University of Alberta

Methods for determining whether subscore reporting is warranted in large-scale achievement assessments

by

Oksana Illivna Babenko

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Measurement, Evaluation and Cognition

Educational Psychology

Fall 2011
Edmonton, Alberta

Dedication

I dedicate this work to my parents Illya and Tetyana Babenko for their unconditional love, support, and beautiful spirits.

Abstract

Officials of large-scale assessment programs often want to report subscale scores in addition to the total test score. However, in addition to the reliability of reported scores, evidence that subscales reveal real differences in student performances must be obtained in order to support reporting of subscale scores. In this study, two correlational methods, including correlations corrected for attenuation, *r*', and the proportional reduction of the mean squared error, *PRMSE* (Haberman, 2005; Sinharay et al., 2007), and the agreement method (Kelley, 1923) for determining whether subscore reporting is warranted in large-scale achievement assessments were examined. Whereas correlation-based methods consider student performances on pairs of measures in terms of ranked positions, the agreement method takes into account actual differences between students' standard scores on the pairs of measures being compared. The correlational methods revealed that with one possible subscale difference, the subscales did not differ among themselves and from the total test for the English Reading (*N* = 128,089) and Mathematics (*N* = 127,596) assessments considered in this study. In contrast, Kelley's agreement method one to five percent students had differences between their scores on the English Reading subscales that were greater than the difference expected due to the chance. However, with two exceptions for the Mathematics assessment, the results of the agreement method were uninterpretable. In agreement with Sinharay, et al. (2007), it was concluded that for the detection methods to work, three conditions need to be met, one substantive (multidimensional construct for which scores are wanted for each dimension), and two statistical (high reliabilities of and low intercorrelations among subscales). The results for replicated random samples (*n* = 250, 500, 1,000, 2,000, and 5,000) revealed that the statistics for the three detection methods were accurate and precise estimators of the corresponding population parameters.

Acknowledgement

I would like to express my sincere gratitude to those who have supported me in the completion of my degree. In particular, I would like to thank my supervisor, Dr. W. Todd Rogers, for his energy, encouragement throughout the dissertation process, and for providing countless opportunities for professional development and academic growth. I appreciate Dr. Rogers' attention to detail ensuring that the quality of the research met high standards.

I am also thankful to the members of my dissertation committee, Dr. Ying Cui, Dr. Rauno Parrila, Dr. Martin Mrazik, Dr. Stephen Norris, and Dr. John Anderson, for their feedback, interest, and thought-provoking questions. Special thanks go to Andrea Gotzmann who introduced me to SAS and enabled me to perform statistical analyses, and Hollis Lai for his prompt feedback and help in solving computer-related problems. I would like to acknowledge Dr. Jacqueline P Leighton, Dr. Cheryl Poth, Dr. Mark Gierl, and Dr. Bob Frender professors which I had the pleasure of learning from, and the CRAME students who have made my time here very enjoyable and enriching. I look forward to our future encounters.

My heartfelt thanks go to all my friends in Canada and the USA – I felt your support and encouragement since my very first day in North America.

Table of Contents

LIST OF TABLES

CHAPTER I: INTRODUCTION

In a large-scale assessment of student achievement, test items should be referenced to a curriculum that is multidimensional in composition, with each dimension characterized by specific content and/or cognitive skills. For example, items on a Mathematics achievement test can be referenced to (a) content areas, such as *number sense and numeration*, *measurement*, *geometry and spatial sense*, *patterning and algebra*, and *data management and probability*, and/or (b) cognitive skills, such as *knowledge and understanding*, *application*, and *problem-solving* as specified in a Mathematics curriculum. In test development, the table of specifications or test blueprint serves to ensure that the test reflects the multidimensionality of the curriculum and that the number of test items reflects the proportional weighting to be given to each cell within the table. Most often the number of items in each cell is such that the total test can be administered in a reasonable amount of time and that the internal consistency (reliability) of the total test is adequate for reporting purposes. However, often there is a desire to report the scores for the content areas and/or cognitive skills identified in the table of specifications. Teachers use these scores to identify areas of strength and areas that need to be addressed for individual students and/or to alter their instruction in ways to maintain strength and address issues at the class level, thereby improving their students' learning and achievement. Consequently, in addition to reporting the *total test score*, officials of large-scale assessment programs want to report *subscale scores*, with the subscales corresponding to the different dimensions of the curriculum as reflected in the table of specifications.

However, before reporting such information, large-scale assessment agencies should determine whether or not subscales are both reliable and distinct to warrant the reporting of scores on each. Determining whether or not the subscales are reliable and distinct complies with the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1990). Two standards apply here:

> Standard 1.12: When interpretation of subscores, score differences, or profile is suggested, the rationale and relevant evidence in support of such interpretation should be provided. Where composite scores are developed, the basis and rationale for arriving at the composites should be given. (p. 20)

> Standard 5.12: When group-level information is obtained by aggregating the results of partial tests taken by individuals, validity and reliability should be reported for the level of aggregation at which results are reported. Scores should not be reported for individuals unless the validity, comparability, and reliability of such scores have been established. (p. 65)

Taken together, these two standards imply that before test developers or practitioners decide to report subscale scores, they must gather reliability and validity evidence in support of such a provision. The evidence may consist of logical evidence, procedural evidence, and empirical evidence. The focus of the present study is on the empirical sources of evidence, in particular on methods for determining whether reporting student performances on the subscales and the total test composed of the subscales is warranted.

Evidence for reliability consists of determining the internal consistency of the test items using Cronbach’s alpha (Cronbach, 1951) and Cronbach’s stratified alpha (1965) for the total test if the reporting of subscale scores is found to be warranted. Various methods for determining whether subscale scores are distinct among each other and, in some cases, from the total test have been developed. These include the use of the standard error of the difference between subscale scores (Gulliksen, 1951; Kelley, 1923; Lord & Novick, 1968; Ryan, 2003), correlations corrected for attenuation due to unreliability of the measures (Haladyna & Kramer, 2004; Harris & Hanson, 1991; McPeek, Altman, Wallmark, & Wingersky, 1976), factor analytic procedures (Grandy, 1992; McPeek et al., 1976), augmented scores (Edwards & Vevea, 2006; Wainer, Sheehan, & Wang, 2000), objective performance index (Yen, 1987), statistical model fit (Harris & Hanson, 1991), and proportional reduction of the mean squared error (Haberman, 2005, 2008; Sinharay, Haberman, & Puhan, 2007). However, the methods used most recently in large-scale assessment are correlation-based methods, namely correlations corrected for attenuation due to unreliability of the measures (*r*´) and proportional reduction of the mean squared error (*PRMSE*) (Haberman, 2005, 2008; Haberman et al., 2009; Haladyna & Kramer, 2004; Harris & Hanson, 1991; Lyren, 2009; McPeek et al., 1976; Sinharay et al., 2007; Sinharay et al., 2009; Sinharay, 2010).

For the $r´$ method, if correlation coefficients corrected for attenuation due to unreliability in measures are equal to or greater than 0.90 (McPeek et al., 1976), then it is concluded that students' performances in terms of their ranked positions on pairs of subscales and/or subscale–total test pairs are not different and thus, reporting of subscale scores is not warranted. For example, Haladyna and Kramer (2004) used the $r´$ method to determine whether subscale scores on a basic biomedical science test revealed any differences in examinees' performances. They found that the corrected correlations were higher than 0.90, suggesting a high degree of similarity in examinees' performances on the subscales of the test. However, a problem with this method is the use of the observed subscale score on its own and as a part of the total test score to estimate the correlation between the two true scores.

For the *PRMSE* method (Haberman, 2005, 2008), if the $PRMSE_{s_s}$ for the true subscore when linearly predicted from the observed subscore is greater than the $PRMSE_{s_x}$ for the true subscore when linearly predicted from the observed total score, then it is concluded that the true student performances on the subscale are better predicted by the observed subscores than by the observed total scores. In this case, student performances on the subscale are concluded to differ from student performances on the total test. Otherwise, student performances on both the subscale and the total test are concluded to be comparable. For example, Haberman (2008) used the *PRMSE* method to determine whether or not the subscale scores on SAT I "had added value over and above the value of the total score." He reported that "none of the section scores of SAT I math or SAT I verbal provide any appreciable information concerning an examinee that is not already provided by the math or verbal total score” (p. 221). Haberman’s procedure was developed to determine if a subscale “added value” over the total test. As with the correlations, a problem with this procedure is the use of the observed subscale score on its own and as a part of the total test score to predict the true subscale score. Hence, as with the *r´* method, there is an overlap between the two separate predictors in the *PRMSE* method. This raises the question of whether, in addition to predicting value, the *PRMSE* method can be used for determining whether subscale scores have diagnostic value.

As mentioned earlier, both the *r´* and *PRMSE* methods are group level methods that are based on correlations. Correlations are high when examinees’ scores on two measures – two subscales or a subscale and the total test – *rank* examinees similarly. Correlations do not reflect whether or not the scores on two measures *agree*. Given this, a measure that takes into account actual differences in individual student performances on subscales should be used to determine whether or not reporting of subscores is warranted.

Kelley (1923) developed a two-step agreement method that takes into account score differences and allows determining ‘probable errors’ of the judgements made about score differences for individual students. First, working with standardized scores (i.e., *z*-scores (μ = 0 and σ = 1)) to account for differences in the means and standard deviations of two measures, Kelley defined the difference between two observed standard scores for an individual student as:

$$d_i = s_{ij} - s_{ij\prime} ,$$

where s_{ij} is the observed standard score of student *i* on subscale *j*,

$s_{ij\prime}$ is the observed standard score of student *i* on subscale *j'*, and

d_i is the difference between the two observed standard scores of student *i*.

Recognizing that the two values (i.e., s_{ij} and $s_{ij\prime}$) would differ by chance, Kelley developed the formula for the standard error of the difference due to measurement error present in each of the two subscales. The standard error of the difference, $\sigma_{d,\infty\omega}$, is given by:

$$\sigma_{d,\infty\omega} = \sqrt{2 - \alpha_{s_j} - \alpha_{s_{j\prime}}},$$

where α_{s_j} and $\alpha_{s_{j\prime}}$ are reliabilities (Cronbach's alpha) of subscales *j* and *j'*, and

∞ and ω are the true scores of student *i* on subscales *j* and *j'*, respectively.

According to Kelley (1923), "this formula fills a long felt need since it makes possible the determination of the probable errors of our judgments of difference of abilities within the individual" (p. 325). Noting that the difference scores were essentially normally distributed, Kelley defined the probable error of an individual difference as:

$$PE = 0.6745\sqrt{2 - \alpha_{s_j} - \alpha_{s_{j\prime}}},$$

where the terms are defined as above. The value of 0.6745 in the *PE* formula refers to the 75th percentile when positive and the 25th percentile when negative in a normal distribution. Thus, the probable error captures the middle 50% of the scores in a normal distribution. If the difference between the two standard scores of student *i* (i.e., d_i) exceeds the probable error, then some sort of intervention to increase learning for the subject area measured by the subscale with the lower score would be warranted.

Next, Kelley (1923) demonstrated that “if the distribution of differences for the entire population of students should have the same standard deviation as this [the standard error of the difference], then, obviously, the obtained differences are no greater than chance indicates” (p. 329). In order to determine this, the ratio of the standard error of the difference, $\sigma_{d,\infty\omega}$, to the standard deviation of differences, σ_d, is computed and subsequently used to determine the proportion of differences in excess of the chance (Table IV in Kelley (1923), p. 330):

$$Kelley's\ ratio = \frac{\sigma_{d,\infty\,\omega}}{\sigma_d} = \frac{\sqrt{2-\alpha_{s_j}-\alpha_{s_{j'}}}}{\sqrt{2-2\rho_{s_j s_{j'}}}},$$

where $\rho_{s_j s_{j'}}$ is the correlation between measures *j* and *j’*, and

α_{s_j} and $\alpha_{s_{j'}}$ are reliabilities (Cronbach’s alpha) of measures *j* and *j’*.

To summarize, with the mandate expressed in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) on determining the distinctiveness of scores, assessment programs must gather validity evidence in support of their decisions to report subscale scores on their assessments in addition to the total score. Although correlational and agreement methods have been used by large-scale assessment programs for this purpose, these methods have not been studied systematically to determine the properties and performances of these methods while keeping certain conditions (i.e., data source, sample size, subject area, number of subscales and their psychometric properties) constant.

Purpose of the Study

The purposes of the present study are to determine whether

1. the correlations corrected for attenuation, proportional reduction of the mean squared error, and the agreement methods lead to the same or different decision regarding the reporting of subscale scores; and
2. the statistics used for each method are accurate and precise.

Delimitations of the Study

The data used in the present study were obtained from the Education Quality and Accountability Office (EQAO) in the province of Ontario. The data were students' scores on the items included in the Junior (Grade 6) Reading and Mathematics 2009 assessments. The EQAO assessments administered at other grade levels (i.e., Primary Division (Grades 1– 3), Grade 9) and assessments administered by other agencies were not considered in this study. Therefore, the results obtained in the present study apply to the Junior Reading and Mathematics assessments administered in 2009.

Definition of Terms

Achievement test. A test used to evaluate the extent of knowledge or skill attained by a test taker in a content domain in which he/she has received instruction (adapted from AERA, APA, & NCME, 1999).

Domain. A set of knowledge and skills to be measured by an achievement test, often organized into categories (i.e., sub-domains) to which test items are referenced (adapted from AERA, APA, & NCME, 1999).

Subscale, subscale score. In an achievement test, the set of items referenced to a sub-domain is usually referred to as a subtest or subscale, with the scores derived from a subscale called subscale scores. In this study, subscale scores and subscores are used interchangeably to distinguish these scores from the total test score. Similarly, subtests and subscales are used interchangeably.

Dimensionality. Dimensionality refers to the conceptual homogeneity or heterogeneity of the content being measured. If the content is unidimensional, a single test score derived from a relatively homogeneous set of items is reported. Multidimensional content often consists of several sub-domains, for which separate scores can be reported (adapted from Schmeiser & Welch, 2006, p. 318).

Validity. Validity is an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment (Messick, 1989, p. 13).

Content-related evidence. “Evidence based on test content” is referred to as content-related evidence (AERA, APA, & NCME, 1999). Such evidence supports the test content as being representative of the important aspects of the curriculum being taught (Phillips, 1993) and the test as a representative sample of the content domain (Schmeiser & Welch, 2006, p. 313).

Reliability. The degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure; the degree to which scores are free of errors of measurement for a given group (adapted from AERA, APA, & NCME, 1999).

Standard error of measurement (SEM). The standard error of measurement is the standard deviation of errors of measurement, with the error of measurement being the difference between an obtained score and its theoretical true score counterpart.

Organization of the Dissertation

The background of the problem, the purpose of the study, delimitations of the study, and definition of terms were outlined in Chapter I. Chapter II contains a review of the methods used for determining whether subscore reporting is warranted on large-scale assessments, followed by a review of research studies in which these methods were used. A discussion of the agreement method concludes the chapter. The methods used to address the problem, including the description of the data, design, and analyses, are outlined in Chapter III. The results and discussion of the results are provided in Chapter IV for the English Reading assessment and Chapter V for the Mathematics assessment. Chapter VI includes the summary of the purposes, research method, and analyses, summary of findings, explanation of findings, conclusions, implications for practice, and recommendations for future research.

CHAPTER II: LITERATURE REVIEW

The literature reviewed in this chapter has specific relevance to the theoretical framework established in this study to compare different methods for determining whether reporting of subscale scores is warranted in large-scale assessments. The chapter is organized in two main sections. In the first section, the definition of a subscale score is provided and the purposes and uses of subscale scores are considered. The development of subscale scores is outlined next, followed by the psychometric criteria used in subscore reporting. The second section of this chapter contains a critical review of research pertaining to methods considered in this dissertation for determining whether reporting of subscale scores is warranted in large-scale assessments. Within this section, the research related to the use of correlations corrected for attenuation is provided first, followed by the use of the proportional reduction of the mean squared error and the agreement method.

Defining Subscale Scores

Assessment results can be reported in the form of either one total test score or a set of scores composed of subscale scores and the total test score. The common practice has been to report only the total score as a summary of proficiency with respect to an entire domain of knowledge and skills. However, in light of increasing interest in diagnostic assessments, large-scale assessment agencies have turned their attention to the generation and reporting of subscale scores in addition to the total score (Bock, Thissen, & Zimowski, 1997; Gessaroli, 2004; Kahraman & Kamata, 2004; Sinharay, Haberman, & Puhan, 2007;

Sinharay, 2010; Tate, 2004; Wainer et al., 2000; Yao & Boughton, 2007; Yen, 1987).

A subscale is typically defined by a group or subset of test items that measure one attribute or trait among the number of attributes or traits measured by the total test. Examples include subscales that are based on content categories or strands within one subject matter area (e.g., subscales for algebra and geometry in a mathematics test) and subtests in a test battery (e.g., subtests in intelligence batteries). Scores derived from subscales and subtests are commonly referred to as subscale scores. Other names found in literature include subtest scores, subscores, profile scores, diagnostic scores, and trait scores. For example, in assessments of writing, subscores are also referred to as trait scores (Dorans, 2005).

Purposes and Uses of Subscale Scores

Wainer et al. (2001) stated that the two most common uses of assessments are ranking students and diagnosing students' strengths and weaknesses. In addition to reporting the total test score, reporting subscale scores is desirable for a number of important reasons (Haberman, Sinharay, & Puhan, 2006). First, while total scores provide information for the total domain of interest, scores on identifiable subsections of a test (i.e., subscales) may be used to provide feedback specific to the corresponding content areas that together comprise the total domain. These subscores are included to provide detailed diagnostic information that may be useful in making individual instructional placement and remediation decisions as well as for improving curriculum, instruction, and learning. For example, unsuccessful candidates or failing students want to know their strengths and weaknesses in different content areas so that they, together with their teacher and parents, can plan future remedial work. Similarly, academic institutions such as schools, colleges, and universities want a profile of performance for their graduates so that they can evaluate their training programs and curriculum effectiveness and, thus, better focus their efforts on areas that need instructional improvement (Haladyna & Kramer, 2004).

Colleges and universities also use course marks and subscale scores to help make admission decisions, particularly when there are a number of applicants with identical or almost identical total scores on admission tests and who are similar on other factors considered for admission purposes (i.e., GPA, educational background). Similarly, employers want to be able to use subscores when hiring employees based on individual skills and proficiencies. They may also use subscores from tests administered to their employees to identify areas in need of remediation and/or professional development.

Finally, there is substantial pressure from the public to limit the number of tests students take, so that more time is devoted to instruction (Monaghan, 2006). Further, educational authorities want to see a reduction in the resources and expenses associated with the administration of assessments. Consequently, obtaining as much information as possible from an assessment is a desired outcome. As Monaghan (2006) noted, “the thinking is that assessment organizations obtain a vast amount of data from their tests that they can then compartmentalize to report on the individual skills of a test taker” (p. 2). This being the case, subscore reporting is often seen as one of the ways to extract maximum information with minimum testing time.

Development of Subscale Scores

Whether or not to generate and report subscale scores should ideally be established at the beginning of the test development process, when a test developer or a client asserts that the content domain of the test is intended to be multidimensional or unidimensional (Haladyna & Kramer, 2004). At the stage of defining the construct or domain to be measured by the test, the test developer or client decides whether subdomains are real and important (i.e., a multidimensional view) or irrelevant (i.e., a unidimensional view).

Clearly, subscores only have meaning in the multidimensional case. Luecht et al. (2006) argued that “inherently unidimensional item and test information cannot be decomposed to produce useful multidimensional score profiles – no matter how well intentioned or which psychometric model is used to extract the information” (p. 6). In the multidimensional case, the identification and definition of the subdomains for which subscores are to be reported must first be established. The subsequent activities in developing a test (i.e., task analysis, test specifications, test design, item development) should reflect the multidimensional nature of what is to be assessed and reported. According to Haladyna and Kramer (2004), the logical and procedural evidence for supporting the argument for test multidimensionality must be supported by empirical evidence to ensure the valid interpretation and use of reliable subscores.

Empirical evidence to support the reporting of subscores is also required for already existing assessment programs, which at the time of their development had no explicit goal to report subscores to examinees (Monaghan, 2006). Since the primary purpose of these assessment programs was not necessarily to provide diagnostic information but rather an overall indication of performance to examinees, reporting of subscores was not included in the design specifications of such assessment programs. However, given the emphasis on accountability and the need to improve curriculum and instruction, assessment agencies feel increased pressure or are, in fact, required “to report subscores with these programs regardless of the primary purpose of the assessments” (Monaghan, 2006, p. 2).

Psychometric Conditions and Criteria for Reporting Subscale Scores

Although subscale scores may be desired, there are important conditions and criteria that should be considered before deciding to report subscores. In particular, three psychometric conditions must be satisfied:

1. The test items in each subscale must be relevant to and representative of the construct being measured by the corresponding subscale (Messick, 1989).
2. The reliability and the standard error of measurement must be adequate for each subscale (AERA, APA, & NCME, 1999; Tate, 2004; Wainer et al., 2001).

3. The scores that capture student performance on each subscale (i.e., subscores) must provide different information among themselves and add information over and above the information that is summarized by the total score (Dorans, 2005; Haberman, 2005).

The first condition, that the items in a subscale must be relevant to and representative of the corresponding subscale, is addressed by using the professional judgment of qualified experts such as teachers, principals, curriculum specialists, and subject matter experts. Their qualifications include: knowledge of the curriculum, including both the content to be learned and the knowledge and cognitive skills to be acquired by the students; and the knowledge of the nature of the students who are expected to learn the curriculum and will be assessed by the assessment instrument. Once selected, the committees of qualified experts are asked to assess (a) the relevancy of the items referenced to each subdomain and (b) the representativeness of the relevant items for each subdomain to ensure that the subscale scores can be validly interpreted. Following this, test items should be sent to teachers who administer the items in their classes to determine if the items work as intended (i.e., students understand the items and their responses correspond to the behaviours called for when the items were developed).

The second condition pertaining to the adequate reliability of each subscale is addressed by including a sufficient number of test items that assess the given knowledge or cognitive skill to produce a stand-alone score for each subdomain. Wainer et al. (2001) and Tate (2004) emphasized the importance of ensuring that subscale scores be sufficiently reliable or, alternatively, that the standard error of measurement (in classical test theory) or the standard error of estimate (in item response theory) be small in order to minimize incorrect decisions. Given that tests are usually administered on one occasion, measures of internal consistency are used to estimate the reliability of scores. When the classical test score model is used, Cronbach's alpha is computed together with the standard error of measurement. When the item response models, both unidimensional and multidimensional, are used, the standard error of estimate for the ability estimate, θ_i, for each student or around the cut-score are considered. In 1972, when the Educational Testing Service (ETS) began reporting subscale scores for the Graduate Record Examinations (GRE), only subscores that attained a reliability of at least 0.80 were reported to examinees for admission purposes (McPeek et al., 1976). However, McPeek et al. (1976) suggested that if subscores were to "be used only for guidance and placement purposes, the statistical standards for reliability could be greatly reduced" (p. 3) because "guidance and placement decisions are perceived reversible, whereas admissions decisions generally are not" (p. 1). Salvia and Ysseldyke (2001) recommended that the minimum reliability value be set at 0.60 for reporting subscores at the group level.

The third condition that subscale scores provide different information among themselves and between each subscale and the total test is addressed by determining whether there are real differences in student performances on the pairs of subscales and subscale–total test pairs. Dorans (2005) and Haberman et al. (2006) considered the second and third conditions together, and indicated that frequently subscale scores provide information that is not reliable and/or is redundant with the information captured by the total score, with the problem being even more serious in cases when assessments were not specifically designed to report subscores.

Further, Standard 1.12 of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) states:

> When interpretation of subscores, score differences, or profile is suggested, the rationale and relevant evidence in support of such interpretation should be provided. Where composite scores are developed, the basis and rationale for arriving at the composites should be given. (p. 20)

In other words, if more than one score is to be reported on an assessment, the distinctiveness of the separate scores must be demonstrated. In addition, Standard 5.12, which also applies to subscale scores, states:

> When group-level information is obtained by aggregating the results of partial tests taken by individuals, validity and reliability should be reported for the level of aggregation at which results are reported. Scores should not be reported for individuals unless the validity, comparability, and reliability of such scores have been established. (p. 65)

What these two standards imply is that before test developers or practitioners decide on what scores to report they must gather validity evidence in support of such a provision. Such validity evidence may consist of logical evidence, procedural evidence, and empirical evidence (see Haladyna & Kramer, 2004).

Accordingly, the focus of the present study is on the empirical sources of evidence, in particular on the methods for determining whether reporting student performances on the subscales and the total test composed of the subscales is warranted.

Methods for Determining Whether Reporting of Subscale Scores is Warranted in Large-scale Assessments

Different statistical procedures have been developed for determining whether reporting of subscale scores is warranted in large-scale assessments. In the framework of the classical test theory (CTT), such methods include:

- zero-order correlations for the pairs of subscales and subscale–total test pairs (Haladyna & Kramer, 2004; Sinharay, Haberman, & Puhan, 2007; Tate, 2004);
- correlations corrected for attenuation due to unreliability for the pairs of subscales and subscale–total test pairs (Gulliksen, 1967; Haladyna & Kramer, 2004; Harris & Hanson, 1991; Lord & Novick, 1968; McPeek et al., 1976);
- proportional reduction of the mean squared error (*PRMSE*) when predicting the true subscale score from the observed subscale score and the observed total score using linear regression for approximation (Haberman, 2005, 2008; Lyren, 2009; Sinharay et al., 2007; Sinharay et al., 2009; Sinharay, 2010);

- agreement method based on the ratio of the standard error of the difference due to measurement error and the standard deviation of the difference (Gulliksen, 1951; Haladyna & Kramer, 2004; Kelley, 1923; Lord & Novick, 1968; Ryan, 2003);
- factor analysis, including both exploratory and confirmatory factor analytic procedures (Grandy, 1992; Haladyna & Kramer, 2004; McPeek et al, 1976);
- reliabilities of augmented scores compared to the reliability of the total score (Edwards & Vevea, 2006; Wainer, Sheehan, & Wang, 2000); and
- fitting a statistical model (Harris & Hanson, 1991).

In the present study, the following methods were considered:

- correlations corrected for attenuation due to unreliability (*r*´) for the pairs of subscales and subscale–total test pairs;
- proportional reduction of the mean squared error (*PRMSE*) when predicting the true subscale score from the observed subscale score and the observed total score; and
- agreement method based on the ratio of the standard error of the difference due to measurement error and the standard deviation of the difference.

The two correlational methods were selected because they are the most common methods used by large-scale assessment agencies, both when the test is explicitly constructed to be multidimensional and when the test is implicitly considered multidimensional (Haberman, 2005, 2008; Haberman et al., 2009; Haladyna & Kramer, 2004; Harris & Hanson, 1991; Lyren, 2009; McPeek et al., 1976; Sinharay et al., 2007; Sinharay et al., 2009). These methods are conceptually similar in that they are based on correlations, which focus on the similarity of students' rankings on the two measures being correlated.

In contrast to the correlational methods, the agreement method focuses on actual differences between standardized scores on any two measures. Given this, the two correlational methods and the agreement method can be contrasted for their capability in detecting differences in student performances on subscales that otherwise would not be captured if only the total score was used in reporting the general level of student performance. What follows next is a description of the selected correlational methods (i.e., *r´* and *PRMSE*) and the agreement method and a review of the research using these methods with large-scale assessments.

Correlations Corrected for Attenuation (r´)

In the context of subscore reporting, the method of correlations corrected for attenuation involves, first, computing zero-order correlation coefficients for the pairs of subscales and subscale–total test pairs, and then, using Spearman's (1904) formula, correcting the observed correlation coefficients for attenuation due to unreliability in subscales and the total test. The basic formula for computing a correlation coefficient corrected for attenuation is given by:

$$\rho_{\tau_j \tau_{j\prime}} = \frac{\rho_{jj\prime}}{\sqrt{\rho_j \rho_{j\prime}}},$$

where $\rho_{\tau_j \tau_{j'}}$ is the correlation between the true scores on measures *j* and *j'* or the correlation corrected for attenuation between two measures (i.e., two subscales or a subscale and the total test),

$\rho_{jj'}$ is the uncorrected (i.e., observed) correlation between the two measures, and

$\rho_j$and $\rho_{j'}$ are the reliabilities of the *j* and *j'*, respectively.

Given the subscales and the total test are administered on one occasion, reliability estimates are determined for one occasion using Cronbach's alpha coefficient.

If the corrected correlations for the pairs of subscales and subscale–total test pairs are high, then it is concluded that students' performances in terms of their ranked positions on the pairs of measures are not different and thus, reporting of subscale scores is not warranted. However, in the case of the correlation between subscale scores and the total test scores, a certain degree of correlation is expected, given that each subscale consists of a subset of items that are part of the total test. As Monaghan (2006) noted, "the total score and subscores often share such a high degree of correlation that one could more reasonably predict the subscores a person would receive on different forms of the test from the score on the whole test than from the test taker's subscore" (p. 3). Consequently, high subscale–total test correlations are expected. Longford (1990) provided the following recommendation for evaluating the usefulness of reporting subscores when using the method of corrected correlation coefficients:

> The "true" subscores underlying the observed subscores are often highly correlated. If two true subscores, related to different

> domains of ability (subtests), are perfectly correlated, then the corresponding observed scores are merely two less reliable versions of the true score underlying the aggregate of the two subtests. Then it is preferable to provide only the observed score for the aggregate, thus simplifying the format of the score report. (p. 92)

When the Educational Testing Service (ETS) began reporting subscale scores for the Graduate Record Examinations (GRE) in 1972, corrected correlations between the scores on each pair of subscales had to be less than 0.90 to warrant reporting subscale scores (Chalifour & Powers, 1988; McPeek et al., 1976). Harris and Hanson (1991) used the *r´* method with the P-ACT+ (American College Testing, 1989), which measures students' proficiency in English, including usage/mechanics and rhetorical skills, and Mathematics, including geometry and pre-algebra/algebra, to determine if the English and Mathematics subscores provided "different and better information for examinee-level [placement] decisions" than the total score. Three forms of the P-ACT+ were administered to randomly equivalent groups of Grade 10 examinees, with approximately 2,000 examinees administered each form. Harris and Hanson (1991) reported that the values of uncorrected zero-order correlations ranged between 0.67 and 0.79 and the values of the corrected correlations ranged from 0.94 to 0.98. They concluded that "the fact that the disattenuated correlations range from 0.94 to 0.98 suggest that the subscores are not measuring distinct constructs" (Harris & Hanson, 1991, p. 7).

In a later study, Haladyna and Kramer (2004) used the *r´* method to determine whether reporting subscale scores was warranted for a basic biomedical science test that was included as a part of the examination program for dentists (Joint Commission on National Dental Examinations, 2004). The number of examinees was 6,390. The values of the uncorrected correlation coefficients ranged between 0.76 and 0.87 and the values of the corrected correlation coefficients ranged from 0.83 to 0.94. Haladyna and Kramer (2004) concluded that high correlations indicated a high degree of common variance among the subscale scores. Attempting to explain the high degree of common variance, they noted that "…all cognitive tests tend to tap general intelligence, which may, in part, account for high correlations among all cognitive measures" (p. 361). They added that the fact that all the candidates received comparable intensive instruction also contributed to the high correlations. However, if McPeek et al.'s (1976) recommendation was taken, the scores on three pairs of subscales, namely the pairs of the dental anatomy and occlusion subscale with the anatomic anatomy sciences, biochemistry and physiology, and microbiology and pathology subscales, would have been determined to be distinct because the corrected correlations for these pairs were below 0.90.

Although computationally simple and relatively easy to explain to non-measurement audience, the method of correlations corrected for attenuation for the pairs of subscales and subscale–total test pairs has a major limitation. As a group level statistics, a correlation coefficient, either uncorrected or corrected, indicates the extent to which the rankings of students on two measures are consistent. It does not, however, reveal whether the actual scores on two measures are similar or different in value.

Proportional Reduction of the Mean Squared Error (PRMSE)

The *PRMSE* method is a more recent correlation-based procedure that has been extensively used with large-scale assessments. Taking the classical test theory (CTT) perspective that the true subscale score, s_t, can be estimated from the observed subscale score, *s*, or the total score, *x*, Haberman (2005, 2008) developed the *PRMSE* method for determining if subscale and total test scores differed.

The *PRMSE* method (Haberman, 2005, 2008) involves the following computations for each subscale. First, two estimates of the true subscale score, s_t, are obtained:

(i) an estimate based on the observed subscale score, *s*, where the estimated true subscale score, s_s, is predicted by its regression on the observed subscale score:

$$s_s = \bar{s} + \alpha(s - \bar{s}),$$

where $\bar{s}$ is the average subscale score for the group of examinees, and α is the reliability of the subscale.

(ii) an estimate based on the observed total score, *x*, where the estimated true subscale score, s_x, is predicted by its regression on the observed total score:

$$s_x = \bar{s} + c(x - \bar{x}),$$

where $\bar{x}$ is the average total score, and *c* is a constant that depends on the reliabilities and standard deviations of the subscale and the total test and the correlation of the subscale and the total test (for the computation of *c* see Haberman (2005)).

Next, the proportional reduction of the mean squared error (*PRMSE*) is computed for each estimate. For the estimate based on the observed subscale score s_s, the $PRMSE_{s_s}$ is given by:

$$PRMSE_{s_s} = \frac{\sigma^2(s_t) - E(s_s - s_t)^2}{\sigma^2(s_t)},$$

where $E(s_s - s_t)^2$ is the mean squared error *(MSE*) for the estimate s_s and defined as:

$$E(s_s - s_t)^2 = \sigma^2(s_t)[1 - \rho^2(s_t, s)],$$

where $\sigma^2(\mathrm{s_t})$ is the variance of the true subscale score, s_t, and which is a product of the observed subscale score variance and the subscale reliability (see Sinharay et al., 2007), and

$\rho^2(s_t, s)$ is the subscale reliability.

The *PRMSE* for the estimate based on the observed total score, $PRMSE_{s_x}$, is defined similarly:

$$PRMSE_{s_x} = \frac{\sigma^2(s_t) - E(s_x - s_t)^2}{\sigma^2(s_t)},$$

where $E(s_x - s_t)^2$ is the mean squared error *(MSE*) for the estimate s_x and defined as:

$$E(s_x - s_t)^2 = \sigma^2(s_t)[1 - \rho^2(s_t, x)],$$

where $\sigma^2(\mathrm{s_t})$ is the variance of the true subscale score, s_t, and which is a product of the observed subscale score variance and the subscale reliability (see Sinharay et al., 2007), and

$\rho^2(s_t, x)$ is the reliability of the total test.

Haberman (2005, 2008) showed that the value of $PRMSE_{s_s}$ is equal to the subscale reliability. He further explained that for a subscale to have added value, it should provide a more accurate prediction of the construct it purports to measure than the total test. If the $PRMSE_{s_s}$ is less than the $PRMSE_{s_x}$, then the subscale score does not provide added value over the total test score because, in this case, the total test score is a more accurate estimate of the true subscale score than the subscale score.

Sinharay et al. (2007) used the *PRMSE* method to determine whether reporting of subscale scores was warranted for a basic skills test administered to prospective and practicing teacher's aides. The numbers of examinees for the two test forms were 3,240 and 2,331. Since the $PRMSE_{s_s}$ was consistently smaller than the $PRMSE_{s_x}$, Sinharay et al. concluded that reporting of either the subscale scores (i.e., reading skills, reading application, mathematics skills, mathematics application, writing skills, and writing application) or the combined subscores for reading, mathematics, and writing was not warranted. In a later study, Haberman (2008) applied the *PRMSE* method to the subscores on the SAT I, which is used for college admission. The sample size of the SAT I test-takers was not indicated in the study. Haberman (2008) reported that "none of the section scores of SAT I math or SAT I verbal provide any appreciable information concerning an examinee that is not already provided by the math or verbal total score" (p. 221). Similarly, Puhan, Sinharay, Haberman, and Larkin (2008) used the *PRMSE* method to examine subscores for eight teacher certification tests that represented a broad range of subject and skill areas, including elementary education, mathematics, social studies, science, and foreign languages. The test-takers were prospective and beginning teachers, entry-level principals, and other school leaders. The total number of examinees for each of the eight tests ranged from 2,154 to 31,001. Like Sinharay et al. (2007) and Haberman (2008), Puhan et al. (2008) concluded that the subscale scores provided no information beyond what was already captured by the total score.

Lyren (2009) examined subscores on a Swedish college admission test (SweSAT). The data consisted of examinees' scores from five consecutive administrations of the SweSAT, a norm-referenced, multiple-choice test with five subscales: vocabulary, Swedish reading comprehension, English reading comprehension, data sufficiency, and diagrams, tables and maps. The number of test-takers ranged from 26,610 to 41,530 for the five administrations. Using the *PRMSE* method, Lyren (2009) found that, for the SweSAT, all but one of the subscores added value beyond the value provided by the total score. In particular, except for the Swedish reading comprehension, the observed subscale scores were better estimates of the true subscale scores than the observed total score. Lyren (2009) concluded that the total score, being a composite of the subscale scores, was a more reliable measure of Swedish reading comprehension than the corresponding subscale score. As a possible explanation, Lyren (2009) suggested that this result could be attributed to the fact that the rest of the SweSAT subscales, in addition to measuring distinct sub-domains, required a certain degree of reading comprehension, making the total score a better estimate of the true subscore of reading comprehension. This was confirmed by a study of the latent structure of the SweSAT (Lyren, 2009).

Based on their work, Haberman (2005) and Sinharay et al. (2007) concluded that “subscores are most likely to have value if they have relatively high reliability by themselves and if the true subscale score and the true total score have only a moderate correlation. Both conditions are important” (Sinharay et al., 2007, p. 28). Given this, Sinharay et al. (2007) noted that the *PRMSE* method is likely to provide support for the reporting of subscale scores “for tests with reasonably large number of items in each subcategory and composed of distinct subcategories” (p. 28). The former condition ensures higher subscore reliabilities, while the second condition ensures moderate correlation of each subscale with the total test. However, reliability is contingent upon the number of items included in a test and item discriminations, with higher values contributing to higher reliability. Despite this, including items with lower discrimination may be required to fill in gaps in the test specifications as dictated by the curriculum or due to a low inventory in the item bank (Haladyna & Kramer, 2004). Further, in the *PRMSE* method, the focus is on the added value of each subscore over the total score. Given this, this method fails to address the question of whether or not subscores are distinct among each other. Yet another potential problem with the *PRMSE* method is the use of the observed subscale score on its own and as a part of the total test score to predict the true subscale score. This creates an overlap between the two separate predictors. Finally, the *PRMSE* method is based on correlations and, as mentioned earlier, correlations are high when the scores on two measures rank students similarly. That is, the actual differences between the scores on each subscale and the total test are not taken into account in this method. Given the limitations associated with the two correlational methods (i.e., *r´* and *PRMSE*), it is proposed that Kelley's (1923) agreement method that takes into account the actual agreement between score values might be better used to determine whether subscore reporting is warranted. This method is reviewed next.

Agreement Method

Kelley (1923) developed a two-step agreement method. As the first step, Kelley proposed that the standard error of the difference be used to determine if the score difference for an individual student was greater than what he called the "probable error" (p. 325) to conclude that the student's performances on a pair of subscales differ beyond the difference that can be expected due to the errors of measurement. As the second step, Kelley demonstrated that "if the distribution of differences for the entire population of students should have the same standard deviation as this [the standard error of the difference], then, obviously, the obtained differences are no greater than chance indicates" (p. 329). These two steps are discussed in detail next.

Step 1

As mentioned earlier, Kelley's method is based on the differences between each student's observed standard scores on two measures. The use of standard scores accounts for the differences in the means and standard deviations of each subscale and the total test at the group level. If ignored, these differences would serve to magnify student differences among subscales. Different standard (scaled) scores can be used. Kelley (1923) adopted z-scores. What a standard score represents is a student's deviation from the common mean for each subscale and the total test, where the deviation is a function of the student's level of performance on each subscale or a subscale and the total test. If the student's standard scores differ between two subscales or a subscale and the total test, then the student's performance on the measure (i.e., a subscale or the total test) with a higher standard score is greater than the student's performance on the measure with a lower standard score. Kelley (1923) defined the difference between two observed standard scores for an individual student as:

$$d_i = s_{ij} - s_{ij'} \ ,$$

where s_{ij} is the observed standard score of student *i* on subscale *j*,

$s_{ij'}$ is the observed standard score of student *i* on subscale *j'*, and

d_i is the difference between the two observed standard scores of student *i*.

If s_{ij} is close in value to$s_{ij'}$ for all *n* students, then the differences (i.e.,d_i) will be close to zero. As the values of s_{ij} and $s_{ij'}$ become more and more distinct (for some if not all students), d_i will increase in value, with larger values implying larger differences in student performances on the two measures.

Rather than working with observed scores, Gulliksen (1950) suggested using estimated true scores, with the difference between the two estimated true scores in standard-score form given as:

$$d_{\tau_i} = \alpha_{s_j} s_{ij} - \alpha_{s_{j'}} s_{ij'} = \tau_{ij} - \tau_{ij'} \; ,$$

where s_{ij} and $s_{ij'}$ are defined as above,

α_{s_j} and $\alpha_{s_{j'}}$are reliabilities (Cronbach's alpha) of measures *j* and *j'*,

τ_{ij} and $\tau_{ij'}$ are estimated true scores of student *i* on measures *j* and *j'*, and

d_{τ_i} is the difference between the two estimated true scores of student *i*.

It should be noted that when reliabilities of the two measures are low, the true score estimates obtained using Gulliksen's method tend to regress to the mean. However, as the reliabilities of subscales increase, observed scores and true score estimates become close in value, leading to no practical difference between the results of Kelley's and Gulliksen's methods.

Next, recognizing that the two observed values (i.e., s_{ij} and $s_{ij'}$) would differ by chance, Kelley developed the formula for the standard error of the difference due to measurement error present in each of the two subscales when the scores were expressed as *z*-scores (μ_z = 0; σ_z = 1). The standard error of the difference, $\sigma_{d,\infty\omega}$, is given by:

$$\sigma_{d,\infty\omega} = \sqrt{2 - \alpha_{s_j} - \alpha_{s_{j'}}},$$

where α_{s_j} and $\alpha_{s_{j'}}$are reliabilities (Cronbach's alpha) of subscales *j* and *j'*, and

∞ and ω are the true scores of student *i* on subscales *j* and *j'*, respectively.

According to Kelley (1923), “this formula fills a long felt need since it makes possible the determination of the probable errors of our judgements of difference of abilities within the individual” (p. 325). The probable error of individual difference is given by:

$$PE = 0.6745\sqrt{2 - \alpha_{s_j} - \alpha_{s_{j'}},}$$

where the terms are defined as above. To be considered a real difference, the difference between the two standard scores of student *i* (i.e., d_i) would have to exceed the probable error, which means that the difference is above the 75th percentile if positive or below the 25th percentile if negative. However, when subscale reliabilities are high, it is common to use a minimum difference of one standard error of the difference (i.e., 68% confidence level) to identify systematic differences for an individual (personal communication, Dr. Troy Janzen, November 24, 2010). The lower percent of confidence is adopted in low-stake assessments such as those used for instructional guidance, placement, and curriculum purposes, because “guidance and placement decisions are perceived reversible, whereas admission decisions generally are not” (McPeek et al., 1976, p. 1). For the latter type of decisions, that is those made for admission and employment purposes, a higher percent of confidence may be adopted to reflect the seriousness of the consequences of wrongly concluding a score difference when there really is not one.

It should be noted that in Kelley’s formula for the standard error of the difference, the errors of measurement in the two measures are assumed to be independent. Kelley developed the formula based on the argument that errors of measurement in the classical test theory are considered random, uncorrelated to the true score, and uncorrelated to each other (see also Gulliksen, 1950; Lord & Novick, 1968). However, questions have been raised about the assumption that the errors of measurement on the two measures for which differences are being interpreted are independent. That is, the errors of measurement may be correlated (Zimmerman, Brotohusodo, & Williams, 1981; Zimmerman & Williams, 1982; Rogosa & Willett, 1983). Zimmerman et al. (1981) provided the following formula to estimate the correlation between the errors of measurement for two measures j and j', $\rho(E_{s_j}\ E_{s_{j'}})$:

$$\left|\rho_{s_j s_{j'}} - \rho(E_{s_j}\ E_{s_{j'}})\sqrt{(1-\alpha_{s_j})(1-\alpha_{s_{j'}})}\right| \leq \sqrt{\alpha_{s_j}\alpha_{s_{j'}}}\ ,$$

where $\rho_{s_j s_{j'}}$ is the correlation between measures j and j',

α_{s_j} and $\alpha_{s_{j\prime}}$are reliabilities (Cronbach's alpha) of measures j and j', and

$\rho(E_{s_j}\ E_{s_{j'}})$ is the correlation between the errors of measurement (i.e., correlated error) for measures j and j' (p. 182).

Although the concern with correlated errors has been addressed in the measurement theory, "there has not been practical work done in attempts to eliminate them [correlated errors] or correct for them in testing. When writers acknowledge the possible existence of correlated errors, they tend to downplay them or to argue that their effects are not large" (personal communication, Dr. Donald Zimmerman, June, 14, 2011)[1]. In the context of the present study,

[1] Theoretical work on score differences has been primarily done in the measurement of change (Cronbach & Furby, 1970; O'Connor, 1972; Rogosa & Willett, 1983), and research on gain scores (Zimmerman & Williams, 1982, 1998).

correlated errors are likely to be more of a concern when a subscale is included as part of another scale, such as when a subscale is part of the total test, than when the two subscales do not overlap.

Step 2

Next, Kelley (1923) sought to determine the proportion of cases in the population in which the score difference, d_i, is so large as to be significant. He argued that "if the distribution of differences for the entire population of students should have the same standard deviation as this [the standard error of the difference, $\sigma_{d,\infty\omega}$], then, obviously, the obtained differences are no greater than chance indicates" (p. 329). In order to determine this, the ratio of the standard error of the difference, $\sigma_{d,\infty\omega}$, to the standard deviation of differences, σ_d, is computed and subsequently used to determine the proportion of differences in excess of the chance (Table IV in Kelley (1923), p. 330):

$$Kelley's\ ratio = \frac{\sigma_{d,\infty\omega}}{\sigma_d} = \frac{\sqrt{2-\alpha_{s_j}-\alpha_{s_{j\prime}}}}{\sqrt{2-2\rho_{s_j s_{j'}}}},$$

where $\rho_{s_j s_{j'}}$ is the correlation between measures *j* and *j'*, and

α_{s_j} and $\alpha_{s_{j\prime}}$are reliabilities (Cronbach's alpha) of measures *j* and *j'*.

In order for Kelley's ratio to work, the mean of the reliabilities has to be greater than the correlation between the two subscales being compared. The reliabilities of subscales that Kelley worked with on the Stanford Achievement battery were moderate to high (0.67–0.95), while the intercorrelations for all possible pairs of subscales were small to moderate (0.02–0.76). Using the ratio of standard deviations, Kelley (1923) determined that, depending on a pair of subscales, the percentages of differences in individual test scores in excess of the chance for Grade 8 students (N = 96) ranged from as low as 10% and as high as 44%.

With respect to practical application of score differences, they have been mostly used in psychological and intelligence testing (Anastasi, 1988; Gulliksen, 1967; Lord & Novick, 1968), with a few studies conducted in the context of large-scale achievement testing (Haladyna & Kramer, 2004; Ryan, 2003). However, rather than following Kelley's approach (i.e., computing the standard error of the difference and Kelley's ratio), the approach adopted by Ryan (2003) and Haladyna and Kramer (2004) involved computing the reliability of the difference which was then used to determine the standard error of the difference. With the assumption made in the classical test theory about the errors on the two measures being independent from each other, the reliability of the difference, ρ_d, is given by:

$$\rho_d = \frac{\frac{\alpha_{s_j} - \alpha_{s_{j'}}}{2} - \rho_{s_j s_{j'}}}{1 - \rho_{s_j s_{j'}}},$$

where $\rho_{s_j s_{j'}}$ is the correlation between measures *j* and *j'*, and

α_{s_j} and $\alpha_{s_{j'}}$ are reliabilities (Cronbach's alpha) of measures *j* and *j'*.

An artefact is associated with the formula for the reliability of the difference. The reliability of the difference can be quite low and may result in a negative value even though the reliability of each subscale is quite high. This situation occurs when the correlation between the two measures, $\rho_{s_j s_{j'}}$, is quite high.

Ryan (2003) examined the subscales in a state assessment of Mathematics at Grade 3 and English Language Arts at Grade 8. The reliabilities of the five subscales in the Mathematics assessment were between 0.44 and 0.83, whereas the reliabilities of the differences in students' performances (N ≈ 48,000) on the subscales ranged from –0.06 to 0.35. Similarly, the reliabilities of Grade 8 English/Language Arts four subscales ranged from 0.47 to 0.88, whereas the reliabilities of the differences were determined to be between 0.05 and 0.43. Ryan concluded that the reliabilities of the differences were well below the level acceptable for making inferences about students, which meant that claims about students being stronger or weaker in various strands assessed by the two tests would be based on differences that were generally not much greater than random variation. Nevertheless, although low, the reliabilities of the differences were then used to compute standard errors of the differences and construct 95% confidence interval around each difference. Ryan (2003) reported that, depending on the pair of subscales, approximately 4% to 12% of students had differences that fell beyond the 95% confidence interval.

In a later study, Haladyna and Kramer (2004) examined subscore differences using Ryan's (2003) approach to determine whether reporting of subscale scores was warranted for a basic biomedical science. With the reliabilities of four subscales ranging from 0.92 to 0.94, which were much higher than those reported in Ryan's study, and the reliabilities of the differences being between 0.47 and 0.69, Haladyna and Kramer (2004) determined that, depending on a subscale pair, approximately 35.6% to 70.7% of examinees had subscore differences greater than two standard errors.

To summarize, the expectation in this study, based on the work of Sinharay et al. (2007, 2009) and Sinharay (2010), was that the use of the agreement method would lead to decisions different from those made using the *r´* and the *PRMSE* methods. Given that the ultimate interest as seen in the section on the purposes and uses of subscale scores (see pp. 12-14) is in informing individual students on their possible strengths and weaknesses to improve learning, the agreement method may be superior in that it looks at the differences in performances of individual students, which is not possible with either *r´* or *PRMSE* methods. Such information can be useful in planning remedial instruction for individual students identified as having differences beyond the chance level.

CHAPTER III: METHOD

This chapter describes the method used to determine whether the correlations corrected for attenuation, proportional reduction of the mean squared error, and the agreement methods lead to the same decision or different decisions regarding the reporting of subscale scores, and whether the statistics used for each method are accurate and precise. Given the data for this study are obtained from the Education Quality and Accountability Office (EQAO), a description of the EQAO assessments is provided first, followed by the description of data, research design, and statistical analyses.

Assessments

The purpose of the EQAO is to ensure greater accountability and better quality of education in the schools in Ontario that are publicly funded. As an agency at arm's length to the Ontario provincial government, the EQAO aims to provide parents, teachers, and the public with reliable information that can be validly interpreted about student achievement. The EQAO also makes recommendations for improvement that policy makers at the provincial and board levels as well as educators in schools and parents can use to improve learning and teaching. To address these purposes, the EQAO assesses student achievement at the end of the Primary (Grades 1–3) and Junior (Grades 4–6) divisions (Reading, Writing, and Mathematics), Grade 9 (Mathematics), and Grade 10 (Ontario Secondary School Literacy Test (OSSLT – Reading and Writing)). These assessments, which are in a paper-and-pencil format, are administered in the English- and French-languages once a year, except for the Grade 9 assessments, which are administered at the end of each semester in both languages. Approximately 120,000 English students and 6,500 French students are administered the Primary assessments. The corresponding numbers are approximately 128,000 and 6,200 students for the Junior assessments and approximately 144,000 and 5,500 students for the Grade 9 assessments. The numbers of first-time eligible English-language and French-language students are approximately 145,500 and 5,500 for the OSSLT[2]. For the Primary, Junior, and Grade 9 assessments, the percentages of students at each of five levels of achievement (i.e., below 1, 1, 2, 3, and 4) and the percentages of students who achieve the acceptable standard (Levels 3 and 4) are reported for each subscale at the provincial, board, and school levels. For the OSSLT, the scale scores for students performing below the passing score are reported both to schools and students in addition to the percentages of students who did and did not achieve the passing score. The students who score below the passing score are also provided with suggestions about how to improve their performance. This type of information is not provided for the other assessments.

The EQAO assessments were selected for this study for two reasons. First, like other large-scale assessment agencies, the EQAO is contemplating the reporting of subscale scores. As indicated above, the EQAO presently reports the percentage of students that achieve the standard of acceptability on each subscale at the provincial, board, and school levels. The EQAO’s consideration of score reporting by subscale is in response to the feedback from the field and relates to

[2] Successful completion of OSSLT is one of 32 requirements for the Ontario Secondary School Diploma (OSSD). Students who fail can take OSSLT again the following year.

teachers doing the same when interpreting class reports they receive containing item responses for the students in their classes. However, the EQAO constructed its assessments and tests with the focus on the total score and there has been little empirical evidence to support the EQAO decision to report subscale scores. That is, no systematic studies have been completed to determine if there is value added by subscale scores beyond the reporting of the percentages of students who met the acceptable standard. Given this, the EQAO provided the data for the present study.

The EQAO uses both multiple-choice and open-response items in each of its assessments. These items are developed by practicing teachers separately in English and in French using the corresponding test blueprints for each language group. In the test blueprints, measurable curriculum expectations set out in the *Ontario Curriculum* are clustered by topic, and test items are developed for each cluster. The teachers are brought together in one location for a two-day item writing session for each assessment. The first half day is devoted to training, with the training completed by EQAO staff members responsible for the development of the assessment instrument. The teachers then write items and share them with other teachers writing items referenced to the cluster and the members of the EQAO staff responsible for the assessment. Following the writing session, the EQAO staff review first-draft items and revise them as needed. The items are then sent to the corresponding teachers who wrote the items. The teachers then administer the items in their own classroom as part of a cognitive lab. These cognitive labs allow item writers to see if the items they wrote are working as intended (i.e., students understand the items and their responses correspond to the behaviours called for by each item). Students' responses are then used to inform the editing and refining of the stem and options for multiple-choice items and prompts and item-specific scoring rubrics for open-response items. Following this, an Assessment Committee, composed of teachers, curriculum specialists and subject matter experts selected from across the province, evaluates the revised test items for their (a) relevance to and (b) representativeness of each construct subdomain to which the items are referenced. Hence, the premise here is that the assessments accurately reflect the multidimensionality of the curriculum. However, there is no deliberate attempt to ensure that the number of items referenced to each subdomain is sufficient to have adequate reliability for the corresponding subscales (see pp. 16-17). At the same time, a separate Sensitivity Committee, composed of teachers and subject experts representative of the province, reviews the items to check for gender bias, reading difficulty, and possible offensiveness with respect to a particular sub-population within the entire population of students. The final revised items are field-tested as embedded items in operational test forms, with a set of field-test items appearing in the same places within each test form. Field-test items with acceptable psychometric properties and as a set representative of the full domain become operational items in the next year and are used to equate the next year's assessments and the present year's assessments to allow a measure of change between the two years.

The present study focused on the EQAO Junior assessments of English Reading and Mathematics. These annual assessments measure the reading and mathematics skills students are expected to learn by the end of Grade 6. The information obtained from these assessments is used to inform educators on student achievement in terms of the curriculum expectations outlined in the *Ontario Curriculum* that were assessed by the EQAO. The number of subscales varies with each content area. In particular, there are three subscales in English Reading and five content subscales and three knowledge and skills subscales in Mathematics.

English Reading

In the table of test specifications, the items included in the English Reading assessment are divided into three subscales: *explicit information* (6 items), *implicit information* (18 items) and *connections* (12 items). While the items in the explicit information subscale require students to detect and understand information and ideas stated explicitly in a variety of text types required by the curriculum, the items in the implicit information subscale probe students' understanding of the implicitly stated information and ideas. In the case of the items in the connections subscale, the students are expected to demonstrate their understanding of text passages by connecting, comparing, and contrasting the ideas presented in passages to their own knowledge, experience, and insights as well as other texts and the world around them. Thus, the three subscales can be ordered in terms of complexity, with the explicit information and connections subscales at the lowest and the highest levels of complexity, respectively. The total number of items on the English Reading assessment is 36, including 10 open-response items scored using a four-point scoring rubrics and 26 dichotomously scored multiple-choice items.

Mathematics

In contrast to the English Reading assessment, the items on the Mathematics assessment are referenced by (a) content areas (i.e., strands) and (b) knowledge and skills as specified in the mathematics curriculum. Consequently, the items on the Mathematics assessment can be arranged into content-based subscales or into knowledge and skills subscales. The five content areas include: *number sense and numeration* (8 items involving estimation, rate, ratio, and use of fractions), *measurement* (8 items involving the use of area relationships, understanding of the dimensions of the shapes needed to calculate their areas, and the conversion of metric area units), *geometry and spatial sense* (6 items dealing with the identification, performance and description of transformations, the identification of angles, and accurate use of rulers and protractors), *patterning and algebra* (7 items dealing with growing patterns, use of diagrams, tables and number sequences to represent the stages of patterns), and *data management and probability* (7 items involving concepts of probability, predicting and representing the probability of an outcome, comparing probabilities using common representations (e.g., common denominators, percents or decimals), and interpreting graphs). The five content areas are not ordered in terms of complexity.

Knowledge and skills are divided into three categories: *knowledge and understanding* (8 items), *application* (15 items), and *thinking-problem solving* (13 items). The items referenced to the knowledge and understanding category require students to demonstrate subject specific content (knowledge) and the comprehension of its meaning and significance (understanding). The application items require students to select and fit an appropriate mathematical tool or get the necessary information. The thinking-problem solving items require students to select and sequence a variety of tools to solve a problem and demonstrate a critical-thinking process. That is, to answer thinking-problem solving items, students need to make a plan. In contrast to the content subscales, the knowledge and skills subscales can be ordered in terms of complexity, with the knowledge and understanding subscale and the thinking-problem solving subscale being at the lowest and the highest levels of complexity, respectively. The total number of items on the Mathematics assessment is 36, including 8 open-response items scored using a four-point scoring rubric and 28 dichotomously scored multiple-choice items.

Data

The operational data for the 2009 Junior (Grade 6) assessments obtained from the Education Quality and Accountability Office (EQAO) were used in the present study. The total numbers of students assessed in the English Reading and Mathematics were 128,089 and 127,596, respectively.

Analyses

The purpose of the present study was to evaluate three methods for determining whether subscale scores are distinct among each other and from the total score in terms of their differential ability to detect differences among subscale scores and the total score and whether or not the methods were equally accurate and precise. The three detection methods considered were:

- correlations corrected for attenuation due to unreliability *(r´)* for each pair of subscales and subscale–total test pairs;
- proportional reduction of the mean squared error (*PRMSE*) when predicting the true subscale score from the observed subscale score and from the observed total score for each subscale; and
- agreement method based on the standard error of the differences (Kelley, 1923) between standardized scores on two measures.

The detection methods were first used with the entire population of students assessed in reading and mathematics and then with replicated random samples drawn from the population. In the case of replicated random samples, five levels of sample size (*n* = 250, 500, 1,000 2,000, and 5,000) were considered, with 1,000 replications for each sample size, to determine the effect of sample size on ability of the different detection methods to recover the parameter values obtained from the analyses at the population level and to determine their accuracy and precision. What follows next is the description of the decision rules used to determine whether subscale scores are distinct among each other and/or from the total score for each of the three detection methods.

Decision Rules

Correlations Corrected for Attenuation (r´)

The decision rule for the *r´* method adopted for this study was based on the rule proposed by McPeek et al. (1976): a value *less* than 0.90 indicates that the rank order of the scores on the two measures being correlated is sufficiently different to warrant reporting the scores of each (i.e., D = 1 if $r´$ < 0.90; D = 0 if $r´$ ≥ 0.90, where D stands for the decision made and $r´$ is the value of the corrected correlation). In the present study, this rule was used with the population and the replicated sampling data.

Proportional Reduction of the Mean Squared Error (PRMSE)

The decision rule for the *PRMSE* method adopted for this study was the rule proposed by Haberman (2005, 2008) and later used by Sinharay et al. (2007, 2009) and Sinharay (2010): if the $\mathrm{PRMSE}_{\mathrm{s_s}}$> $PRMSE_{s_x}$, then students' true performances on the subscale are concluded to be better predicted by the observed subscale score than by the observed total test score (i.e., D = 1 if $\mathrm{PRMSE}_{\mathrm{s_s}}$ > $PRMSE_{s_x}$; D = 0 if $PRMSE_{s_s} \leq PRMSE_{s_x}$, where D stands for the decision made, $PRMSE_{s_s}$ is the proportional reduction of the mean squared error when predicting the true subscale score from the observed subscale score by itself, and $PRMSE_{s_x}$ is the proportional reduction of the mean squared error when predicting the true subscale score from the observed total score).

Agreement Method

Several decisions had to be made when the agreement method was used with the two assessments considered in this study. First, the agreement method was used with the observed standardized scores (Kelley, 1923) rather than with the estimated true scores (Gulliksen, 1950). Had the Gulliksen's approach been used, the estimates of the true scores would have regressed to the means of corresponding subscales due to the somewhat low reliabilities of some of the subscales. Second, given the artefact associated with the reliability of the difference formula (see p. 36), the agreement method proposed by Kelley (1923) and not the approach adopted by Ryan (2003) and Haladyna and Kramer (2004) was used in the present study.

Third, although, in the classical test theory, the errors of measurement are assumed to be uncorrelated and thus, were not originally considered by Kelley (1923), the correlations of the errors on the pairs of subscales and subscale–total test pairs were examined to determine the tenability of the zero correlation assumption. In the context of the present study, correlated errors were likely to be more of a concern when a subscale was included as part of the other scale, such as when a subscale was part of the total test, than when the two subscales did not overlap. For the subscale–total test pairs, if the correlated errors were to be determined to be large due to the overlap of two measures, these pairs would have been excluded altogether from the subsequent analyses. For the pairs of subscales, if the correlated errors were to be determined to be negligible (i.e., close to zero), Kelley's agreement method would have been used; otherwise, Zimmerman et al.'s (1981) recommendation to include the correlated errors when examining score differences would have been followed. Finally, Kelley's ratio of the standard error of the difference, $\sigma_{d,\infty\omega}$, to the standard deviation of differences, σ_d, was computed (see p. 35) to determine the proportion of differences in excess of the chance (Table IV in Kelley (1923), p. 330).

The results of the analyses for each of the three detection methods are presented in the next two chapters. The results for the English Reading assessment are provided in Chapter IV; the results for the Mathematics assessment are presented in Chapter V. For each detection method, the consistency of the decisions was evaluated using the percentage of samples that led to the same decision made at the population level. Means and standard deviations of the distributions of sample statistics (i.e., $r\acute{}$, $PRMSE_{s_s}$ and $PRMSE_{s_x}$, Kelley’s ratio) were used to evaluate the three detection methods with respect to their accuracy and precision.

CHAPTER IV: RESULTS AND DISCUSSION – ENGLISH READING

For the English Reading assessment, the analyses were performed first for the original three subscales – explicit information (E), implicit information (I), and connections (C) – as specified in the table of test specifications. The two information subscales were then combined given the small number of explicit information items and, consequently, low internal consistency (i.e., reliability) of this subscale (see below). This combination was confirmed with the item developers responsible for Reading at the EQAO, the agency that provided the data. The analyses were then repeated for the new subscale, information (IN). The results for the population are provided first, followed by the results for the replicated random samples drawn from the population.

Psychometric Properties of the English Reading Assessment

The number of items, maximum score, mean, standard deviation, skewness, kurtosis, internal consistency, and standard error of measurement for the explicit information, implicit information, combined information, and connections subscales and the total test are reported in Table 1 for the population of students. The means and standard deviations are reported in the observed score units and as percentages (in parentheses). The means (percentages) revealed that students' performance declined on the original three subscales as the complexity of the constructs increased. The standard deviations (percentages) were essentially the same for the implicit information and connections subscales, which are the two higher levels of complexity, but smaller than for the explicit information subscale, likely because of the smaller number of items in the explicit subscale.

Table 1

Psychometric Properties: English Reading, N = 128,089

Subscale	*k/ms*	$\bar{X}_.$	s_X	*sk*	*ku*	α_X	s_e
Explicit Information	6/6	4.59 (76.5)	1.28 (21.3)	-0.81	0.14	0.47	0.93
Implicit Information	18/30	20.92 (69.7)	4.42 (14.7)	-0.82	0.72	0.76	2.17
Information	24/36	25.52 (70.9)	5.27 (14.6)	-0.85	0.69	0.80	2.36
Connections	12/30	17.16 (57.2)	4.33 (14.4)	-0.13	-0.12	0.74	2.21
Total Test	36/66	42.68 (64.7)	8.97 (13.6)	-0.54	0.23	0.87	3.23

Note. *k* is number of items in a subscale or the total test and *ms* is the maximum score greater than or equal to *k* given the use of dichotomously scored multiple-choice items and polytomously scored open-response items; $\bar{X}.$ – the mean; s_X – standard deviation; *sk* – skewness; *ku* – kurtosis; a_X – internal consistency (Cronbach's alpha); s_e – standard error of measurement.
Means and standard deviations expressed as percents are shown in parentheses.

The mean and standard deviation of the combined information subscale were close to the mean and standard deviation of the implicit information subscale, which is attributable to the larger number of implicit items included in the combined information subscale. The distributions of scores on the explicit information, implicit information, and information subscales were more negatively skewed than the distribution of scores on the connections subscale, meaning that, as a group, students performed higher on the explicit information, implicit information and information subscales than on the connections subscale.

Next, as shown in Table 1, the values of internal consistency (i.e., reliability) of each subscale were not the same. The internal consistency (Cronbach's alpha) of the explicit subscale, 0.47, was much lower than the internal consistency of the implicit information, combined information, and connections subscales, which are 0.76, 0.80, and 0.74, respectively. Only one of these values satisfied the criterion of 0.80 used by ETS (McPeek, et al., 1976).

However, as indicated in the review of literature, the internal consistencies of the implicit information and connections subscales are congruent with the notion that subscale reliabilities may be lower if subscores were to be used for guidance and placement purposes where the decisions are perceived reversible as opposed to admission decisions (McPeek et al., 1976, p. 3). Such was the case for the EQAO, where the results were to be used for instructional and curriculum purposes. The low reliability of the explicit information is likely due to the relatively small number of items in this subscale in comparison to the other subscales (6 items vs. at least 12 items). However, the explicit subscale was retained to see what effect the low level of reliability would have on the results. The estimate of the internal consistency of the total test was 0.87, with the total number of items (both multiple-choice and open response) being 36.

Detection of Performance Differences

Correlations Corrected for Attenuation (r´)

The zero-order correlations for the population – uncorrected, *r*, and corrected for attenuation due to unreliability, *r´*, – among the subscales and between each subscale and the total test are reported in Table 2. The uncorrected correlations are reported in the upper triangle and the corrected correlations are reported in the lower triangle. The uncorrected correlations are provided because they serve as input in the *r´* and the *PRMSE* methods. Correlation coefficients shown in italics are between two measures where one measure is a part of the second measure. For example, each subscale is a part of the total test. Similarly, the explicit information and implicit information subscales are parts of the

Table 2

Detection of Performance Differences: Uncorrected and Corrected Correlations, English Reading, N = 128,089

Subscale	Explicit Information	Implicit Information	Information	Connections	Total Test
Explicit Information	-	0.59	*0.74*	0.52	*0.69*
Implicit Information	0.98 (0)	-	*0.98*	0.74	*0.93*
Information	*1.21*	*1.26*	-	0.75	*0.95*
Connections	0.89 (1)	0.98 (0)	0.97 (0)	-	*0.92*
Total Test	*1.07* (0)	*1.14* (0)	*1.13* (0)	*1.15* (0)	-

Note. Uncorrected correlations are in the upper triangle and corrected correlations are in the lower triangle. Correlation coefficients, both uncorrected and corrected shown in italic are between two measures, with one of the measures being a part of the other. Decision made with respect to the scores on a pair of measures (0 – not different; 1 – different) is shown in parentheses.

combined information subscale. As a result, the observed correlations are higher due to the presence of common items in both measures, which likely led to the values of the corrected correlations being greater than one.

As described in Chapter III, the decision rule for the method of correlations corrected for attenuation is: a value less than 0.90 (McPeek, et al., 1976) indicates that the ranks on the two measures being correlated are sufficiently different to warrant reporting the scores on each (i.e., $D = 1$ if $r' < 0.90$; $D = 0$ if $r' \geq 0.90$, where D stands for the decision made and r' is the value of the corrected correlation). As shown in Table 2, the only performance difference found is between the explicit information and connections subscales (E_C) when the r' method is used. Although below the decision value of 0.90, the value of the corrected correlation of 0.89 for the E_C is still very close to 0.90.

Interestingly, the corrected correlation of 0.98 for the explicit information and implicit information pair (E_I) is much higher than the corrected correlation for the E_C pair. Given the reliabilities of the implicit information and connections subscales are approximately the same (0.74 and 0.76, respectively), the corrected correlation of 0.89 for the E_C pair is likely due to the difference in the complexity levels between what is called for when using the information that is explicitly stated in the text and when making connections between what is read and one's own personal experience. The values of the corrected correlations for the remaining pairs of the original subscales all exceed 0.90, thus indicating that students' performances as reflected by their rank order on the measures in each pair are similar.

Somewhat disturbing, but not unexpected, are the values of the corrected correlations of each subscale with the total test (i.e., E_T, I_T, IN_T and C_T) and the explicit and implicit information subscales with the combined information subscale (i.e., E_IN and I_IN). All are greater than one. As indicated above, this is likely due to the common items present in the two measures. In particular, the presence of common items inflates the uncorrected correlations between the measures in each of these pairs, and subsequently leads to the corrected correlations being greater than one. Given these findings, the *r´* method was used only with the pairs of the original subscales and the combined information–connections pair in the replicated sampling component of this study.

The results for the *r´* method when used with the replicated random samples are presented in Table 3. In the table, the values in the first row for each pair of subscales are, respectively, the frequency and percentage (in parentheses) of random samples (out of 1,000) with the same decision as in the population at a given sample size. The values in the second row for each pair of subscales are the mean and standard deviation (in parentheses) of the sampling distribution of 1,000 corrected correlations at a given sample size. The last column contains the values of the corrected correlations in the population.

As shown in Table 3, the consistency of the decisions made using sample data increased as (a) the population value of the corrected correlation increased, (b) the sample size increased, (c) the reliability increased, and (d) the difference in the complexity levels between two subscales became smaller. First, with the exception of the E_C pair, the percentage agreement between the decisions made using sample data and the decisions made in the population was high if not 100% for the pairs that were determined to be highly correlated in the population (*r*´ greater or equal to 0.97). Second, the lowest percentage agreement for each pair occurred when *n* = 250 and increased as the sample size increased, reaching 100% when *n* = 2,000 for the E_I pair and *n* = 500 for the I_C and IN_C pairs. Third, while the percentage agreement increased with the increase in the sample size for the E_C pair, it never exceeded 77%. As it was the case for the population, this finding is most likely attributable to the difference in complexity between the explicit information and connections subscales, which are, respectively, of the lowest and highest levels of complexity. Further, with the value of the corrected

Table 3

Frequency of Decisions that Agreed with Decisions Made in the Population, Means and Standard Deviations of the Sampling Distributions of Corrected Correlations at Five Levels of Sample Size for English Reading

	Sample Size					
Pair	250	500	1,000	2,000	5,000	Pop.
E_I (0)[a]	905 (90.5)[b]	951 (95.1)	991 (99.1)	1000 (100)	1000 (100)	
	0.99 (.078)[c]	0.99 (.051)	0.99 (.037)	0.98 (.026)	0.98 (.016)	0.98
E_C (1)	514 (51.4)	554 (55.4)	581 (58.1)	664 (66.4)	768 (76.8)	
	0.90 (.081)	0.89 (.056)	0.89 (.039)	0.89 (.028)	0.89 (.017)	0.89
I_C (0)	996 (99.6)	1000 (100)	1000 (100)	1000 (100)	1000 (100)	
	0.99 (.031)	0.99 (.022)	0.98 (.015)	0.98 (.011)	0.98 (.007)	0.98
IN_C (0)	993 (99.3)	1000 (100)	1000 (100)	1000 (100)	1000 (100)	
	0.97 (.029)	0.97 (.021)	0.97 (.015)	0.97 (.011)	0.97 (.007)	0.97

Note. E – explicit information; I – implicit information; C – connections; IN – information.

[a] Decision (0 – not different; 1 – different) made in the population.

[b] The first value is the frequency and the value in parenthesis is the percentage of random samples (out of 1,000) with the same decision as in the population at a given sample size.

[c] The first value is the mean and the value in parentheses is the standard deviation of the sampling distribution (i.e., standard error) of 1,000 corrected correlations at a given sample size.

correlation for the E_C pair in the population being 0.89, which is 0.01 below the decision value of 0.90, and larger sampling variability due principally to the explicit information subscale, the sample estimates of the corrected correlation for the E_C pair varied more, resulting in the low decision consistency.

With respect to the accuracy of sample estimates, the corresponding values of the corrected correlations in the population were recovered well. The sample estimators of population parameters of the corrected correlations were accurate and precise. Namely, for all the pairs of subscales and sample size levels, the means of the sampling distributions of the corrected correlations were within 0.01 of the corresponding population values. The standard errors of sample estimators decreased as the sample size increased. For *n* = 250, the standard errors ranged between 0.029 and 0.081, whereas for *n* = 5,000, the standard errors were as low as 0.007 and as high as 0.017. Given the low reliability of the explicit information subscale, the standard errors for the pairs involving this subscale (i.e., E_I and E_C) were consistently higher than the standard errors for the remaining pairs.

Proportional Reduction of the Mean Squared Error (PRMSE)

The results for the *PRMSE* method are reported in Table 4 for the population (the last column) and the replicated sample data (columns 3 through 7). In the table, the values in the first row for each subscale are, respectively, the frequency and percentage (in parentheses) of random samples (out of 1,000) with the same decision as in the population. The values in the second row for each subscale are the mean and standard deviation (in parentheses) of the sampling distributions of the $PRMSE_{s_s}$ at a given sample size. Similarly, the values in the third row for each subscale are the mean and standard deviation (in parentheses) of the sampling distributions of the $PRMSE_{s_x}$ at a given sample size.

The decision rule for this method is: if the $PRMSE_{s_s}$ > $PRMSE_{s_x}$, then students’ true performances on the subscale are concluded to be better predicted by the observed subscale score ($PRMSE_{s_s}$) than by the observed total test score ($PRMSE_{s_x}$). In this case, student performances on the subscale are said to be different from student performances on the total test. Otherwise, student performances on both the subscale and the total test are comparable (i.e., D = 1 if

Table 4

Frequency of Decisions that Agreed with Decisions Made in the Population, Means and Standard Deviations of the Sampling Distributions of $PRMSE_{s_s}$ and $PRMSE_{s_x}$ *at Five Levels of Sample Size for English Reading*

		Sample Size					
Subscale		250	500	1,000	2,000	5,000	Pop.
Explicit Information (0)[a]		999 (99.9)[b]	1000 (100)	1000 (100)	1000 (100)	1000 (100)	
	$PRMSE_{s_s}$	0.47 (.054)[c]	0.47 (.038)	0.47 (.027)	0.47 (.020)	0.47 (.012)	0.47
	$PRMSE_{s_x}$	0.80 (.100)	0.79 (.065)	0.79 (.047)	0.79 (.033)	0.79 (.020)	0.79
Implicit Information (0)		1000 (100)	1000 (100)	1000 (100)	1000 (100)	1000 (100)	
	$PRMSE_{s_s}$	0.76 (.023)	0.76 (.016)	0.77 (.011)	0.77 (.008)	0.77 (.005)	0.77
	$PRMSE_{s_x}$	0.87 (.017)	0.87 (.013)	0.87 (.008)	0.87 (.006)	0.87 (.004)	0.87
Information (0)		1000 (100)	1000 (100)	1000 (100)	1000 (100)	1000 (100)	
	$PRMSE_{s_s}$	0.80 (.019)	0.80 (.014)	0.80 (.009)	0.80 (.007)	0.80 (.004)	0.80
	$PRMSE_{s_x}$	0.86 (.016)	0.86 (.011)	0.86 (.008)	0.86 (.006)	0.86 (.003)	0.86
Connections (0)		1000 (100)	1000 (100)	1000 (100)	1000 (100)	1000 (100)	
	$PRMSE_{s_s}$	0.73 (.023)	0.73 (.016)	0.73 (.011)	0.74 (.008)	0.74 (.005)	0.74
	$PRMSE_{s_x}$	0.85 (.020)	0.85 (.014)	0.85 (.010)	0.85 (.007)	0.85 (.004)	0.85

Note. $PRMSE_{s_s}$ denotes the proportional reduction of the mean squared error for the estimated subscale true score that is linearly predicted from the observed subscale score; $\mathrm{PRMSE}_{\mathrm{s_x}}$ denotes the proportional reduction of the mean squared error for the estimated subscale true score that is linearly predicted from the observed total test score.

[a] Decision (0 – not different; 1 – different) made in the population.

[b] The first value is the frequency and the value in parentheses is the percentage of random samples (out of 1,000) with the same decision as in the population.

[c] The first value is the mean and the value in parentheses is the standard deviation (i.e., standard error) of the sampling distributions of the corresponding *PRMSE* at a given sample size.

$PRMSE_{s_s} > PRMSE_{s_x}$; $D = 0$ if $PRMSE_{s_s} \leq PRMSE_{s_x}$, where D stands for the decision made). According to Haberman (2005, 2008), the *PRMSE* is conceptually similar to reliability, and the value of $PRMSE_{s_x}$ is exactly equal to the reliability of a subscale (cf. the last column in Table 4 and column 7, Table 1). As shown in

Table 4, regardless of the complexity and reliability of each subscale, $PRMSE_{s_s} < PRMSE_{s_x}$ for all subscales, meaning that the true subscale scores are better predicted by the observed total scores than by the observed subscale scores. That is, students' performances on each subscale are no different from their performances on the total test. Further, regardless of the sample size, reliability and complexity of each subscale, the decisions made using sample data were 100% consistent with the population decisions. The means of the distributions of sample estimators of $PRMSE_{s_x}$ and $PRMSE_{s_s}$ were within 0.01 of the corresponding population values for all four subscales. The standard errors of sample estimators were the largest when the explicit information subscale was considered (e.g., the standard error for $PRMSE_{s_x}$ = 0.100 for n = 250) but decreased as the sample size increased, ranging between 0.003 and 0.020 n = 5,000. The smallest standard errors were for the combined information subscale for all the sample size levels, given that the internal consistency of this subscale was the largest (0.80).

The results for the *PRMSE* are not unexpected given the low values for reliability. As pointed out in Chapter II (see pp. 27-28), subscales are most likely to have value over and above the total test if the subscales have relatively high reliability and if the true subscale score and the true total score have only a moderate correlation. As seen from Tables 1 and 2, the reliabilities are not especially high and the correlations corrected for attenuation due to unreliability are high.

Agreement Method

The results for the agreement method are reported in Tables 5 and 6. First, the correlations between the errors on the pairs of subscales and subscale–total test pairs were obtained to determine the tenability of the zero correlation assumption. The results of this analysis are provided in Table 5. As was done in Table 2, the correlation coefficients shown in italics in Table 5 are between two measures, with one of the measures being a part of the other.

The examination of the correlation coefficients revealed low, if not close to zero, correlated errors for the pairs of the original subscales and high correlated errors for the remaining pairs (i.e., E_IN, I_IN, E_T, I_T, IN_T, and C_T). The high correlated errors are not unexpected, given the presence of the common items. The presence of common items inflates the correlations between the measures in each of these pairs since the errors for the common items are identical. Given this finding as well as the fact that previous studies (Haladyna & Kramer, 2004; Ryan, 2003) examined score differences for the pairs of subscales and not for the subscale–total test pairs, the agreement method (Kelley, 1923) was, therefore, used only with the pairs of the original subscales in this study. Further, what curriculum specialists and teachers will look at are the students' performances on the pairs of subscales and not the students' performances on subscale–total test pairs.

Next, Kelley's ratio, defined as the ratio of the standard error of the difference due to measurement error, $\sigma_{d,\infty\omega}$, to the standard deviation of the difference between observed scores, σ_d, was computed for the pairs of the original

Table 5

Correlated Errors, English Reading, N = 128,089

Subscale	Explicit Information	Implicit Information	Information	Connections
Explicit Information	-			
Implicit Information	-.021	-		
Information	*.389*	*.914*	-	
Connections	-.188	-.040	-.085	-
Total Test	*.193*	*.662*	*.718*	*.640*

Note. Correlation coefficients shown in italic are between two measures, with one of the measures being a part of the other.

subscales in each of the 1,000 replicated samples. The mean value of this ratio and its standard error are provided for each pair of subscales in Table 6. As shown in the table, the mean of Kelley’s ratio in the random samples was within 0.01 of the corresponding population value for each subscale pair and sample size. The standard deviation of the sampling distribution (i.e., the standard error) of Kelley’s ratio decreased as the sample size increased. Namely, for *n* = 250 the standard errors were between 0.043 and 0.049, whereas for *n* = 5,000 the standard errors were between 0.009 and 0.011. That is, the sample estimators of Kelley’s ratio were accurate and precise.

Somewhat disturbing, however, are the values of Kelley’s ratio, which are all close to one. As mentioned in Chapter II, for Kelley’s procedure to work, the mean of the two reliabilities has to be greater than the correlation between the two subscales being compared. Given that the reliabilities of the English Reading

Table 6

Means and Standard Deviations of Kelley's Ratio at Five Levels of Sample Size and Percentage of Differences in Excess of the Chance for English Reading

	Sample Size						
	250	500	1,000	2,000	5,000	Pop.	%
E_I	0.97 (.049)[a]	0.97 (.035)	0.97 (.024)	0.97 (.017)	0.97 (.011)	0.97	≈1.5%
E_C	0.92 (.046)	0.91 (.032)	0.91 (.022)	0.91 (.015)	0.91 (.010)	0.91	≈5.0%
I_C	0.98 (.044)	0.98 (.032)	0.98 (.022)	0.98 (.016)	0.98 (.010)	0.98	≈1.0%
IN_C	0.96 (.043)	0.95 (.032)	0.95 (.021)	0.95 (.015)	0.95 (.009)	0.96	≈2.0%

Note. E – explicit information; I – implicit information; C – connections; IN – information.
[a] The first value is the mean and the value in parentheses is the standard deviation (i.e., standard error) of the sampling distribution of the corresponding statistics at a given sample size.
% – percentage of differences in excess of the chance in the population, determined using Table IV in Kelley (1923, p. 330).

subscales and the correlations between them were very close in value, high values of Kelley's ratio were observed for these pairs of subscales. Finally, using Table IV in Kelley (1923, p. 330), the proportion of score differences in excess of the chance was determined to be between 0.01 and 0.05, meaning that depending on a pair of subscales, between one to five percent of students had score differences that could not be attributed to the chance. Given that the means of Kelley's ratio at the five levels of sample size were within 0.01 of the corresponding population values for each pair, with the standard error being very low, the proportion of score differences in excess of the chance was not computed in the replicated samples.

Summary

To summarize, both correlational methods (i.e., *r´* and *PRMSE*) led to the same decisions. Namely, student performances on the pairs of the original subscales and subscale–total test pairs were determined to be no different. The agreement method revealed that, depending on a subscale pair, there were between one to five percent of students showing differences in their subscale scores sufficiently great that they could not be attributed to chance and thus, were likely due to something systematic. The statistics used for each method were determined to be accurate and precise. Sample estimators were within 0.01 of the corresponding population parameters. Standard errors of sample estimators decreased as the sample size increased from $n = 250$ to $n = 5{,}000$.

CHAPTER V: RESULTS AND DISCUSSION – MATHEMATICS

This chapter describes the results of the analyses conducted to determine whether the correlations corrected for attenuation, proportional reduction of the mean squared error, and the agreement methods led to the same decision or different decisions regarding the reporting of subscale scores on the Mathematics assessment, and whether the statistics used for each method were accurate and precise. The analyses were performed first for the three knowledge and skills subscales: knowledge and understanding (K), application (A), and thinking-problem solving (P). Following this, the analyses were performed for the five content subscales: number sense and numeration (N), measurement (M), geometry and spatial sense (G), patterning and algebra (A), and data management and probability (P). Presentation of the Mathematics results mirrors the presentation of the English Reading results.

Psychometric Properties of the Knowledge and Skills Subscales

The psychometric properties of the knowledge, application, and problem solving subscales and the total test are reported in Table 7 for the population of students. The means and standard deviations are reported in the observed score units and as percentages (in parentheses). As with the English Reading, the means (percentages) revealed that students’ performance on the three mathematics cognitive subscales declined as the level of required thinking increased from knowledge to application to problem solving. The standard deviations (percentages) were essentially the same for the application and problem solving subscales, which are two higher levels of complexity, but smaller than the

Table 7

Psychometric Properties: Knowledge and Skills Subscales, Mathematics, N = 127,596

Subscale	k/ms	$\bar{X}_{.}$	s_X	*sk*	*ku*	α_X	s_e
Knowledge	8/8	5.45 (68.0)	1.87 (23.4)	-0.49	-0.50	0.60	1.18
Application	15/24	14.98 (62.4)	4.91 (20.5)	-0.15	-0.83	0.75	2.46
Problem Solving	13/28	15.36 (54.8)	5.42 (19.4)	-0.12	-0.71	0.78	2.54
Total Test	36/60	35.79 (59.7)	11.20 (18.7)	-0.07	-0.81	0.89	3.71

Note. *k* is number of items in a subscale or the total test and *ms* is the maximum score greater than or equal to *k* given the use of dichotomously scored multiple-choice items and polytomously scored open-response items; $\bar{X}.$ – the mean; s_X – standard deviation; *sk* – skewness; *ku* – kurtosis; a_X – internal consistency (Cronbach’s alpha); s_e – standard error of measurement.
Means and standard deviations expressed as percents are shown in parentheses.

standard deviation for the knowledge subscale, likely because of the smaller number of items in the knowledge subscale. The distribution of scores on the knowledge subscale was more negatively skewed than the distributions of scores on the application and problem solving subscales, indicating that again, as a group, students performed higher on the knowledge subscale than on the application and problem solving subscales.

Next, as shown in Table 7, the values of internal consistency (i.e., reliability) of each subscale were not the same. The internal consistency (Cronbach’s alpha) of the knowledge subscale, 0.60, was lower than the internal consistencies of the application and problem solving subscales, which were more alike, 0.75 and 0.78, respectively. The somewhat low value of reliability for the knowledge subscale is likely due to the relatively smaller number of items (8) in this subscale as compared to the numbers of items in the other two subscales (15 for the application and 13 for the problem solving subscales). With the results of the assessment being used by teachers and education authorities for instructional and curriculum purposes, the low reliability values are perceived to be congruent with the spirit that the reliability values may be reduced because the decisions associated with instruction and curriculum are perceived reversible (McPeek et al., 1976). The estimate of internal consistency of the total test was 0.89, with the total number of items (both multiple choice and open response) being 36.

Detection of Performance Differences

Correlations Corrected for Attenuation (r´)

The results for the *r´* method when used with the population data are reported in Table 8. As shown in the lower triangle, regardless of the complexity and reliability of subscales in each pair, no performance difference was found either for pairs of subscales or subscale–total test pairs, thus indicating that students' performances as reflected by their rank order were similar on the subscales and the total test. As with some pairs of the English Reading subscales, the values of corrected correlations for the Mathematics subscales, except for the K_P pair, were greater than one. For the K_A and A_P pairs, the corrected correlations greater than one were observed because their uncorrected correlations were close in value to the reliabilities of subscales in each pair. Despite this finding, the decision was made to proceed with the analyses using the *r´* method to examine the distribution of the *r´* statistics in replicated random samples. For the subscale–total test pairs (i.e., K_T, A_T, and P_T), a factor contributing to the corrected correlations being greater than one was the presence of common items in the two measures. Given this, the *r´* method was not used with the subscale–total test pairs in the replicated sampling component of this study.

Table 8

Detection of Performance Differences: Uncorrected and Corrected Correlations, Mathematics Knowledge and Skills Subscales, N = 127,596

Subscale	Knowledge	Application	Problem Solving	Total Test
Knowledge	-	0.70	0.67	*0.80*
Application	1.04 (0)	-	0.79	*0.94*
Problem Solving	0.98 (0)	1.03 (0)	-	*0.94*
Total Test	*1.09* (0)	*1.14* (0)	*1.13* (0)	-

Note. Uncorrected correlations are in the upper triangle and corrected correlations are in the lower triangle. Correlation coefficients, both uncorrected and corrected, shown in italic are between two measures, with one of the measures being a part of the other. Decision made with respect to the scores on a pair of measures (0 – not different; 1 – different) is shown in parentheses.

The results for replicated random samples are presented in Table 9. As shown in this table, the consistency of decisions made using the *r´* method and sample data increased as (a) the population value of the corrected correlation increased, (b) the sample size increased, (c) the reliability increased, and (d) the difference in the complexity levels between two subscales became smaller. First, the percentage agreement between the decisions made using sample data and the decisions made in the population was high if not 100%. Second, the lowest percentage agreement occurred when *n* = 250 and increased as the sample size increased. Third, given the greater difference in complexity between the constructs measured by the knowledge and problem solving subscales than the difference between the constructs measured by the knowledge and application subscales, the lowest percentage agreement was observed for the K_P pair. With respect to the accuracy of sample estimators, the corresponding *r´* values in the

Table 9

Frequency of Decisions that Agreed with Decisions Made in the Population, Means and Standard Deviations of the Sampling Distributions of Corrected Correlations at Five Levels of Sample Size for Mathematics Knowledge and Skills Subscales

	Sample Size					
Pair	250	500	1,000	2,000	5,000	Pop.
K_A (0)[a]	998 (99.8)[b] 1.04 (.047)[c]	1000 (100) 1.04 (.031)	1000 (100) 1.04 (.023)	1000 (100) 1.04 (.016)	1000 (100) 1.04 (.010)	1.04
K_P (0)	972 (97.2) 0.99 (.048)	998 (99.8) 0.99 (.031)	1000 (100) 0.98 (.023)	1000 (100) 0.98 (.016)	1000 (100) 0.98 (.010)	0.98
A_P (0)	1000 (100) 1.03 (.026)	1000 (100) 1.03 (.018)	1000 (100) 1.03 (.013)	1000 (100) 1.03 (.009)	1000 (100) 1.03 (.006)	1.03

Note. K – knowledge; A – application; P – problem solving.
[a] Decision (0 – not different; 1 – different) made in the population.
[b] The first value is the frequency and the value in parenthesis is the percentage of random samples (out of 1,000) with the same decision as in the population at a given sample size.
[c] The first value is the mean and the value in parentheses is the standard deviation of the sampling distribution of 1,000 corrected correlations at a given sample size.

population were recovered very well (within 0.01); the standard errors of sample estimators were small and, as expected, decreased with an increase in the sample size. For n = 250, the standard errors ranged between 0.026 and 0.048, whereas for n = 5,000, the standard errors were as low as 0.006 and as high as 0.010. Given the relatively low reliability of the knowledge subscale, the standard errors of sample estimates for the K_A and K_P pairs were consistently higher than the standard errors of sample estimates for the A_P pair for all the levels of sample size. Overall, although the results for the two pairs of subscales (i.e., K_A and A_P) were uninterpretable, r' was determined to be an accurate and precise estimator of the corresponding population parameter.

Proportional Reduction of the Mean Squared Error (PRMSE)

As shown in the last column in Table 10, regardless of the complexity and reliability of each subscale, $PRMSE_{s_s} < PRMSE_{s_x}$ for all subscales, meaning that the true subscale scores were better predicted by the observed total scores than by the observed subscale scores. As with the English Reading, students' performances on each of the three knowledge and skills subscales in the Mathematics assessment were no different from their performances on the total test. The results for replicated samples are presented in columns 3 through 7 in Table 10. As shown in this table, regardless of the sample size, reliability and complexity of each subscale, the decisions made using the *PRMSE* method and sample data were 100% consistent with the decisions made in the population. The means for the sample estimators of $PRMSE_{s_x}$ and $PRMSE_{s_s}$ were within 0.01 of their corresponding population values for all three subscales and five levels of sample size. The standard errors of sample estimators were the largest when the knowledge subscale was considered (e.g., the standard error for $PRMSE_{s_x}$ = 0.100 for n = 250) but decreased as the sample size increased, ranging between 0.003 and 0.012 for n = 5,000. The largest standard errors were for the knowledge subscale for all the sample size levels, given that the internal consistency of this subscale was the smallest (0.60).

Agreement Method

The results for the agreement method are reported in Tables 11 and 12. First, the correlations of the errors on the pairs of subscales and each subscale–

Table 10

Frequency of Decisions that Agreed with Decisions Made in the Population, Means and Standard Deviations of the Sampling Distributions of $PRMSE_{s_s}$ and $PRMSE_{s_x}$ *at Five Levels of Sample Size for Mathematics Knowledge and Skills Subscales*

		Sample Size					
Subscale		250	500	1,000	2,000	5,000	Pop.
		1000 (100)[b]	1000 (100)	1000 (100)	1000 (100)	1000 (100)	
Knowledge (0)[a]	$PRMSE_{s_s}$	0.60 (.038)[c]	0.60 (.026)	0.60 (.019)	0.60 (.013)	0.60 (.008)	0.60
	$PRMSE_{s_x}$	0.90 (.055)	0.89 (.036)	0.89 (.027)	0.89 (.018)	0.89 (.012)	0.89
		1000 (100)	1000 (100)	1000 (100)	1000 (100)	1000 (100)	
Application (0)	$PRMSE_{s_s}$	0.75 (.019)	0.75 (.013)	0.75 (.010)	0.75 (.007)	0.75 (.004)	0.75
	$PRMSE_{s_x}$	0.91 (.016)	0.91 (.011)	0.91 (.008)	0.91 (.005)	0.91 (.004)	0.91
		1000 (100)	1000 (100)	1000 (100)	1000 (100)	1000 (100)	
Problem Solving (0)	$PRMSE_{s_s}$	0.78 (.017)	0.78 (.012)	0.78 (.009)	0.78 (.006)	0.78 (.004)	0.78
	$PRMSE_{s_x}$	0.89 (.014)	0.90 (.010)	0.90 (.007)	0.90 (.005)	0.90 (.003)	0.90

Note. $PRMSE_{s_s}$ denotes the proportional reduction of the mean squared error for the estimated subscale true score that is linearly predicted from the observed subscale score; $PRMSE_{s_x}$ denotes the proportional reduction of the mean squared error for the estimated subscale true score that is linearly predicted from the observed total test score.

[a] Decision (0 – not different; 1 – different) made in the population.

[b] The first value is the frequency and the value in parentheses is the percentage of random samples (out of 1,000) with the same decision as in the population.

[c] The first value is the mean and the value in parentheses is the standard deviation of the sampling distributions of the corresponding *PRMSE* at a given sample size.

total test pair were obtained to determine the tenability of the zero correlation assumption. The results of this analysis are provided in Table 11. The correlation coefficients shown in italics in Table 11 are between two measures, with one of the measures being a part of the other. Examination of the correlation coefficients revealed low, if not close to zero, correlated errors for the subscale pairs and high correlated errors for the subscale–total test pairs. Given this finding as well as the approach adopted when the subscales in the English Reading were examined, the

Table 11

Correlated Errors, Mathematics Knowledge and Skills Subscales, N = 127,596

Subscale	Knowledge	Application	Problem Solving
Knowledge	-		
Application	.092	-	
Problem Solving	-.048	.107	-
Total Test	*.330*	*.742*	*.687*

Note. Correlation coefficients shown in italic are between two measures, with one of the measures being a part of the other.

agreement method (Kelley, 1923) was used only with the pairs of the subscales.

Next, Kelley's ratio was computed for each subscale pair using first the entire population of students and then replicated random samples of different sizes. The results for the population are shown in the second last column on the right, and the results for the samples, including the mean of Kelley's ratio and its standard error across random samples, are provided in columns three to seven in Table 12. As shown in this table, the values of the ratio were greater than one for the two of the three pairs of subscales. As mentioned in Chapter II, for Kelley's procedure to work, the mean of the two reliabilities has to be greater than the correlation between the two subscales being compared. Given that the reliabilities of the Mathematics subscales and the correlations between them were very close in value, the values of Kelley's ratio greater than one or close to one were observed. However, as was the case when the *r´* method was used with the Mathematics subscales, the decision was made to use the agreement method with all the pairs of subscales in the replicated sampling component of this study, although the ratio for the two pairs were found to be uninterpretable in the

Table 12

Means and Standard Deviations of Kelley's Ratio at Five Levels of Sample Size and Percentage of Differences in Excess of the Chance

	Sample Size						
	250	500	1,000	2,000	5,000	Pop.	%
K_A	1.04 (.052)[a]	1.04 (.035)	1.04 (.026)	1.03 (.017)	1.04 (.011)	1.04	-
K_P	0.98 (.049)	0.98 (.031)	0.98 (.024)	0.97 (.016)	0.97 (.011)	0.97	≈1.5%
A_P	1.05 (.048)	1.05 (.035)	1.05 (.024)	1.05 (.018)	1.05 (.011)	1.05	-

Note. K – knowledge; A – application; P – problem solving.
[a] The first value is the mean and the value in parentheses is the standard deviation (i.e., standard error) of the sampling distribution of the corresponding statistics at a given sample size.
% – percentage of differences in excess of the chance in the population, determined using Table IV in Kelley (1923, p. 330).

population. This allowed determining whether or not Kelley's ratio was accurate and precise. The mean value of this ratio and its standard error are provided for each pair of subscales in Table 12. As shown in the table, the mean of Kelley's ratio in the random samples was within 0.01 of the corresponding population value for each subscale pair and sample size. The standard deviation of the sampling distribution (i.e., the standard error) of Kelley's ratio decreased as the sample size increased. Namely, for n = 250 the standard errors were between 0.048 and 0.052, whereas for n = 5,000 the standard errors were 0.011 for the three pairs. That is, the sample estimators of Kelley's ratio were accurate and precise.

Finally, using Table IV in Kelley (1923, p. 330), the proportion of score differences in excess of the chance was determined only for the K_P pair and not for the other two pairs due to the uniterpretability of Kelley's ratio for these two pairs. Approximately 1.5% of students in the population were determined to have differences between their scores on the knowledge and problem solving subscales that could not be attributed to the chance. Given that the means of Kelley's ratio at the five levels of sample size were within 0.01 of the corresponding population values, with the standard error being very low, the proportion of score differences in excess of the chance was not computed in the replicated random samples.

Psychometric Properties of the Content Subscales

The psychometric properties of the five content subscales and the total test are reported in Table 13 for the population of students. The means and standard deviations are reported in the observed score units and as percentages (in parentheses). The means (percentages) revealed that student performance was the highest on the algebra subscale, 66.3%, and the lowest on the probability subscale, 55.4%. The corresponding numbers for the measurement, geometry, and numeration subscales are 57.6%, 59.8%, and 60.3%. The standard deviations (percentages) were somewhat larger for the measurement and geometry subscales, 24.3% and 23.3%, respectively, than the standard deviations for the numeration, algebra, and probability subscales, which were essentially the same, 21.6%, 21.4% and 20.8%, respectively. The values of skewness further reinforce what has been stated with respect to the means of subscales. In particular, out of the five subscales, the distribution of scores on the algebra subscale was the most negatively skewed, *sk* = -0.39, meaning that the majority of students tended to perform high on this subscale; the values of skewness for the remaining subscales were close to zero. Given the numbers of items in each subscale do not differ much as they did in the case of reading and mathematics knowledge and skills,

Table 13

Psychometric Properties: Content Subscales, Mathematics, N = 127,596

Subscale	k/ms	$\overline{X}_.$	s_X	*sk*	*ku*	α_X	s_e
Numeration	8/14	8.44 (60.3)	3.02 (21.6)	-0.06	-0.80	0.63	1.84
Measurement	8/11	6.34 (57.6)	2.67 (24.3)	0.03	-1.02	0.63	1.62
Algebra	7/10	6.63 (66.3)	2.14 (21.4)	-0.39	-0.38	0.58	1.39
Probability	7/13	7.20 (55.4)	2.71 (20.8)	0.12	-0.68	0.61	1.69
Geometry	6/12	7.18 (59.8)	2.80 (23.3)	-0.08	-0.92	0.60	1.77
Total Test	36/60	35.79 (59.7)	11.20 (18.7)	-0.07	-0.81	0.89	3.71

Note. *k* is number of items in a subscale or the total test and *ms* is the maximum score greater than or equal to *k* given the use of dichotomously scored multiple-choice items and polytomously scored open-response items; $\overline{X}.$ – the mean; s_X – standard deviation; *sk* – skewness; *ku* – kurtosis; a_X – internal consistency (Cronbach's alpha); s_e – standard error of measurement.
Means and standard deviations expressed as percents are shown in parentheses.

the internal consistencies (Cronbach's alpha) of the five content subscales were essentially the same, ranging from 0.58 to 0.63. The estimate of the internal consistency of the total test was 0.89.

Detection of Performance Differences

Correlations Corrected for Attenuation (r´)

The results for the *r´* method are reported in Table 14 for the population and in Table 15 for the random sampling data. As shown in Table 14, no performance differences were found, thus indicating that students' performances as reflected by their rank order on two measures were similar for each of the pairs of subscales and subscale–total test pairs. As with the Mathematics knowledge and skills subscales, the values of the corrected correlations for the pairs of content subscales were greater than one when the uncorrected correlations were close in value to subscale reliabilities. However, the analyses for these pairs were

Table 14

Detection of Performance Differences: Uncorrected and Corrected Correlations, Content Subscales, Mathematics, N = 127,596

Subscale	N	M	A	P	G	Total Test
Numeration	-	0.66	0.62	0.67	0.63	*0.87*
Measurement	1.04 (0)	-	0.59	0.63	0.63	*0.84*
Algebra	1.03 (0)	0.97 (0)	-	0.62	0.59	*0.80*
Probability	1.09 (0)	1.02 (0)	1.04 (0)	-	0.62	*0.85*
Geometry	1.03 (0)	1.03 (0)	0.99 (0)	1.03 (0)	-	*0.83*
Total Test	*1.16* (0)	*1.12* (0)	*1.11* (0)	1.15 (0)	1.14 (0)	-

Note. N – numeration; M – measurement; A – algebra; P – probability; G – geometry; T – total test. Uncorrected correlations are in the upper triangle and corrected correlations are in the lower triangle. Correlation coefficients, both uncorrected and corrected, shown in italic are between two measures, with one of the measures being a part of the other. Decision made with respect to the scores on a pair of measures (0 – not different; 1 – different) is shown in parentheses.

still conducted using replicated random samples to examine the psychometric properties of the *r´* statistics with the Mathematics content subscales. As with the English Reading and Mathematics knowledge and skills subscales, no analyses were conducted for the subscale–total test pairs, given the presence of common items on the two measures.

The results for the *r´* method for the random samples are presented in Table 15. As shown, the accuracy of sample estimates and the consistency of the decisions made using sample data increased as (a) the population value of the corrected correlation increased and (b) the sample size increased. That is, the percentage agreement between the decisions made using sample data and

Table 15

Frequency of Decisions that Agreed with Decisions Made in the Population, Means and Standard Deviations of the Sampling Distributions of Corrected Correlations at Five Levels of Sample Size for Mathematics Content Subscales

	Sample Size					
Pair	250	500	1,000	2,000	5,000	Pop.
N_M	995 (99.5)[b]	1000 (100)	1000 (100)	1000 (100)	1000 (100)	
(0)[a]	1.05 (.054)[c]	1.05 (.037)	1.05 (.026)	1.05 (.018)	1.04 (.012)	1.04[d]
N_A	989 (98.9)	1000 (100)	1000 (100)	1000 (100)	1000 (100)	
(0)	1.04 (.062)	1.04 (.042)	1.04 (.031)	1.03 (.020)	1.03 (.014)	1.03
N_P	1000 (100)	1000 (100)	1000 (100)	1000 (100)	1000 (100)	
(0)	1.09 (.051)	1.09 (.038)	1.09 (.027)	1.09 (.018)	1.09 (.012)	1.09
N_G	989 (98.9)	999 (99.9)	1000 (100)	1000 (100)	1000 (100)	
(0)	1.03 (.058)	1.03 (.040)	1.03 (.029)	1.03 (.021)	1.03 (.013)	1.03
M_A	895 (89.5)	959 (95.9)	988 (98.8)	998 (99.8)	1000 (100)	
(0)	0.98 (.062)	0.97 (.045)	0.97 (.032)	0.97 (.022)	0.97 (.014)	0.97
M_P	986 (98.6)	999 (99.9)	1000 (100)	1000 (100)	1000 (100)	
(0)	1.02 (.055)	1.02 (.038)	1.02 (.028)	1.02 (.020)	1.02 (.013)	1.02
M_G	970 (97.0)	1000 (100)	1000 (100)	1000 (100)	1000 (100)	
(0)	1.02 (.062)	1.03 (.041)	1.03 (.031)	1.03 (.020)	1.03 (.013)	1.03
A_P	986 (98.6)	1000 (100)	1000 (100)	1000 (100)	1000 (100)	
(0)	1.04 (.064)	1.04 (.043)	1.04 (.030)	1.04 (.022)	1.04 (.013)	1.04
A_G	924 (92.4)	985 (98.5)	998 (99.8)	1000 (100)	1000 (100)	
(0)	0.99 (.066)	1.00 (.046)	0.99 (.033)	0.99 (.023)	0.99 (.014)	0.99
P_G	990 (99.0)	999 (99.9)	1000 (100)	1000 (100)	1000 (100)	
(0)	1.04 (.059)	1.03 (.042)	1.03 (.029)	1.03 (.021)	1.03 (.014)	1.03

Note. N – numeration; M – measurement; A – algebra; P – probability; G – geometry.

[a] Decision (0 – not different; 1 – different) made in the population.

[b] The first value is the frequency and the value in parenthesis is the percentage of random samples (out of 1,000) with the same decision as in the population at a given sample size.

[c] The first value is the mean and the value in parentheses is the standard deviation of the sampling distribution of 1,000 corrected correlations at a given sample size.

[d] The value of the corrected correlation in the population.

decisions made using sample data increased as (a) the population value of the corrected correlation increased and (b) the sample size increased. That is, the percentage agreement between the decisions made using sample data and thedecisions made in the population was high if not 100%. With respect to the accuracy of sample estimates, the corresponding r' values in the population were recovered very well (within 0.01); the standard errors of sample estimates were small and, as expected, decreased with an increase in the sample size. For n = 250, the standard errors ranged between 0.051 and 0.066, whereas for n = 5,000, the standard errors were between 0.012 and 0.014. Given the reliabilities of the Mathematics content subscales did not differ as much as the reliabilities of the English Reading and Mathematics knowledge and skills subscales, the standard errors for the pairs of content subscale were comparable across the pairs at each level of sample size. Overall, although the r' statistic was determined to be an accurate and precise estimator of the population parameter, the results were uninterpretable (i.e., r' > 1.00).

Proportional Reduction of the Mean Squared Error (PRMSE)

The results for the *PRMSE* method are reported in Table 16 for the population (the last column) and the replicated sampling data (columns 3 through 7). First, as shown in the last column of the table, regardless of the reliability and complexity of each subscale, $PRMSE_{s_s} < PRMSE_{s_x}$ for all subscales, meaning that the true subscale scores were better predicted by the observed total scores than by the observed subscale scores. That is, students' performances on each subscale were no different from their performances on the total test. Next, regardless of the

Table 16

Frequency of Decisions that Agreed with Decisions Made in the Population, Means and Standard Deviations of the Sampling Distributions of $PRMSE_{s_s}$ and $PRMSE_{s_x}$ at Five Levels of Sample Size for Mathematics Content Subscales

		Sample Size					
Subscale		250	500	1,000	2,000	5,000	Pop.
		1000 (100)[b]	1000 (100)	1000 (100)	1000 (100)	1000 (100)	
Numeration (0)[a]	$PRMSE_{s_s}$	0.63 (.028)[c]	0.63 (.019)	0.63 (.015)	0.63 (.010)	0.63 (.006)	0.63[d]
	$PRMSE_{s_x}$	0.94 (.037)	0.94 (.025)	0.94 (.019)	0.94 (.012)	0.94 (.008)	0.94
		1000 (100)	1000 (100)	1000 (100)	1000 (100)	1000 (100)	
Measurement (0)	$PRMSE_{s_s}$	0.63 (.029)	0.63 (.024)	0.63 (.015)	0.63 (.010)	0.63 (.006)	0.63
	$PRMSE_{s_x}$	0.90 (.042)	0.90 (.031)	0.90 (.021)	0.89 (.014)	0.89 (.009)	0.89
		1000 (100)	1000 (100)	1000 (100)	1000 (100)	1000 (100)	
Algebra (0)	$PRMSE_{s_s}$	0.58 (.038)	0.58 (.025)	0.58 (.019)	0.58 (.013)	0.58 (.008)	0.58
	$PRMSE_{s_x}$	0.89 (.055)	0.88 (.039)	0.88 (.027)	0.88 (.019)	0.88 (.012)	0.88
		1000 (100)	1000 (100)	1000 (100)	1000 (100)	1000 (100)	
Probability (0)	$PRMSE_{s_s}$	0.61 (.032)	0.61 (.022)	0.61 (.019)	0.61 (.011)	0.61 (.007)	0.61
	$PRMSE_{s_x}$	0.94 (.042)	0.93 (.029)	0.93 (.020)	0.93 (.015)	0.93 (.009)	0.93
		1000 (100)	1000 (100)	1000 (100)	1000 (100)	1000 (100)	
Geometry (0)	$PRMSE_{s_s}$	0.60 (.033)	0.60 (.024)	0.60 (.017)	0.60 (.011)	0.60 (.007)	0.60
	$PRMSE_{s_x}$	0.90 (.045)	0.90 (.031)	0.90 (.023)	0.90 (.016)	0.90 (.010)	0.90

Note. $PRMSE_{s_s}$ denotes the proportional reduction of the mean squared error for the estimated subscale true score that is linearly predicted from the observed subscale score; $PRMSE_{s_x}$denotes the proportional reduction of the mean squared error for the estimated subscale true score that is linearly predicted from the observed total test score.

[a] Decision (0 – not different; 1 – different) made in the population.

[b] The first value is the frequency and the value in parentheses is the percentage of random samples (out of 1,000) with the same decision as in the population.

[c] The first value is the mean and the value in parentheses is the standard deviation of the sampling distributions of the corresponding *PRMSE* at a given sample size.

[d] The value of the corresponding *PRMSE* in the population.

sample size, the decisions made using sample data were 100% consistent with the decisions made in the population. The means of the distributions of sample estimators of $PRMSE_{s_x}$ and $PRMSE_{s_s}$ were within 0.01 of the corresponding population values for all the five subscales. The standard errors of sample estimators were the largest when the algebra subscale was considered (e.g., the standard error for $PRMSE_{s_x}$ = 0.055 for *n* = 250) but decreased as the sample size increased, ranging between 0.006 and 0.012 for *n* = 5,000. The smallest standard errors were for the numeration and measurement subscales, given that the internal consistency for these subscales was slightly larger (both 0.63) than for the other content subscales in the Mathematics assessment.

Agreement Method

The results for the agreement method are reported in Tables 17 and 18. First, the correlations of the errors on the pairs of subscales and each subscale–total test pair were obtained. The results of this analysis are provided in Table 17. As shown in this table, whereas the correlated errors for the pairs of subscales were low, the correlated errors for the subscale–total test pairs were high due to the common items present in the two measures. Given these findings, the agreement method was used only with the pairs of subscales.

Next, Kelley's ratio was computed for each subscale pair using first the entire population of students and then replicated random samples of different sizes. The results for the population are shown in the second last column on the right, and the results for the samples, including the mean of Kelley's ratio and its standard error across random samples, are provided in columns three to seven in

Table 17

Correlated Errors, Mathematics Content Subscales, N = 127,596

Subscale	N	M	A	P	G
Numeration	-				
Measurement	.081	-			
Algebra	.039	-.037	-		
Probability	.132	.027	.062	-	
Geometry	.039	.039	.000	.038	-
Total Test	*.601*	*.452*	*.379*	*.546*	*.473*

Note. Correlation coefficients shown in italic are between two measures, with one of the measures being a part of the other.

Table 18. As shown in the table, the values of Kelley's ratio were greater than one for all but one pair of subscales. Given that the reliabilities of the Mathematics content subscales and the intercorrelations among the subscales were very close in value, the values of Kelley's ratio greater than one were observed for these pairs. The results for the agreement method when used with random samples were congruent with the results at the population level. The mean value of Kelley's ratio and its standard error are provided for each pair of subscales in Table 18. As shown in the table, the mean was within 0.01 of the corresponding population value for each subscale pair and sample size. The standard error of Kelley's ratio decreased as the sample size increased. Namely, for *n* = 250 the standard errors were between 0.045 and 0.051, whereas for *n* = 5,000 the standard errors were no greater than 0.011 for all the pairs of subscales. That is, the sample estimators of Kelley's ratio were accurate and precise.

Table 18

Means and Standard Deviations of Kelley's Ratio at Five Levels of Sample Size and Percentage of Differences in Excess of the Chance for Mathematics Content Subscales

	Sample Size						
	250	500	1,000	2,000	5,000	Pop.	%
N_M	1.04 (.051)[a]	1.04 (.035)	1.04 (.025)	1.04 (.017)	1.04 (.011)	1.04	-
N_A	1.03 (.050)	1.03 (.034)	1.03 (.025)	1.03 (.016)	1.03 (.011)	1.02	-
N_P	1.09 (.050)	1.09 (.037)	1.08 (.026)	1.08 (0.18)	1.08 (.011)	1.07	-
N_G	1.03 (.049)	1.03 (.037)	1.03 (.025)	1.02 (.017)	1.02 (.011)	1.02	-
M_A	0.98 (.045)	0.98 (.033)	0.98 (.023)	0.98 (.016)	0.98 (.010)	0.98	≈1.0%
M_P	1.02 (.046)	1.02 (.032)	1.02 (.024)	1.02 (.017)	1.02 (.010)	1.01	-
M_G	1.02 (.052)	1.03 (.035)	1.02 (.026)	1.02 (.017)	1.02 (.011)	1.02	-
A_P	1.03 (.049)	1.03 (.033)	1.03 (.024)	1.03 (.017)	1.03 (.010)	1.03	-
A_G	1.00 (.047)	1.00 (.033)	1.00 (.023)	1.00 (.017)	1.00 (.010)	1.00	-
P_G	1.03 (.048)	1.03 (.034)	1.03 (.023)	1.03 (.017)	1.03 (.011)	1.02	-

Note. N – numeration; M – measurement; A – algebra; P – probability; G – geometry.
[a] The first value is the mean and the value in parentheses is the standard deviation (i.e., standard error) of the sampling distribution of the corresponding statistics at a given sample size.
% – percentage of differences in excess of the chance in the population, determined using Table IV in Kelley (1923, p. 330).

Finally, using Table IV in Kelley (1923, p. 330), the proportion of score differences in excess of the chance was determined only for the M_A pair and not for the remaining pairs due to the uniterpretability of Kelley's ratio for these pairs. Approximately 1% of the students in the population were determined to have differences in their scores on the measurement and algebra subscales that could not be attributed to the chance.

Summary

To summarize, as with the English Reading, both correlational methods (i.e., *r´* and *PRMSE*) led to the same decisions when used with the Mathematics knowledge and skills subscales and the Mathematics content subscales. Namely, student performances on the pairs of subscales and subscale–total test pairs were determined to be no different. Using the agreement method with the Mathematics knowledge and skills subscales, it was determined that there were 1.5% of students with differences between their scores on the knowledge and problem solving subscales greater than the chance. In case of the Mathematics content subscales, 1.0% of students had differences between their scores on the measurement and algebra subscales greater than what could be attributable to the chance. For the remaining pairs of subscales, the differences in subscale scores were unlikely due to anything systematic when the agreement method was used. The statistics used for each method were determined to be accurate and precise. Sample estimators were within 0.01 of the corresponding population parameters. Standard errors of sample estimators decreased as the sample size increased from *n* = 250 to *n* = 5,000.

CHAPTER VI: SUMMARY AND CONCLUSIONS

The final chapter is organized in six sections. A brief summary of the purposes of the present study, the research method and analyses conducted to determine if the three detection methods identified subscale scores as distinct among each other and from the total score is provided in the first section. A summary of findings for each detection method is presented in the second section, followed by the explanation of findings. The limitations of the study are identified in the third section followed by the conclusions formulated from the findings and taking into account the limitations. The last two sections contain, respectively, the implications for practice and recommendations for future research.

Summary of the Purposes, Research Method, and Analyses

The purposes of the present study were to determine whether

1. the correlations corrected for attenuation, proportional reduction of the mean squared error, and the agreement methods led to the same or different decision regarding the reporting of subscale scores; and
2. the statistics used for each method were accurate and precise.

The three detection methods were:

- correlations corrected for attenuation due to unreliability (*r´*) (Haladyna & Kramer, 2004; Harris & Hanson, 1991; McPeek, et al., 1976);
- proportional reduction of the mean squared error (*PRMSE*) (Haberman, 2005, 2008; Lyren, 2009; Sinharay et al., 2007; Sinharay, 2010); and

- agreement method based on the ratio of the standard error of the difference due to measurement error and the standard deviation of the difference (Gulliksen, 1951; Kelley, 1923; Lord & Novick, 1968).

Whereas the *r´* and, especially, the *PRMSE* methods have been extensively used with large-scale assessments, the agreement method introduced by Kelley in 1923 is a method that has not received much attention in achievement testing and, therefore, was re-visited in the present study.

The data were provided by the Education Quality and Accountability Office (EQAO), which, as an agency at arm's length to the Ontario provincial government, aims to provide parents, teachers, and the public with reliable information that can be validly interpreted about student achievement. Like other large-scale assessment agencies, the EQAO considers score reporting by subscale on their achievement assessments in response to the feedback from teachers doing the same when interpreting reports they receive on their students' achievement. However, empirical evidence is required for assessment programs that consider subscore reporting in addition to reporting total test scores on their assessments to be gathered to support their decisions.

The EQAO assessments considered in this study included (a) the 2009 Junior English Reading with three process subscales and (b) the 2009 Junior Mathematics with three knowledge and skills subscales and five content subscales. These large-scale assessments are developed to reflect *The Ontario Curriculum* and administered annually to students in Grade 6 in all publically funded schools in the province of Ontario. Using the three detection methods with these assessments, the analyses were conducted first using the population of students assessed in 2009, with N = 128,089 for the English Reading assessment and N = 127,596 for the Mathematics assessment. Following this, the analyses were repeated for five different sample sizes – 250, 500, 1,000 2,000, and 5,000 – to examine the effect of sample size on each of the three detection methods considered in this study in terms of how well the three methods agree when detecting score differences and how accurate and precise the statistics used with each method are. A replicated sampling method, in which 1,000 samples were randomly drawn from the population with replacement at each of the five levels of sample size, was used to examine decisions made at different levels of sample size for the three detection methods. For each detection method, the consistency of the decisions was evaluated using the percentage of samples that led to the same decision made at the population level. Means and standard deviations (i.e., standard errors) of the distributions of sample estimators (i.e., $r´$, $PRMSE_{s_s}$ and $PRMSE_{s_x}$, Kelley's ratio) were used to evaluate the three detection methods in terms of their accuracy and precision.

Summary of Findings

Correlations Corrected for Attenuation (r´)

When the $r´$ method was used with the English Reading, only one pair of subscales was determined to be distinct at the population level, meaning that students' performances in terms of their rank-ordered positions on the two subscales were different. In particular, the explicit information and connections subscales, which were of the lowest and the highest levels of complexity respectively, had the corrected correlation, 0.89, slightly below the value specified by the decision rule, 0.90. For the Mathematics assessment, none of the pairs of subscales was determined to be distinct at the population level. Further, the corrected correlations for the two out of three pairs of the knowledge and skills subscales and the eight out of ten pairs of the content subscales were greater than one. This was likely due to low reliabilities of subscales and/or the mean of the reliabilities and correlations being of similar magnitude. The corrected correlations greater than one were also observed for all the subscale–total test pairs both for the English Reading and Mathematics assessments. This result was due to the presence of common items on the two measures. Given this, the *r´* method was not used with these pairs in the replicated sampling component of this study. The consistency of the decisions made using sample data for the English Reading subscales varied between 51% and 100% with the lower values for the pairs involving the explicit information subscale, which had the lowest reliability (0.47). In the case of the Mathematics assessment, the consistency of the decisions varied between 97% and 100% for the knowledge and skills subscales and between 89% and 100% for the content subscales. For all three sets of subscales, the consistency of decisions increased as (a) the population value of the corrected correlation increased, (b) the sample size increased, (c) the reliability increased, and (d) the difference in the complexity levels between two subscales became smaller. The sample estimators of the corrected correlations were accurate, within 0.01 of the population values for the three sets of subscale pairs. The standard errors of sample estimators decreased as the sample size increased. For n = 250, the standard errors ranged from 0.029 to 0.081 for the English Reading subscales, 0.026–0.048 for the Mathematics knowledge and skills subscales, and 0.051–0.066 for the Mathematics content subscales, whereas for *n* = 5,000, the standard errors ranged from 0.007 to 0.017 for the English Reading, 0.006–0.010 for the Mathematics knowledge and skills subscales, and 0.012–0.014 for the Mathematics content subscales. Given the low reliability of the explicit information subscale (0.47) in the English Reading, the standard errors for the pairs involving this subscale (i.e., E_I and E_C) were consistently higher than the standard errors for the other pairs of subscales.

Proportional Reduction of the Mean Squared Error (PRMSE)

When the *PRMSE* method was used with the population data, students' true performances on the subscales in the English Reading and Mathematics assessments were determined to be better predicted from students' performances on the total tests than from their performances on the corresponding subscales. Regardless of the complexity and reliability of subscales, $PRMSE_{s_s}$was smaller than $PRMSE_{s_x}$ for each subscale.The decisions made using sample data were 100% consistent with the decisions made in the population, irrespective of the (a) sample size, (b) reliability, and (c) complexity of each subscale. For both the English Reading and Mathematics assessments, the sample estimators of $PRMSE_{s_s}$ and $PRMSE_{s_x}$were accurate, within 0.01 of the corresponding population values. The standard errors of sample estimators were the largest when the explicit information subscale in the English Reading was considered (e.g., the standard error for $PRMSE_{s_x}$ = 0.100 for *n* = 250) but decreased as the sample size increased.

The smallest standard errors across the five levels of sample size were for the combined information subscale in the English Reading, given that the internal consistency for this subscale was the largest (0.80). The standard errors of sample estimators for the Mathematics content subscales were comparable across the five levels of sample size, given the values of internal consistency for these subscales (0.59–0.63) did not vary as much as they did for the English Reading subscales (0.47–0.80) and the Mathematics knowledge and skills subscales (0.60–0.78). Overall, the statistics for the *PRMSE* method were determined to be accurate and precise.

However, the fact that the true subscale score were determined to be better predicted by the observed total score that contains information on students' performances on each of the subscales, questions the appropriateness of using the *PRMSE* method for determining the "added value" of subscale scores over the total score. As indicated by Sinhary et al. (2007), the finding that the *PRSME* method did not reveal that the subscale scores had added value over the total score is due to the low reliability estimates and the high correlations between the true scores on each subscale and the total test. However, it is not clear whether low correlations between true scores on a subscale and the total test can be realized given the total test contains the subscale with which it is being correlated.

Agreement Method

First, in response to issues regarding the need to consider the correlation between error scores for the pair of subscales or each subscale with the total test (Zimmerman, 1981), the correlations for each pair were examined. Whereas the correlations between the errors on the pairs of subscales were determined to be low ($< |0.20|$), the correlations between the errors for the subscale–total test pairs were higher (> 0.33, with the exception of the explicit information subscale for the English Reading (0.19)), due to the presence of common items. Therefore, as with the *r´* method, the agreement method was used with the pairs of subscales given the low correlated errors (less than 10% shared variance) and not with the subscale–total test pairs.

For both the English Reading and Mathematics assessments, the sample estimators of Kelley's ratio were determined to be accurate. For each pair of subscales, the sample estimators were within 0.01 of the population values across the five levels of sample size. The standard error of Kelley's ratio decreased as the sample size increased. In particular, for *n* = 250 the standard errors ranged from 0.043 to 0.049 for the English Reading, 0.048–0.052 for the Mathematics knowledge and skills subscales, and 0.045–0.052 for the Mathematics content subscales, whereas for *n* = 5,000, the standard errors ranged from 0.009 to 0.011 for the English Reading, the Mathematics knowledge and skills subscales, and the Mathematics content subscales. That is, Kelley's ratio was accurate and precise.

Turning to the agreement among the subscale scores, the values of Kelley's ratio were high. For the English Reading, the values ranged from 0.91 to 0.98. Depending on a subscale pair, between one to five percent of students had differences in subscale scores greater than the chance. That is, the score differences for these students were attributable to something systematic. The largest percentage (i.e., ≈5.0%) of students was observed for the explicit information and connections subscales, which were, respectively, of the lowest and the highest levels of complexity. For the Mathematics assessment, the values of Kelley's ratio exceeded 1.00 for all but one pair of knowledge and skills subscales and one pair of content subscales. Values for Kelley's ratio equal to or greater than one are not interpretable – the standard error of the difference due to measurement error theoretically cannot be equal to or exceed the standard deviation of differences. For the two pairs of Mathematics subscales with interpretable ratios, 1.5% of students had differences in their scores on the knowledge and problem solving subscales greater than the chance, and 1.0% of students had greater-than-the-chance differences between their scores on the measurement and algebra subscales. For the remaining pairs of Mathematics subscales, the score differences were unlikely due to anything systematic.

To summarize, both correlational methods (i.e., *r´* and *PRMSE*) led to the same decisions when used with the English Reading, Mathematics knowledge and skills subscales, and the Mathematics content subscales. Namely, student performances on the pairs of subscales and subscale–total test pairs were determined to be no different. Further, most of the corrected correlations could not be meaningfully interpreted due to their values being greater than one. For the *PRMSE* method, none of the true subscale scores were determined to be better predicted by the corresponding observed subscale score than by the observed total score. In contrast, the agreement method revealed that between 1.0% to 5% of students had subscore differences greater in absolute value than the difference expected due to the chance when used with the English Reading assessment.

Using the agreement method with the Mathematics knowledge and skills subscales, it was determined that there were 1.5% of students with differences between their scores on the knowledge and problem solving subscales greater than the chance. For the Mathematics content subscales, 1.0% of students had differences between their scores on the measurement and algebra subscales greater than what could be attributable to the chance. For the remaining pairs of Mathematics subscales, the results of the agreement method were uninterpretable and thus, differences in subscale scores were unlikely due to anything systematic. Overall, although the results of the analyses for the three methods could not be meaningfully interpreted in some cases, the decisions made in replicated samples were consistent with the decisions made in the population, with the statistics for each method being accurate and precise estimators of the corresponding population parameters.

Explanation of Findings

The reliabilities of and intercorrelations among the subscales in the two assessments used to examine the different detection methods were close in value. Based on their work, Haberman (2005) and Sinharay et al. (2007) concluded that “subscores are most likely to have value if they have relatively high reliability by themselves and if the true subscale score and the true total score have only a moderate correlation. Both conditions are important” (Sinharay et al., 2007, p. 28). Such was not the case for most of the pairs in the present study. With the reliabilities of 0.47–0.80 for the subscales, and 0.87 for the English Reading and 0.89 for the Mathematics total tests, and the corrected correlations between each subscale and the total test being all greater than one, the two conditions stated by Sinharay et al. (2007) were not realized in the two assessments. Further, Sinharay et al. (2007) noted that the *PRMSE* method is likely to provide support for the reporting of subscale scores “for tests with reasonably large number of items in each subcategory and composed of distinct subcategories” (p. 28). The former condition ensures higher subscore reliabilities, while the second condition ensures moderate correlations of each subscale with the total test. However, reliability is contingent upon the number of items included in each subscale, with higher numbers contributing to higher reliability. The number of items included in each subscale in the two assessments considered in this study was as low as 6 and as high as 18 items, and thus, the reliabilities for some of the subscales were low to moderate (0.47–0.80).

Similarly, Kelley’s agreement method will only work if the mean of the reliabilities of subscales is greater than the correlation between the two subscales, with the greater the difference leading to the identification of the number of students with systematic differences in their subscale scores. Again, this statistical condition is most likely to be satisfied if subscales were specifically developed to measure a multidimensional construct or domain, with clearly defined subdomains (substantive condition). As mentioned earlier, given the English Reading and Mathematics assessments were developed to report total scores, the achievement domain assessed by each assessment was likely conceived as unidimensional rather than multidimensional, with no clear subdomains, and thus, moderate to high inter-correlations (0.59–0.79) were observed. Clearly, subscores have meaning only in the multidimensional case. The results of this study provide support for Luecht et al.'s (2006) contention that "inherently unidimensional item and test information cannot be decomposed to produce useful multidimensional score profiles – no matter how well intentioned or which psychometric model is used to extract the information" (p. 6).

Limitations of the Study

This study was limited by the assessments examined. As mentioned earlier, the data used in the study were obtained from the Education Quality and Accountability Office (EQAO) in the province of Ontario. The data consisted of students' scores on the Junior (Grade 6) English Reading and Mathematics assessments administered in June 2009. The EQAO administers assessments at other grade levels (i.e., Primary Division (Grades 1–3), Grade 9) that were not considered in the present study. Likewise, assessments administered by other large-scale agencies were not considered.

The EQAO assessments were initially developed to yield total scores and not specifically for diagnosing students' strengths and weaknesses in the subdomains specified in the table of specifications for each assessment. Given the original purpose of the two assessments, the reliabilities of the subscales used in this study (0.47–0.80) were borderline, if not below the minimum reliability required for subscore reporting. But the minimum is somewhat controversial. According to McPeek et al. (1976), only subscores that have a reliability of at least 0.80 should be reported to examinees for high-stake purposes such as admission. However, he goes on to say if subscores are to "be used only for guidance and placement purposes, the statistical standards for reliability could be greatly reduced" (p. 3) because "guidance and placement decisions are perceived reversible, whereas admissions decisions generally are not" (p. 1). Salvia and Ysseldyke (2001) recommended that the minimum reliability value for reporting subscale scores be set at 0.60. In this study, the examination of the effect of reliability on the different detection methods was limited to the values of reliabilities of subscales (0.47–0.80) in the English Reading and Mathematics assessments, with several of the reliabilities being below the minimum reliability recommended by Salvia and Ysseldyke (2001). Finally, given the English Reading and Mathematics assessments were developed to report total scores, the achievement domain assessed by each assessment was likely conceived to be unidimensional rather than multidimensional, thus leading to moderate to high inter-correlations (0.59–0.79). Consequently, the examination of the different detection methods was limited to these correlations and reliabilities of subscales in the English Reading and Mathematics.

Conclusion

Based on the results of the study, the following is concluded:

1. the correlational methods agreed;
2. the agreement method did not agree with the correlational methods when the correlations between the subscales was lower than the mean of the reliabilities;
3. the statistics used for the three methods were accurate and consistent;

and

4. three conditions need to be met, one substantive (multidimensional construct for which scores are wanted for each dimension), and the other two statistical (high reliabilities of and low intercorrelations among subscales). In agreement with Sinharay et al. (2007), the results of this study clearly show that, although the estimates are accurate and consistent, “subscores are most likely to have value if they have relatively high reliability by themselves and if the true subscale score and the true total score have only a moderate correlation. Both conditions are important” and “...for tests with reasonably large number of items in each subcategory and composed of distinct subcategories” (Sinharay et al., 2007, p. 28).

Implications for Practice

The rationale for this study was predicated on the notion that reporting of subscale scores allows extracting information about students’ strengths and weaknesses on the assessed subdomains with minimum testing time. However, a number of issues need to be considered when reporting subscale scores in addition to the total scores. The implications for practice include those for test development and use of assessment results by teachers in the field.

Whether or not to generate and report subscale scores should be established at the beginning of the test development process. At the stage of defining or prescribing the domain to be measured by a test, which should be the beginning activity, effort should be devoted to determining if the domain is unidimensional or multidimensional. If the domain is found to be unidimensional, then the question of subdomains and subscores is irrelevant. On the other hand, if the domain is found to be multidimensional and scores for the subdomains are to be reported, then the items and final test form should be developed accordingly (Haladyna & Kramer, 2004; Luecht et al., 2006). That is, domain clarity should be established in the first place. In the multidimensional case, the subdomains to be assessed by the subscales should be clearly distinguishable and the items included in each subscale should be relevant to and representative of the subdomain to which they are referenced and not to the other subdomains assessed by the test. Further, a sufficient number of items should be included in each subscale so that the reliabilities of subscales are high. In sum, to warrant reporting of subscale scores, tests must be built to satisfy three conditions, one substantive (multidimensional construct for which scores are wanted for each dimension), and the other two statistical (high reliabilities of and low intercorrelations among subscales).

With respect to the use of assessment results by teachers in the field, the following implication is in place. Given that teachers use scores to identify areas of strength and areas that need to be addressed for individual students and/or to alter their instruction to improve their students' learning and achievement, only Kelly's (1923) agreement method provides such information. The agreement method works with actual score differences on pairs of subscales, thereby providing information on the magnitude of score differences for individual students. Neither the *r*´ nor the *PRMSE* methods are able to provide directly such information to teachers. If the correlational methods suggest that there are differences, teachers would then look at an individual students' scores on the pairs of subscales for which distinctiveness was found. And this is what the agreement method is able to provide to teachers directly, thus, meeting teachers' need for useful and interpretable assessment results.

Recommendations for Future Research

Based on the findings and in the light of limitations of this study, the following recommendations for future research are in place.

1. Examination of the performance of different detection methods with assessments specifically designed to have subscales for diagnostic purposes (i.e., assessments of constructs and achievement areas that are multidimensional) is needed.
2. Simulation studies are needed to systematically examine the effects of reliability of subscales and correlations among subscales and each subscale with the total test for each detection method given the domain is multidimensional. This will allow determining the minimum magnitude of the difference between the mean of reliabilities and correlations required for differences in subscale scores for individual students to be detected.

The results from such work will allow making generalizations with respect to the use of the detection methods in a variety of assessment contexts and situations.

Bibliography

American College Testing. (1989). *P-ACT+ Program technical manual*. Iowa City, IA: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Anastasi, A. (1988). *Psychological testing (6th ed.).* New York: Macmillan Publishing Company, Inc.

Bock, R. D., Thissen, D., & Zimowski, M. F. (1997). IRT estimation of domain scores. *Journal of Educational Measurement, 34,* 197-211.

Chalifour, C. & Powers, D. E. (1988). Content characteristics of GRE Analytical Reasoning items (ETS Research Report No. RR-88-07). Princeton, NJ: Educational Testing Service.

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

Cronbach, L. J., Schönemann, P., & McKie, T. D. (1965). Alpha coefficients for stratified-parallel tests. *Educational and Psychological Measurement,* 291-312.

Cronbach, L. & Furby, L. (1970). How we should measure “change”? – or should we? *Psychological Bulletin, 74,* 68-80.

Dorans, N. J. (2005, March). Why trait scores can be problematic. Presentation to the ETS Visiting Panel on Research, Princeton, NJ: ETS.

Dwyer, A., Boughton, K.A., Yao, L., Lewis, D., & Steffen, M. (2006). A comparison of subscore augmentation methods using empirical data. Paper presented at the National Council on Measurement in Education, San Francisco, CA, USA.

Edwards, M. C. & Vevea, J. L. (2006). An empirical Bayes approach to subscore augmentation: How much strength can we borrow? *Journal of Educational and Behavioral Statistics, 31*, 241–259.

Gessaroli, M. E. (2004, April). Using hierarchical multidimensional item response theory to estimate augmented subscores. Paper presented at the annual meeting of the National Council for Measurement in Education, San Diego, CA.

Glass, G. V. & Stanley, J. C. (1970). *Statistical methods in education and psychology*. Prentice-Hall, Inc., Englewood Cliffs, NJ.

Glass, G. V. & Hopkins, K. D. (1996). *Statistical methods in education and psychology*. Needham, MN: Allyn & Bacon.

Grandy, J. (1992). Construct validity study of the NTE core battery using confirmatory factor analysis. (ETS Research Report No. RR-92-03). Princeton, NJ: Educational Testing Service.

Gulliksen, H. (1950, 1967). *Theory of mental tests.* New York: John Wiley & Sons, Inc.

Haberman, S. J. (2005). When can subscores have value? (ETS Research Report No. RR-05-08). Princeton, NJ: Educational Testing Service.

Haberman, S. J. (2008). Subscores and validity. (ETS Research Report No. RR-08-64). Princeton, NJ: Educational Testing Service.

Haberman, S. J., Sinharay, S., & Puhan, G. (2009). Reporting subscores for institutions. *British Journal of Mathematical and Statistical Psychology, 62,* 79–95.

Haladyna, T. M. & Kramer, G. A. (2004). The validity of subscores for a credentialing test. *Evaluation and the Health Professions, 27,* 349–368.

Hanson, B. A. (1989). Scaling the P-ACT+. In R. L. Brennan (Ed.) *Methodology used in scaling the ACT Assessment and P-ACT+*. Iowa City, IA: American College Testing.

Harris, D. J. & Hanson, B. A. (1991, April). Methods of examining the usefulness of subscores. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Joint Commission on National Dental Examinations (2004). *National board dental examinations technical report.* Chicago: American Dental Association.

Kahraman, N. & Kamata, A. (2004). Increasing the precision of subscale scores by using out-of-scale information. *Applied Psychological Measurement, 28,* 407-426.

Kelley, T. L. (1923). A new method for determining the significance of differences in intelligence and achievement scores. *Journal of Educational Psychology, 14,* 300-303.

Kelley, T.L. (1947). *Fundamentals of statistics.* Cambridge, MA: Harvard University Press.

Kuder, G. & Richardson, M. (1937). The theory of estimation of test reliability. *Psychometrika*, *2,* 151-160.

Linn, R. (1989). *Educational measurement* (3rd ed.). New York: American Council on Education and Macmillan.

Longford, N. T. (1990). Multivariate variance component analysis: An application in test development. *Journal of Educational and Behavioral Statistics, 15,* 91-112.

Lord, F. M. (1965). A strong true-score theory with applications. *Psychometrika, 30,* 239-270.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. New York: Addison-Wesley.

Luecht, R. M., Gierl, M. J., Tan, X., & Huff, K. (2006, April). Scalability and the development of useful diagnostic scales. Paper presented at the annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Lyrén, P. E. (2009). Reporting Subscores from College Admission Tests. *Practical Assessment, Research and Evaluation, 14(4),* 1-10.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education and Macmillan.

McPeek, M., Altman, R., Wallmark, M., & Wingersky, B. C. (1976). An investigation of the feasibility of obtaining additional subscores on the GRE Advanced Psychology Test (GRE Board Professional Report No. 74-4P). Princeton, NJ: Educational Testing Service. (ERIC Document No. ED163090)

Monaghan, W. (2006). The facts about subscores (ETS R&D Connections No. 4). Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/ RD_Connections4.pdf.

O'Connor, E. (1972). Extending classical test theory to the measurement of change. *Review of Educational Research, 42,* 73-97.

Puhan, G. (2003). Evaluating the effectiveness of two-stage testing for English and French examinees on the SAIP science 1996 and 1999 tests. (Unpublished doctoral dissertation). University of Alberta, Edmonton, Canada.

Puhan, G., Sinharay, S., Haberman, S. J., & Larkin, K. (2008). Comparison of subscores based on classical test theory methods (ETS Research Report No. RR-08-54). Princeton, NJ: Educational Testing Service.

Rogosa, D. & Willett, J. (1983). Demonstrating the reliability of the difference score in the measurement of change. *Journal of Educational Measurement, 20,* 335-343.

Ryan, J. (2003). *An analysis of item mapping and test reporting strategies.* Greensboro, NC: South Carolina Department of Education.

Salvia, J. & Ysseldyke, J. (2001). *Assessment.* (8th ed.). New York: Houghton Mifflin Company.

Schmeiser, C. B. & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307-354). Washington, DC: American Council on Education.

Sinharay, S., Haberman, S. J., & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice, 26,* 21-28.

Sinharay, S., Puhan, G., & Haberman, S. (2009). Reporting diagnostic scores: Temptations, pitfalls, and some solutions. Paper presented at the National Council on Measurement in Education, San Diego, CA, USA.

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement, 47,* 150-174.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15,* 72-101.

Tate, R. L. (2004). Implications of multidimensionality for total score and subscore performance. *Applied Measurement in Education, 17,* 89-112.

Thorndike, R. (1971). *Educational measurement* (2rd ed.). Washington, D.C.: American Council on Education.

Wainer, H., Sheehan, K. M., & Wang, X. (2000). Some paths toward making Praxis scores more useful. *Journal of Educational Measurement, 37,* 113-140.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., Rosa, K., Nelson, L., Swygert, K. A., & Thissen, D. (2001). Augmented scores –"borrowing strength" to compute scores based on small numbers of items. In *Test Scoring* (pp. 343-387). Mahwah, NJ: Lawrence Erlbaum Associates.

Yao, L. & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement, 31,* 83-105.

Yen, W. M. (1987, June). A Bayesian/IRT index of objective performance. Paper presented at the annual meeting of the Psychometric Society, Montreal, Québec, Canada.

Zimmerman, D., Brotohusodo, T., & Williams, R. (1981). The reliability of sums and differences of test scores: Some new results and anomalies. *Journal of Experimental Education, 49*, 177-186.

Zimmerman, D. & Williams, R. (1982). Gain scores in research can be highly reliable. *Journal of Educational Measurement, 19*, 149-154.

Zimmerman, D. & Williams, R. (1998). Reliability of gain scores under realistic assumptions about properties of pretest and posttest scores. *British Journal of Mathematical and Statistical Psychology, 51*, 343-351.

Zimmerman, D. (2009). The reliability of difference scores in populations and samples. *Journal of Educational Measurement, 46,* 19-42.