite Score (Y) = X_1 + X_2,

\[ Y = mc(2.2) + cr(.24) \]

Method Four: Equal Weight (EQUAL)

The fourth method of forming a composite score was equal weight. For booklet eight, 50% of the composite score came from the multiple-choice component score, and 50% came from the constructed-response component score. Thus,

\[ \frac{mc}{22} = X_1 / 50 \quad \frac{cr}{211} = X_2 / 50 \]

\[ X_1 = mc (50/22) \quad X_2 = cr (50/211) \]
\[ X_1 = mc(2.27) \quad X_2 = cr(.24) \]

Composite Score \( (Y) = X_1 + X_2 \)

\[ Y = mc(2.27) + cr(.24) \]

**Method Five: Equal Variance (Z-SC)**

The fifth method of forming a composite score was equalizing the variances of the multiple-choice and constructed-response components. This was done for all eight booklets by computing the z-scores for each raw component score, multiplying it by the respective raw component score (either multiple-choice or constructed-response), and then linearly adding the two weighted scores to form a composite score. Thus, the composite score = \((\text{raw multiple-choice score}) \times (\text{standardized value of raw multiple-choice score}) + (\text{raw constructed-response score}) \times (\text{standardized value of raw constructed-response score})\).

Z-scores are continuous and have equal units. To calculate the z-scores, a researcher must first subtract the mean score of the total group from the individual’s raw score \((X - \overline{X})\) where \(X\) is the individual’s raw score, and \(\overline{X}\) is the mean score or the total group. Next, a researcher must divide the previous result by the standard deviation of the group’s scores,

\[
 sd_x = \hat{s} = \frac{\sum (x - \overline{x})^2}{N - 1} \tag{16}
\]

where \(sd_x\) is the standard deviation of the group’s scores. This will result in a z-score \(z = (X - \overline{X}) / sd_x\). \tag{17}
3.2.2 Dependent Variables

There were two dependent variables in this study: (a) the calculated reliability estimates, and (b) the effect sizes between the un-weighted mean reliability estimates and the corresponding weighted composite score mean reliability estimates. The Feldt-Raju technique (see formula 15 on page 52 of this thesis) was used to calculate the reliability estimates of both the un-weighted and the weighted composite scores for each condition dictated by sample size, test mixture, and weighting method.

3.3 Criteria for Comparison

The mean reliability estimates of the un-weighted composite scores were used as the benchmark from which the deviation of the corresponding mean reliability estimates of the five ad-hoc weighting methods was calculated. Symbolically, the reliability estimates of the un-weighted scores were subtracted from the reliability estimates of the weighted scores, meaning that positive deviations signified artificially inflated reliability estimates, and negative deviations signified artificially deflated reliability estimates.

Thus,

$$F-R_{\text{weighted}} - F-R_{\text{un-weighted}} = \Delta$$

where $F-R$ represents the Feldt-Raju reliability estimates, and $\Delta$ represents the difference between the two. In actuality, it was "effect sizes" that were calculated and used to signify inflation and deflation of the reliability estimates. This was done by first calculating the t-ratio for a sample population mean,

$$t = \frac{\bar{X} - \mu_{hp}}{S_x}$$

and then solving for the effect size in terms of Cohen's $d$, 

18

19
\[ d = t \times \frac{z}{\sqrt{df}} \] (Rosenthal, 1991).

The comparison of weighting methods was made by examining the mean, median, and range of the resulting effect size calculations for each weighting method. The closer the mean/median effect size was to zero, the less biased the corresponding method. Also, the smaller the range of effect sizes, the more consistent the corresponding method.

Minimal deviation from the un-weighted mean reliability estimates was most desirable because both artificial inflation and deflation are undesirable characteristics of a weighting method. Signed deviations were reported for greater precision in the discussion of inflated and deflated reliability estimates.

### 3.4 Analysis and Procedures

I obtained the TIMSS data set from their data compact disc and selected the Population 2 science-achievement data. For each of the eight test booklets, or test mixtures, I calculated the number of items, score points, and time-to-complete for both the multiple-choice and constructed-response components (see Table 3). Then I calculated the mc-cr item ratios, score point ratios, and time ratios (see Table 2).

Using a data-sampling program (designed by Dr. Nand Kishor), I created a pool of random samples of the various proposed sample sizes (15, 30, and 45) from the real TIMSS data set. From these samples, I calculated the Feldt-Raju reliability estimates for the un-weighted and the weighted composite scores of each condition. I checked for a statistically significant sample size effect using a one-way ANOVA; I decided a priori that if there was no statistically significant effect, then I would just use data generated using the largest sample size (n=45) to take advantage of the more stable variability.
I calculated the mean reliability estimates for each weighting method condition and then compared them to the corresponding mean un-weighted reliability estimates in an effort to evaluate the degree of method-related artificial inflation or deflation.

I checked for a "methods" effect using a one-way repeated-measures ANOVA. I made an a priori decision that if there were a statistically significant omnibus F test result, I would run a pairwise comparison on each of the ten possible pairs using the Holm's sequential Bonferroni procedure (Howell, 1997).

While in search for a method to measure the amount each weighting method would distort the rank ordering of examinees scores, I realized that the rank order is strongly influenced by the actual reliability estimate of the unweighted composite scores. With this, I realized that I could not separate the reliability of unweighted composite scores from the correlation between unweighted and weighted composite scores for each weighting method in an effort to report findings about a weighting method's resulting distortion of examinees' scores rank order.

Finally, I ranked the weighting methods in terms of their reliability estimates' deviations from the corresponding un-weighted reliability estimates. This ranking was done based on the calculated effect sizes.
CHAPTER FOUR

Results

This chapter provides a detailed description of the results of the present study. Most of the results shown here are directly related to the research question and the procedures of analysis, but some are shown for interest sake and will be discussed in greater detail in chapter five.

The calculated reliability estimates for each component score, shown separately for each booklet, is presented below in Table 4. The contents of this table would enable a researcher to further examine the competing reliability theories proposed by Wainer and Thissen (1993), and Birenbaum and Tatsuoka (1987) as described in chapter two of the present study.

Table 4

Reliability Estimates of Component Scores

<table>
<thead>
<tr>
<th>Item Ratio (Booklet)</th>
<th>mc KR-20</th>
<th># mc items</th>
<th>cr alpha</th>
<th># cr items</th>
<th># cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>33:1 (1)</td>
<td>.74</td>
<td>33</td>
<td>n/a</td>
<td>1</td>
<td>2095</td>
</tr>
<tr>
<td>16:1 (3)</td>
<td>.78</td>
<td>32</td>
<td>.22</td>
<td>2</td>
<td>2073</td>
</tr>
<tr>
<td>8.7:1 (5)</td>
<td>.75</td>
<td>26</td>
<td>.41</td>
<td>3</td>
<td>2069</td>
</tr>
<tr>
<td>5:1 (4)</td>
<td>.77</td>
<td>30</td>
<td>.41</td>
<td>6</td>
<td>2071</td>
</tr>
<tr>
<td>4:1 (7)</td>
<td>.75</td>
<td>28</td>
<td>.51</td>
<td>7</td>
<td>2060</td>
</tr>
<tr>
<td>3.9:1 (2)</td>
<td>.81</td>
<td>31</td>
<td>.54</td>
<td>8</td>
<td>2084</td>
</tr>
<tr>
<td>3.4:1 (6)</td>
<td>.76</td>
<td>31</td>
<td>.51</td>
<td>9</td>
<td>2065</td>
</tr>
<tr>
<td>2.2:1 (8)</td>
<td>.63</td>
<td>22</td>
<td>.63</td>
<td>10</td>
<td>2064</td>
</tr>
</tbody>
</table>

Note: 'KR-20' refers to the Kuder-Richardson technique for the estimation of reliability, and 'alpha' refers to Coefficient Alpha technique for the estimation of reliability. "Mc" and "cr" represent "multiple-choice" and "constructed-response" respectively.
For example, in Table 4 for booklet one, the KR-20 reliability estimate for the multiple-choice component score was .74; the number of multiple choice items was 33; the Coefficient Alpha reliability estimate for the constructed-response items was not available because the number of constructed-response items was 1; finally for this example, the number of cases for booklet one was 2095. This will be discussed in greater detail in chapter five.

The Analysis of Variance for sample size effect produced a statistically non-significant effect: F(2,241) = .516, Mse = 0.103, p = .598. Therefore, as decided upon a priori in chapter three of the present study, only data generated using the largest sample size (n=45) was used in an effort to use data with the most stable variability. The reliability estimates for both the un-weighted (raw) composite scores and the weighted composite scores are presented in Table 5.

Table 5
Mean Reliability Estimates for Raw Score and Weighted Composites Sorted by Booklet

<table>
<thead>
<tr>
<th>Booklet</th>
<th><em>raw_scores</em></th>
<th>Raw</th>
<th>Weighted Via</th>
<th>Method</th>
<th># of reps</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>SCPT (1)</td>
<td>ITEM(2)</td>
<td>TIME (3)</td>
</tr>
<tr>
<td>1</td>
<td>.71(.07)</td>
<td>.52(.12)</td>
<td>.98(.01)</td>
<td>.53(.12)</td>
<td>.56(.12)</td>
</tr>
<tr>
<td>3</td>
<td>.58(.10)</td>
<td>.49(.13)</td>
<td>.79(.13)</td>
<td>.47(.13)</td>
<td>.48(.13)</td>
</tr>
<tr>
<td>5</td>
<td>.62(.09)</td>
<td>.63(.13)</td>
<td>.75(.10)</td>
<td>.59(.14)</td>
<td>.61(.14)</td>
</tr>
<tr>
<td>4</td>
<td>.56(.13)</td>
<td>.79(.07)</td>
<td>.81(.06)</td>
<td>.72(.08)</td>
<td>.68(.09)</td>
</tr>
<tr>
<td>7</td>
<td>.62(.09)</td>
<td>.83(.05)</td>
<td>.81(.05)</td>
<td>.73(.07)</td>
<td>.72(.08)</td>
</tr>
<tr>
<td>2</td>
<td>.65(.09)</td>
<td>.83(.06)</td>
<td>.82(.06)</td>
<td>.77(.07)</td>
<td>.73(.08)</td>
</tr>
<tr>
<td>6</td>
<td>.62(.10)</td>
<td>.82(.07)</td>
<td>.80(.07)</td>
<td>.75(.09)</td>
<td>.70(.10)</td>
</tr>
<tr>
<td>8</td>
<td>.68(.07)</td>
<td>.91(.03)</td>
<td>.76(.06)</td>
<td>.73(.07)</td>
<td>.73(.07)</td>
</tr>
</tbody>
</table>

Note: Parentheses indicates standard deviations. Each mean was calculated from the corresponding number of replications ( # of reps) taken from a booklet. Sample size was 45.
For example, in Table 5 for booklet eight, the sample size was 45, and the number simulated replications (simulated random samples) was 46. For this example, the \textit{unweighted} composite score mean reliability estimate was $F-R = 0.68$ and the standard deviation of this estimate was 0.07. The \textit{weighted} composite score mean reliability estimate for weighting-method one was $F-R = 0.91$, with a standard deviation of 0.03.

A One-Way Repeated Measures ANOVA was performed for each booklet to find out whether there was a \textit{weighting methods} effect on the mean reliability estimates (see Table 6). Notice that for every booklet, the F value was statistically significant at the $p<0.01$ level.

Table 6

\textbf{One-Way Repeated-Measures ANOVA Results for the Methods Effect}

<table>
<thead>
<tr>
<th>Booklet</th>
<th>Wilks' Lambda</th>
<th>F</th>
<th>Hypothesis df</th>
<th>Error df</th>
<th>P (2-tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.012</td>
<td>830.10</td>
<td>4</td>
<td>41</td>
<td>0.000</td>
</tr>
<tr>
<td>2</td>
<td>0.018</td>
<td>566.57</td>
<td>4</td>
<td>42</td>
<td>0.000</td>
</tr>
<tr>
<td>3</td>
<td>0.028</td>
<td>358.79</td>
<td>4</td>
<td>41</td>
<td>0.000</td>
</tr>
<tr>
<td>4</td>
<td>0.012</td>
<td>872.99</td>
<td>4</td>
<td>42</td>
<td>0.000</td>
</tr>
<tr>
<td>5</td>
<td>0.016</td>
<td>644.48</td>
<td>4</td>
<td>41</td>
<td>0.000</td>
</tr>
<tr>
<td>6</td>
<td>0.019</td>
<td>521.41</td>
<td>4</td>
<td>41</td>
<td>0.000</td>
</tr>
<tr>
<td>7</td>
<td>0.041</td>
<td>239.71</td>
<td>4</td>
<td>41</td>
<td>0.000</td>
</tr>
<tr>
<td>8</td>
<td>0.020</td>
<td>502.52</td>
<td>4</td>
<td>41</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Because there was a statistically significant omnibus F, I performed Holm's Sequential Bonferroni procedure to examine the pairwise comparisons and found that for every booklet, every possible pair was statistically significantly different (see Table 7).
For example in Table 7, the pairwise comparison of all weighting-method combinations yielded statistically significant t-values for each booklet, e.g. for booklet one, for method pair SCPT/ITEM, $t = -27.4$ is statistically significant with $p<.05$, with 44 or 45 degrees of freedom.

Table 7

<table>
<thead>
<tr>
<th>Method Pair</th>
<th>t-value for each Booklet</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCPT/TIME</td>
<td>-7.07</td>
<td>15.74</td>
<td>7.55</td>
<td>16.55</td>
<td>23.94</td>
<td>12.30</td>
<td>23.77</td>
<td>28.46</td>
<td></td>
</tr>
<tr>
<td>SCPT/EQUAL</td>
<td>-18.31</td>
<td>24.36</td>
<td>16.00</td>
<td>26.46</td>
<td>27.99</td>
<td>22.79</td>
<td>27.26</td>
<td>28.42</td>
<td></td>
</tr>
<tr>
<td>SCPT/Z-SC</td>
<td>2.56</td>
<td>24.60</td>
<td>9.87</td>
<td>26.86</td>
<td>19.59</td>
<td>24.16</td>
<td>27.30</td>
<td>28.31</td>
<td></td>
</tr>
<tr>
<td>ITEM/TIME</td>
<td>26.60</td>
<td>25.21</td>
<td>28.88</td>
<td>27.46</td>
<td>22.85</td>
<td>25.39</td>
<td>25.38</td>
<td>14.75</td>
<td></td>
</tr>
<tr>
<td>ITEM/EQUAL</td>
<td>26.47</td>
<td>23.38</td>
<td>27.29</td>
<td>26.16</td>
<td>19.69</td>
<td>24.24</td>
<td>22.91</td>
<td>15.41</td>
<td></td>
</tr>
<tr>
<td>ITEM/Z-SC</td>
<td>27.42</td>
<td>23.09</td>
<td>28.53</td>
<td>25.88</td>
<td>24.56</td>
<td>23.02</td>
<td>22.85</td>
<td>16.73</td>
<td></td>
</tr>
<tr>
<td>TIME/Z-SC</td>
<td>9.68</td>
<td>19.03</td>
<td>2.15</td>
<td>19.10</td>
<td>6.26</td>
<td>19.20</td>
<td>12.42</td>
<td>3.15</td>
<td></td>
</tr>
<tr>
<td>EQUAL/Z-SC</td>
<td>16.14</td>
<td>1.95</td>
<td>6.16</td>
<td>1.92</td>
<td>14.23</td>
<td>3.85</td>
<td>.90</td>
<td>2.26</td>
<td></td>
</tr>
</tbody>
</table>

Note: $p< .05$ (2-tailed) applies to every t-value in this table. Sample size ranged from 44-45.

The zero-order correlation coefficients between the mean reliability estimates of the different weighted composite scores and the corresponding mean reliability estimates of the un-weighted composite scores were calculated and found to be statistically significant for every booklet (see Tables 8). This strength of association indicated that the rank order for the composite scores would be similar for each method in every booklet.
Table 8

Correlation of Raw (Un-weighted) Composite Scores for Each Booklet with Corresponding Weighted Composite Scores, By Methods

<table>
<thead>
<tr>
<th>Booklet (N)</th>
<th>Method</th>
<th>SCPT (1)</th>
<th>ITEM (2)</th>
<th>TIME (3)</th>
<th>EQUAL (4)</th>
<th>Z-SC (5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (46)</td>
<td></td>
<td>.909</td>
<td>.887</td>
<td>.931</td>
<td>.873</td>
<td>.914</td>
</tr>
<tr>
<td>2 (47)</td>
<td></td>
<td>.662</td>
<td>.480</td>
<td>.522</td>
<td>.590</td>
<td>.594</td>
</tr>
<tr>
<td>3 (46)</td>
<td></td>
<td>.534</td>
<td>.322</td>
<td>.542</td>
<td>.545</td>
<td>.544</td>
</tr>
<tr>
<td>4 (47)</td>
<td></td>
<td>.595</td>
<td>.400</td>
<td>.456</td>
<td>.515</td>
<td>.519</td>
</tr>
<tr>
<td>5 (46)</td>
<td></td>
<td>.704</td>
<td>.651</td>
<td>.729</td>
<td>.736</td>
<td>.731</td>
</tr>
<tr>
<td>6 (46)</td>
<td></td>
<td>.731</td>
<td>.577</td>
<td>.602</td>
<td>.657</td>
<td>.669</td>
</tr>
<tr>
<td>7 (46)</td>
<td></td>
<td>.774</td>
<td>.607</td>
<td>.671</td>
<td>.708</td>
<td>.708</td>
</tr>
<tr>
<td>8 (46)</td>
<td></td>
<td>.695</td>
<td>.565</td>
<td>.627</td>
<td>.625</td>
<td>.620</td>
</tr>
</tbody>
</table>

Note: Correlations are all statistically significant at the .01 level (2-tailed); 'N' represents the number of random samples. Sample size = 45.

Just as a reminder, there were two dependent variables in this study: (a) the reliability estimates and (b) the deviation (effect sizes) between the un-weighted composite score mean reliability estimates and the corresponding weighted composite score reliability estimates. The mean reliability estimates of the un-weighted composite scores were used as the benchmark from which the deviation of the corresponding mean reliability estimates of the five weighted composite scores was calculated. The reliability estimates of the weighted composite scores were subtracted from the reliability estimates of un-weighted composite scores, meaning that positive deviations signified artificially inflated reliability estimates, and negative deviations signified artificially deflated
reliability estimates. Thus,

$$F-R_{\text{un-weighted}} - F-R_{\text{weighted}} = \Delta$$

Where $F-R$ represents the Feldt-Raju reliability estimates, and $\Delta$ represents the difference between the two. The actual effect sizes were calculated by first calculating the t-ratio for a single population mean,

$$t = \frac{\bar{X} - \mu_{\text{hyp}}}{S_x}$$

and then solving for the effect size in terms of Cohen's $d$,

$$d = t \times \frac{2}{\sqrt{df}}.$$ 

See Table 9 for a summary of the effect size calculations. For booklet one, for example, 46 samples were drawn; the calculated t-value for the difference in mean reliability estimates of the weighted and un-weighted composite scores for SCPT method was 2.71 and the resulting effect size was -.81.
Table 9

Deviation of Mean Reliability Estimates for Each Composite Method Sorted By Booklet

<table>
<thead>
<tr>
<th>Booklet</th>
<th>runs</th>
<th>df</th>
<th>t</th>
<th>d</th>
<th>t</th>
<th>d</th>
<th>t</th>
<th>d</th>
<th>t</th>
<th>d</th>
<th>t</th>
<th>d</th>
<th>t</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCPT (1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ITEM (2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TIME (3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EQUAL (4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z-SC (5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>46</td>
<td>45</td>
<td>2.71</td>
<td>.81</td>
<td>-3.9</td>
<td>1.15</td>
<td>2.57</td>
<td>.75</td>
<td>2.14</td>
<td>.62</td>
<td>2.71</td>
<td>.79</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>46</td>
<td>45</td>
<td>.90</td>
<td>.27</td>
<td>-2.1</td>
<td>-.63</td>
<td>1.10</td>
<td>.32</td>
<td>1.00</td>
<td>.29</td>
<td>1.10</td>
<td>.32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>46</td>
<td>45</td>
<td>-.11</td>
<td>.03</td>
<td>-1.4</td>
<td>-.43</td>
<td>.33</td>
<td>.10</td>
<td>.11</td>
<td>.03</td>
<td>.33</td>
<td>.10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>46</td>
<td>45</td>
<td>-1.77</td>
<td>-.53</td>
<td>-1.9</td>
<td>-.57</td>
<td>-1.23</td>
<td>.36</td>
<td>-.92</td>
<td>-.27</td>
<td>-.92</td>
<td>-.27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>46</td>
<td>45</td>
<td>-2.33</td>
<td>-.69</td>
<td>-2.1</td>
<td>-.61</td>
<td>-1.22</td>
<td>.35</td>
<td>-1.11</td>
<td>-.32</td>
<td>-1.11</td>
<td>-.32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>47</td>
<td>46</td>
<td>-2.00</td>
<td>-.59</td>
<td>-1.9</td>
<td>-.55</td>
<td>-1.33</td>
<td>-.39</td>
<td>-.89</td>
<td>-.18</td>
<td>-.89</td>
<td>-.26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>46</td>
<td>45</td>
<td>-2.00</td>
<td>-.60</td>
<td>-1.8</td>
<td>-.52</td>
<td>-1.30</td>
<td>-.38</td>
<td>-.80</td>
<td>-.23</td>
<td>-.80</td>
<td>-.23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>46</td>
<td>45</td>
<td>-3.29</td>
<td>-.98</td>
<td>-1.1</td>
<td>-.33</td>
<td>-.71</td>
<td>-.21</td>
<td>-.71</td>
<td>-.21</td>
<td>-.71</td>
<td>-.21</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: 'runs' represents the number of random samples used via the simulation program; 'df' represents the degrees of freedom; 't' represents the t-value (2-tailed); and 'd' represents Cohen's d, the effect size.

The comparison of weighting methods was made by examining both the mean and the range of the resulting effect sizes for each weighting method. The closer the mean effect size was to zero, the less biased the corresponding method was judged to be. Also, the smaller the range of effect sizes, the more consistent the corresponding method was judged to be. Minimal deviation (from the un-weighted mean reliability estimates) in either direction of the effect sizes was most desirable, because both artificial inflation and deflation are undesirable characteristics. In Figure 1, a "boxplot" graph presented a comparison of the mean effect sizes for all deviations examined. Also, corresponding descriptive statistics are shown in Table 10.

An explanation of "Box-Plot and Whisker" graphs are given: Each "Box" in the
graph extends from the 25th to the 75th percentile. The thick line inside the box represents the median. The thin lines extending from the ends of the boxes called "whiskers" extend to the largest and smallest observed values within 1.5 box lengths. If there are any outlying values, they will be reported.

![Box plot](image)

**Figure 1.** Effect sizes (vertical axis) corresponding to the weighting methods (horizontal axis)

**Table 10**

<table>
<thead>
<tr>
<th>Method</th>
<th>N</th>
<th>Min</th>
<th>Max</th>
<th>Range</th>
<th>Mean</th>
<th>sd</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCPT (1)</td>
<td>8</td>
<td>-.98</td>
<td>.81</td>
<td>1.79</td>
<td>-.29</td>
<td>.59</td>
<td>-.57</td>
</tr>
<tr>
<td>ITEM (2)</td>
<td>8</td>
<td>-1.15</td>
<td>-.33</td>
<td>.82</td>
<td>-.60</td>
<td>.24</td>
<td>-.57</td>
</tr>
<tr>
<td>TIME (3)</td>
<td>8</td>
<td>-.39</td>
<td>.75</td>
<td>1.14</td>
<td>-.07</td>
<td>.41</td>
<td>-.30</td>
</tr>
<tr>
<td>EQUAL (4)</td>
<td>8</td>
<td>-.32</td>
<td>.62</td>
<td>.94</td>
<td>-.03</td>
<td>.33</td>
<td>-.20</td>
</tr>
<tr>
<td>Z-SC (5)</td>
<td>8</td>
<td>-.32</td>
<td>.79</td>
<td>1.11</td>
<td>-.01</td>
<td>.39</td>
<td>-.23</td>
</tr>
</tbody>
</table>
The distribution of the *equal weight* (EQUAL) weighting-method effect sizes had
the second least variability, the median closest to zero, and the mean second closest to
zero. The distribution of *item weight* (ITEM) weighting-method effect sizes had the least
variability, but its median and mean were furthest from zero. Firstly, from the results of
Table 10, I declared *equal weight* (EQUAL) weighting method as the best method overall
because it resulted in reliability estimates that were the least inflated or deflated.
Secondly, from the results of Table 10, I declared *item weight* (ITEM) weighting method
as the worst method overall because it resulted in reliability estimates that were the most
deflated. A greater discussion of the results follows in chapter five of the present study.

In summary of the results, there were three main points to highlight. First, the
Analysis of Variance for "sample size" effect produced a statistically non-significant
effect, therefore only the largest sample size (n=45) was used in an effort to use data with
the most stable variability. Second, because there was a statistically significant omnibus
F, Holm's Sequential Bonferroni procedure was performed to examine the pairwise
comparisons and found that for every booklet, every possible pair was statistically
significantly different. Third, based on the deciding criteria of mean reliability estimate
inflation/deflation, I declared *equal weight* (EQUAL) weighting method as the best
method overall, and *item weight* (ITEM) weighting method as the worst method overall.
CHAPTER FIVE

Discussion

5.1 An Introduction to the Conclusion

Welcome to chapter five. You made it through the first four chapters and you are still reading. You should be delighted to know that things do get a lot less complicated in this fifth and final chapter. The purpose of this final chapter was to bring all of the research-related questions that were developed in chapter one of the present study to a well-developed and well-discussed easy-to-understand close. This was mainly done so that teachers and administrators, the practitioners who were targeted as the main beneficiaries of the present study, would be able to understand and reap the rewards of the research findings as well as understand their limitations.

The format of this chapter is straightforward. The educational-research-based problem identified in chapter one, and the purpose for seeking its solution is restated in lay terms. Then, the research question born from this problem is answered and discussed in lay terms. Finally, the limitations of the study as well as ideas for further research are discussed.

5.2 Research Problem and Purpose

The research problem identified in chapter one of the present study was based on the fact that teachers and administrators often weight and combine component scores from varied item formats to create composite scores without knowing anything about the research-based merit of doing so. Sometimes teachers and administrators weight the component scores before combining them in an effort to increase the validity of the inferences made from the resulting composite scores. Increasing the validity of the
resulting composite scores is a good thing, but not if it is at the expense of creating composite scores that are artificially deflated or inflated in terms of reliability. This deflation or inflation in reliability estimates could lead to composite scores that are distorted. This in turn could lead to unfair composite-score related consequences for all examinees involved.

The purpose of the present study was to discover which component-score weighting method produced the least amount of artificial deflation or inflation in the reliability estimates of the resulting composite test scores. This was examined in the specific case when two component scores were weighted and combined, where one component was made of multiple-choice items and the other was made of constructed-response items. Once the best component-score weighting method was identified, it could be shared with the teaching and administrative communities in an effort to increase the accuracy of composite scores reported for examinees.

5.3 Answers to Research Question

The answer to the research question will come after some discussion of the independent variables (sample size, test mixture, and weighting method), used in the present study. This is done in an effort to reveal the depth and breadth of the answer to such a complex research question.

5.3.1 Independent Variables

Sample Size

For the present study, sample sizes 15, 30, and 45 were selected to represent typical class sizes in Canada. One of the reasons for examining three different sample sizes was to explore the hypothesis that this independent variable would have influence
on the composite score reliability estimates, meaning that the reliability of test scores will
be a function of the number of students in a classroom taking the test. Contrary to my
initial intuitive-based instinct, no statistically significant difference in the composite
score reliability estimates was found between the different sample sizes. This means that
for all three sample sizes explored, the reliability estimates were almost the same. The
one difference that was found between the different sample sizes was that the standard
deviation for the sample sizes got smaller as the sample size increased. This indicated
that a more stable variability existed as the class size increased. Because of this, the
composite score reliability estimates calculated with the largest of the three sample sizes,
45, were used for the analysis in an effort to use data with the most stable variability.

**Test mixture**

In the present study, "test mixture" referred to the numerical combination of
multiple-choice and constructed-response items that formed each composite score.
Because TIMSS data was used, it wasn’t possible for me to impose combinations.
Instead, each item combination was dictated by the "test mixtures" corresponding to each
of the eight TIMSS booklets. Hence, the concept of "test mixture" was referred to in the
present study as **test mixture, booklet and/or item ratio**. From Tables 2 to 5, the test
mixtures were presented in item-ratio rank order, starting with booklet one with an item
ratio of 33 multiple-choice items to 1 constructed-response item (33:1), ending with
booklet eight with an item ratio of 2.2 multiple-choice items to 1 constructed-response
item (2.2:1).

The various test mixtures examined in this study did have an effect on the range
of deviation scores calculated from the F-R reliability estimates of the un-weighted and
weighted test scores. This means that for each composite score weighting method, the reliability estimates were influenced by the test mixture. Because the purpose of the present study was to discover which of the teacher-designed ad-hoc methods of weighting and combining multiple-choice and constructed-response component scores was best, it was the range of reliability estimates across a variety of test mixtures that was considered a part of the evaluative nature of the weighting methods. The smallest range would be indicative of the greatest resistance to the influences of test mixture; that was a positive thing. From Figure 1, it could be seen that the range was the greatest for the score points as weights (SCPT) weighting method, and least for the number of items as weight (ITEM) weighting method. But this wasn’t enough to make a decision on which weighting method was the best because the effect-size deviation from zero of each weighting method also has to be factored in the conclusion about best weighting method.

To understand why score points as weights (SCPT) weighting method had the largest range in reliability estimates and number of items as weight (ITEM) weighting method had the smallest range, is an interesting task that will not be examined in the present study. No hypotheses were made about such queries and from the existing calculated results, not explanation can be made. Further research designed to examine this query is justified.

In the results section of the present study, the reliability estimates of the multiple-choice and constructed-response component scores were calculated separately before they were weighted and/or combined (see Table 4). It was found that the reliability estimates calculated for each of the eight test mixtures were dependent on the number of items in the particular component. This finding simply replicates a well documented
finding that reliability estimates are a function of the *number of items* (Lord & Novick, 1968).

**Weighting Method**

All possible composite score weighting-method comparisons were found to be statistically significantly different from each other in terms of the mean reliability estimates that were produced. This means that for every test mixture (booklet), the choice of weighting method had enough influence on the resulting composite score reliability estimates to produce statistically significantly different reliability estimates. For this reason, it was safe to assume that I could examine each test mixture for each weighting method as a separate and unique data point. Using statistical procedures, it was established that the choice of weighting method used to create the composite scores did have an effect on artificially inflating or deflating the weighted composite score reliability estimates.

For the *equal weight* (EQUAL) weighting-method, weighted composite scores were produced that had reliability estimates that were the least inflated or deflated from the original un-weighed composite score F-R reliability estimates. The EQUAL method also had nearly the smallest range of effect-size deviation scores. Because the best method overall was judged by amount of inflation/deflation and the degree of range in effect-size deviation scores, the EQUAL method came out on top.

For the *score points as weight* (SCPT) weighting-method, weighted composite scores were produced that had reliability estimates that were the most inflated or deflated from the original un-weighted composite score F-R reliability estimates. The SPCT
method also had the largest range of effect-size deviation scores. Hence, it was found to be the worst method overall.

Comments on the reasons for the ITEM weighing method being the best and SCPT weighting method being the worst are minimal. The present study was an exploratory one. No hypothesis was made as to which method would be the best. A solid rationale for undertaking this study was proposed without offering a hypothesis about the findings. Further research with fewer limitations may allow for a greater understanding. Regardless of the mystery that remains, the findings that were reported (with limitations made clear) do offer teachers and test administrators some research-based information from which to make decisions regarding the choice of a teacher-designed ad-hoc weighting method for the combining and weighting of mixed-item component scores.

5.3.2 Findings in Context

The primary finding of the present study was simply the answer to the research question posed in chapter one. The secondary findings are the pieces of evidence that either support or defeat the related research findings of others in the field of educational measurement that were described in chapter two of the present study.

Primary Findings

The development of the present study was an evolutionary consequence of my thoughts about what was important in educational research, and the base of related research literature. The research question that was posed in the present study was an original one rather than a replication of another’s research. I chose to answer a research question not yet answered in the research literature. Like all primary research studies, the
findings of my study will eventually need replicating before they may be deemed as theoretically sound by my peers in the field of educational researcher.

The primary findings presented in the present study offer new insight to teachers and administrators. When these teachers or administrators are faced with validity-related reasons to weight and combine component scores of different item types in an effort to form a more meaningful composites score, they can now base their decisions on some research findings rather than on intuition. If teachers and administrators examine and adapt the findings of the present study before they impose an ad-hoc weighting method, they will likely make weighting method decisions that will result in composite scores that are near in reliability to the un-weighted composite score. This will not only affect the reliability of the reported weighted composite scores, but also the rank order of the examinees' scores. The adaptation of the findings would be most desirable because the calculated reliability estimate of the weighted composite score should not deviate from the calculated reliability estimate of the un-weighted composite score if one is to claim that the weighting method does not artificially inflate or deflate the reliability estimates of the composite scores.

Teachers might also find it interesting that the size of the class (15, 30, or 45) did not have a statistically significant effect on the results, meaning that the results were the same for all three class sizes examined. It should be noted once again that the findings of the present study were all based on scores from the academic domain of science, and transferability of findings across academic domains has yet to be explored. With regards to the ad-hoc weighting method for an academic domain other than science, I would recommend that teachers and administrators stick with their intuition rather than make
un-warranted generalizations from science to other academic domains. I am hopeful that further research examining other academic domains in a similar manner will eventually yield generalizable results.

**Secondary Findings**

Some *secondary* findings of this present study did not offer support for the hypotheses of some of the researchers discussed in the present study. For example, Birenbaum and Tatsuoka's (1987) had hypothesized that the reliability of a composite score would be at a maximum if the multiple-choice and constructed-response components were near equal and as high as possible with regards to number of items. If this were true, the raw-score reliability estimates would perform best when the multiple-choice: constructed-response item ratio revealed near-equality between the two item formats (booklet 8, mc: cr =2.2:1) and worst when the ratio reveals the greatest disequilibrium (booklet 1, mc: cr =33:1). The raw-score reliability estimate results as seen in Table 5 show that this hypothesis is not supported because booklet 8, which was hypothesized to perform the best, actually performed second best with a mean reliability estimate of .68. And booklet 1, hypothesized to perform worst, actually performed the best with a mean reliability estimate of .71. The hypothesis set forth by Birenbaum and Tatsuoka was not supported by the findings of the present study. These secondary findings of the present study created doubt about the accuracy of the findings of Birenbaum and Tatsuoka.

Other *secondary* findings of this present study were found to be in support of the related hypotheses of another pair of researchers, Wainer and Thissen (1993). From the research of Wainer and Thissen, it was hypothesized that for each booklet, the highest
mean reliability estimates would be calculated for the method that gave the most weight to the multiple-choice component of a composite score. For all booklets, as seen in Table 3, it was the ITEM weighing method that gave the most weight to the multiple-choice component. It was shown in Table 5 that for five of the booklets, ITEM method did produce the highest mean reliability estimates. For the other three booklets, ITEM method produced the second highest mean reliability estimates. These secondary findings of the present study created support for the accuracy of the findings of Wainer and Thissen.

5.3.3 Implications of Findings

The implications of the primary findings of the present study offer teachers and administrators some practical information with regards to research-based guidance when pondering the selection of a composite score weighting method within the academic domain of science. It should be noted that all the pondering for best weighting method will not lead to scores that are highly valid unless the unweighted composite scores are highly reliable. Implications of the secondary findings of the present study speak more to the theory of educational measurement. Theoretically, the relevant foundations laid down by Birenbaum and Tatsuoka (1987) with regards to reliability estimates of mixed item-type component scores appear to be flawed, while those of Wainer and Thissen (1993) do not. Although the primary findings do have a direct impact on the decisions of teachers, the secondary findings are more theoretical and need more attention in the educational research community before teachers and administrators are advised to pay heed.
5.4 Limitations of Study

One limitation of the present study was primarily related to the data used for the calculations of the reliability estimates. For the present study, a simulation program was run with real data to compute mean reliability estimates that were based on many runs or trials. As explained in great detail in this thesis, the real data used were from the TIMSS study. The limitation of this study was recognized when I attempted to make inferences about "teachers"' weighting methods from the TIMSS large-scale assessment data. I did not actually use data collected from teachers who used the various weighing methods examined in this present study.

Other limitations of the study were its inclusion of only one academic domain (science), its inclusion of only one grade (grade eight), its inclusion of only three possible class sizes (15, 30, and 45), and its inclusion of test mixtures in which the number of multiple-choice items was always greater than the number of constructed-response items.

5.5 Future Research

Although the present study did answer the main research question posed in chapter one, it did this with certain limitations that do limit the generalizability of the findings. Directions that interested researchers should pursue in an effort to fully understand the nature of the research question and its answer are as follows:

i) replicate the study in subject areas other than science, such as social studies or mathematics;

ii) replicate the study using different populations;

iii) explore other ad-hoc weighting methods used by teachers and/or administrators; and
iv) explore test mixtures that were not included in the TIMSS data set such as equal number of items per component, or a dominance in the number of constructed-response items in each set of multiple-choice and constructed-response components.

5.6 A Summary of the Conclusion

The main research task posed by the present study was to find out which teacher-designed ad-hoc component-score weighting method, designed for the combining of two component scores (one made of multiple-choice items and the other of constructed-response items), was the best. By examining a variety of aspects of three factors (class size, test mixture, and weighting method), an evaluative look at the general tendency for each weighing method to cause artificially deflated/inflated component-score reliability estimates was examined. It was found that the component-score weighting method that treated the components equally, called the equal weights (EQUAL) weighing method in the present study, lead to the least amount of deflation/inflation in the composite score reliability estimates. Please remember the limitations of the study: the results were calculated using test scores from the academic domain of science only, and these test scores were only gathered from Canadian grade eight students.
REFERENCES


APPENDIX A

Weighting Coefficients for Each Method with Each Test Mixture

**Booklet One**

<table>
<thead>
<tr>
<th></th>
<th>Multiple-Choice</th>
<th>Constructed-Response</th>
<th>Composite</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Score Points:</td>
<td>33 (63%)</td>
<td>19 (37%)</td>
<td>52 (100%)</td>
</tr>
<tr>
<td>Number of Items:</td>
<td>33 (97%)</td>
<td>1 (3%)</td>
<td>34 (100%)</td>
</tr>
<tr>
<td>Time (minutes):</td>
<td>33 (73%)</td>
<td>12 (27%)</td>
<td>45 (100%)</td>
</tr>
</tbody>
</table>

*Note: Raw multiple-choice scores and raw constructed-response scores are denoted as "mc" and "cr" respectively.*

**Method 1 (score points as weights):**

\[
\frac{mc}{33} = X_1/63 \quad \text{cr}/19 = X_2/37
\]

\[
X_1 = mc (63/33) \quad X_2 = cr (37/19)
\]

\[
X_1 = mc (1.9) \quad X_2 = cr (1.9)
\]

\[
\text{Composite Score} = X_1 + X_2
\]

\[
= mc(1.9) + cr(1.9)
\]

**Method 2 (number of items as weights):**

\[
\frac{mc}{33} = X_1/97 \quad \text{cr}/19 = X_2/3
\]

\[
X_1 = mc (97/33) \quad X_2 = cr (3/19)
\]

\[
X_1 = mc (2.9) \quad X_2 = cr (.16)
\]

\[
\text{Composite Score} = X_1 + X_2
\]

\[
= mc(2.9) + cr(.16)
\]

**Method 3 (time per item as weights):**

\[
\frac{mc}{33} = X_1/73 \quad \text{cr}/19 = X_2/27
\]

\[
X_1 = mc (73/33) \quad X_2 = cr (27/19)
\]

\[
X_1 = mc (2.2) \quad X_2 = cr (1.4)
\]

\[
\text{Composite Score} = X_1 + X_2
\]

\[
= mc(2.2) + cr(1.4)
\]

**Method 4 (equal weighs):**

\[
\frac{mc}{33} = X_1/50 \quad \text{cr}/19 = X_2/50
\]

\[
X_1 = mc (50/33) \quad X_2 = cr (50/19)
\]

\[
X_1 = mc (1.5) \quad X_2 = cr (2.6)
\]

\[
\text{Composite Score} = X_1 + X_2
\]

\[
= mc(2.9) + cr(2.6)
\]

**Method 5 (z-scores as weights):**

Calculated in SPSS using Statistics/Descriptives/select "save as standardized variable"/ OK.

\[
\text{Composite Score} = mc(\text{standardized value of 'rawmc'}) + cr (\text{standardized value of 'rawcr'})
\]
**Booklet Two**

<table>
<thead>
<tr>
<th></th>
<th>Multiple-Choice</th>
<th>Constructed-Response</th>
<th>Composite</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Score Points</td>
<td>31 (18%)</td>
<td>145 (82%)</td>
<td>176 (100%)</td>
</tr>
<tr>
<td>Number of Items</td>
<td>31 (79%)</td>
<td>8 (21%)</td>
<td>39 (100%)</td>
</tr>
<tr>
<td>Time (minutes)</td>
<td>31 (69%)</td>
<td>14 (31%)</td>
<td>45 (100%)</td>
</tr>
</tbody>
</table>

**Method 1** (score points as weights):

\[
\begin{align*}
X_1 &= \text{mc} \times \frac{18}{31} \\
X_2 &= \text{cr} \times \frac{82}{145} \\
\text{Composite Score} &= X_1 + X_2 \\
&= \text{mc}(0.58) + \text{cr}(0.57)
\end{align*}
\]

**Method 2** (number of items as weights):

\[
\begin{align*}
X_1 &= \text{mc} \times \frac{79}{31} \\
X_2 &= \text{cr} \times \frac{21}{145} \\
\text{Composite Score} &= X_1 + X_2 \\
&= \text{mc}(2.5) + \text{cr}(0.14)
\end{align*}
\]

**Method 3** (time per item as weights):

\[
\begin{align*}
X_1 &= \text{mc} \times \frac{69}{31} \\
X_2 &= \text{cr} \times \frac{31}{145} \\
\text{Composite Score} &= X_1 + X_2 \\
&= \text{mc}(2.2) + \text{cr}(0.21)
\end{align*}
\]

**Method 4** (equal weights):

\[
\begin{align*}
X_1 &= \text{mc} \times \frac{50}{31} \\
X_2 &= \text{cr} \times \frac{50}{145} \\
\text{Composite Score} &= X_1 + X_2 \\
&= \text{mc}(1.6) + \text{cr}(0.34)
\end{align*}
\]

**Method 5** (z-scores as weights):

Calculated in SPSS using Statistics/Descriptives/select "save as standardized variable"/ OK.

\[
\text{Composite Score} = \text{mc}(\text{standardized value of 'rawmc'}) + \text{cr} (\text{standardized value of 'rawcr')}
\]
**Booklet Three**

<table>
<thead>
<tr>
<th></th>
<th>Multiple-Choice</th>
<th>Constructed-Response</th>
<th>Composite</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Score Points:</td>
<td>32 (46%)</td>
<td>38 (54%)</td>
<td>70 (100%)</td>
</tr>
<tr>
<td>Number of Items:</td>
<td>32 (94%)</td>
<td>2 (6%)</td>
<td>34 (100%)</td>
</tr>
<tr>
<td>Time (minutes):</td>
<td>32 (71%)</td>
<td>13 (27%)</td>
<td>45 (100%)</td>
</tr>
</tbody>
</table>

**Method 1** (score points as weights):

\[
\frac{mc}{32} = X_1 \times \frac{46}{32} \\
X_1 = mc \times \frac{46}{32} \\
X_1 = mc (1.44) \\
\text{Composite Score} = X_1 + X_2 \\
= mc (1.44) + cr (1.42)
\]

**Method 2** (number of items as weights):

\[
\frac{mc}{32} = X_1 \times \frac{94}{32} \\
X_1 = mc \times \frac{94}{32} \\
X_1 = mc (2.84) \\
\text{Composite Score} = X_1 + X_2 \\
= mc (2.84) + cr (.16)
\]

**Method 3** (time per item as weights):

\[
\frac{mc}{32} = X_1 \times \frac{71}{32} \\
X_1 = mc \times \frac{71}{32} \\
X_1 = mc (2.22) \\
\text{Composite Score} = X_1 + X_2 \\
= mc (2.2) + cr (.4)
\]

**Method 4** (equal weights):

\[
\frac{mc}{32} = X_1 \times \frac{50}{32} \\
X_1 = mc \times \frac{50}{32} \\
X_1 = mc (1.56) \\
\text{Composite Score} = X_1 + X_2 \\
= mc (1.56) + cr (1.32)
\]

**Method 5** (z-scores as weights):

Calculated in SPSS using Statistics/Descriptives/select "save as standardized variable"/OK.

\[
\text{Composite Score} = mc(\text{standardized value of 'rawmc'}) + cr (\text{standardized value of 'rawcr'})
\]
**Booklet Four**

<table>
<thead>
<tr>
<th></th>
<th>Multiple-Choice</th>
<th>Constructed-Response</th>
<th>Composite</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Score Points:</td>
<td>30 (19%)</td>
<td>125 (81%)</td>
<td>155 (100%)</td>
</tr>
<tr>
<td>Number of Items:</td>
<td>30 (83%)</td>
<td>6 (17%)</td>
<td>36 (100%)</td>
</tr>
<tr>
<td>Time (minutes):</td>
<td>30 (67%)</td>
<td>15 (33%)</td>
<td>45 (100%)</td>
</tr>
</tbody>
</table>

Method 1 (score points as weights):

\[
\frac{mc}{30} = \frac{X_1}{19} \quad \frac{cr}{125} = \frac{X_2}{81}
\]

\[
X_1 = mc (19/30) = mc (.63)
\]

\[
X_2 = cr (81/125) = cr (.65)
\]

Composite Score = \( X_1 + X_2 \)

\[
= mc (.63) + cr (.65)
\]

Method 2 (number of items as weights):

\[
\frac{mc}{30} = \frac{X_1}{83} \quad \frac{cr}{125} = \frac{X_2}{17}
\]

\[
X_1 = mc (83/30) = mc (2.77)
\]

\[
X_2 = cr (17/125) = cr (.14)
\]

Composite Score = \( X_1 + X_2 \)

\[
= mc (2.77) + cr (.14)
\]

Method 3 (time per item as weights):

\[
\frac{mc}{30} = \frac{X_1}{67} \quad \frac{cr}{125} = \frac{X_2}{33}
\]

\[
X_1 = mc (67/30) = mc (2.23)
\]

\[
X_2 = cr (33/125) = cr (1.4)
\]

Composite Score = \( X_1 + X_2 \)

\[
= mc (2.23) + cr (1.4)
\]

Method 4 (equal weighs):

\[
\frac{mc}{30} = \frac{X_1}{50} \quad \frac{cr}{125} = \frac{X_2}{50}
\]

\[
X_1 = mc (50/30) = mc (1.67)
\]

\[
X_2 = cr (50/125) = cr (.40)
\]

Composite Score = \( X_1 + X_2 \)

\[
= mc (1.67) + cr (.40)
\]

Method 5 (z-scores as weights):

Calculated in SPSS using Statistics/Descriptives/select "save as standardized variable"/ OK.

Composite Score = \( mc(\text{standardized value of 'rawmc'}) + cr(\text{standardized value of 'rawcr'}) \)
**Booklet Five**

<table>
<thead>
<tr>
<th>Max Score Points:</th>
<th>Multiple-Choice</th>
<th>Constructed-Response</th>
<th>Composite</th>
</tr>
</thead>
<tbody>
<tr>
<td>26 (44%)</td>
<td>33 (56%)</td>
<td>59 (100%)</td>
<td></td>
</tr>
<tr>
<td>Number of Items:</td>
<td>26 (90%)</td>
<td>29 (100%)</td>
<td></td>
</tr>
<tr>
<td>26 (58%)</td>
<td>19 (42%)</td>
<td>45 (100%)</td>
<td></td>
</tr>
</tbody>
</table>

Method 1 (score points as weights):

\[
\frac{mc}{26} = \frac{X_1}{44} \quad \frac{cr}{33} = \frac{X_2}{56}
\]

\[
X_1 = mc \left( \frac{44}{26} \right) \quad X_2 = cr \left( \frac{56}{33} \right)
\]

Composite Score = \( X_1 + X_2 \)

\[= mc(1.69) + cr(1.70)\]

Method 2 (number of items as weights):

\[
\frac{mc}{26} = \frac{X_1}{90} \quad \frac{cr}{33} = \frac{X_2}{10}
\]

\[
X_1 = mc \left( \frac{90}{26} \right) \quad X_2 = cr \left( \frac{10}{33} \right)
\]

Composite Score = \( X_1 + X_2 \)

\[= mc(3.46) + cr(.30)\]

Method 3 (time per item as weights):

\[
\frac{mc}{26} = \frac{X_1}{58} \quad \frac{cr}{33} = \frac{X_2}{42}
\]

\[
X_1 = mc \left( \frac{73}{33} \right) \quad X_2 = cr \left( \frac{42}{33} \right)
\]

Composite Score = \( X_1 + X_2 \)

\[= mc(2.2) + cr(1.27)\]

Method 4 (equal weighs):

\[
\frac{mc}{26} = \frac{X_1}{50} \quad \frac{cr}{33} = \frac{X_2}{50}
\]

\[
X_1 = mc \left( \frac{50}{26} \right) \quad X_2 = cr \left( \frac{50}{33} \right)
\]

Composite Score = \( X_1 + X_2 \)

\[= mc(1.9) + cr(1.5)\]

Method 5 (z-scores as weights):

Calculated in SPSS using Statistics/Descriptives/select "save as standardized variable"/ OK.

Composite Score = \( mc(\text{standardized value of 'rawmc'}) + cr \)

(standardized value of 'rawcr')
### Booklet Six

<table>
<thead>
<tr>
<th></th>
<th>Multiple-Choice</th>
<th>Constructed-Response</th>
<th>Composite</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Score Points:</td>
<td>31 (15%)</td>
<td>172 (85%)</td>
<td>203 (100%)</td>
</tr>
<tr>
<td>Number of Items:</td>
<td>31 (78%)</td>
<td>9 (22%)</td>
<td>40 (100%)</td>
</tr>
<tr>
<td>Time (minutes):</td>
<td>31 (69%)</td>
<td>14 (31%)</td>
<td>45 (100%)</td>
</tr>
</tbody>
</table>

**Method 1** (score points as weights):

\[
\text{mc/31}=X_1/15 \quad \text{cr}/172=X_2/85
\]

\[
X_1=\text{mc} (15/31) \quad X_2=\text{cr} (85/172)
\]

\[
X_1=\text{mc} (.48) \quad X_2=\text{cr} (.49)
\]

Composite Score = \( X_1 + X_2 \)

\[
= \text{mc}(.48) + \text{cr}(.49)
\]

**Method 2** (number of items as weights):

\[
\text{mc}/31=X_1/78 \quad \text{cr}/172=X_2/22
\]

\[
X_1=\text{mc} (78/31) \quad X_2=\text{cr} (22/172)
\]

\[
X_1=\text{mc} (2.5) \quad X_2=\text{cr} (.13)
\]

Composite Score = \( X_1 + X_2 \)

\[
= \text{mc}(2.5) + \text{cr}(.13)
\]

**Method 3** (time per item as weights):

\[
\text{mc}/31=X_1/69 \quad \text{cr}/172=X_2/31
\]

\[
X_1=\text{mc} (69/31) \quad X_2=\text{cr} (31/72)
\]

\[
X_1=\text{mc} (2.2) \quad X_2=\text{cr} (.18)
\]

Composite Score = \( X_1 + X_2 \)

\[
= \text{mc}(2.2) + \text{cr}(.18)
\]

**Method 4** (equal weighs):

\[
\text{mc}/31=X_1/50 \quad \text{cr}/172=X_2/50
\]

\[
X_1=\text{mc} (50/31) \quad X_2=\text{cr} (50/172)
\]

\[
X_1=\text{mc} (1.6) \quad X_2=\text{cr} (.29)
\]

Composite Score = \( X_1 + X_2 \)

\[
= \text{mc}(1.6) + \text{cr}(.29)
\]

**Method 5** (z-scores as weights):

Calculated in SPSS using Statistics/Descriptives/select "save as standardized variable"/ OK.

Composite Score = \( \text{mc}($\text{standardized value of 'rawmc'}$) + \text{cr}($\text{standardized value of 'rawcr'}$) \)
**Booklet Seven**

<table>
<thead>
<tr>
<th></th>
<th>Multiple-Choice</th>
<th>Constructed-Response</th>
<th>Composite</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Score Points:</td>
<td>28 (17%)</td>
<td>134 (83%)</td>
<td>162 (100%)</td>
</tr>
<tr>
<td>Number of Items:</td>
<td>28 (80%)</td>
<td>7 (20%)</td>
<td>35 (100%)</td>
</tr>
<tr>
<td>Time (minutes):</td>
<td>28 (62%)</td>
<td>17 (38%)</td>
<td>45 (100%)</td>
</tr>
</tbody>
</table>

**Method 1** (score points as weights):

\[
\frac{mc}{28} = \frac{X_1}{17} \\
X_1 = mc \left( \frac{17}{28} \right) \\
X_1 = mc \left( 0.61 \right) \\
\text{Composite Score} = X_1 + X_2 \\
= mc(0.61) + cr(0.62)
\]

**Method 2** (number of items as weights):

\[
\frac{mc}{28} = \frac{X_1}{80} \\
X_1 = mc \left( \frac{80}{28} \right) \\
X_1 = mc \left( 2.9 \right) \\
\text{Composite Score} = X_1 + X_2 \\
= mc(2.9) + cr(1.5)
\]

**Method 3** (time per item as weights):

\[
\frac{mc}{28} = \frac{X_1}{62} \\
X_1 = mc \left( \frac{62}{28} \right) \\
X_1 = mc \left( 2.21 \right) \\
\text{Composite Score} = X_1 + X_2 \\
= mc(2.2) + cr(1.4)
\]

**Method 4** (equal weights):

\[
\frac{mc}{28} = \frac{X_1}{50} \\
X_1 = mc \left( \frac{50}{28} \right) \\
X_1 = mc \left( 1.79 \right) \\
\text{Composite Score} = X_1 + X_2 \\
= mc(1.79) + cr(0.37)
\]

**Method 5** (z-scores as weights):

Calculated in SPSS using Statistics/Descriptives/select "save as standardized variable"/ OK. 

\[
\text{Composite Score} = mc(\text{standardized value of 'rawmc'}) + cr \\
\text{(standardized value of 'rawcr')}
\]
**Booklet Eight**

<table>
<thead>
<tr>
<th></th>
<th>Multiple-Choice</th>
<th>Constructed-Response</th>
<th>Composite</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Score Points:</td>
<td>22 (9%)</td>
<td>211 (91%)</td>
<td>233 (100%)</td>
</tr>
<tr>
<td>Number of Items:</td>
<td>22 (69%)</td>
<td>10 (31%)</td>
<td>32 (100%)</td>
</tr>
<tr>
<td>Time (minutes):</td>
<td>22 (73%)</td>
<td>23 (27%)</td>
<td>45 (100%)</td>
</tr>
</tbody>
</table>

**Method 1** (score points as weights):
\[
\frac{mc}{22} = \frac{X_1}{9} \quad \frac{cr}{211} = \frac{X_2}{91} \\
X_1 = mc \left(\frac{9}{22}\right) \quad X_2 = cr \left(\frac{91}{211}\right) \\
X_1 = mc \left(0.41\right) \quad X_2 = cr \left(0.43\right) \\
\text{Composite Score} = X_1 + X_2 \\
\quad = mc(0.41) + cr(0.43)
\]

**Method 2** (number of items as weights):
\[
\frac{mc}{22} = \frac{X_1}{69} \quad \frac{cr}{211} = \frac{X_2}{31} \\
X_1 = mc \left(\frac{69}{22}\right) \quad X_2 = cr \left(\frac{31}{211}\right) \\
X_1 = mc \left(3.1\right) \quad X_2 = cr \left(0.15\right) \\
\text{Composite Score} = X_1 + X_2 \\
\quad = mc(3.1) + cr(0.15)
\]

**Method 3** (time per item as weights):
\[
\frac{mc}{22} = \frac{X_1}{49} \quad \frac{cr}{211} = \frac{X_2}{51} \\
X_1 = mc \left(\frac{49}{22}\right) \quad X_2 = cr \left(\frac{51}{211}\right) \\
X_1 = mc \left(2.2\right) \quad X_2 = cr \left(0.24\right) \\
\text{Composite Score} = X_1 + X_2 \\
\quad = mc(2.2) + cr(0.24)
\]

**Method 4** (equal weighs):
\[
\frac{mc}{22} = \frac{X_1}{50} \quad \frac{cr}{211} = \frac{X_2}{50} \\
X_1 = mc \left(\frac{50}{22}\right) \quad X_2 = cr \left(\frac{50}{211}\right) \\
X_1 = mc \left(2.27\right) \quad X_2 = cr \left(0.24\right) \\
\text{Composite Score} = X_1 + X_2 \\
\quad = mc(2.27) + cr(0.24)
\]

**Method 5** (z-scores as weights):
Calculated in SPSS using Statistics/Descriptives/select "save as standardized variable"/ OK.
\[
\text{Composite Score} = mc(\text{standardized value of 'rawmc'}) + cr(\text{standardized value of 'rawcr'})
\]