This is to certify that the

dissertation entitled

Evaluating Content Validity
in Cross-National Achievement Tests
presented by

Pamela Marie Jakwerth

has been accepted towards fulfillment
of the requirements for

Ph.D degree in Measurement & Quantitative Methods

William A Schmidt
Major professor

Date 11-14-96

EVALUATING CONTENT VALIDITY IN CROSS-NATIONAL ACHIEVEMENT TESTS

By

Pamela M. Jakwerth

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Counseling, Educational Psychology, and Special Education

1996

ABSTRACT

The main purpose of this study was to use the results of an extensive multi-national curriculum analysis to analyze the content of a cross-national mathematics achievement test. A second purpose was to determine the impact on national scores and ranks that would result from altering test content to improve curricular match. The ultimate goal was to use this information to enhance the validity of cross-national comparisons of student achievement.

I compared data on the mathematics curriculum of 17 nations to the content of the TIMSS mathematics field trial instrument for 13 year old students. Three different data sources from the curriculum analysis component of the Third International Mathematics and Science Study (TIMSS) were used to describe the intended mathematics curriculum of each country. I also used the curriculum data to develop several sets of test specifications based on different methods of summarizing the mathematics curricula in the 17 countries. Using the country performance data from the field trial, I calculated country mean scores on each of the specified "tests." I then ranked each country on each test and compared the country scores and ranks across the different tests

The content of the mathematics curriculum varied across and within the 17 nations involved in this study. Consequently, the content of the field-trial instrument matched the content of the curriculum of some of the countries better than others. This variation in curriculum and differential match has implications

for the validity of inferences made from the test, but a final conclusion of test validity will depend on the purpose for which the test will be used.

Variation in county scores and ranks on the different tests I developed was minimal; however, some isolated variations did exist. Patterns suggest that, at the total score level, the impact of test-curriculum mis-match is likely to be minimal. However, the presence of variation in performance across topics and performance expectations indicate that total scores may be reflecting a general math ability, rather than achievement of a particular curriculum. The implication is that the concept of test-curriculum match is more complex than merely matching on topic coverage.

for my mother.

ACKNOWLEDGMENTS

I can't believe I am finally writing this. So many people have helped to make this accomplishment possible. First, I need to thank my advisor Dr. Betsy Becker and my dissertation supervisor Dr. William Schmidt. Dr. Becker has worked with me for over eight years. She has read, edited, and re-read countless pieces of work, and is responsible for many of my skills as a writer and researcher. Dr. Schmidt employed me - twice - and supplied the topic of this dissertation. He was the person that finally pushed me enough to get this thing finished. I appreciate all the input and time from the rest of my committee, Dr. Bill Mehrens, Dr. Richard Houang, and Dr. Sandra Wilcox. Each has contributed a different perspective to this work, and I thank them for enlightening me. I also have to thank Bill Frey for refusing to let me give up. He has served many roles in my life over the past eight years - from surrogate parent to mentor - and has taught me so much. Thanks to my family - who can finally stop asking when I will finish school, my friends - who have been absolutely encouraging, and my coworkers past and present - who bore the brunt of a lot of my stress. I am so grateful you all have stood by me. Finally, I would not be here today if it were not for my mother. Over eight years ago, she challenged me into pursuing this degree. She was my sounding-board for many years. I only can hope that she somehow knows what I have accomplished and is proud of what I have become. I never imagined what a struggle it would be to get here. Thank you all.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

Introduction and Study Focus

> Interest has never been higher in comparable information about education internationally, both for noble and ignoble reasons. In certain hands, such information opens a window to a whole new world, one becoming increasingly smaller in this information-technology-driven age of global communication and conversation. In other hands, the same information can serve as a sword to slay imagined enemies and vanquish challengers to the power and status of nations.
>
> We cannot escape the ideological use and misuse of cross-national data for political purposes. We can only hope to overwhelm the most base misrepresentations with the wealth of knowledge and understanding international studies can provide. These are the motivations that historically have led scholars world-wide to engage in cross-national studies through IEA and that have convinced enlightened government and non-governmental officials to support these efforts. (Burstein, 1993, p. xxxi)

Introduction

Studies comparing the structure of educational systems and the performance of students in nations across the world have been a reality for over 30 years. Educators, policy-makers, and researchers maintain that comparative cross-national studies provide nations with a broad perspective for ascertaining the effectiveness of their educational systems (Linn & Baker, 1995; Mislevy, 1995; Porter, 1990; Robitaille, McKnight, Schmidt, Britton, Raizen & Nicol, 1993; Schmidt & Valverde, 1995). Information from these studies can be used as input for policy decisions aimed at educational improvement. Comparative studies also are conducted *within* nations to monitor educational effectiveness. Within the United States, for example, such studies may use results of student achievement testing to compare states (e.g., National Assessment of Educational

Progress - NAEP), districts (e.g., Michigan Educational Assessment Program, MEAP; California Learning Assessment System, CLAS; Kentucky Instructional Results Information System, KIRIS), or programs within districts (LaPointe, 1991).

Researchers conducting comparative-education studies typically collect a wide array of information from participating educational systems. In addition to collecting student performance data, comparative researchers may collect descriptive information related to the structure and processes of each educational system or attitudinal information from stakeholders such as students, teachers, or administrators. Despite the availability of descriptive information, however, the public, educators, and policy-makers focus much of their attention on student performance results, and, often, these results receive the primary emphasis in reporting and analysis (Husen, 1987; Linn, 1988).

One popular approach for reporting student performance results in cross-national studies is to rank countries using total scores, or selected sub-scores, on tests presumed to measure student achievement in various subject areas. The common interpretation of these rankings is that students in nations ranking at or near the "top" are achieving, or have learned, more than students in nations ranking lower. The implication is that the nations at the top have more effective educational systems, at least in particular subject areas, than do the nations at the bottom. The accuracy and meaningfulness of these interpretations, however, depend on the ability of the test that was used to obtain the rankings to measure what it was intended to measure (i.e., the validity of the test). At issue, though, when evaluating the validity of achievement tests used in comparative studies of educational systems, is determining exactly what a particular study was

intending to compare before evaluating how well a particular test measures the variables (e.g., skills, knowledge) needed to make the comparisons.

Generally, the primary goal of researchers who conduct comparative-education studies is not to highlight differences in student performance in and of themselves (Burstein, 1992; Husen, 1982; Schmidt & McKnight, 1995). Rather, the goal is to do so in a way that accounts for the differences in educational contexts, inputs, and processes across and within nations (McDonnell, 1995; Robitaille et al., 1993; Schmidt & McKnight, 1995). Simply finding out that the students of one nation perform better on a set of items than do students of another nation is not meaningful to educational improvement if student performance cannot be not linked to some characteristic of a particular educational system. Therefore, the value of many comparative achievement studies depends upon the extent to which student test performance reflects achievement that can be attributed to the student's educational experiences (Airasian & Madaus, 1983; Linn, 1987; Mislevy, 1995; Nitko, 1989; Schmidt & McKnight, 1995). According to Airasian and Madaus (1983),

> When a standardized achievement test is used to compare achievement differences among schools or programs, the presumption is that the test taps characteristics specific to the schools or programs....If we want to make inferences about differential school, program or instructional effectiveness, then the processes underlying performance on the achievement measures need to be closely linked to instruction....If the issue is how effective are schools in developing general, transferable skills, traditional achievement tests may be fine. But if we are interested in whether schools develop the specific skills and knowledge they set out to develop, then such general tests are not valid. (p. 106)

The "specific skills and knowledge" educational systems "set out to develop" are articulated in the curriculum of that nation. Therefore, many comparative studies focus

on the success with which educational systems impart to their students a certain defined curriculum. The tests developed for these studies are designed to measure student attainment of this curriculum. A key component to evaluating the validity of these tests is determining how representative the test content is of the corresponding curriculum. Often, measurement specialists refer to this particular component of validity as content validity. Schmidt (1983) refers to the lack of content validity as content bias and considers this to be one cause of test invalidity.

Statement of the Problem

Difficulties in Domain Identification and Specification

Domain identification. The content validity of a test is evaluated in relation to the specific domain (in this case, a specific curriculum) about which test scores are used to make inferences (Crocker, Miller, & Franks, 1989; Fitzpatrick, 1983; Messick, 1989). The more representative the items are of the domain of interest, the greater is the chance that student performance on the sample of items will mirror their performance within the entire domain (Messick, 1989). A test may have high (content) validity in relation to one domain but low (content) validity in relation to another, and all persons who use the results of a particular test, however, may not be interested in the same domain, and.

Different curricula (i.e., domains), or components of a curriculum, may be of interest to educators and researchers who conduct cross-national studies (Schmidt & McKnight, 1995). For example, aside from the particular subject matter of interest, researchers may be interested in the curriculum as laid out in official documents (e.g.,

curriculum guides, national goals statements) or as laid out in textbooks and other instructional materials. Additionally, some researchers may be interested in the curriculum that is actually delivered by teachers. A crucial, and often ignored, issue in the development of cross-national achievement tests is determining what specific component of a curriculum (i.e., domain) is of particular interest (Airasian & Madaus, 1983; Mislevy, 1995) and, therefore, whether achievement results should reflect what students are intended to learn, what is in text books, what is delivered in the classroom, what the students of most nations achieve, or something else (Airasian & Madaus, 1983).

Domain specification. Even when a specific domain is identified, cross-national researchers still face challenges in writing test specifications for that domain. For example, a test could consist of only those topics that all countries include in their curriculum, topics that most countries include in their curriculum, or all topics included in the curriculum of any country (Linn, 1988; Linn & Baker, 1995; Porter, 1990). Generally, however, cross-national achievement tests are comprised of items that represent an internationally negotiated set of content (Linn & Baker, 1995). Critics of cross-national achievement studies often argue that the tests used in these studies provide, at best, an abstract definition of achievement in a particular subject area and may not adequately represent the curriculum of any participating nation (Linn & Baker, 1995; Mislevy, 1995; Porter, 1990; Westbury, 1992, 1993).

The accuracy and meaningfulness of interpretations of cross-national achievement results are impacted by the degree to which the test used in a particular cross-national study, reflects the curriculum of each country in the study (Guiton & Oakes, 1995; Linn & Baker, 1995; McDonnell, 1995; Romberg & Wilson, 1992). Performance results on a

test that is not based on a clearly defined domain provides little more than the knowledge of who outperforms who on a specific set of items (Airasian & Madaus, 1983; Robitaille et al., 1993). Interpretations of educational effectiveness or explanations of cross-national differences that are based on such results are questionable, if not invalid (Airasian & Madaus, 1983; Berliner, 1993; Guiton & Oakes, 1995; Guskey & Kifer, 1990; McDonnell, 1995; Stedman, 1994; Westbury, 1992, 1993). Therefore, in order to validly interpret comparative-cross-national-achievement data, it is important to understand the relationship of the test items used to obtain these data to the curricula of each participating nation (Airasian & Madaus, 1983; Linn & Baker, 1995; Schmidt & McKnight, 1995; Schmidt, McKnight, Valverde, Houang, & Wiley, 1996).

Constraints in Item Development

Two prevailing constraints on cross-national achievement-test construction exist. One of these constraints stems from the politics of item negotiation. Decisions about the specific content of cross-national achievement tests evolve through years of negotiation. Reaching even a minimal level of consensus from participating nations demands sensitivity to the unique concerns and political realities of each nation. Often, reaching consensus entails cutting corners in test development and adding or deleting certain items or topics despite specifications to the contrary.

A second constraint on cross-national achievement test construction relates to the adequacy of the item pool available to test developers. Item writing is an arduous and costly process. It is even more difficult in the cross-national arena as it involves developing items that transcend cultures and translations. Often, researchers will draw

from existing item pools when constructing large-scale achievement tests (Garden & Orpwood, 1996; Husen, 1983). However, the existing item pool may not always adequately represent the range of topics and behaviors included in the curricula of all nations. Items, especially those measuring higher-order thinking or complex reasoning, may be sparse, and resources may prohibit the development of enough items to overcome the deficits.

The reality of these constraints may mean that cross-national tests will never allow for a perfect match to all potential curricula. Therefore, researchers must continue to explore ways to use the information available on cross-national curricular differences to aid in the interpretation of cross-national-achievement results (Linn & Baker, 1995; Porter, 1990). A key question remaining to be answered is: what methods for selecting test content and strategies for analyzing and presenting test results provide the most valid basis for comparing student achievement across nations?

Purpose

The purpose of this study was to use the results of an extensive multi-national curriculum analysis to analyze the content of a cross-national mathematics achievement test in relation to the curriculum of nations administering the test. A second purpose was to determine if altering the content of the test to better match the countries' mathematics curricula has an impact on national performance and to evaluate the subsequent consequences of such content alterations on test validity. The ultimate goal was to use this information to enhance the validity of cross-national comparisons of student

achievement. My primary focus was on the relationship between test items and curriculum as a key element of test validity.

This study is one of the first applications of the results from an extensive multi-national curriculum analysis undertaken as a part of the Third International Mathematics and Science Study (TIMSS). Preliminary results of the curriculum analysis and TIMSS achievement testing are due for release in late fall of 1996. The curriculum analysis entailed an exhaustive review of curricular intentions for math and science in 50 countries (Robitaille et al., 1993; Schmidt & McKnight, 1995; Schmidt et al., 1996). It necessitated the development of a curriculum framework describing subject-area content, performance expectations, and perspectives (i.e., attitudes; Robitaille et al., 1993). The framework was subsequently used to guide the construction of student achievement tests. The data provide the opportunity for using a common framework to link student attainment with the results of curricular intentions across nations. Additionally, information on curricular intentions obtained using the framework can be used to guide the development of future cross-national achievement tests.

The results of this study may be applicable to intra-national comparative achievement studies in addition to cross-national studies. As mentioned earlier, interest in the ability to compare student achievement across states, districts, and schools continues to grow in the United States (Linn & Baker, 1995; Mislevy, 1995; Porter, 1990). Calls continue for a national system of assessments that recognizes the individuality of states while measuring progress toward common standards. The diversity of the American educational system introduces many of the same problems encountered

when conducting cross-national studies (Linn, 1988). The results of the present study will apply to these situations as well.

CHAPTER II

Review of Related Literature

Comparative-Achievement Studies - Growth, Rationale, and Impact

> Comparisons are fascinating and they make juicy items of gossip, but they do not necessarily lead to improvement. The penchant for comparing is taken for granted with little thought as to what is gained by such comparisons. (Maeroff, 1991, p. 92)

The Growth of Comparative-Achievement Studies

Many nations have demonstrated a long-term interest in comparing the achievement of their students with that of the students in other nations (Linn & Baker, 1995; Pelgrum, 1989; Porter, 1991). The concept for "a study of cognitive competence in children belonging to different national systems of education" (Husen, 1982, p. 6) was being discussed as early as 1958 at a meeting of the UNESCO Institute for Education in Hamburg. It was not until 1961, however, that researchers established an organization aimed at achieving this goal. The International Association for the Evaluation of Educational Achievement (IEA) was founded "to promote research aimed at examining educational problems common to many countries and thereby devise evaluative procedures which can provide facts which can be useful in the ultimate improvement of educational systems" (Husen, 1987, p. 30). The IEA completed a preliminary study of 12 countries in 1961. The First International Math Study (FIMS) took place between 1962 and 1965 (Husen, 1987).

Today the interest in cross-national achievement studies continues. The IEA studies have expanded from their original focus on mathematics to include studies in science, reading, literature, writing, civics, French and English as foreign languages, computers, and preprimary education (Linn & Baker, 1995; Pelgrum, 1989). Among recent comparative studies are the 1990-91 IEA study in reading literacy and the 1991 International Assessment of Educational Progress (IAEP) studies in math and science. Additionally, the IEA, together with other researchers around the world, is preparing the results of the Third International Math and Science Study (TIMSS) for release in late 1996 through 1997. This most recent cross-national study involved testing three populations of students from approximately 50 nations.

Additionally, within the United States current educational reform movements highlight the need for comparative data to insure that American students remain competitive with other major industrialized nations. Supporters of these reforms encourage policy-makers to develop high standards for education that are "benchmarked" against the achievement of students in other nations (Linn & Baker, 1995; Resnick, Nolan & Resnick, 1995). Groups like the National Education Goals Panel (NEGP), the National Council on Education Standards and Testing (NCEST), and the National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessments have begun work in this area (Linn, 1988; Linn & Baker, 1995; Schmidt & Valverde, 1995). These and other groups, such as the Council of Chief State School Officers (CCSSO), also advocate state-by-state comparisons of educational achievement (Bracey, 1995; LaPointe, 1991; Linn, 1987, 1988; Porter, 1991; Postlethwaite, 1987).

The Rationale for Comparative-Achievement Studies

Husen (1987) stated that the first international studies "were inspired by expanding international communication, trade, and military competition" (p. 43). In a world of increasing global competition and interdependence, many nations still have a desire (or need) to know where they stand in comparison to other nations (Berliner, 1993; Guthrie, 1986; Mislevy, 1995). This desire to know who is "first" or "best" is sometimes referred to as the "cognitive Olympics" or "international horse race" (Husen, 1987; Schmidt & Valverde, 1995). An example of this competition is evident in the *Goals 2000, Educate America Act* (H.R. 1804 and S.R. 846, 1993). One goal in this policy statement declared that "U.S. students will be first in the world in science and mathematics achievement" by the year 2000. As a result, U.S. educators and the public are eagerly awaiting the results of TIMSS to determine the progress they are making toward this goal and to compare the ranking of American students with their poor standing in past studies.

Another reason for conducting comparative-achievement studies is to determine priorities for expenditures and resource allocation within educational systems (Guthrie, 1986; Mislevy, 1995). Budget cuts and shortages of resources such as computers, textbooks, and qualified teaching staff in some educational systems have led to a need to closely examine educational priorities. Comparative achievement studies can help educators determine areas of strength and weakness in their educational system in relation to other educational systems. This can result in more informed decision making regarding budgeting and resource allocation. Additionally, comparative studies can

increase public awareness of the standing of their own educational system in relation to others which, in turn, may lead to support for increases in funding or a re-focusing of systemic priorities (Cohen, 1988).

Another, perhaps most important, reason for conducting comparative-achievement studies is school improvement (Guthrie, 1986; Mislevy, 1995). Comparative studies provide researchers and policy-makers with information than cannot be gained from single-system studies (Postlethwaite, 1987; Robitaille et al., 1993; Schmidt & Valverde, 1995). Comparisons with systems both different from and similar to ones own broadens knowledge about what is and is not possible. These comparisons also can provide greater opportunities for reviewing the impact of educational interventions.

> By looking at the educational systems of the world we challenge our own conceptions, gain new and objective insights into education in our own country, and are thus empowered with fresh vision with which to formulate effective educational policy and new tools to monitor the effects of these new policies. (Schmidt & Valverde, 1995, p. 7)

The Impact of Comparative-Achievement Studies

Comparative studies have had a significant impact on the U.S. educational system. Results of past international studies have led to the ruin of "new math" (Husen, 1987; Schmidt & Valverde, 1995) and have led to questions about classroom-grouping and school-tracking policies (Husen, 1987; Schmidt & Valverde, 1995). Such studies have highlighted the inadequacy of the American curriculum in math and science and have resulted in nationwide curricular reform (McKnight, Crosswhite, Dossey, Kifer, Swafford, Travers, & Cooney, 1987). Furthermore, one of the most significant influences

on the school-reform movement of the past decade, *A Nation at Risk* (National Commission on Excellence in Education, 1983), was written partly in response to poor U.S. performance on cross-national studies (Kaestle, 1985).

Results of achievement testing within nations also significantly impact educational systems. In the U.S., for example, funding, endorsements, or program continuation may be tied to comparisons of student performance results. Performance rankings factor into real-estate prices and the attractiveness of certain districts or states. Furthermore, U.S. educational systems, programs, and teachers have received substantial criticism from the media, public, and researchers as a result of performance in national and cross-national comparative studies (Bracey, 1995).

Validity and Comparative-Achievement Studies

Accusations of Invalidity

Controversy over the use of country ranks. Cross-national studies provide stakeholders and consumers with a variety of results. However, the results that historically have received the most attention are rankings of countries on national mean-achievement scores (Husen, 1987; Linn, 1988; Schmidt & McKnight, 1995). Policy-makers encourage such rankings because they provide a simple yardstick for gauging educational health (Postlethwaite, 1987). Many researchers, on the other hand, discourage such rankings because of problems reaching valid interpretations for all countries (Berliner, 1993; Husen, 1987; Linn & Baker, 1995; Mislevy, 1995; Porter, 1990; Postlethwaite, 1987; Stedman, 1994; Westbury, 1992, 1993).

Criticisms of the validity of cross-national achievement results come from many parties. First, some critics maintain that cross-national-achievement results have historically been based on poor sampling methodology (Bracey, 1995; Linn & Baker, 1995; Porter, 1991). Some of the countries involved in past cross-national studies tested populations that were not representative of the entire population to which the results were intended to generalize. For example, some countries tested only higher achieving students or native-language speakers. Stedman (1994), however, maintained that these problems are becoming fewer and more isolated. Furthermore, countries that do not employ adequate sampling procedures are being identified in the reporting of TIMSS results.

Differences inherent in the test populations of each nation are sometimes cited as reasons for invalidity (Berliner, 1993; Linn & Baker, 1995). For instance, test populations may differ in the total years of schooling students have received prior to the testing age. Students in some countries begin school at earlier ages than students in others. Additionally, critics also point out that differences in tracking practices across nations sometimes result in comparisons of elite populations of students with more comprehensive populations. These differences, too, are less extreme today than they were in the past (Linn & Baker, 1995).

Another concern about using ranks relates to cross-national differences in student motivation to do well (Berliner, 1993; Porter, 1991; Stedman, 1994). One often cited example is that of Korean students being applauded by their classmates as they leave the classroom to take an achievement test for a cross-national study (Berliner, 1993; Mislevy, 1995).

Differences in the focus and priorities of comparative education studies. Some of the most serious criticisms about the validity of cross-national achievement testing relate to the differing curricula of the nations involved in the studies and the problems that arise in test development and reporting as a result of these differences (Berliner, 1993; Linn & Baker, 1995; Stedman, 1994; Westbury, 1992, 1993). According to Husen (1983), "comparing the outcomes of learning in different countries is in several respects an exercise in comparing the incomparable" (p. 455). The difficulty stems from the fact that educational systems are unique to the culture of each country (Passow, 1984; Purves, 1987). They are based upon differing views of development and childhood (Berliner, 1993). They have differing goals which reflect differing social, political, economic, and resource needs and priorities (Schmidt & McKnight, 1995; Schmidt & Valverde, 1995). The time available for formal education is limited, making it impossible to teach everything. It is highly unlikely that different nations will choose to fill this limited time in exactly the same ways (Schmidt & McKnight, 1995). Therefore, the degree of variability in curricular goals and offerings across differing educational systems has a direct impact on the interpretation of results from comparative studies of these systems (Berliner, 1993; Linn & Baker, 1995; Mislevy 1995; Stedman, 1994; Westbury, 1992, 1993).

A Definition of Validity

Categories of test validity. The 1985 *Standards for Educational and Psychological Testing* opens with the following:

> Validity is the most important consideration in test evaluation. The concept refers to the appropriateness, meaningfulness, and usefulness of

> the specific inferences made from test scores. Test validation is the process of accumulating evidence to support such inferences. (AERA, APA, & NCME, 1985, p. 9)

In describing validity, Cronbach (1971) refers to "accuracy," Messick (1989) refers to "adequacy and appropriateness," and Mehrens and Lehmann (1991) refer to "truthfulness." Test validation is the process of evaluating the accuracy, adequacy, appropriateness, truthfulness, or usefulness of inferences made from test results, as opposed to evaluating the test itself. A test is never valid in and of itself; however, it may be valid for a certain purpose (Cronbach, 1971; Mehrens & Lehmann, 1991; Messick 1989).

Historically, three categories of validity evidence have been described: construct-related, content-related, and criterion-related (AERA, APA, & NCME, 1985; Cronbach, 1971; Mehrens & Lehmann, 1991; Messick, 1989). Some measurement specialists consider all validity evidence to be construct-related (Messick, 1989); others have challenged the notion of or the usefulness of content validity (Fitzpatrick, 1983; Guion, 1978; Messick, 1989); still others have discussed additional categories of validity such as consequential (Messick, 1989, 1994; Moss, 1992) and systemic (Frederikson & Collins, 1989). Linn, Baker, and Dunbar (1991) presented alternative criteria for evaluating the validity of assessments that are more performance oriented. These criteria are consequences, fairness, transfer and generalizability, cognitive complexity, content quality, content coverage, meaningfulness, and cost and efficiency.

Content validity. Test validation is a process in which evidence is gathered about the accuracy of test inferences. The process is never complete, as one can continually collect evidence which supports or disputes the validity of test inferences for different

purposes. The validity evidence most often sought when evaluating content-oriented achievement tests, including those used for comparative purposes, is evidence of content validity (Cronbach, 1971; Mehrens & Lehmann, 1991).

> Content validity is particularly important for achievement tests. Typically, we wish to make an inference about a student's degree of attainment of the universe of situations and/or subject-matter domain. The test behavior serves as a sample, and the important question is whether the test items do, in fact, constitute a representative sample of behavioral stimuli. (Mehrens & Lehmann, 1991, p. 267)

Content validity is the extent to which test items constitute an adequate sample of the content domain about which inferences are intended (AERA, APA, & NCME, 1985; Anastasi, 1982; Cronbach, 1971; Mehrens & Lehmann, 1991; Messick, 1989). Evaluations of content validity typically rely on judgments about test content as opposed to empirical analyses of test results (Messick, 1989). As such, Messsick (1989) prefers to speak of content *relevance* (i.e., the degree to which *each* item reflects the content domain) and content *representation* (i.e., the degree to which *all* items adequately represent the domain and any sub-domains) as opposed to content *validity*. Some authors, though, (e.g., Airasian & Madaus, 1983; Schmidt, Porter, Schwille, Floden, & Freeman, 1983) further sub-divide content validity into *curricular* and *instructional* validity depending upon the specific domain of reference. Messick (1989), however, referred to these two concepts as curricular relevance and representation and instructional relevance and representation. In any sense, "content validity" is one necessary but not sufficient condition for test validity (Guion, 1978; Messick, 1989; Schmidt, 1983).

The content validity of any test is evaluated in light of the specific purposes for which the test is to be used and the specific domain(s) it is intended to represent (Messick, 1989). An evaluation of content validity first requires a clear and operational definition of the content domain (Cronbach, 1971; Haertel & Calfee, 1983; Mehrens & Lehmann, 1991; Messick, 1989; Millman & Greene, 1989).

> The nature of the behavioral domain about which inferences are to be drawn or predictions made becomes especially important at two points in the measurement process: first, at the stage of test construction, where domain specifications serve as a blueprint or guide for what kinds of items should be constructed or selected for inclusion in the test; second, at the stage of test use, where the relevance and coverage of the constructed test must be evaluated for applicability to a specific, possibly different applied domain. The central problem at either stage, of course, is determining how to conceptualize the domain. (Messick, 1989, p. 37)

Domain Specification in Comparative-Achievement Tests

The range of domain possibilities. Domain specification is often one of the most difficult aspects of test development (Cronbach, 1971; Messick, 1989). A testing domain defines the parameters from which the content of test items can be drawn and sets limits to the inferences that can be made from test results. If not clearly defined and articulated by the test developer, the testing domain can be defined only in terms of the set of items that comprise the test. Too often, the testing domain in comparative studies of educational achievement is not always clear, and, in actuality, several different domains may be of interest to researchers conducting, or others using the results of, such studies.

Three categories of testing domains are relevant as sources of content for the achievement tests used in cross-national-education studies. The first of these testing domains relates to the *a priori* (Schmidt, 1983) or intentional achievement goals of a

nation, referred to as the *intended curriculum* in IEA studies (Schmidt & McKnight, 1995). The content of such domains is specified in formal statements of educational goals and curricular objectives. The second category of testing domains is defined in the content of curricular or instructional materials. This domain is sometimes considered the curricular domain (Schmidt, 1983). The third category of testing domain is based on the content of the actual instruction delivered by teachers. This *instructional* domain (Schmidt, 1983) corresponds to what IEA studies term the *implemented curriculum.* In addition, Schmidt et al. (1996) considered textbooks to be a bridge between the intended and implemented curriculum, that is, an articulation of the *potentially implemented curriculum* of a nation.

Researchers conducting comparative-achievement studies need to determine precisely in which curriculum domain they are interested before selecting or developing the tests they will use. They must determine if they are interested in student achievement of what the students' educational systems intended they learn, of what is contained in actual instructional materials, or of what they were taught in the classroom. Schmidt (1983) stated that researchers sometimes use tests that have been developed in reference to one domain and make inferences about student achievement in reference to another domain. He considered this to be one source of the content bias of a test.

Mislevy (1989) further identified an additional distinction that must be made in domain specification. This is the distinction between the concept of *immediate* curriculum (or instruction) and *ultimate* curriculum (or instruction). The immediate curriculum (instruction) relates to what was actually included in a *specific* curriculum or actually addressed in the classroom during a particular period of time (e.g., a school year).

Ultimate curriculum (instruction) relates to the final objectives of a curriculum or instruction, or those objectives generally desired for similar groups of students. Some researchers (Cronbach, 1971, Mehrens & Lehmann, 1991; Millman & Greene, 1989) have considered achievement results based on a general or ultimate curricular (instructional) domain to be more meaningful for comparative purposes than those based on the immediate domain.

A particular difficulty, then, in domain specification is determining exactly which domain is of primary interest. The resolution of this difficulty lies in the purpose of the test. If the purpose is to evaluate student achievement of subject-matter knowledge and skills most members of society deem important, one would be interested in a domain reflecting ultimate implicit educational goals. If the purpose is to evaluate student achievement of what was presented in textbooks, one would be interested in the immediate curricular domain. If the purpose of a test is to evaluate student achievement of what they were taught, one would be interested in the immediate instructional domain. Sometimes more than one domain may be of interest.

The purpose of cross-national achievement studies. Often conflicting purposes for conducting cross-national comparative-achievement studies exist. Policy makers may be interested only in student achievement comparisons in and of themselves. However, most cross-national studies typically have a purpose beyond merely ranking countries on student-test performance (Burstein, 1992; Husen, 1983; Postlethwaite, 1987). Some researchers (Bracey, 1991; Burstein, 1992; Linn & Baker, 1995) find it valuable to know how students within a nation perform on test content that is unique to their particular educational system and to compare this performance to the performance of students in

other nations on content that is unique to their system. These researchers are less interested in performance differences due to "student attributes" (Burstein, 1991, p. 50) or ability than they are interested in detecting differences due to schooling and determining how and why these differences arise (Burstein, 1991; Husen, 1983). Burstein (1993), in the prologue to his edited volume on SIMS results, recounts the historical purpose behind IEA testing. In it, he quotes from Husen's preface to the 1967 volume on the First International Mathematics Study:

> ...the overall aim is, with the aid of psychometric techniques, to compare outcomes in different educational systems. The fact that these comparisons are cross-national should not be taken as an indication that the primary interest was, for instance, national means and dispersions in school achievements at certain age or school levels.
> ...the main objective of the study is to investigate the "outcomes" of various school systems by relating as many as possible of the relevant input variables (to the extent that they could be assessed) to the output assessed by international test instruments...In discussions at an early stage in the project, education was considered as a part of a larger social-political-philosophical system. In most countries, rapid changes are occurring...Any fruitful comparison must take account of how education responded to changes in the society. One aim of this project is to study how mathematics teaching and learning have been influenced by such development.(p. 30)
> ...The IEA study was not designed to compare countries; needless to say, it is not to be conceived of as an "international contest" ...its main objective is to test hypotheses which have been advanced within a framework of comparative thinking in education. Many of the hypotheses cannot be tested unless one takes into consideration cross-national differences related to the various school systems operating within the countries participating in this investigation. (in Burstein, 1993, p. xxxii)

Complexities in study purposes and conflicting priorities only add to the difficulty of domain specification. However, the important point in domain specification is that it cannot begin without first clearly defining ones purpose(s) for the testing program (Millman & Greene, 1989) even if the purposes are many.

Determining test content. Once a particular domain is chosen as the focus of a test, the test developer must determine the exact content that will be included on the test and the proportion of items that will be allocated to each content area or topic (Messick, 1989; Millman & Greene, 1989; Postlethwaite, 1987). Much debate exists over the exact method of specifying the desired domain. For example, suppose one were to develop a test to measure student achievement of the curriculum included in math textbooks. Cross-national studies of textbook content (Schmidt et al., 1996) have found considerable variability in the content of these textbooks across nations. Different methods exist for determining the exact topics to include on such tests and the proportions of items to allocate to each topic.

> For any given target population, no two countries have exactly the same curriculum. Is it then possible to make valid comparisons of student achievement? The way in which international tests are currently constructed consists of first undertaking in each nation a content analysis of what is meant to have been learned by the end of a given period of time by a target population...It would seem reasonable to make comparisons about mathematics achievement in general if 80 percent or more of the content is the same between countries (and if the target populations are very similar and if the standard errors of sampling are small). What about 79 percent? What is a reasonable cutoff point? (Postlethwaite, 1987, p. 153)

Linn (1988) described three methods of domain specification first proposed by Seldon (in Linn, 1988). These are identifying a "least-common denominator" of content (e.g., content that is common to all textbooks across the nations), an "optimal" set of content (e.g., content found in a large number of textbooks), or an "inclusive" set of content (e.g., content that is found in any of the textbooks). Linn believed that the least

common denominator may appear most fair, but would tend to favor those systems that are narrow in their curriculum.

Linn and Baker (1995) and Porter (1991) stressed the need for a more inclusive approach to test development to ensure that cross-national achievement studies provide U.S. educators with adequate data on how well their students perform on educational goals specific to the U.S. Linn and Baker proposed that the tests be developed in such a way that a subset of content could be "mapped" onto specific national standards. This would entail developing a comprehensive assessment that "assesses the union rather than only the intersection of content standards of participating countries." Linn and Baker and Porter acknowledged, however, the potential political difficulties inherent in negotiating test content. Garden and Orpwood (1996) detailed these difficulties in their technical report on the development of the TIMSS achievement tests. Additionally, an inclusive approach to test development would demand large amounts of testing time from students unless complex matrix sampling designs were employed. However, the results of cross-national achievement studies may be limited without the ability to match national goals or practices to test results (Linn, 1987; Linn & Baker, 1995; Porter, 1991).

Other issues related to test content. Other difficulties in developing tests with high content validity relate to issues such as the balance of breadth versus depth in content coverage (Burstein, 1986). Should a limited number of items be used to measure a large number of topics superficially or should some topics be measured in depth at the expense of others? Another issue relates to the adequacy of the item pool for test development. Items that measure integrated topics and higher-order thinking processes

are in limited supply, are difficult to write, and encounter more resistance in country negotiations (Garden & Orpwood, 1996; Linn, 1987).

Additionally, the increasing complexity of subject matter calls into question the unidimensionality of test domains. Lack of unidimensionality raises questions about the meaning of total scores used in country ranks and subsequent analyses (Airasian & Madaus, 1983; Maeroff, 1983). Researchers (Burstein, 1991; Kupermintz, Ennis, Hamilton, Talbert, & Snow, 1995; Maeroff, 1983; Muthen et al., 1995) have suggested that mathematics scores aggregated over different topics represent general-math ability rather than math achievement that can be linked to curriculum or instruction. Student performance varies, sometimes significantly, across sub-topics (Ariasian & Madaus, 1983). This general-math factor may be so strong that it masks any correlation between curriculum and achievement (Burstein, 1991). Better linkage between tests and curriculum is obtained at the sub-topic level (Airasian & Madaus, 1983; Burstein 1991; Mislevy, 1995); although, some researchers suggest that the most useful performance results are at the item level (Guskey & Kifer, 1990; Mislevy, 1995). As Mislevy (1995) has stated, "The outcome for every individual task in an international assessment tells a story in its own right. Assessments with hundreds of tasks, like those of IEA and IAEP, tell hundreds of stories" (p. 426).

Additionally, domain specification also must consider what students are expected to do with test content (Airasian & Madaus, 1983; Linn, 1983; Mislevy, 1995; Snow & Lohmann, 1989; Walker & Schafarzick, 1974). New cognitive theories have resulted in increased attention to expectations for student performance. Often, these expectations

vary within and across educational systems; always, they add to the complexity of the domain.

Finally, the level of specificity of domain definition needs to be determined. Burstein (1986) found differing levels of domain specification across different tests and curricular documents. Mehrens and Phillips (1987) have shown that the level of specificity of domain definition has an impact on the degree of test to domain match. According to Schmidt et al. (1981), "The domain should be at a fine enough level to make important distinctions but not such a fine level of detail so as to classify everything within the subject matter as being important" (p. 136).

Evaluating Content Validity

Two primary approaches exist for evaluating test-content validity (Airasian & Madaus, 1983; Leinhardt & Seewald, 1981). The first approach uses test results to compare the performance of individuals who have been exposed to curricular content with the performance of those who have not. The intent is either to determine if test scores discriminate between these two groups or to find items that do (Airasian & Madaus, 1983; Burstein, 1991; Muthen et al. 995). This approach includes the use of IRT, intra-class correlations, factor analysis, and generalizability theory. The methodology is used *post hoc* and does not directly evaluate the content being measured by test items (Airasian & Madaus, 1983).

The second approach to evaluating test to curriculum match relies on a judgment of the overlap between a test and a domain (Airasian & Madaus, 1983; Crocker et al., 1989; Leinhardt, 1983; Leinhardt & Seewald, 1981; Messick, 1989). Generally, a

taxonomy to which the domain and test are matched is developed (e.g., Burstein, 1986; Gamoran, Porter, Smithson & White, 1996; Schmidt et al., 1983). This taxonomy may include only topics or a matrix of topics and cognitive processes. In some cases (e.g., Leinhardt, 1983; Leinhardt & Seewald, 1981; Schmidt & McKnight, 1995), actual test items are matched to textbooks or teacher coverage.

Several methods have been used to quantify overlap and results often depend upon the specific method used. Crocker et al. (1989) reviewed a series of methods for evaluating the overall fit between items and a content domain. Many of the procedures involved using judges to rate the proportion of items that assess what is in a curriculum or what is deemed to be an important learning objective. Judges will typically rate the relevance or value of items and these ratings are averaged across judges.

Concepts in profile analysis also may be useful to consider when evaluating content validity, especially for cross-national purposes. A profile is a vector of k elements where each element could correspond to the proportion of test items in a given topic area, the proportion of time spent teaching a topic, the proportion of a textbook allocated to a topic, or the weight a topic is given in curricular intentions. Profiles of topic areas on a test could be compared to the curriculum profiles to determine the degree of similarity between the two.

A profile has three main properties: shape, elevation, and scatter (Cattell, 1949). Elevation is the mean of all the profile elements; scatter (dispersion) is the standard deviation of all profile elements from the mean (elevation); shape (configuration) is the relative highs and lows (or rank correlation) of profile elements. Differences of opinion exist as to which elements should be considered when assessing profile similarity.

Indices based on correlation look almost entirely at shape without regard to the other two properties. Euclidean distance measures (D) utilize all three factors of profile similarity (Skinner, 1978). D^2 is the sum of the squared distances between corresponding elements in two profiles; D is the square root of this measure. Cronbach and Gleser (1953) recommend the use of D as opposed to D^2, as it tends to exaggerate large differences.

Schmidt (1983) uses a similar concept to define the content bias of a test as the following:

$$\text{Total bias} = \Sigma(W_j^T - W_j^D),$$

where W_j^T is the weight for a topic, often defined by a proportion of items, for the test and W_j^D is the weight for the topic in the domain of interest (e.g., proportion of a textbook or proportion of instructional time devoted to a topic). Schmidt's formula is similar to the Euclidean distance formulas used in profile analysis (Cronbach & Gleser, 1953). Gamoran et al. (1996) in a recent study also drew upon profile analysis to measure content coverage. They developed an indicator that combined the "proportion of instructional time spent covering tested material (level of coverage), and the match of relative emphases of types of content between instruction and the test (configuration of coverage)" (p. 12). The formula for the configuration of coverage was

$$1 - (\Sigma \mid W_j^T - W_j^D \mid / 2)$$

where W_j^T is the proportion of items in each tested area and W_j^D is the proportion of instructional time spent on each tested area. The final index was the product of the level of coverage and the configuration of coverage.

The Impact of Low Content Validity

Considerable disagreement exists as to the impact of the lack of fit between a test and a domain. One impact of the lack of fit is the perceived importance of the test to stakeholders. Linn (1987) stated, "If a test does not measure the outcomes that correspond to important program goals, the evaluation will surely be considered unfair" (p. 6), especially if it better measures the goals of another program in the study.

Studies have shown that results on tests not well-matched to a domain can be misleading (Berliner, 1993; Linn, 1988; Stedman, 1994; Westbury, 1992, 1993). Others have found that ranks on total scores are unstable, may result in unfair comparisons (Guskey & Kifer, 1991; Linn, 1987; Mislevy, 1995), and are dependent on the relative weighting of sub-topic areas (Cronbach, 1971). IEA studies introduced the notion of opportunity to learn (OTL) as a means of ensuring the technical validity of their findings (McDonnel, 1995). Researchers have shown that opportunity-to-learn the skills being tested is a significant explanatory variable of student performance (Berliner, 1993; Burstein, 1992; Burstein et al., 1990; Husen, 1983; Kupermintz et al., 1995; McDonnell, 1995; Muthen, Huang, Jo, Khoo, Goff, Novak, & Shi, 1995; Purves, 1987; Walker & Schaffarzick, 1974).

Additionally, Westbury (1993) found that differences between the scores of American and Japanese students on SIMS decreased when controlling for curriculum. Studies by Raizen and Jones (1985) found a correlation between mathematics achievement and the number of math courses students take. One particular critic of cross-national studies has stated

> We make curricular decisions different from those that other countries make. Thus differences in achievement are most parsimoniously explained as differences in national curricula, rather than differences in the efficiency or effectiveness of a particular national system of education. (Berliner , 1993, p.),

Differing opinions about the impact of curriculum on student achievement also exist. In a reanalysis of the Westbury data, Baker (1993) still found large differences between American and Japanese scores even when accounting for opportunity to learn. Furthermore, although he did find some curricular impact on test results, Stedman (1994) found that curriculum was just one of many variables having an impact. Phillips and Mehrens (1988) maintained that studies comparing test-to-curriculum match "have not provided any evidence regarding the impact of the mismatch" (p. 34). Mehrens (1984), Mehrens and Phillips (1987), and Phillips and Mehrens (1988) felt that impact of mismatch on achievement would be minimal in norm-referenced testing situations where the curriculum is basically homogenous. However, they surmised that the results could be quite different if comparing "two totally different curricula" (Mehrens & Phillips, 1987, p. 368) or when comparing "countries in which textbooks are not as homogeneous as those in the United States" (Phillips & Mehrens, 1988, p.50).

> It is reasonable to assume that the more different the curricula, the more likely those differences will have an impact on the test scores. Thus if differences in curricula between, for example, the United States and Japan are great, those differences may indeed impact scores on a common test. Examining score differences across countries, we could make incorrect inferences about the quality of the instruction or the quality of the students rather than making correct inferences about the impact of curricular differences on test scores. (Mehrens & Phillips, 1987, p. 358)

Recent Advances

In TIMSS, the IEA has collected information that may provide a means to overcome some of the difficulties in the domain specification of cross-national achievement tests (Schmidt et al., 1996). Two methodological innovations in particular relate to domain specification. The first of these was the development of a detailed curriculum framework used to code all the content of materials and instruments in the study (Robitaille et al., 1993). The second was an exhaustive analysis of the content of the intended curricula of participating nations (Schmidt et al., 1996). In addition, in order to obtain measures of the implemented curriculum within each nation, the IEA revised questionnaires used in previous studies (IEA, 1994a). These questionnaires asked teachers to identify from a list of mathematics topics those that they taught during the school year and the amount of time allocated to each.

The TIMSS curriculum frameworks, document analyses, and teacher questionnaires provide educators and researchers with the tools for reducing the content bias of cross-national achievement tests and increasing test validity. Information from these materials provides a window into the unique educational experiences confronted by students across the world and provides a framework for domain specification. However, many issues still remain to be resolved. For example, researchers will still need to determine which types of domains are of particular interest. They then must determine how information across countries will be combined in domain specification.

CHAPTER III

Study Design and Procedures

Purpose and Questions

One purpose of this study was to use data on curricular intentions from a review of mathematics curriculum in 17 countries to evaluate the content of a mathematics assessment being developed for cross-national comparisons. A second purpose was to explore ways of using the curriculum data to improve test-to-curriculum match. A final purpose was to investigate the relationship between student performance results and test-to-curriculum *mis-match*, and the subsequent implications for test validity.

I compared the mathematics-curriculum data collected through the TIMSS document analyses to the content of the TIMSS mathematics field-trial instrument for 13-year-old students. I also developed several sets of test specifications based on different methods of summarizing the curriculum data. Using the country-performance data from the field trial, I calculated country-mean scores on each of the specified "tests." I then ranked each country on each test and compared the country scores and ranks across the different tests. The questions I attempted to answer were

1. How much variation in content exists across the 17 nations in the mathematics curricula for 13-year-old students? How well does the content of the TIMSS field-trial instrument match these curricula?
2. What test specifications provide a good curricular match across countries? How well does the content of the TIMSS field-trial instrument match these test specifications?

3. What test specifications would improve the content match between the TIMSS field-trial instrument and the countries' math curricula? How well do these specifications match the curricula?
4. How stable are country scores and ranks across tests developed using the new test specifications when compared to the total scores and ranks on the field-trial instrument?

A brief description of TIMSS as well as information on the study population, data sources, and methods for answering these questions follow.

The Third International Mathematics and Science Study

The TIMSS is the largest cross-national study of educational systems ever attempted (Robitaille, et al., 1993, Schmidt & McKnight, 1995). Approximately 50 nations have been involved in some aspect of the study. The primary objective of the study was to "contribute to improvements in the teaching and learning of mathematics and science" (Robitaille, et al., 1993, p. 35). The study revolves around three components: A study of the intended curriculum, the implemented curriculum, and the attained curriculum of the nations involved. Data on the intended curriculum were collected through expert questionnaires and document analyses in each country. Data on the implemented curriculum were collected through school, teacher, and student questionnaires. Data on the attained curriculum were collected through student-achievement testing and student questionnaires.

The TIMSS study population included students in the two grades in which most students were 8 years old, students in the two grades in which most students were 13

years old, and students in their final year of schooling. Additionally, a sub-population of students in their final year of schooling specializing in calculus or physics was also tested.

Data collection on curricular intentions began in 1991 and was completed in 1995. Data were cleaned and initial analyses released in 1996. Achievement testing was completed in 1995 and school, teacher, and student questionnaires were completed at the same time. The field-trial data used in this study were collected in May, 1994.

The research questions (Robitaille & Garden, 1996) for TIMSS were:

1. How do countries vary in the intended learning goals for mathematics and science; and what characteristics of educational systems, schools, and students influence the development of these goals? (p. 38)
2. What opportunities are provided for students to learn mathematics and science; how do instructional practices in mathematics vary among nations; and what factors influence these variations? (p. 40)
3. What mathematics and science concepts, processes, and attitudes have students learned; and what factors are linked to students' opportunity to learn? (p. 40)
4. How are the intended, the implemented, and the attained curriculum related with respect to the contexts of education, the arrangements for teaching and learning, and the outcomes of the educational process? (p. 42)

Study Population

I used data from 17 countries to conduct my analyses, with the unit of analysis being the country. The countries included in the study were those that participated in the TIMSS mathematics field trial for 13-year-old students and for which information was available from the TIMSS document analyses. Seven of the original 25 countries participating in the field trial were dropped from planned analyses because they either did not participate in or did not have complete data from the document analyses; one country was dropped because it had incomplete data on the field trial. The study countries

consisted of 2 Asian nations, 2 Eastern European nations, 10 Western European nations, 2 North American nations, and 1 South Pacific nation.

National Research Coordinators (NRCs) in each country were asked to select a "judgment sample" of students for participation in the field trial (IEA, 1994c). NRCs first identified (at least) 12 schools having classes that fit within the specified target populations (i.e., the two adjacent grades that contained the largest proportion of 13-year-old students) in each of their countries. Next, they selected one or more classes within these schools for testing. The sample sizes in each country were to be at least 100 students for each of four test booklets administered; at least 60 of those students were to have been in the upper target grade. The minimum sample size, then, was 400 students for each country (240 at the upper grade and 160 at the lower grade). Each student was given one test booklet, with all four booklets being used within each classroom. The IEA instructed NRCs to use the "best evidence" available to select as wide a range of student ability and educational and socioeconomic settings as possible. Some lack of geographic representation was tolerated for ease of data collection and quick turn-around. Because most of the curriculum information collected for TIMSS corresponds to grades (primarily the upper-target grades) instead of ages, I used data from 13-year-old students in only the upper grade of each country for this study.

Country-sample sizes are reported in Table 1. The IEA provided sample statistics only for the combined-grade samples (i.e., upper and lower grades combined) of each country, so I was unable to determine the sample sizes of the upper grades. Most countries met the requirement of 100 students per test booklet. Of those which did not, all but one had between 90 and 99 students per booklet. Country P had only 86 students

for booklet 8. Distributions of students-to-booklets were fairly uniform within a country. All but four of the countries exceeded the minimum total sample size of 400 students, several by over 100 students. The sample sizes of the remaining countries ranged between 374 and 396 students.

Table 1

Country Sample Size for the Combined Upper and Lower Grades of Each Country

	Test Booklet				
Country	3	5	6	8	Total
A	107	106	115	108	436
B	126	123	129	119	497
C	111	107	102	98	418
D	134	135	143	135	547
E	104	97	94	92	387
F	133	142	143	140	558
G	96	101	99	100	396
H	108	105	106	108	427
I	95	99	96	90	380
J	103	104	107	105	419
K	119	113	107	114	453
L	122	136	133	133	524
M	122	116	116	115	469
N	126	126	122	127	501
O	104	99	99	104	406
P	96	97	95	86	374
Q	178	180	183	183	724
Total	1987	1991	1995	1965	7916

Note. Booklets 1, 2, 4, and 7 contained items for the science-assessment field trial.

Instrumentation

Curriculum Frameworks for Mathematics

All data sets and test items were linked through codes from the mathematics curriculum framework developed for the TIMSS study (Robitaille et al., 1993; see Appendix A). The framework specifies three types of codes relating to three "aspects" of curriculum: content (i.e., topic area), performance expectations (i.e., math-related behaviors), and perspectives (i.e., attitudes or values). I used only the content (here after referred to as "topic") and performance-expectation codes for this study.

At the most general level, the mathematics framework has 10 main-topic categories (e.g., numbers, proportionality) and 5 main-performance-expectation categories (e.g., knowing, communicating). All topic categories have one or two levels of sub-categories (for a total of 44 individual sub-categories at the lowest level), and all performance expectations have one level of sub-categories (see Figure 1) The framework covers most mathematics topics relevant to "K-12" education across nations, and it reflects recent reforms and trends in mathematics education. It is meant to provide researchers with a meaningful description of mathematics content to be used throughout and beyond the duration of this study.

Sample Content Category with Sub-Categories

1.1 Numbers

1.1.1 Whole numbers

1.1.1.1 Meaning

1.1.1.2 Operations

1.1.1.3 Properties of operations

Sample Performance Expectation Category with Sub-Categories

2.1 Knowing

2.1.1 Representing

2.1.2 Recognizing equivalents

2.1.3 Recalling mathematical objects and properties

Figure 1. Example of content and performance-expectation curriculum framework codes for mathematics.

Field-Trial Instrument

The TIMSS mathematics-achievement-item field-trial instrument for 13-year-old students consisted of four booklets containing a series of multiple choice, short-answer, and extended-response items. The test was developed by a multi-national team of national research coordinators, subject-matter specialists, and measurement specialists. Items from past IEA studies and other large-scale achievement studies comprised the initial item pool. Additional items were provided by countries or developed as needed. Test blueprints were not completed prior to the initial stages of item development but were completed prior

to the field trial (see Garden & Orpwood, 1996). The blueprints were based on preliminary data from the document analyses. They also reflected the desire to evaluate "in-depth" performance on a sub-set of topics (Garden & Orpwood; 1996).

The field-trial instrument consisted in approximately twice as many test items as were desired for the final achievement test. Extended-response items were particularly more predominant than indicated by the blueprints due to a greater need for information on the properties of these items. Limitations in items and testing time made it impossible to obtain enough items to report performance on every topic and to include items measuring all topic by performance-expectation intersections. Therefore, topics were limited to six reporting categories (fractions and number sense; geometry; algebra; data representation, analysis, and probability; measurement; proportionality).

All test items were coded with topic and performance-expectation codes from the mathematics framework. Items could receive up to four codes each (two topic and two performance expectation codes). I reviewed the codes of all items prior to undertaking my analyses. I disagreed with the original item coding for some items. As a result, I re-coded these items. I discussed the item codes on a sample of approximately 25 items with a math-content specialist and a senior researcher in the TIMSS study. Additionally, this senior researcher independently re-coded items from the final mathematics-achievement test for TIMSS. I compared my re-coding on items that appeared in both the field-trial instrument and the final TIMSS test with the researcher's codes, and we discussed any

disagreements. We re-coded approximately 40% of the items. Re-coding entailed changing either the topic, performance-expectation, or both codes; adding either topic, performance-expectation, or both codes; or a combination of the two. The topic codes of 6% of the items were changed, and additional topic codes were added to 19% of the items.

Appendix B contains information on each of the four test booklets in the TIMSS mathematics field-trial instrument for 13-year-old students. The full test consisted of 241 unique items (197 multiple choice, 25 short answer, and 19 extended response) dispersed throughout 4 test booklets, two of which had 63 items and two of which had 74. Thirty-three "linked" items appeared in two different test booklets. Two of the linked items were short answer; one was extended response; the remaining were multiple choice.

Fifteen of the 44 framework sub-categories were not represented on the test, and three of the 10 main categories, 1.8 *Elementary Analysis,* 1.9 *Validation and Structure,* and 1.10 *Other Content* were not represented at all on the test. Sixteen of the topics were represented in all booklets; seven topics were represented in three booklets each; four topics were represented in two booklets each; two topics were represented in only one booklet each. Extended-response and short-answer items were evenly distributed across the booklets. However, they were not evenly distributed across topics.

The IEA provided item statistics for all test items on the field-trial instrument. It also provided information on the percent of students at the lower, upper, and combined grades who passed each item in each country. Data on linked items were summarized across the two sub-groups responding to the items and were presented only for both groups as a whole. Although extended response items were scored using a multi-level-

scoring rubric, students were not given partial credit in the international scoring. They either passed or did not pass each item. Percentages of students receiving each rubric point were provided.

Data Sources

I used three data sources from the TIMSS curriculum analysis to identify and describe the curriculum of the 17 study countries. Each source contained data for each country from analyses involving either expert topic mapping, curriculum-guide coding, or textbook coding. The expert-topic-mapping data source described each country's intended coverage of each of the 44 topics on the mathematics framework. The curriculum-guide-data source contained data on topic, performance-expectation, and perspectives coverage in a selection of curriculum guides for each country. The textbook-data source provided the same information for a selection of textbooks in each country. The curriculum-data sources are described in detail below. For a more detailed explanation, refer to Schmidt, et al. (1996).

Expert Topic Mapping

A panel of subject-matter experts familiar with the mathematics curriculum in each country identified the ages in which each topic on the mathematics framework was intended to be introduced, was intended to be taught, and was intended to receive focus (i.e., receive special emphasis or attention in the curriculum relative to other years). The data sets for expert topic mapping consisted of matrices of 0s, 1s, and 2s for each age. Zeros represented topics not intended in the curriculum of a country at a particular age; 1s represented topics that were intended in the curriculum of a country at a particular age,

but were not focused; 2s represented focus topics. The expert topic mapping contained data on only the topics aspect of the mathematics framework and not on the performance expectations or perspectives aspects of the framework. I used only the data for age 13 in this study.

Curriculum-Guide Analyses

Curriculum guides were collected within each country at the national level if they existed or at regional levels if necessary. The collection was to include those curriculum guides pertaining to at least half of the students at the TIMSS testing grades. The collections of curriculum guides were to have represented any major school types or geographic regions. Subject-matter experts within each country participated in a standardized training session on coding the document sample (i.e., curriculum guides and textbooks) from their countries. Once trained, they coded the untranslated versions of all documents from their countries. The curriculum guide coding entailed dividing the documents into conceptual units representing the "smallest functional segment" (e.g., introduction, objectives, pedagogy; Schmidt et al., 1996, p.191) of each guide. Each unit was coded with the appropriate topic, performance-expectation, and/or perspective codes from the mathematics framework.

I used only the data on the topics aspect of the mathematics framework for the grade corresponding to the "upper grade" of age 13 (8^{th} grade in the U.S.). Because curriculum guides did not constitute a random sample, it was difficult to determine exactly what proportion of each country's school population each guide represented. Therefore, the collection of guides within each country was taken as a whole to represent

students in the corresponding grade within a country. Additionally, curriculum guides varied drastically in their unit structure and meaning. For example, some guides included pages of detailed objectives; others contained only a simple list of objectives. Therefore, it was difficult to determine what it meant if a topic was more prevalent in one curriculum guide versus another. As a result, the curriculum guide data consisted of 1s and 0s in each cell of the countries by topics matrix. Ones indicated that a particular topic was included in any of the curriculum guides collected for a country; 0s indicated that it was not included. Most of the 17 countries collected only one curriculum guide. One country collected 5, one collected 6, and one collected 15.

Textbook Analyses

In each country, math and science textbooks corresponding to the same target grades described for curriculum guides were collected. Each country was to collect textbooks used by at least 50% of the students in the country within each target grade. Many countries needed to collect a series of textbooks to meet this 50% criterion while others needed only one. Additionally, some countries found it difficult just to meet the 50% criterion, while others could collect the specific book(s) used by 100% of their students. Seven countries in this study had one textbook in their textbook sample; seven countries had two textbooks; one country had three textbooks; two countries had four textbooks.

Coders divided the textbooks into units representing one to three days of instruction which were further sub-divided into blocks. The content within each block was coded with all topic, performance-expectation, and/or perspectives codes that

applied.[1] Again, country-level aggregates of data were developed for each country. These data indicated the average proportion of blocks across all sampled textbooks that were devoted to a particular topic, performance expectation, or perspective. I used the data on only the topics aspect of the mathematics framework for most of my analyses. I used data on the performance expectations and the content by performance-expectation intersections for selected analyses.

Data Analyses

My data analyses consisted of four primary steps. These were

1. Describe and compare the content of the three curriculum sources, and compare this content to the content of the TIMSS field-trial instrument.
2. Develop 12 test blueprints using 3 methods of summarizing, for each of the 3 curriculum sources taken individually plus 1 overall aggregate (incorporating all three data sources), then calculate the match between the content of the TIMSS field-trial instrument and each of the 12 test blueprints.
3. Identify those topics from the TIMSS field-trial instrument included in each of the 12 test blueprints, and re-write the blueprints using only these topics creating 12 new sets of "inclusive" test blueprints (i.e., the same test blueprint for all countries); write 4 sets of "unique" test blueprints for each country based on the four data sources (3 individual and 1 aggregate) using only the topics included in the field-trial instrument

[1] To evaluate the reliability of the document coding, the following process was used (1) two units were randomly sampled from textbooks selected from different countries, were translated, and were coded by an expert coder; (2) an iterative process was used to match blocks and the coding sequence of the country coders to the standard produced by the expert coder. Forty-five documents from 12 countries were used in the reliability study. The estimated reliability was .80 (see, Schmidt et al., 1995).

(total: 17x4=68); calculate the match between each of the 12 sets of inclusive-test blueprints and each country's corresponding curricula as represented by the three data sources individually plus the aggregate as well as the match between each country's 4 unique-test blueprints and each country's corresponding curricula.

4. Use country-level performance on the items that measured the topics included on each of the test specifications developed in step 3 to compute 32 sets of scores for each country (12 sets of "weighted" scores on inclusive tests, weighting each topic to match curriculum emphasis; 12 sets of "unweighted" scores on inclusive tests; 4 sets of "weighted" scores on unique tests; 4 sets of "unweighted" scores on unique tests). Compare country level results on the TIMSS field trial with results on the new sets of tests (24 inclusive tests and 8 unique tests for each country).

I used the three curriculum sources as different representations of the mathematics curriculum of each country. The expert topic mapping and curriculum guide analyses provided two representations of the curriculum that each country *intended* to be taught by teachers (i.e., attained by students). The textbook analyses provided a representation of the curriculum that was *potentially implemented* (Schmidt et al., 1996) by teachers. Teacher data on the implemented curriculum are not yet available internationally; therefore, textbooks provide the best indication of what may have actually been taught in the classroom. Additionally, the textbook data are much more detailed than the data in the other two curriculum sources and may better represent how topics are treated in the classroom. However, because teachers do not always teach all topics included in their textbooks, I also combined data across the three curriculum sources to obtain a second

estimate of the *potentially implemented* curriculum of each nation. I averaged across only those topics contained in all three data sources within each country. Although the presence of a topic in all three sources of curriculum data does not guarantee the topic will be taught, the potential for a topic to be taught should increase over the potential for those topics included in fewer of the data sources. The aggregate of the data sources, then, should represent a lower bound of the topics taught in the classroom.

I re-scaled the numbers in the cells of the expert-topic-mapping and curriculum-guide data sets so that they summed to one across all topics within each country by summing over all elements (i.e., the 44 topics) in each country vector and dividing each element by this sum. These numbers were estimates of the relative proportion of emphasis for each topic within a country. Countries that included fewer topics in these data sources received higher proportions of emphasis for each topic included than did countries that included more topics. To construct the aggregate-data set I averaged over proportions of emphasis on only those topics that were included in all three data sources for a country.

Compare Curriculum Sources and Compare Match to Field-Trial Instrument

I reviewed the content of each curriculum source and summarized it across countries and across topics. I compared topic inclusion and coverage both across and within countries.

I then evaluated test-curriculum match using several methods. For most analyses, I treated each set of topic proportions (i.e., the proportions of emphasis computed for the expert-topic-mapping, curriculum-guide, and aggregate-data sources and the proportion

of textbook blocks in the textbook-data source) for each country as a different "profile" of the mathematics curriculum for the country. Likewise, topic weights (i.e., proportions of items allocated to each topic) on the field-trial instrument provided a "profile" of test emphasis. Thus, I sought to compare the similarity of the four curriculum profiles for each country to or dissimilarity from the test profile.

I looked at the match between the curriculum profiles and the test profile separately for each country. I conducted six different analyses to estimate test-curriculum match. First, I calculated the proportion of items on the mathematics field-trial instrument that measured topics appearing in each of the four curriculum profiles. Second, I calculated the proportion of each curriculum profile that was tested on the field-trial instrument. Third, I calculated differences between measures of topic inclusion (i.e., presence) on the field-trial instrument and topic inclusion in each of the four curriculum profiles. Fourth, I calculated differences between topic weights (i.e., the proportion of items for each topic) on the field-trial instrument and topic emphasis proportions in each curriculum profile. Finally, I computed correlations and Euclidean-distance measures, $\sqrt{\sum_{j=1}^{44}(W_j^T - W_{ji}^C)^2}$ – where W_i^T is the weight of topic j on the field-trial instrument and W_{ji}^C is the weight of topic j in the curriculum of country i, between the topic weights on the field-trial instrument and topic-emphasis proportions in each of the four curriculum profiles.

Write Test Blueprints and Calculate Match between Blueprints and Field-Trial Instrument

I wrote test blueprints for three "inclusive" tests (i.e., the same test for each country, combining curriculum information across countries) for each curriculum-data source and the aggregate-data source (for a total of 12 blueprints) using the following methods:

1. a *strict intersection* (SI) method that included only the topics in all countries' curriculum profiles within each of the four data sources,
2. a *70% intersection* (7I) method that included only the topics common to at least 70% of the countries' curriculum profiles within each of the four data sources, and
3. a *union* (UN) method that included all topics in any of the countries' curriculum profiles within each of the four data sources.

I averaged across each country's proportion of emphasis for each topic included in each blueprint to obtain weights for each topic on each of the 12 sets of test blueprints. Each set of weights was re-scaled to sum to 1 across all topics. I then repeated the same analyses described in step 1 comparing the "profile" of topic weights for each of the 12 sets of test specifications with the field-trial instrument's "profile" of topic weights.

Write Test Blueprints to Improve Match with Field-Trial Instrument and Calculate Match to Curricula

My intention was to use the test blueprints and field-trial data to compute country scores on each of the 12 new "tests" and compare these scores to country performance on

the total field-trial instrument. However, the field-trial instrument did not cover all topics in the mathematics framework, so I re-wrote the 12 test blueprints developed in step 2 using only the topics for which items were included on the field-trial instrument. I used the same aggregate methods described earlier (i.e., strict intersection, 70% intersection, union). In addition, I also wrote four sets of specifications for "unique" tests for each country, using only the topics that appeared in both the country profiles for each curriculum-data source and the field-trial instrument. The topic weights on the test blueprints were scaled to sum to one across topics.

I then conducted the same comparisons outlined in step 1 to evaluate the match between each set of the 12 inclusive-test specifications and each country's corresponding curriculum profile as well as each country's 4 unique-test specifications and the country's corresponding curriculum profile (i.e., the test specifications using the expert mapping data were compared to each country's profile of expert mapping data, etc.). The comparisons between the unique-test specifications and the curriculum profiles provided an estimate of "best possible match" between the curriculum profiles and any test developed using the field-trial topics.

Evaluate Country Performance across the New Tests

I calculated scores for each country using the topics on each set of specifications I developed. I calculated both weighted and unweighted scores. To calculate unweighted scores, I computed the average percent of students passing items with a particular topic code and averaged across all topics included on each "test." To calculate weighted scores, I multiplied the average percent of students passing items within a topic by the

corresponding weight on a particular test specification. I then summed these numbers over topics. I ranked each country on each measure.

I compared all country scores and ranks with their scores and ranks on the field trial. First, I compared an average of scores and ranks across all new tests with the field-trial scores and ranks. I also looked at country variability across all scores and ranks. Next, I computed differences between the field-trial total scores and ranks and each new score and rank.

Finally, I looked at between-countries variation in scores and ranks on each topic. Then I calculated country-level scores using performance-expectation codes and compared country results across these measures.

The results of all analyses follow.

CHAPTER IV

Results

Curriculum Comparisons

Description of the Mathematics Curricula

Expert topic mapping. Summary statistics for the expert topic mapping are contained in Tables 2 (for topics) and 3 (for countries). Table C-1 in Appendix C presents the full set of data.

The columns in Table 2 represent (1) the average across countries of the proportions of emphasis[2] for each topic, (2) the standard deviations of the proportions of emphasis, (3) the median across countries of the proportions of emphasis for each topic, (4) the maximum proportions of emphasis across countries for each topic (minimum proportions were 0 for all but topic 1.3.2 *Basic 2D Geometry* with a minimum of .024), (5) the number of countries in which the topic was intended for coverage in the curriculum (whether focused or unfocused), and (6) the number of countries in which the intended topic was a focus topic.

Table 2 reveals that three topics (1.1.4.3 *Complex Numbers*, 1.8.1 *Infinite Processes*, 1.8.2 *Change*) were not intended for coverage in the curriculum of any country; one topic (1.3.5 *Vectors*) was intended for coverage by only one country; and

[2] "Emphasis" in the expert-topic-mapping data source was calculated by adding up all 0s, 1s, and 2s for each topic for a country and dividing each number by the total; emphasis in the curriculum-guide-data source was calculated by adding up all 0s and 1s for each topic for a country and dividing each number by the total; emphasis in the textbook data source corresponds to the proportion of textbook blocks associated with each topic for a country.

Table 2

Summary of Expert-Topic-Mapping Proportions for Each Math Topic across all 17 Countries

Topic Code	Topic	Ave. Prop. of Emphasis	*SD*	Median Prop. of Emphasis	Max. Prop.of Emphasis	Num. of Countries Including Topic	Num. of Countries That Focus
1.1.1.1	Wh.Num.-Meaning	0.012	0.013	0	0.034	8	1
1.1.1.2	Wh.Num.-Oper.	0.017	0.020	0.020	0.071	9	3
1.1.1.3	Prop. of Oper.	0.017	0.022	0	0.071	8	4
1.1.2.1	Common Fractions	0.032	0.025	0.027	0.105	14	6
1.1.2.2	Decimal Fractions	0.028	0.023	0.027	0.105	14	4
1.1.2.3	Relat. of Fractions	0.025	0.024	0.025	0.105	12	4
1.1.2.4	Percentages	0.025	0.019	0.027	0.054	12	6
1.1.2.5	Prop. of Frac.	0.025	0.026	0.024	0.105	12	3
1.1.3.1	Negative Numbers	0.032	0.019	0.028	0.071	14	8
1.1.3.2	Rational Numbers	0.030	0.025	0.025	0.105	14	5
1.1.3.3	Real Numbers	0.021	0.018	0.020	0.054	11	3
1.1.4.1	Binary Arithmetic	0.003	0.008	0	0.032	2	0
1.1.4.2	Exponents	0.036	0.018	0.039	0.069	15	9
1.1.4.3	Complex Numbers	0	0	0	0	0	0
1.1.4.4	Number Theory	0.024	0.016	0.027	0.056	13	4
1.1.4.5	Counting	0.004	0.011	0	0.041	2	1
1.1.5.1	Estim. Quant.& Size	0.014	0.013	0.020	0.036	9	1
1.1.5.2	Rounding	0.030	0.018	0.027	0.065	14	7
1.1.5.3	Estim. Comput.	0.023	0.018	0.025	0.054	12	5
1.1.5.4	Exponents&Mag.	0.027	0.021	0.027	0.069	12	7
1.2.1	Measurement Unit	0.029	0.015	0.027	0.054	15	5
1.2.2	Per.,Area,Volume	0.029	0.014	0.027	0.054	15	5
1.2.3	Estim. Errors	0.018	0.016	0.020	0.065	11	2
1.3.1	2D Geo:Coordinate	0.029	0.018	0.027	0.065	14	5
1.3.2	2D Geo:Basics	0.039	0.014	0.034	0.071	17	8
1.3.3	2D Geo: Polygons	0.034	0.016	0.030	0.065	16	7
1.3.4	3D Geo	0.034	0.017	0.030	0.069	15	7
1.3.5	Vectors	0.002	0.007	0	0.028	1	0
1.4.1	Geo. Transform.	0.033	0.019	0.027	0.069	15	7
1.4.2	Cong. & Sim.	0.031	0.021	0.027	0.069	14	7
1.4.3	Constructions	0.024	0.018	0.024	0.061	13	3
1.5.1	Proport. Concepts	0.030	0.015	0.028	0.054	15	6
1.5.2	Proport. Prob.	0.041	0.019	0.041	0.071	16	11
1.5.3	Slope & Trig.	0.021	0.022	0.020	0.065	10	4
1.5.4	Lin. Interp.	0.011	0.017	0	0.061	6	2
1.6.1	Pat., Rel., Func.	0.032	0.018	0.027	0.065	15	6
1.6.2	Equat. & Formulas	0.041	0.017	0.041	0.069	16	11
1.7.1	Data Rep. & Anal.	0.039	0.018	0.039	0.069	16	9
1.7.2	Uncer. & Prob.	0.015	0.016	0.014	0.053	9	1
1.8.1	Infinite Process.	0	0	0	0	0	0
1.8.2	Change	0	0	0	0	0	0
1.9.1	Val. & Just.	0.013	0.018	0	0.065	7	2
1.9.2	Struc. & Abs.	0.012	0.014	0	0.041	8	2
1.10.1	Other	0.017	0.015	0.020	0.048	10	2
Average		0.023	0.016	0.020	0.060	11	4

one topic (1.1.4.1 *Binary Arithmetic*) was intended for coverage by two countries. Only topic 1.3.2 *Basic 2D Geometry* was intended for coverage by all countries. The average number of countries intending a topic be included in the curriculum at age 13 was 11 (65%). The number of countries that intended that a topic be a focus topic at age 13 ranged from 0 to 11. Eleven of the 17 countries intended focus on topics 1.5.2 *Proportionality Problems* and 1.6.2 *Equations and Formulas*. Thirty-nine of the 44 topics were intended as focus topics at age 13 by at least one country. The average number of countries that intended focus on any given topic was 4 (24%).

Of those topics being intended for coverage in the curriculum of at least one of the countries, the average proportion of emphasis ranged from .002 (1.3.5. *Vectors*) to .041 (1.5.2 *Proportionality Problems*, 1.6.2 *Equations and Formulas*). Lower average proportions of emphasis mean that (1) few countries intended coverage of the topic, (2) few or no countries intended focus on the topic, or (3) the topic was intended for instruction in countries that intended a large number of topics (therefore, each topic would receive a lower proportion in those countries). Topics intended and/or focused on by a large number of countries and intended by countries with a narrow curriculum would receive higher proportions of emphasis. To better interpret the proportions of emphasis, on can treat them as the percent of mathematics class periods allocated to particular topics over the course of a school year. Out of a school year with 180 mathematics periods, for example, a proportion of .002 would represent less than one class period and .041 would represent seven class periods. Most standard deviations of the proportions were between .01 and .02 (two to four class periods). Medians were generally within three hundredths

of the means (five class periods). Maximum proportions ranged from around .03 (five class periods) to .10 (18 class periods).

Table 3 summarizes the expert-topic-mapping data for each country. The second column indicates the average proportion of topic emphasis for each country across all topics with non-zero proportions (i.e., all topics of which the country intends coverage at age 13), the next column indicates standard deviations of topic proportions of emphasis across all topics (including those with 0s), and the final two columns indicate the number of topics intended for coverage at age 13 as well as the number of intended topics that were also intended for focus in the country at that age.

Table 3 shows variation in intended topic coverage across the 17 countries. The column of average proportions was calculated across topics only with non-zero proportions (i.e., only across topics that were intended in the curriculum of a country). Averaging across all topics would have generated identical values for all countries because the proportions of emphasis sum to one within all countries. Similarly, the averages of non-zero proportions within a country were simply a factor of the number of topics intended in the curriculum of that country. Countries that intended the same number of topics in their curriculum had the same average proportion, regardless of the ratio of focused to intended topics. The numbers are presented in the table as indications of the magnitude of topic intention differences.

Average proportions of topic emphasis ranged from .026 to .071 (5 class periods to 13 class periods). Standard deviations of proportions ranged for most countries from .01 to .02. Country N had the smallest standard deviation (.009), and country G had the largest (.036). The numbers of topics intended for coverage at age 13 ranged from 14 to

Table 3

Summary of Expert-Topic-Mapping Proportions for Each Country across Topics

Country	Ave.[a] Prop. of Emphasis	*SD*[b]	Number of Topics Intended to be Included	Number of Topics Intended to be Focused
A	0.030	0.016	33	7
B	0.032	0.018	31	11
C	0.031	0.016	32	5
D	0.027	0.013	37	14
E	0.042	0.023	24	9
F	0.036	0.020	28	9
G	0.071	0.036	14	5
H	0.037	0.020	27	14
I	0.032	0.017	31	18
J	0.053	0.028	19	12
K	0.037	0.021	27	10
L	0.029	0.014	35	6
M	0.037	0.021	27	9
N	0.026	0.009	39	35
O	0.048	0.026	21	8
P	0.048	0.026	21	7
Q	0.029	0.015	35	14
Average	0.038	0.020	28	11

[a]Average of non-zero numbers. [b]*SD* of non-zero numbers.

39 topics, and countries intended focus on 5 to 18 of the intended topics. The proportions of intended topics within a country that were also intended for focus ranged from .16 to .90 with an average of .40 and a standard deviation of .17.

Curriculum-guide analyses. Tables 4 and 5 present the curriculum-guide data, and Table C-2 in Appendix C presents the full data set.

The data in Table 4 are by topic, as are the data in Table 2 for the expert topic mapping. The only difference is the absence of a count of focused topics. All topics were included in the curriculum guides of at least one country. Two topics (1.8.1 *Infinite Processes* and 1.1.4.3 *Complex Numbers*) were included in the guides of only two countries. The average proportions of emphasis for these two topics were .004 and .003 respectively. Two topics (1.3.4 *3D Geometry*, 1.6.2 *Equations and Formulas*) were included in the curriculum guides of all countries. The average proportions of emphasis for these two topics were .04.

Most standard deviations of the proportions of emphasis were around .01 to .02 and most medians were within a few hundredths of the mean. Maximum proportions ranged from .023 to .091, with a mean of .06. If these were thought of as proportions of a 180 period school year, this range would be about 4 to 16 class periods, with a mean of 11 class periods.

Table 5 presents the curriculum-guide data summarized for each country. It reveals that the number of topics included in a country's curriculum guide ranged from 11 to 44, with an average of 27 topics. Average proportions in this table were merely a function of the number of topics included in a country's curriculum guide(s): Countries with the same number of topics had the same average proportion, and average proportions

Table 4

Summary of Curriculum-Guide-Topic Proportions for Each Topic across Countries

Topic Code	Topic	Ave. Prop.of Topic Emphasis	*SD*	Median Prop. of Emphasis	Max. Prop.of Emphasis	# of Countries Including Topic
1.1.1.1	Wh.Num.-Meaning	0.023	0.025	0.026	0.091	10
1.1.1.2	Wh.Num.-Oper.	0.027	0.023	0.029	0.091	12
1.1.1.3	Prop. of Oper.	0.020	0.024	0.023	0.091	9
1.1.2.1	Common Fractions	0.017	0.016	0.023	0.045	9
1.1.2.2	Decimal Fractions	0.020	0.018	0.026	0.056	10
1.1.2.3	Relat. of Fractions	0.021	0.020	0.026	0.059	10
1.1.2.4	Percentages	0.022	0.017	0.028	0.048	11
1.1.2.5	Prop. of Frac.	0.008	0.013	0	0.034	5
1.1.3.1	Negative Numbers	0.033	0.016	0.033	0.059	15
1.1.3.2	Rational Numbers	0.028	0.018	0.032	0.059	13
1.1.3.3	Real Numbers	0.028	0.023	0.029	0.091	12
1.1.4.1	Binary Arithmetic	0.009	0.015	0	0.040	5
1.1.4.2	Exponents	0.023	0.019	0.028	0.059	11
1.1.4.3	Complex Numbers	0.003	0.008	0	0.026	2
1.1.4.4	Number Theory	0.026	0.019	0.029	0.059	12
1.1.4.5	Counting	0.007	0.012	0	0.034	4
1.1.5.1	Estim. Quant.& Size	0.011	0.016	0	0.042	6
1.1.5.2	Rounding	0.020	0.017	0.026	0.048	10
1.1.5.3	Estim. Comput.	0.015	0.016	0	0.042	8
1.1.5.4	Exponents&Mag.	0.016	0.017	0	0.042	8
1.2.1	Measurement Unit	0.027	0.017	0.032	0.056	13
1.2.2	Per.,Area,Volume	0.031	0.017	0.032	0.059	14
1.2.3	Estim. Errors	0.023	0.019	0.028	0.056	11
1.3.1	2D Geo:Coordinate	0.024	0.017	0.029	0.048	12
1.3.2	2D Geo:Basics	0.030	0.017	0.032	0.059	14
1.3.3	2D Geo: Polygons	0.038	0.019	0.034	0.091	16
1.3.4	3D Geo	0.041	0.016	0.034	0.091	17
1.3.5	Vectors	0.018	0.017	0.023	0.048	9
1.4.1	Geo. Transform.	0.035	0.020	0.033	0.091	15
1.4.2	Cong. & Sim.	0.032	0.018	0.033	0.059	14
1.4.3	Constructions	0.024	0.017	0.029	0.050	12
1.5.1	Proport. Concepts	0.030	0.017	0.032	0.056	14
1.5.2	Proport. Prob.	0.032	0.015	0.033	0.056	15
1.5.3	Slope & Trig.	0.021	0.027	0	0.091	8
1.5.4	Lin. Interp.	0.013	0.018	0	0.050	6
1.6.1	Pat., Rel., Func.	0.038	0.019	0.034	0.091	16
1.6.2	Equat. & Formulas	0.041	0.016	0.034	0.091	17
1.7.1	Data Rep. & Anal.	0.035	0.020	0.033	0.091	15
1.7.2	Uncer. & Prob.	0.025	0.018	0.029	0.056	12
1.8.1	Infinite Process.	0.004	0.012	0	0.045	2
1.8.2	Change	0.005	0.011	0	0.034	3
1.9.1	Val. & Just.	0.008	0.016	0	0.059	4
1.9.2	Struc. & Abs.	0.015	0.020	0	0.059	7
1.10.1	Other	0.031	0.017	0.033	0.059	14
Average		0.023	0.017	0.021	0.060	11

Table 5

Summary of Curriculum-Guide-Topic Proportions for Each Country across Topics

Country	Ave.[a] Prop.of Topic Emphasis	*SD*[b]	# of Topics Included in Guide
A	0.034	0.016	29
B	0.042	0.021	24
C	0.026	0.008	39
D	0.029	0.012	35
E	0.045	0.023	22
F	0.028	0.011	36
G	0.056	0.027	18
H	0.091	0.039	11
I	0.032	0.015	31
J	0.040	0.020	25
K	0.050	0.025	20
L	0.033	0.016	30
M	0.032	0.015	31
N	0.059	0.029	17
O	0.048	0.024	21
P	0.034	0.016	29
Q	0.023	0.000	44
Average	0.041	0.019	27

[a]Average of non-zero numbers. [b]*SD* of non zero numbers.

were larger as fewer topics were included in a curriculum guide. Again, these proportions indicate the magnitude of differences in inclusion of topics. The average of the average proportions of emphasis was .04 (7 class periods). The range of proportions of emphasis was .023 (about 4 class periods) to .091 (over 3 weeks of classes).

Textbook analyses. Tables 6 and 7 present summaries of the textbook-data sources. Table C-3 presents the full data set. These analyses were conducted using only the topic codes, even though textbooks were also coded with performance-expectation codes. Other analyses will make use of performance-expectation codes.

Table 6 presents textbook summaries over topics. Only one topic (1.8.2 *Change*) did not appear in the textbook-data source of any country. Three topics (1.3.2 *Basic 2D Geometry*; 1.6.1 *Patterns, Relations, and Functions*; 1.6.2 *Equations and Formulas*) appeared in the textbook data sources of all countries. Overall, the highest proportion of textbook blocks was devoted to topic 1.6.2. This topic, on average, appeared in 21% of textbook blocks. The next most emphasized topic, 1.3.3 *Polygons and Circles*, appeared in an average of 10% of textbook blocks.

Standard deviations were larger than in the expert topic mapping and curriculum-guide-data sources suggesting greater variation of topic coverage patterns. Two topics (1.3.4 *3D Geometry*, 1.6.2 *Equations and Formulas*) had standard deviations of over .10 (10% of text blocks). For some topics, the medians were quite different from the means indicating skewed distributions. Some topics had proportions at or near 0 in many countries, but may also have had a few large proportions. Such distributions impact the mean more than the median, making the median a better measure of central tendency.

Table 6

Summary of Textbook Proportions for Each Topic across Countries

Topic Code	Topic	Ave. Prop. of Text Blocks	*SD*	Median Prop. of Text Blocks	Max. Prop. of Text Blocks	# of Countries Including Topic
1.1.1.1	Wh.Num.-Meaning	0.015	0.026	0.004	0.106	11
1.1.1.2	Wh.Num.-Oper.	0.040	0.049	0.010	0.184	15
1.1.1.3	Prop. of Oper.	0.021	0.023	0.009	0.069	15
1.1.2.1	Common Fractions	0.041	0.034	0.036	0.126	16
1.1.2.2	Decimal Fractions	0.024	0.024	0.014	0.065	15
1.1.2.3	Relat. of Fractions	0.013	0.010	0.010	0.031	15
1.1.2.4	Percentages	0.035	0.034	0.035	0.129	14
1.1.2.5	Prop. of Frac.	0.006	0.010	0.001	0.042	11
1.1.3.1	Negative Numbers	0.041	0.036	0.040	0.110	15
1.1.3.2	Rational Numbers	0.028	0.071	0.010	0.306	12
1.1.3.3	Real Numbers	0.026	0.064	0.002	0.278	11
1.1.4.1	Binary Arithmetic	0.001	0.003	0	0.012	4
1.1.4.2	Exponents	0.041	0.038	0.034	0.117	14
1.1.4.3	Complex Numbers	0.000	0.001	0	0.002	3
1.1.4.4	Number Theory	0.016	0.022	0.007	0.072	11
1.1.4.5	Counting	0.002	0.006	0	0.025	7
1.1.5.1	Estim. Quant.& Size	0.002	0.003	0.001	0.011	9
1.1.5.2	Rounding	0.007	0.008	0.007	0.028	10
1.1.5.3	Estim. Comput.	0.008	0.009	0.006	0.032	12
1.1.5.4	Exponents&Mag.	0.007	0.015	0.000	0.062	7
1.2.1	Measurement Unit	0.040	0.042	0.031	0.167	15
1.2.2	Per.,Area,Volume	0.071	0.057	0.075	0.164	13
1.2.3	Estim. Errors	0.002	0.003	0.000	0.009	7
1.3.1	2D Geo:Coordinate	0.034	0.032	0.032	0.112	14
1.3.2	2D Geo:Basics	0.055	0.042	0.043	0.142	17
1.3.3	2D Geo: Polygons	0.098	0.054	0.093	0.202	16
1.3.4	3D Geo	0.068	0.121	0.019	0.469	13
1.3.5	Vectors	0.005	0.013	0	0.053	7
1.4.1	Geo. Transform.	0.056	0.064	0.052	0.243	13
1.4.2	Cong. & Sim.	0.040	0.060	0.012	0.231	11
1.4.3	Constructions	0.008	0.012	0.002	0.035	9
1.5.1	Proport. Concepts	0.008	0.010	0.004	0.028	10
1.5.2	Proport. Prob.	0.020	0.023	0.017	0.095	12
1.5.3	Slope & Trig.	0.014	0.025	0.000	0.083	6
1.5.4	Lin. Interp.	0.002	0.004	0	0.014	4
1.6.1	Pat., Rel., Func.	0.060	0.054	0.049	0.208	17
1.6.2	Equat. & Formulas	0.205	0.118	0.174	0.388	17
1.7.1	Data Rep. & Anal.	0.048	0.032	0.057	0.099	14
1.7.2	Uncer. & Prob.	0.003	0.008	0.000	0.034	6
1.8.1	Infinite Process.	0.001	0.001	0	0.004	4
1.8.2	Change	0	0	0	0	0
1.9.1	Val. & Just.	0.022	0.072	0.002	0.309	10
1.9.2	Struc. & Abs.	0.021	0.034	0.007	0.117	9
1.10.1	Other	0.036	0.062	0.006	0.223	11
Average		0.029	0.032	0.020	0.119	11

The maximum proportions of textbook blocks for many of the topics were around .10 or more. Some of the topics with larger maximum proportions were 1.3.4 *3D Geometry* (.47), 1.6.2 *Equations and Formulas* (.39), 1.9.1 *Validation and Justification* (.31), and 1.1.3.2 *Rational Numbers* (.31). Topics with some of the lowest maximum proportions were 1.1.4.1 *Binary Arithmetic*, 1.1.4.3 *Complex Numbers*, 1.2.3 *Measurement Estimation and Error*, 1.5.4 *Linear Interpolation*, and 1.8.1 *Infinite Processes* (all around .01 or less).

Table 7 contains the summary of textbook data for each country. It shows (1) the average of the proportions of textbook blocks devoted to a topic across all 44 topics for each country, (2) the standard deviation of proportions across all topics, (3) the average proportion of textbook blocks across only topics included in the textbook(s) of each country, (4) the maximum proportion of textbook blocks devoted to each topic, (5) the number of topics included in each country's textbook(s), and (6) the number of topics within a country's textbook(s) that appeared in at least 10% of the textbook blocks.

The numbers of topics included in country textbooks varied. The average number of topics included in a textbook was 28. One country included only 11 topics while another included 40 topics. Less variation existed in the average proportion of blocks devoted to any topic (average of .03) most likely due to the fact that most proportions summed to around 1 (proportions could sum to more than one due to the potential for the presence of multiple-topic codes within each block). One country (N) did, however, have an average proportion of .06 with a standard deviation of .10. The average of the standard deviations was .05 (5% of text blocks).

Table 7

Summary of Textbook Proportions for Each Country across Topics

Country	Ave. Prop. of Text Blocks	*SD*	Ave. Prop. for Included Topics	Max. Prop.of Text Blocks	# of Topics Included by Country	# of Topics with Prop. >.1
A	0.024	0.035	0.035	0.145	30	2
B	0.023	0.057	0.039	0.282	26	3
C	0.033	0.037	0.038	0.163	39	3
D	0.028	0.027	0.034	0.087	36	0
E	0.025	0.053	0.067	0.243	16	3
F	0.022	0.037	0.047	0.148	21	3
G	0.032	0.075	0.056	0.374	25	4
H	0.026	0.039	0.039	0.141	30	2
I	0.033	0.037	0.044	0.123	33	4
J	0.023	0.070	0.091	0.388	11	4
K	0.028	0.063	0.045	0.356	28	3
L	0.025	0.036	0.034	0.174	32	2
M	0.035	0.063	0.045	0.323	35	5
N	0.061	0.107	0.084	0.469	32	8
O	0.029	0.058	0.059	0.296	22	5
P	0.022	0.041	0.038	0.184	26	2
Q	0.029	0.039	0.032	0.236	40	1
Average	0.029	0.051	0.048	0.243	28	3

When averaging only proportions of topics included in a country's textbook(s), an increase of approximately 2% of text blocks was seen (average .048). These proportions ranged from .032 to .091. Also, the range of maximum proportions and the number of topics with proportions over .10 showed that some textbooks devoted a lot of space to a few topics while others spread their space over many topics. The maximum amount of textbook space devoted to a single topic ranged from 9% of a textbook to almost half of a textbook. The data indicated that in one country no topic received over 10% of the space, while in another county eight topics received over 10% of the space. In most countries, however, between two and four topics received over 10% of the textbook space.

Aggregate-data source. The results of the three curriculum-data sources were combined to obtain a composite picture of mathematics curriculum in each country. Table 8 presents data on the agreement of topic inclusion across the three data sources. Table C-4 includes the full set of data. Table 8 shows the average number of each countries' three data sources in which each topic was included, the number of countries in which the topic appeared in all three of the data sources, and the number of countries in which the topic appeared in none of the data sources. Additionally, I calculated the proportion of countries that had agreement of topic inclusion across the three data sources (i.e., the proportion of countries in which the topic either appeared in all three data sources or none of the data sources). Table 8 also presents summaries of proportions of emphasis. Within each country, proportions of topic emphasis were averaged for only those topics appearing in all three data sources for that country. Other topics were given 0s. These averages were scaled to sum to 1.00 across the included topics within a country. Table 8 presents the average of these proportions for each topic.

Table 8

Agreement of Topic Inclusion across Expert-Mapping-, Curriculum-Guide-, and Textbook-Data Sources Presented for Topics across Countries

Topic Code	Topic	Ave. # of Sources[a]	# of Cntrys: 3 Sources[b]	# of Cntrys: 0 Sources[c]	Agreement[d]	Ave.[e] Prop.of Emphasis	*SD*	Median[e] Prop. of Emphasis	Max.[e] Prop. of Emphasis
1.1.1.1	Wh.Num.-Meaning	1.7	3	1	0.24	0.006	0.014	0	0.053
1.1.1.2	Wh.Num.-Oper.	2.1	8	1	0.53	0.024	0.032	0	0.121
1.1.1.3	Prop. of Oper.	1.9	6	2	0.47	0.016	0.023	0	0.070
1.1.2.1	Common Fractions	2.3	8	1	0.53	0.025	0.027	0	0.072
1.1.2.2	Decimal Fractions	2.3	9	1	0.59	0.025	0.031	0.023	0.122
1.1.2.3	Relat. of Fractions	2.2	9	1	0.59	0.017	0.017	0.018	0.048
1.1.2.4	Percentages	2.2	8	2	0.59	0.024	0.030	0	0.107
1.1.2.5	Prop. of Frac.	1.6	4	3	0.41	0.007	0.014	0	0.046
1.1.3.1	Negative Numbers	2.6	12	0	0.71	0.031	0.027	0.028	0.083
1.1.3.2	Rational Numbers	2.3	8	0	0.47	0.032	0.076	0	0.327
1.1.3.3	Real Numbers	2.0	6	0	0.35	0.017	0.028	0	0.093
1.1.4.1	Binary Arithmetic	0.6	0	9	0.53	0	0	0	0
1.1.4.2	Exponents	2.4	10	2	0.71	0.026	0.026	0.030	0.089
1.1.4.3	Complex Numbers	0.3	0	14	0.82	0	0	0	0
1.1.4.4	Number Theory	2.1	8	2	0.59	0.017	0.019	0	0.050
1.1.4.5	Counting	0.8	1	9	0.59	0.002	0.007	0	0.031
1.1.5.1	Estim. Quant.& Size	1.4	4	4	0.47	0.006	0.010	0	0.031
1.1.5.2	Rounding	2.0	7	2	0.53	0.011	0.014	0	0.037
1.1.5.3	Estim. Comput.	1.9	7	4	0.65	0.011	0.014	0	0.040
1.1.5.4	Exponents&Mag.	1.6	3	1	0.24	0.005	0.011	0	0.032
1.2.1	Measurement Unit	2.5	11	0	0.65	0.032	0.029	0.029	0.099
1.2.2	Per.,Area,Volume	2.5	12	0	0.71	0.045	0.041	0.043	0.146
1.2.3	Estim. Errors	1.7	4	2	0.35	0.005	0.010	0	0.028
1.3.1	2D Geo:Coordinate	2.4	8	0	0.47	0.022	0.027	0	0.087
1.3.2	2D Geo:Basics	2.8	14	0	0.82	0.043	0.028	0.041	0.114
1.3.3	2D Geo: Polygons	2.8	15	0	0.88	0.065	0.037	0.063	0.131
1.3.4	3D Geo	2.6	11	0	0.65	0.047	0.053	0.035	0.161
1.3.5	Vectors	1.0	1	6	0.41	0.002	0.009	0	0.037
1.4.1	Geo. Transform.	2.5	11	0	0.65	0.049	0.061	0.036	0.242
1.4.2	Cong. & Sim.	2.3	9	0	0.53	0.037	0.053	0.020	0.183
1.4.3	Constructions	2.0	5	0	0.29	0.008	0.013	0.000	0.034
1.5.1	Proport. Concepts	2.3	9	0	0.53	0.017	0.017	0.021	0.051
1.5.2	Proport. Prob.	2.5	10	0	0.59	0.027	0.029	0.028	0.095
1.5.3	Slope & Trig.	1.4	3	4	0.41	0.010	0.028	0	0.116
1.5.4	Lin. Interp.	0.9	0	7	0.41	0	0	0	0
1.6.1	Pat., Rel., Func.	2.8	15	0	0.88	0.059	0.039	0.053	0.135
1.6.2	Equat. & Formulas	2.9	16	0	0.94	0.133	0.087	0.106	0.338
1.7.1	Data Rep. & Anal.	2.6	13	0	0.76	0.051	0.038	0.051	0.133
1.7.2	Uncer. & Prob.	1.6	4	4	0.47	0.006	0.012	0	0.035
1.8.1	Infinite Process.	0.4	0	12	0.71	0	0	0	0
1.8.2	Change	0.2	0	14	0.82	0	0	0	0
1.9.1	Val. & Just.	1.2	3	4	0.41	0.008	0.024	0	0.101
1.9.2	Struc. & Abs.	1.4	5	5	0.59	0.011	0.019	0	0.063
1.10.1	Other	2.1	8	2	0.59	0.022	0.032	0	0.129
Average		1.9	7	2.7	0.57	0.02	0.03	0.01	0.09

[a]The average number of data sources (out of 3) in a country in which the topic appears. [b]The number of countries in which the topic appears in all 3 data sources. [c]The number of countries in which the topic appears in no data sources. [d]The proportion of countries in which the topic appears in all 3 or none of the data sources. [e]Within each country, the average, median, or maximum proportions for topics included in all 3 data sources.

Seven topics (1.1.3.1 *Negative Numbers and Integers*; 1.2.2 *Perimeter, Area, Volume*; 1.3.2 *Basic 2D Geometry;* 1.3.3 *Polygons and Circles*; 1.6.1 *Patterns, Relations, Functions*; 1.6.2 *Equations and Formulas*; 1.7.1 *Data Representation and Analysis*) appeared in all three curricular data sources of at least 70% of the countries. Topics 1.1.4.1 *Binary Arithmetic*, 1.1.4.3 *Complex Numbers*, 1.5.4 *Linear Interpolation*, 1.8.1 *Infinite Processes*, and 1.8.2 *Change* did not appear in all three data sources for any country. Topics 1.1.4.3, 1.8.1, and 1.8.2 also appeared in none of the data sources for at least 70% of the countries. The average proportion of agreement across the data sources (i.e., topics either appeared in all three or none of the data sources within a country) was almost 60%. Topic 1.6.2 *Equations and Formulas* had agreement across all three data sources in 94% of the countries while topics 1.1.1.1 *Whole Number Meanings* and 1.1.5.4 *Exponents* had agreement across the data sources in less than 25% of the countries.

The aggregate of the proportions of emphasis in each data source (i.e., an average of the proportions of emphasis across all three data sources for topics that appeared only in all three sources in a country) ranged from .002 (1.1.4.5 *Systematic Counting* and 1.3.5 *Vectors*) to .133 (1.6.2 *Equations and Formulas*; this was .07 more emphasis than the next closest topic had). Most standard deviations of these proportions were around .03; although, a few were larger (1.1.3.1 *Negative Numbers and Integers*, .076; 1.4.1 *Transformations*, .061; 1.6.2 *Equations and Formulas*, .087). Medians of the aggregate proportions of emphasis for many topics differed from the means due to the high proportions of zeros in the data source (i.e., any topic not appearing in all three data sources for a country received 0 as the proportion of emphasis in the aggregate-data source). Maximums of the aggregate proportions for each topic averaged .09. Several of

the maximum proportions were quite small (e.g., 1.1.4.5 *Systematic Counting*, 1.1.5.1 *Estimating Quantity and Size*, 1.2.3 *Measurement Estimation and Errors*). The largest were around .30 (1.1.3.2 *Real Numbers*, 1.6.2 *Equations and Formulas*), indicating topics that received approximately 1/3 of the emphasis within a country.

Table 9 shows agreement and proportion summaries for each country. It indicates the number of topics within a country that either appeared in all or none of the data sources and the proportion of the 44 topics this represents. None of the countries had 100% agreement across data sources. The lowest amount of agreement was 34% (country G), and the highest was 80% (country D), with an average of 57%. Within a country, the numbers of topics appearing in all three data sources ranged from 5 to 32 with an average of 18. Five countries had at least 10 topics appearing in none of the data sources. For country Q, all topics appeared in at least one of the data sources.

Average emphasis across topics ranged from .03 (countries C, D, Q) to .20 (country G). Standard deviations of the proportions averaged .04 but ranged from .02 to .73 (country G). Country G had the highest average emphasis but the largest standard deviation because so few topics appeared in all three data sources. Most of the maximum proportions were at least .10. Differences between means and medians reflected the number of 0s in the data set.

Analyses of Match between the Field-Trial Instrument and the Curricula

I evaluated the match between the field-trial instrument and the curriculum-data sources in several ways. First, I compared the number of countries including each field-trial topic in each data source, the number of countries including topics not on the field-

Table 9

Agreement of Topic Inclusion across Expert-Mapping-, Curriculum-Guide-, and Textbook-Data Sources Presented for Countries across Topics

Country	Ave. # Sources Topics Appear	# Topics in All 3 Sources	# Topics in 0 Sources	Prop. Topics in 3 or 0 Sources	Ave.[a] Prop. of Emphasis	*SD*	Median[a] Prop. of Emphasis	Max.[a] Prop. of Emphasis
A	2.1	19	5	0.55	0.053	0.028	0.000	0.094
B	1.8	19	9	0.64	0.053	0.035	0.000	0.154
C	2.5	30	3	0.75	0.033	0.019	0.024	0.070
D	2.5	32	3	0.80	0.031	0.016	0.025	0.049
E	1.4	7	10	0.39	0.143	0.057	0.000	0.242
F	1.9	17	6	0.52	0.059	0.032	0.000	0.107
G	1.3	5	10	0.34	0.200	0.073	0.000	0.338
H	1.5	8	9	0.39	0.125	0.049	0.000	0.161
I	2.2	25	6	0.70	0.040	0.022	0.027	0.065
J	1.3	9	15	0.55	0.111	0.055	0.000	0.269
K	1.7	11	8	0.43	0.091	0.049	0.000	0.251
L	2.2	25	5	0.68	0.040	0.023	0.024	0.088
M	2.1	23	4	0.61	0.043	0.030	0.021	0.136
N	2.0	17	4	0.48	0.059	0.036	0.000	0.142
O	1.5	11	11	0.50	0.091	0.044	0.000	0.186
P	1.7	18	11	0.66	0.056	0.031	0.000	0.121
Q	2.7	32	0	0.73	0.031	0.020	0.022	0.106
Average	1.9	18.1	7.0	0.57	0.07	0.04	0.01	0.15

[a]Within each country, the average, median, or maximum proportions for topics included in all 3 data sources. Average shows the average of non-zero numbers.

trial instrument in each data source, and the average proportions of emphasis for each topic within each curriculum source (including the aggregate of the curriculum sources) with the proportion of items for each topic (i.e., topic weight) in the field-trial instrument. Second, I calculated the proportion of items on the field-trial instrument that measured topics included in each of the curriculum-data sources for each country and the proportion of each country's curricula (according to the four data sources) that was tested by the field-trial instrument. Third, I calculated differences between topic inclusion on the field-trial instrument and topic inclusion in each of the four data sources for each country, and I did the same for a comparison of topic weight (i.e., proportion of items) on the field-trial instrument and proportion of emphasis for each topic in each of the curriculum-data sources. Fourth, I computed correlations and Euclidean-distance measures between the field-trial instrument topic "profiles" (i.e., patterns of topic weights) and the four curriculum "profiles" (i.e., patterns of proportions of topic emphasis) for each country. The results of each of these analyses are presented below.

Summary comparison. Table 10 provides a summary of (1) the numbers of countries that included each topic within each data source, (2) the average proportions of emphasis for topics across all countries for all data sources, and (3) the numbers and proportions of items for each topic on the field-trial instrument. The proportions of items on the field-trial instrument sum to more than one because many items had more than one content code.

The higher frequencies of items on the field-trial instrument were for topics 1.1 .2.1 *Common Fractions*, 1.5.2 *Proportionality Problems*, 1.6.2 *Equations and Formulas*, and 1.7.1 *Data Representation and Analysis*. Most of the topics with higher

Table 10

Document and Field-Trial Proportion Comparisons

		Number of Countries				Ave. Prop. of Emphasis				Field Trial	
Topic Code	Topic	Expert Map.	Curr. Guide	Text	Aggre-gate	Expert Map.	Curr. Guide	Text	Aggre-gate	# Items	Prop. Items
1.1.1.1	Wh.Num.-Meaning	8	10	11	3	0.012	0.023	0.015	0.006	4	0.017
1.1.1.2	Wh.Num.-Oper.	9	12	15	8	0.017	0.027	0.040	0.024	14	0.058
1.1.1.3	Prop. of Oper.	8	9	15	6	0.017	0.020	0.021	0.016	2	0.008
1.1.2.1	Common Fractions	14	9	16	8	0.032	0.017	0.041	0.025	34	0.141
1.1.2.2	Decimal Fractions	14	10	15	9	0.028	0.020	0.024	0.025	17	0.071
1.1.2.3	Relat. of Fractions	12	10	15	9	0.025	0.021	0.013	0.017	11	0.046
1.1.2.4	Percentages	12	11	14	8	0.025	0.022	0.035	0.024	7	0.029
1.1.2.5	Prop. of Frac.	12	5	11	4	0.025	0.008	0.006	0.007	0	0
1.1.3.1	Negative Numbers	14	15	15	12	0.032	0.033	0.041	0.031	3	0.012
1.1.3.2	Rational Numbers	14	13	12	8	0.030	0.028	0.028	0.032	0	0
1.1.3.3	Real Numbers	11	12	11	6	0.021	0.028	0.026	0.017	0	0
1.1.4.1	Binary Arithmetic	2	5	4	0	0.003	0.009	0.001	0	0	0
1.1.4.2	Exponents	15	11	14	10	0.036	0.023	0.041	0.026	3	0.012
1.1.4.3	Complex Numbers	0	2	3	0	0	0.003	0	0	0	0
1.1.4.4	Number Theory	13	12	11	8	0.024	0.026	0.016	0.017	1	0.004
1.1.4.5	Counting	2	4	7	1	0.004	0.007	0.002	0.002	0	0
1.1.5.1	Estim. Quant.& Size	9	6	9	4	0.014	0.011	0.002	0.006	9	0.037
1.1.5.2	Rounding	14	10	10	7	0.030	0.020	0.007	0.011	8	0.033
1.1.5.3	Estim. Comput.	12	8	12	7	0.023	0.015	0.008	0.011	7	0.029
1.1.5.4	Exponents&Mag.	12	8	7	3	0.027	0.016	0.007	0.005	1	0.004
1.2.1	Measurement Unit	15	13	15	11	0.029	0.027	0.040	0.032	18	0.075
1.2.2	Per.,Area,Volume	15	14	13	12	0.029	0.031	0.071	0.045	16	0.066
1.2.3	Estim. Errors	11	11	7	4	0.018	0.023	0.002	0.005	3	0.012
1.3.1	2D Geo:Coordinate	14	12	14	8	0.029	0.024	0.034	0.022	6	0.025
1.3.2	2D Geo:Basics	17	14	17	14	0.039	0.030	0.055	0.043	7	0.029
1.3.3	2D Geo: Polygons	16	16	16	15	0.034	0.038	0.098	0.065	8	0.033
1.3.4	3D Geo	15	17	13	11	0.034	0.041	0.068	0.047	4	0.017
1.3.5	Vectors	1	9	7	1	0.002	0.018	0.005	0.002	0	0
1.4.1	Geo. Transform.	15	15	13	11	0.033	0.035	0.056	0.049	10	0.041
1.4.2	Cong. & Sim.	14	14	11	9	0.031	0.032	0.040	0.037	14	0.058
1.4.3	Constructions	13	12	9	5	0.024	0.024	0.008	0.008	0	0
1.5.1	Proport. Concepts	15	14	10	9	0.030	0.030	0.008	0.017	8	0.033
1.5.2	Proport. Prob.	16	15	12	10	0.041	0.032	0.020	0.027	23	0.095
1.5.3	Slope & Trig.	10	8	6	3	0.021	0.021	0.014	0.010	0	0
1.5.4	Lin. Interp.	6	6	4	0	0.011	0.013	0.002	0	0	0
1.6.1	Pat., Rel., Func.	15	16	17	15	0.032	0.038	0.060	0.059	12	0.050
1.6.2	Equat. & Formulas	16	17	17	16	0.041	0.041	0.205	0.133	33	0.137
1.7.1	Data Rep. & Anal.	16	15	14	13	0.039	0.035	0.048	0.051	27	0.112
1.7.2	Uncer. & Prob.	9	12	6	4	0.015	0.025	0.003	0.006	11	0.046
1.8.1	Infinite Process.	0	2	4	0	0	0.004	0.001	0	0	0
1.8.2	Change	0	3	0	0	0	0.005	0	0	0	0
1.9.1	Val. & Just.	7	4	10	3	0.013	0.008	0.022	0.008	0	0
1.9.2	Struc. & Abs.	8	7	9	5	0.012	0.015	0.021	0.011	0	0
1.10.1	Other	10	14	11	8	0.017	0.031	0.036	0.022	0	0
Average		11	11	11	7	0.023	0.023	0.029	0.023	7	0.030

numbers of items also had high rates of inclusion and emphasis in all data sources. The exceptions were 1.1.5.1 *Estimating Quantity and Size* and 1.7.2 *Uncertainty and Probability.* Nine items measured 1.1.5.1 and 11 measured 1.7.2. However, both topics were included in each of the curriculum sources of less than 70% of the countries and topic 1.1.5.1 had an average of only .002 blocks across all country textbooks while topic 1.7.2 had an average of .003. On the other hand, topics 1.1.3.1 *Negative Numbers,* 1.1.4.2 *Exponents,* and 1.1.4.4 *Number Theory* were measured by three or fewer items. However, topic 1.1.3.1 was included in each of the curriculum-data sources of at least 70% of the countries and topics 1.1.4.2 and 1.1.4.4 were included in two of the curriculum sources in at least 70% of the countries. Additionally, average proportions of emphasis for these topics were similar to those proportions for many other topics. No items measured topic 1.1.3.2 *Rational Numbers* which appeared in the three main data sources of at least 70% of the countries or topics 1.1.3.3 *Real Numbers* and 1.4.3 *Geometric Constructions* which also appeared in the data sources of many countries. However, the proportions of textbook blocks and proportions of emphasis in the aggregate of the data sources were low for topic 1.4.3.

Proportions of items/curricula covered. The top half of Table 11 shows the proportions of items on the field-trial instrument that measured topics included in each of the curriculum-data sources for each country (hereafter referred to as covered items). Across data sources and countries, these proportions ranged from .18 to 1.00 (18-100%) with an average of .75. Only one of the averages of the proportions of covered items for each country-data source was below .75. This exception was for the average of the proportions of items measuring topics included in the aggregate of each of the country's

Table 11

Proportions of Items in Field-Trial Instrument that are in Each Country's Curricula and Proportions of Each Country's Curricula Tested in Field-Trial Instrument

	Country																				
Data Source	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	AVE	SD	MIN	MAX
Proportion of Tested Items Covered in Curriculum																					
Expert Mapping	0.97	0.92	0.82	1.00	0.73	0.91	0.51	0.80	0.90	0.44	0.88	0.98	0.71	1.00	0.51	0.63	0.80	0.79	0.18	0.44	1.00
Curr. Guide	0.76	0.81	0.95	0.99	0.46	0.64	0.56	0.80	0.85	0.36	0.86	0.87	0.95	0.72	0.71	0.85	0.98	0.77	0.18	0.36	0.99
Textbook	0.98	0.95	0.82	1.00	0.81	0.94	0.78	0.93	0.91	0.47	0.92	1.00	0.73	1.00	0.90	0.63	0.94	0.87	0.14	0.47	1.00
Aggregate	0.66	0.54	0.85	0.98	0.41	0.64	0.18	0.28	0.71	0.34	0.38	0.82	0.60	0.27	0.36	0.57	0.78	0.55	0.23	0.18	0.98
Average	0.84	0.80	0.86	0.99	0.60	0.78	0.51	0.70	0.84	0.40	0.76	0.92	0.75	0.75	0.62	0.67	0.87	0.75	0.18	0.36	0.99
SD	0.13	0.16	0.05	0.01	0.17	0.14	0.21	0.25	0.08	0.06	0.22	0.07	0.13	0.30	0.20	0.10	0.09	0.12	0.03	0.11	0.01
Min	0.66	0.54	0.82	0.98	0.41	0.64	0.18	0.28	0.71	0.34	0.38	0.82	0.60	0.27	0.36	0.57	0.78	0.55	0.14	0.18	0.98
Max	0.98	0.95	0.95	1.00	0.81	0.94	0.78	0.93	0.91	0.47	0.92	1.00	0.95	1.00	0.90	0.85	0.98	0.87	0.23	0.47	1.00
Proportion of Curriculum Tested																					
Expert Mapping	0.85	0.79	0.78	0.84	0.70	0.95	0.74	0.83	0.88	0.74	0.89	0.80	0.78	0.77	0.86	0.93	0.78	0.82	0.07	0.70	0.95
Curr. Guide	0.86	0.79	0.69	0.83	0.91	0.69	0.89	0.82	0.77	0.72	0.75	0.87	0.74	0.65	0.76	0.83	0.66	0.78	0.08	0.65	0.91
Textbook	0.96	0.77	0.88	0.89	0.91	0.81	0.76	0.92	0.84	0.99	0.94	0.96	0.91	0.70	0.94	0.99	0.83	0.88	0.08	0.70	0.99
Aggregate	0.97	0.83	0.83	0.89	1.00	0.97	0.67	0.81	0.86	0.96	0.93	0.94	0.88	0.67	1.00	0.95	0.79	0.88	0.10	0.67	1.00
Average	0.91	0.80	0.80	0.86	0.88	0.86	0.76	0.84	0.84	0.85	0.88	0.89	0.83	0.70	0.89	0.93	0.76	0.84	0.08	0.68	0.96
SD	0.05	0.02	0.07	0.03	0.11	0.11	0.08	0.04	0.04	0.12	0.08	0.06	0.07	0.05	0.09	0.06	0.06	0.04	0.01	0.02	0.04
Min	0.85	0.77	0.69	0.83	0.70	0.69	0.67	0.81	0.77	0.72	0.75	0.80	0.74	0.65	0.76	0.83	0.66	0.78	0.07	0.65	0.91
Max	0.97	0.83	0.88	0.89	1.00	0.97	0.89	0.92	0.88	0.99	0.94	0.96	0.91	0.77	1.00	0.99	0.83	0.88	0.10	0.70	1.00

data sources (.55). Within-country averages of the proportions of covered items were more variable and ranged from .48 to 1.00. For most countries, the highest proportions of covered items were those measuring topics included in the textbooks. Additionally, for most countries, the lowest proportions of covered items were those measuring topics included in the aggregate of the data sources – the most restricted data source. Standard deviations of proportions of coverage within countries (across data sources) ranged from .01 to .30. When looking across countries (within data sources), the least variability was for the proportion of items measuring topics included in the textbook data sources, and the greatest variability was for the proportion of items measuring topics included in the aggregate of the data sources.

Much less variability existed in the proportions of the curriculum tested within each country. These proportions ranged from .65 to 1.00. Averages for the four data sources (across countries) were around .80 to .90. Most averages across data sources within countries were in that same range with the exception of one average proportion of .70 (country N). On average, 30% of this country's curriculum was not tested. The highest proportions of tested curricula within each country varied across the data sources, with the majority of countries having more of the curriculum as defined by the textbook(s) tested that the curriculum as defined by any of the other data sources. The lowest proportions of tested curriculum were associated with the curriculum guides. All standard deviations were .12 or less.

Differences in topic inclusion and emphasis. Differences for each topic between topic inclusion in each curriculum-data source and topic inclusion in the field-trial instrument are presented in Table 12. The second column of the table indicates which

Table 12

Differences in Topic Inclusion between the Field-Trial Instrument and Each Curriculum Source for Each Topic

		Expert Mapping		Curr. Guide		Textbook		Aggregate			
Topic Code	Test Items	# Mis-Match	Prop. Match	# Mis-Match	Prop. Match	# Mis-Match	Prop. Match	# Mis-Match	Prop. Match	Ave. # Mis-Match	Ave. Prop. Match
1.1.1.1	✓	9	0.47	7	0.59	6	0.65	14	0.18	9.0	0.47
1.1.1.2	✓	8	0.53	5	0.71	2	0.88	9	0.47	6.0	0.65
1.1.1.3	✓	9	0.47	8	0.53	2	0.88	11	0.35	7.5	0.56
1.1.2.1	✓	3	0.82	8	0.53	1	0.94	9	0.47	5.3	0.69
1.1.2.2	✓	3	0.82	7	0.59	2	0.88	8	0.53	5.0	0.71
1.1.2.3	✓	5	0.71	7	0.59	2	0.88	8	0.53	5.5	0.68
1.1.2.4	✓	5	0.71	6	0.65	3	0.82	9	0.47	5.8	0.66
1.1.2.5		12	0.29	5	0.71	11	0.35	4	0.76	8.0	0.53
1.1.3.1	✓	3	0.82	2	0.88	2	0.88	5	0.71	3.0	0.82
1.1.3.2		14	0.18	13	0.24	12	0.29	8	0.53	11.8	0.31
1.1.3.3		11	0.35	12	0.29	11	0.35	6	0.65	10.0	0.41
1.1.4.1		2	0.88	5	0.71	4	0.76	0	1	2.8	0.84
1.1.4.2	✓	2	0.88	6	0.65	3	0.82	7	0.59	4.5	0.74
1.1.4.3		0	1.00	2	0.88	3	0.82	0	1	1.3	0.93
1.1.4.4	✓	4	0.76	5	0.71	6	0.65	9	0.47	6.0	0.65
1.1.4.5		2	0.88	4	0.76	7	0.59	1	0.94	3.5	0.79
1.1.5.1	✓	8	0.53	11	0.35	8	0.53	13	0.24	10.0	0.41
1.1.5.2	✓	3	0.82	7	0.59	7	0.59	10	0.41	6.8	0.60
1.1.5.3	✓	5	0.71	9	0.47	5	0.71	10	0.41	7.3	0.57
1.1.5.4	✓	5	0.71	9	0.47	10	0.41	14	0.18	9.5	0.44
1.2.1	✓	2	0.88	4	0.76	2	0.88	6	0.65	3.5	0.79
1.2.2	✓	2	0.88	3	0.82	4	0.76	5	0.71	3.5	0.79
1.2.3	✓	6	0.65	6	0.65	10	0.41	13	0.24	8.8	0.49
1.3.1	✓	3	0.82	5	0.71	3	0.82	9	0.47	5.0	0.71
1.3.2	✓	0	1.00	3	0.82	0	1.00	3	0.82	1.5	0.91
1.3.3	✓	1	0.94	1	0.94	1	0.94	2	0.88	1.3	0.93
1.3.4	✓	2	0.88	0	1.00	4	0.76	6	0.65	3.0	0.82
1.3.5		1	0.94	9	0.47	7	0.59	1	0.94	4.5	0.74
1.4.1	✓	2	0.88	2	0.88	4	0.76	6	0.65	3.5	0.79
1.4.2	✓	3	0.82	3	0.82	6	0.65	8	0.53	5.0	0.71
1.4.3		13	0.24	12	0.29	9	0.47	5	0.71	9.8	0.43
1.5.1	✓	2	0.88	3	0.82	7	0.59	8	0.53	5.0	0.71
1.5.2	✓	1	0.94	2	0.88	5	0.71	7	0.59	3.8	0.78
1.5.3		10	0.41	8	0.53	6	0.65	3	0.82	6.8	0.60
1.5.4		6	0.65	6	0.65	4	0.76	0	1	4.0	0.76
1.6.1	✓	2	0.88	1	0.94	0	1.00	2	0.88	1.3	0.93
1.6.2	✓	1	0.94	0	1.00	0	1.00	1	0.94	0.5	0.97
1.7.1	✓	1	0.94	2	0.88	3	0.82	4	0.76	2.5	0.85
1.7.2	✓	8	0.53	5	0.71	11	0.35	13	0.24	9.3	0.46
1.8.1		0	1.00	2	0.88	4	0.76	0	1	1.5	0.91
1.8.2		0	1.00	3	0.82	0	1.00	0	1	0.8	0.96
1.9.1		7	0.59	4	0.76	10	0.41	3	0.82	6.0	0.65
1.9.2		8	0.53	7	0.59	9	0.47	5	0.71	7.3	0.57
1.10.1		10	0.41	14	0.18	11	0.35	8	0.53	10.8	0.37
Average		4.64	0.73	5.52	0.68	5.16	0.70	6.20	0.64	5.38	0.68

topics had items included on the field-trial instrument. The "Mis-Match" column under each curriculum source represents either (1) the number of countries including topics that are not on the field-trial instrument in a particular data source or (2) the number of countries not including a topic that was on the field-trial instrument in a particular data source. The columns labeled "Prop. Match" indicate the proportion of countries in which a match occurred (i.e., the topic appeared in both the field-trial instrument and curriculum source or did not appear in either).

The average of the proportions of "match" between the field-trial instrument and the curriculum-data sources in topic inclusion ranged from .64 to .73 for data sources and from .37 to .97 for topics. The topics with the lowest rates of match in topic inclusion were 1.1.3.2 *Rational Numbers* (an average of 5 countries corresponding in topic inclusion in their data source with topic inclusion on the field-trial instrument across data sources), 1.1.3.3 *Real Numbers* (an average of 7 countries corresponding), 1.1.5.1 *Estimating Quantity & Size* (an average of 7 countries corresponding), and 1.10 *Other Content* (an average of 6 countries corresponding). Those with the highest match with the field-trial instrument in topic inclusion were 1.1.4.3 *Complex Numbers*, 1.3.3 *Polygons and Circles*, 1.6.1. *Patterns, Relations, and Functions*, 1.6.2 *Equations and Formulas*, and 1.8.2 *Change*.

As expected, the lowest average proportion of match with the field-trial instrument on topic inclusion was between the field-trial instrument and the aggregate of the three curriculum-data sources for each country. The highest rate of topic-inclusion correspondence was between the field-trial instrument and the expert mapping. On average, though, across topics 60% to 70% of the countries either included topics in the

curriculum sources that were included on the field-trial instrument or did not include topics in the curriculum sources that were not included on the field-trial instrument.

Table 13 shows the summary information on the match of topic inclusion between the field-trial instrument and data sources on topic inclusion across countries. "Prop. Match" is the proportion of topics for which the topic inclusion (or lack of) in a country's data source corresponds with the topic inclusion (or lack of) on the field-trial instrument. "In Curr." is the number of topics included in a particular curriculum for each country that is not included on the field-trial instrument. "Not in Curr." is the number of topics included on the field-trial instrument that is not included in a particular curriculum-data source. Both inclusion and exclusion in the curriculum are important enough to consider separately. High numbers on "In Curr." indicated that countries intend that students be taught more topics than those being tested. High numbers on "Not in Curr." indicated that students were being tested on more topics than those they were intended to be taught.

Average proportions of match with the field-trial instrument on topic inclusion between the curriculum-data sources and the field-trial instrument ranged within countries from .52 to .84. Averages within data sources were the same as those in Table 12. The average numbers of topics included in a country's curriculum but not on the field-trial instrument was five, and the average number of topics included on the field-trial instrument but not in the curriculum was nine. The highest numbers of non-tested topics included in a curriculum source were for topics included in the textbooks and curriculum guides. The highest numbers of tested topics not in a curriculum source were for topics not included in the aggregate-data source. The best correspondence overall was for the expert mapping. Within countries, average rates of non-tested topics included in

Table 13

Differences in Topic Inclusion between the Field-Trial Instrument and Each Curriculum Source for Each Country

	Data Source													
	Expert Mapping			Curriculum Guide			Textbook			Aggregate					
Country	Prop. Match	In Curr.	Not in Curr.	Prop. Match	In Curr.	Not in Curr.	Prop. Match	In Curr.	Not in Curr.	Prop. Match	In Curr.	Not in Curr.	Ave. Prop. Match	Ave. in Curr.	Ave. Not in Curr.
A	0.82	6	2	0.82	4	4	0.70	7	6	0.73	1	11	0.77	4.5	5.8
B	0.77	6	4	0.66	5	10	0.75	4	7	0.68	2	12	0.72	4.3	8.3
C	0.80	6	3	0.68	12	2	0.68	12	2	0.75	6	5	0.73	9.0	3.0
D	0.82	8	0	0.86	6	0	0.80	8	1	0.89	4	1	0.84	6.5	0.5
E	0.57	7	12	0.75	2	9	0.43	6	19	0.50	0	22	0.56	3.8	15.5
F	0.89	2	3	0.66	11	4	0.59	5	13	0.68	1	13	0.70	4.8	8.3
G	0.52	3	18	0.66	2	13	0.64	6	10	0.41	1	25	0.56	3.0	16.5
H	0.73	5	7	0.50	2	20	0.70	7	6	0.43	2	23	0.59	4.0	14.0
I	0.82	5	3	0.73	7	5	0.73	8	4	0.73	4	8	0.75	6.0	5.0
J	0.50	6	16	0.59	7	11	0.50	2	20	0.50	1	21	0.52	4.0	17.0
K	0.82	3	5	0.57	5	14	0.75	5	6	0.55	1	19	0.67	3.5	11.0
L	0.82	7	1	0.84	4	3	0.80	6	3	0.82	2	6	0.82	4.8	3.3
M	0.64	7	9	0.68	8	6	0.82	7	1	0.68	4	10	0.70	6.5	6.5
N	0.77	10	0	0.45	6	18	0.66	9	6	0.45	6	18	0.59	7.8	10.5
O	0.64	4	12	0.59	5	13	0.80	1	8	0.59	0	18	0.65	2.5	12.8
P	0.77	1	9	0.77	5	5	0.80	3	6	0.70	1	12	0.76	2.5	8.0
Q	0.68	10	4	0.66	15	0	0.70	12	1	0.70	8	5	0.69	11.3	2.5
Average	0.73	5.6	6.4	0.68	6.2	8.1	0.70	6.4	7.0	0.64	2.6	13.5	0.68	5.2	8.7

curriculum sources ranged from 2.5 to 11, and overall average rates of tested topics not in curriculum sources ranged from approximately 1 to 17.

Tables 14 and 15 show the differences between proportions of emphasis for topics in each of the curriculum-data sources and the topic weight (i.e., number of items for each topic) on the field-trial instrument. Table 14 highlights differences across topics, and Table 15 highlights differences across countries. Positive differences occur when topics receive a higher emphasis in a curriculum-data source than on the field-trial instrument, and negative differences occur when topics receive a higher emphasis on the field-trial instrument than in a curriculum-data source. Again, both indices are important. The tables show standard deviations of differences in topic emphasis for each topic or country within each curriculum-data source as well as the averages within data sources of the positive and negative differences in topic emphasis for each topic or each country. The tables also show averages of these numbers across data sources.

Across data sources, the topic with the largest negative average difference between field-trial weight and curriculum emphasis (Table 14) was topic 1.1.2.1 *Common Fractions*. This topic, on average, was emphasized more on the field-trial instrument that in the curriculum-data sources. The topics with the largest positive average differences varied. Some of the larger differences were for topics 1.3.4 *3D Geometry*, 1.4.2 *Congruence and Similarity*, and 1.6.2 *Equations and Formulas*. The largest differences in emphasis were between the field-trial instrument and the textbooks.

On average across all curriculum sources, three topics (1.1.2.1 *Common Fractions*, 1.5.2 *Proportionality Problems*, 1.7.1 *Data Representation and Analysis*) had 0 as an average positive difference meaning that the topics were not emphasized more in

Table 14

Difference in Topic Emphasis between the Field-Trial Instrument and Each Curriculum-Source for Each Topic

		Expert Mapping			Curriculum Guide			Textbook		
Topic Code	Prop. Items	*SD* of all Dif.	Ave. Pos. Dif.	Ave. Neg. Dif.	*SD* of all Dif.	Ave. Pos. Dif.	Ave. Neg. Dif.	*SD* of all Dif.	Ave. Pos. Dif.	Ave. Neg. Dif.
1.1.1.1	0.017	0.013	0.009	-0.017	0.025	0.023	-0.017	0.026	0.035	-0.013
1.1.1.2	0.058	0.020	0.013	-0.044	0.023	0.033	-0.035	0.049	0.047	-0.046
1.1.1.3	0.008	0.022	0.028	-0.008	0.024	0.030	-0.008	0.023	0.029	-0.006
1.1.2.1	0.141	0.025	0	-0.109	0.016	0	-0.124	0.034	0	-0.100
1.1.2.2	0.071	0.023	0.035	-0.047	0.018	0	-0.051	0.024	0	-0.047
1.1.2.3	0.046	0.024	0.060	-0.026	0.020	0.009	-0.029	0.010	0	-0.032
1.1.2.4	0.029	0.019	0.016	-0.017	0.017	0.009	-0.019	0.034	0.032	-0.023
1.1.2.5	0	0.026	0.035	0	0.013	0.028	0	0.010	0.009	0
1.1.3.1	0.012	0.019	0.026	-0.012	0.016	0.025	-0.012	0.036	0.045	-0.010
1.1.3.2	0	0.025	0.037	0	0.018	0.037	0	0.071	0.040	0
1.1.3.3	0	0.018	0.032	0	0.023	0.039	0	0.064	0.039	0
1.1.4.1	0	0.008	0.023	0	0.015	0.031	0	0.003	0.004	0
1.1.4.2	0.012	0.018	0.029	-0.012	0.019	0.023	-0.012	0.038	0.045	-0.012
1.1.4.3	0	0	0	0	0.008	0.024	0	0.001	0.001	0
1.1.4.4	0.004	0.016	0.028	-0.004	0.019	0.032	-0.004	0.022	0.026	-0.004
1.1.4.5	0	0.011	0.034	0	0.012	0.028	0	0.006	0.006	0
1.1.5.1	0.037	0.013	0	-0.024	0.016	0.004	-0.028	0.003	0	-0.035
1.1.5.2	0.033	0.018	0.015	-0.016	0.017	0.005	-0.021	0.008	0	-0.026
1.1.5.3	0.029	0.018	0.017	-0.015	0.016	0.006	-0.023	0.009	0.003	-0.023
1.1.5.4	0.004	0.021	0.035	-0.004	0.017	0.029	-0.004	0.015	0.018	-0.004
1.2.1	0.071	0.015	0	-0.046	0.017	0	-0.047	0.042	0.034	-0.050
1.2.2	0.062	0.014	0	-0.038	0.017	0	-0.036	0.057	0.051	-0.048
1.2.3	0.012	0.016	0.015	-0.012	0.019	0.024	-0.012	0.003	0	-0.011
1.3.1	0.025	0.018	0.015	-0.012	0.017	0.010	-0.021	0.032	0.032	-0.018
1.3.2	0.029	0.014	0.019	-0.003	0.017	0.012	-0.014	0.042	0.052	-0.013
1.3.3	0.033	0.016	0.014	-0.011	0.019	0.015	-0.009	0.054	0.071	-0.033
1.3.4	0.017	0.017	0.021	-0.017	0.016	0.025	0	0.121	0.109	-0.013
1.3.5	0	0.007	0.028	0	0.017	0.033	0	0.013	0.013	0
1.4.1	0.037	0.019	0.016	-0.019	0.020	0.014	-0.016	0.064	0.058	-0.034
1.4.2	0.062	0.021	0.007	-0.034	0.018	0	-0.028	0.060	0.077	-0.047
1.4.3	0	0.018	0.031	0	0.017	0.035	0	0.012	0.016	0
1.5.1	0.033	0.015	0.010	-0.012	0.017	0.010	-0.014	0.010	0	-0.025
1.5.2	0.095	0.019	0	-0.054	0.015	0	-0.063	0.023	0	-0.075
1.5.3	0	0.022	0.035	0	0.027	0.045	0	0.025	0.041	0
1.5.4	0	0.017	0.031	0	0.018	0.036	0	0.004	0.007	0
1.6.1	0.050	0.018	0.007	-0.026	0.019	0.014	-0.019	0.054	0.048	-0.024
1.6.2	0.129	0.017	0	-0.096	0.016	0	-0.096	0.118	0.164	-0.041
1.7.1	0.112	0.018	0	-0.073	0.020	0	-0.077	0.032	0	-0.064
1.7.2	0.046	0.016	0.007	-0.033	0.018	0.007	-0.024	0.008	0	-0.042
1.8.1	0	0	0	0	0.012	0.034	0	0.001	0.003	0
1.8.2	0	0	0	0	0.011	0.028	0	0.000	0	0
1.9.1	0	0.018	0.031	0	0.016	0.034	0	0.072	0.037	0
1.9.2	0	0.014	0.027	0	0.020	0.036	0	0.034	0.040	0
1.10.1	0	0.015	0.028	0	0.017	0.038	0	0.062	0.056	0
Average	0.030	0.016	0.018	-0.019	0.017	0.020	-0.020	0.032	0.029	-0.021

Table 14 (Contd.)

Topic Code	Prop. Items	Aggregate *SD* of all Dif.	Aggregate Ave. Pos. Dif.	Aggregate Ave. Neg. Dif.	Ave. of Pos. Dif.	Ave. of Neg. Dif.	*SD* of Pos. Dif.	*SD* of Neg. Dif.	*SD* of All Dif.
1.1.1.1	0.017	0.014	0.018	-0.017	0.021	-0.016	0.009	0.002	0.020
1.1.1.2	0.058	0.032	0.063	-0.040	0.039	-0.041	0.018	0.004	0.042
1.1.1.3	0.008	0.023	0.036	-0.008	0.031	-0.008	0.003	0.001	0.019
1.1.2.1	0.141	0.027	0	-0.117	0	-0.113	0.000	0.009	0.057
1.1.2.2	0.071	0.031	0	-0.051	0.022	-0.049	0.022	0.002	0.039
1.1.2.3	0.046	0.017	0	-0.031	0.018	-0.029	0.024	0.002	0.029
1.1.2.4	0.029	0.030	0.022	-0.029	0.020	-0.022	0.008	0.005	0.022
1.1.2.5	0	0.014	0.031	0	0.026	0	0.010	0.000	0.015
1.1.3.1	0.012	0.027	0.032	-0.012	0.032	-0.012	0.008	0.001	0.023
1.1.3.2	0	0.076	0.068	0	0.045	0	0.013	0.000	0.025
1.1.3.3	0	0.028	0.048	0	0.040	0	0.006	0.000	0.020
1.1.4.1	0	0	0	0	0.014	0	0.013	0.000	0.012
1.1.4.2	0.012	0.026	0.032	-0.012	0.032	-0.012	0.008	0.000	0.023
1.1.4.3	0	0	0	0	0.006	0	0.010	0.000	0.008
1.1.4.4	0.004	0.019	0.031	-0.004	0.029	-0.004	0.002	0.000	0.017
1.1.4.5	0	0.007	0.031	0	0.025	0	0.011	0.000	0.015
1.1.5.1	0.037	0.010	0	-0.032	0.001	-0.030	0.002	0.004	0.016
1.1.5.2	0.033	0.014	0	-0.028	0.006	-0.023	0.006	0.005	0.015
1.1.5.3	0.029	0.014	0.007	-0.021	0.008	-0.020	0.005	0.003	0.015
1.1.5.4	0.004	0.011	0.025	-0.004	0.027	-0.004	0.006	0.000	0.016
1.2.1	0.071	0.029	0.013	-0.050	0.012	-0.048	0.014	0.002	0.032
1.2.2	0.062	0.041	0.051	-0.037	0.025	-0.040	0.025	0.005	0.037
1.2.3	0.012	0.010	0.010	-0.012	0.012	-0.012	0.009	0.001	0.014
1.3.1	0.025	0.027	0.024	-0.022	0.020	-0.018	0.008	0.004	0.020
1.3.2	0.029	0.028	0.025	-0.022	0.027	-0.013	0.015	0.007	0.023
1.3.3	0.033	0.037	0.044	-0.022	0.036	-0.019	0.023	0.010	0.033
1.3.4	0.017	0.053	0.056	-0.017	0.053	-0.012	0.035	0.007	0.041
1.3.5	0	0.009	0.037	0	0.028	0	0.009	0.000	0.015
1.4.1	0.037	0.063	0.070	-0.025	0.039	-0.024	0.025	0.007	0.036
1.4.2	0.062	0.053	0.083	-0.044	0.042	-0.038	0.038	0.008	0.049
1.4.3	0	0.013	0.027	0	0.027	0	0.007	0.000	0.014
1.5.1	0.033	0.017	0	-0.021	0.009	-0.018	0.005	0.005	0.014
1.5.2	0.095	0.029	0	-0.068	0	-0.065	0.000	0.008	0.033
1.5.3	0	0.028	0.054	0	0.044	0	0.007	0.000	0.022
1.5.4	0	0	0	0	0.018	0	0.015	0.000	0.014
1.6.1	0.050	0.039	0.039	-0.025	0.027	-0.023	0.017	0.003	0.028
1.6.2	0.129	0.087	0.115	-0.053	0.070	-0.071	0.072	0.025	0.089
1.7.1	0.112	0.038	0	-0.067	0	-0.070	0.009	0.005	0.038
1.7.2	0.046	0.012	0	-0.039	0.004	-0.035	0.004	0.007	0.020
1.8.1	0	0	0	0	0.009	0	0.014	0.000	0.011
1.8.2	0	0	0	0	0.007	0	0.012	0.000	0.009
1.9.1	0	0.024	0.047	0	0.037	0	0.006	0.000	0.019
1.9.2	0	0.019	0.037	0	0.035	0	0.005	0.000	0.018
1.10.1	0	0.032	0.046	0	0.042	0	0.010	0.000	0.022
Average	0.030	0.025	0.030	-0.021	0.024	-0.020	0.013	0.003	0.025

any curriculum source than on the field-trial instrument. These three topics had among the highest average negative differences. On average, topic 1.1.2.1 had over 10% more emphasis on the field-trial instrument than in the curriculum sources; topics 1.5.2 and 1.7.1 *Uncertainty and Probability* had around 7% more emphasis. Topic 1.6.2 *Equations and Formulas* had the highest positive difference (approximately 7% more emphasis on average in the curriculum than on the field-trial instrument), it also had a high negative difference (approximately 7% more emphasis on average on the field-trial instrument than in the curriculum). Looking at data sources shows that the topic has a higher negative than positive difference in topic emphasis between the expert-mapping data and the field-trial instrument and also between the curriculum-guide data and the field-trial instrument. It has a higher positive than negative difference for the other two data sources. The only topics with 0 as an average negative difference in emphasis were those averages for topics not included on the field-trial instrument. The average of the positive averages was .024 while the average of the negative averages was -.020.

Table 15 shows more variability in topic-emphasis differences between countries than existed between topics. Again, standard deviations and means of differences in topic emphasis within data sources were highest for the textbooks and the aggregate-data source. Textbooks had a larger average standard deviation of differences within countries than other sources had, and one country's (N) standard deviation of differences across topics was .109. Average positive differences in topic emphasis within countries across topics and data sources ranged from .019 to .071 with an average of around .037. Average negative differences in topic emphasis ranged from -.032 to -.043, with an

Table 15

Differences in Topic Emphasis between the Field-Trial Instrument and Each Curriculum Source for Each Country

	Expert Mapping			Curriculum Guide			Textbook			Aggregate							
Country	*SD* of all Dif.	Ave. Pos. Dif.	Ave. Neg. Dif.	*SD* of all Dif.	Ave. Pos. Dif.	Ave. Neg. Dif.	*SD* of all Dif.	Ave. Pos. Dif.	Ave. Neg. Dif.	*SD* of all Dif.	Ave. Pos. Dif.	Ave. Neg. Dif.	Ave.of Pos. Dif.	Ave. of Neg. Dif.	*SD* of Pos. Dif.	*SD* of Neg. Dif.	*SD* of All Dif.
A	0.034	0.020	-0.037	0.034	0.019	-0.041	0.034	0.024	-0.033	0.030	0.027	-0.037	0.023	-0.037	0.003	0.003	0.030
B	0.036	0.025	-0.035	0.039	0.023	-0.047	0.065	0.072	-0.040	0.046	0.052	-0.036	0.043	-0.040	0.020	0.004	0.044
C	0.038	0.026	-0.036	0.036	0.020	-0.035	0.039	0.027	-0.031	0.034	0.021	-0.038	0.024	-0.035	0.003	0.003	0.029
D	0.032	0.017	-0.040	0.033	0.021	-0.032	0.031	0.019	-0.030	0.030	0.018	-0.033	0.019	-0.034	0.002	0.004	0.027
E	0.034	0.027	-0.032	0.033	0.027	-0.035	0.054	0.047	-0.043	0.051	0.105	-0.036	0.051	-0.036	0.032	0.004	0.050
F	0.028	0.021	-0.033	0.036	0.022	-0.036	0.044	0.042	-0.036	0.033	0.036	-0.031	0.030	-0.034	0.009	0.002	0.033
G	0.042	0.045	-0.039	0.039	0.034	-0.039	0.071	0.061	-0.042	0.072	0.143	-0.042	0.071	-0.041	0.043	0.001	0.063
H	0.034	0.025	-0.038	0.048	0.066	-0.042	0.043	0.037	-0.035	0.053	0.087	-0.039	0.054	-0.038	0.024	0.003	0.049
I	0.033	0.019	-0.039	0.035	0.023	-0.038	0.040	0.033	-0.031	0.033	0.024	-0.035	0.025	-0.036	0.005	0.003	0.031
J	0.041	0.034	-0.042	0.040	0.029	-0.040	0.061	0.075	-0.041	0.050	0.069	-0.040	0.052	-0.041	0.020	0.001	0.049
K	0.036	0.027	-0.036	0.037	0.030	-0.035	0.055	0.052	-0.036	0.041	0.053	-0.034	0.041	-0.035	0.012	0.001	0.039
L	0.037	0.023	-0.038	0.032	0.016	-0.038	0.030	0.019	-0.031	0.028	0.020	-0.033	0.020	-0.035	0.003	0.003	0.028
M	0.036	0.030	-0.034	0.038	0.023	-0.039	0.057	0.056	-0.034	0.036	0.033	-0.032	0.035	-0.035	0.013	0.003	0.036
N	0.035	0.019	-0.035	0.048	0.041	-0.048	0.109	0.096	-0.048	0.050	0.047	-0.043	0.051	-0.043	0.028	0.005	0.051
O	0.038	0.027	-0.045	0.039	0.030	-0.041	0.052	0.078	-0.036	0.042	0.050	-0.039	0.046	-0.040	0.020	0.003	0.046
P	0.036	0.026	-0.039	0.032	0.021	-0.034	0.037	0.029	-0.032	0.033	0.028	-0.033	0.026	-0.035	0.003	0.003	0.030
Q	0.036	0.022	-0.035	0.037	0.019	-0.037	0.034	0.024	-0.026	0.032	0.020	-0.031	0.021	-0.032	0.002	0.004	0.027
Average	0.036	0.026	-0.037	0.037	0.027	-0.039	0.050	0.047	-0.036	0.041	0.049	-0.036	0.037	-0.037	0.014	0.003	0.039

average of -.037. Smaller overall positive and negative differences were noted for countries D and Q, and larger differences were noted for countries G, J, and N.

Correlations and Euclidean-Distance measures. Correlations between the proportions of emphasis patterns for topics in each country's curriculum-data source and topic weights on the field-trial instrument are in Table 16. The correlations ranged from -.064 to .66 with an overall average of .36. The average of the correlations of topic-emphasis patterns with the field-trial instrument was highest for the aggregate of the data sources. However, only five countries had their highest correlations for that data source. The lowest average correlation was for the curriculum guides. Average correlations within countries across data sources varied considerably. These ranged from a low of .111 (country N) to a high of .524 (country D). Thus, some countries had curriculum-topic-emphasis "profiles" that were all, or mostly all, uncorrelated with the field-trial instrument topic-weight profile while others had curriculum-topic-emphasis profiles that were almost all moderately correlated with the field-trial instrument topic-weight profile. Standard deviations of correlations within countries and across data sources varied from .04 (country G) to .23 (country Q).

Euclidean distances are shown in Table 17. These numbers represent the square roots of the sums of squared differences between the proportions of emphasis patterns for topics in a particular curriculum-data source for each country and the proportion of items for each topic on the field-trial instrument. These distances can be used to determine the extent of dissimilarity between each of the country curriculum-topic-emphasis "profiles' and the field-trial instrument topic-weight profile. The larger numbers indicate greater dissimilarity, and, for all practical purposes, the numbers are relative. However, some

Table 16

Correlations between the Proportions of Topic-Emphasis-Profiles for Each Country in Each Curriculum-Data Source and the Topic-Weight Profile for the Field-Trial Instrument

Country	Expert Map	Curr. Guide	Textbook	Aggregate	Average	*SD*
A	0.359*	0.404**	0.556**	0.591**	0.477	0.098
B	0.282	0.150	0.073	0.157	0.165	0.075
C	0.146	0.215	0.434**	0.394**	0.297	0.120
D	0.513**	0.419**	0.552**	0.612**	0.524	0.070
E	0.428**	0.456**	0.324*	0.506**	0.428	0.066
F	0.636**	0.223	0.270	0.548**	0.419	0.176
G	0.331*	0.265	0.361*	0.284	0.310	0.038
H	0.387**	0.201	0.358*	0.270	0.304	0.074
I	0.432**	0.287	0.406**	0.436**	0.390	0.061
J	0.210	0.118	0.487**	0.470**	0.321	0.161
K	0.333*	0.316*	0.493**	0.563**	0.426	0.105
L	0.195	0.498**	0.663**	0.632**	0.497	0.185
M	0.312*	0.073	0.459**	0.412**	0.314	0.149
N	0.352*	-0.064	0.104	0.054	0.111	0.152
O	0.300*	0.215	0.480**	0.471**	0.367	0.113
P	0.374*	0.475**	0.566**	0.525**	0.485	0.072
Q	0.244	0.144	0.587**	0.484**	0.329	0.227
Ave	0.343	0.258	0.422	0.436	0.363	0.114
SD	0.116	0.151	0.157	0.157	0.109	0.051

*$p < .05$. *$p < .01$.

Table 17

Euclidean Distances between the Proportions-of-Topic-Emphasis Profiles in Each Curriculum-Data Source and the the Topic-Weight Profile for the Field-Trial Instrument

Country	Expert Map	Curr. Guide	Textbook	Aggregate	Average	SD
A	0.233	0.228	0.228	0.207	0.224	0.010
B	0.243	0.266	0.435	0.312	0.314	0.074
C	0.256	0.243	0.260	0.231	0.247	0.012
D	0.218	0.228	0.208	0.203	0.214	0.009
E	0.231	0.226	0.360	0.335	0.288	0.060
F	0.195	0.243	0.299	0.223	0.240	0.038
G	0.284	0.267	0.468	0.481	0.375	0.100
H	0.233	0.323	0.285	0.354	0.299	0.045
I	0.225	0.240	0.267	0.228	0.240	0.017
J	0.278	0.267	0.407	0.334	0.321	0.056
K	0.241	0.252	0.366	0.278	0.284	0.049
L	0.248	0.217	0.201	0.195	0.215	0.020
M	0.244	0.260	0.377	0.246	0.282	0.055
N	0.234	0.322	0.752	0.333	0.410	0.201
O	0.258	0.264	0.345	0.281	0.287	0.034
P	0.246	0.220	0.249	0.227	0.235	0.012
Q	0.244	0.248	0.227	0.219	0.234	0.012
Ave	0.24	0.25	0.34	0.28	0.28	0.04
SD	0.020	0.030	0.130	0.073	0.063	0.043

benchmarks can be identified. For example, if the proportions of emphasis of all 44 topics in the curriculum differed from the corresponding topic weight on the field-trial instrument by .01 (and the proportions summed to 1 within the curriculum and the test), the Euclidean distance would be .07; if all topics differed by .10, the Euclidean distance would be .66; if 1/4 of the topics differed by .10 and the other 3/4 were the same, the Euclidean distance would be .33; if 1/2 of the topics differed by .10, the Euclidean distance would be .47. Finally, the field-trial tested only topics not included in the curriculum, the Euclidean distance would be 1.4. Thus, the smallest possible Euclidean distance was 0 and the largest (if both sets of proportions summed to 1) was 1.4.

The Euclidean distances ranged from .195 to .752, with an overall average of .28. The largest average distance was found between the field-trial instrument topic-weight profile and the textbook-topic-emphasis profiles. The smallest was between the field-trial instrument profiles and the expert mapping profiles. For most countries, the largest distance was between the field-trial instrument profiles and the textbook profiles. The exceptions were for countries D, H, L, and Q.

The smallest average Euclidean distance was for country D (.214). The largest was for country N (.410). Average standard deviations across countries within curriculum sources were between .24 and .34. Average standard deviations within countries across curriculum sources ranged from .01 to .20. Country N had by far the most variability in Euclidean distances across curriculum-data sources.

Development of Test Blueprints

For the next sets of analyses, I used the curriculum information from each country to design test blueprints with optimal content match to the curriculum depending upon the test purpose. I focused on three questions when developing the blueprints: What was the purpose of the test (i.e., what inferences will be made), what topics should be included in the test; What proportion of items should be allocated to each included topic?

Determine the Purpose of the Test

I assumed that, at the most general level, the purpose of all tests would be to compare cross-national student achievement of the content included in the mathematics curriculum for 13-year-old students. As seen earlier, the mathematics curriculum for 13-year-old students varied within and across nations. Therefore, I was interested in specifying the exact nature of the curriculum on which students would demonstrate achievement. The specific purpose of the test, therefore, had implications for the topics that would be included on the test.

I focused on two specific purposes for test development. The first purpose was to compare student achievement of the content of the intended mathematics curriculum cross-nationally. The expert-mapping and curriculum-guide analyses served as data sources for these blueprints. The second purpose was to evaluate student achievement of the content of the mathematics curriculum to which the students were likely to have been exposed (i.e., the potentially implemented mathematics curriculum). The results of the textbook analyses served as one the data source for these blueprints; the aggregate of the data sources served as another source for these test blueprints.

Determine Topic Inclusion

Topic inclusion in each test blueprint was the next issue to confront. If the curriculum sources were used to develop separate test blueprints for each country, any topic that appeared in the relevant curriculum source for a country would be included in the blueprint. However, when looking cross-nationally, the decision was not as simple. Topic inclusion varied across the countries, and, although many commonalties existed, many differences existed also.

I used four methods for determining topic inclusion in each test blueprint. The first method was the development of a *unique*-test blueprint for each country, only including those topics that appeared in a particular data source for each country. The other three methods were three different ways of combining each country's curriculum data to develop *inclusive*-test blueprints (i.e., one test for all countries). The first of the inclusive methods was to include on the test a *union* of all topics that any country included in each data source. The second inclusive method was a *70% intersection* method, that is the inclusion of topics that appeared in a particular data source for at least 70% of the countries. Finally, the third inclusive method for determining topic inclusion was to develop a test using a *strict intersection* of topics appearing in the relevant data source of all countries. Table 18 presents information on topic inclusion for the union, 70%-intersection, and strict-intersection methods. One check in a column indicates that the corresponding topic would appear in the union-test blueprint only, two checks indicate that a topic would also appear in the 70%-intersection-test blueprint, and three checks indicate that a topic would appear in the union-, 70%-intersection-, and strict-intersection-test blueprints.

Table 18

Items Included on Test Blueprints

Topic Code	Topic	Curriculum Source: Expert Mapping	Curr. Guide	Textbook	Aggregate
1.1.1.1	Wh.Num.-Meaning	✓	✓	✓	✓
1.1.1.2	Wh.Num.-Oper.	✓	✓✓	✓✓	✓
1.1.1.3	Prop. of Oper.	✓	✓	✓✓	✓
1.1.2.1	Common Fractions	✓✓	✓	✓✓	✓
1.1.2.2	Decimal Fractions	✓✓	✓	✓✓	✓
1.1.2.3	Relat. of Fractions	✓✓	✓	✓✓	✓
1.1.2.4	Percentages	✓✓	✓	✓✓	✓
1.1.2.5	Prop. of Frac.	✓✓	✓	✓	✓
1.1.3.1	Negative Numbers	✓✓	✓✓	✓✓	✓✓
1.1.3.2	Rational Numbers	✓✓	✓✓	✓✓	✓
1.1.3.3	Real Numbers	✓	✓✓	✓	✓
1.1.4.1	Binary Arithmetic	✓	✓	✓	
1.1.4.2	Exponents	✓✓	✓	✓✓	✓
1.1.4.3	Complex Numbers		✓	✓	
1.1.4.4	Number Theory	✓✓	✓✓	✓	✓
1.1.4.5	Counting	✓	✓	✓	✓
1.1.5.1	Estim. Quant.& Size	✓	✓	✓	✓
1.1.5.2	Rounding	✓✓	✓	✓	✓
1.1.5.3	Estim. Comput.	✓✓	✓	✓✓	✓
1.1.5.4	Exponents&Mag.	✓✓	✓	✓	✓
1.2.1	Measurement Unit	✓✓	✓✓	✓✓	✓
1.2.2	Per.,Area,Volume	✓✓	✓✓	✓✓	✓✓
1.2.3	Estim. Errors	✓	✓	✓	✓
1.3.1	2D Geo:Coordinate	✓✓	✓✓	✓✓	✓
1.3.2	2D Geo:Basics	✓✓✓	✓✓	✓✓✓	✓✓
1.3.3	2D Geo: Polygons	✓✓	✓✓	✓✓	✓✓
1.3.4	3D Geo	✓✓	✓✓✓	✓✓	✓
1.3.5	Vectors	✓	✓	✓	✓
1.4.1	Geo. Transform.	✓✓	✓✓	✓✓	✓
1.4.2	Cong. & Sim.	✓✓	✓✓	✓	✓
1.4.3	Constructions	✓✓	✓✓	✓	✓
1.5.1	Proport. Concepts	✓✓	✓✓	✓	✓
1.5.2	Proport. Prob.	✓✓	✓✓	✓✓	✓
1.5.3	Slope & Trig.	✓	✓	✓	✓
1.5.4	Lin. Interp.	✓	✓	✓	
1.6.1	Pat., Rel., Func.	✓✓	✓✓	✓✓✓	✓✓
1.6.2	Equat. & Formulas	✓✓	✓✓✓	✓✓✓	✓✓
1.7.1	Data Rep. & Anal.	✓✓	✓✓	✓✓	✓✓
1.7.2	Uncer. & Prob.	✓	✓✓	✓	✓
1.8.1	Infinite Process.		✓	✓	
1.8.2	Change		✓		
1.9.1	Val. & Just.	✓	✓	✓	✓
1.9.2	Struc. & Abs.	✓	✓	✓	✓
1.10.1	Other	✓	✓✓	✓	✓
Union		41	44	43	39
70% Int.		26	21	21	7
Int.		1	2	3	0

According to Table 18, union-test blueprints would contain between 39 and 44 topics. The numbers of topics on the 70%-intersection-test blueprints ranged from 7 to 26 topics. The strict-intersection-test blueprints would include only from 0 to 3 topics. All but five topics would be included in the union-test blueprints for all data sources, and all topics would be included in the union-test blueprints for at least one of the data sources. Thirty-one topics would be included in at least one of the 70%-intersection-test blueprints, and four topics would be included in at least one of the strict-intersection-test blueprints. Only seven topics would appear in the 70%-intersection blueprints developed for the aggregate of the data sources, and no topics would appear in the strict-intersection blueprints developed for the aggregate of the data sources.

Determine Topic Emphasis

Next, the relative emphasis (i.e., weight) that topics would receive on each test had to be determined. For the inclusive-test blueprints (i.e., union, 70% intersection, strict intersection), I weighted topics according to the average of the proportions of emphasis allocated to each topic for each country within each of the data sources. For the unique tests, I weighted topics differently for each country according to the proportion of emphasis to each topic in the relevant data source for each country.

The topic weights for the two types of intersection-test blueprints are presented in Table 19. The weights for the union tests have been presented earlier in this paper (Table 10), and the weights for the unique tests are presented in Appendix C. The weights for topics on the 70%-intersection blueprints based on the expert mapping and the curriculum

Table 19

Topic Weights on Test Blueprints

Topic Code	Topic	Expert 70%	Expert Strict	Curr. Gd. 70%	Curr. Gd. Strict	Text 70%	Text Strict	Aggre. 70%	AVE
1.1.1.1	Wh.Num.-Meaning	0	0	0	0	0	0	0	0.000
1.1.1.2	Wh.Num.-Oper.	0	0	0.042	0	0.038	0	0	0.011
1.1.1.3	Prop. of Oper.	0	0	0	0	0.020	0	0	0.003
1.1.2.1	Common Fractions	0.040	0	0	0	0.039	0	0	0.011
1.1.2.2	Decimal Fractions	0.035	0	0	0	0.023	0	0	0.008
1.1.2.3	Relat. of Fractions	0.031	0	0	0	0.013	0	0	0.006
1.1.2.4	Percentages	0.032	0	0	0	0.034	0	0	0.009
1.1.2.5	Prop. of Frac.	0.031	0	0	0	0	0	0	0.004
1.1.3.1	Negative Numbers	0.039	0	0.050	0	0.040	0	0.073	0.029
1.1.3.2	Rational Numbers	0.038	0	0.043	0	0.027	0	0	0.015
1.1.3.3	Real Numbers	0	0	0.042	0	0	0	0	0.006
1.1.4.1	Binary Arithmetic	0	0	0	0	0	0	0	0.000
1.1.4.2	Exponents	0.045	0	0	0	0.039	0	0	0.012
1.1.4.3	Complex Numbers	0	0	0	0	0	0	0	0.000
1.1.4.4	Number Theory	0.030	0	0.039	0	0	0	0	0.010
1.1.4.5	Counting	0	0	0	0	0	0	0	0.000
1.1.5.1	Estim. Quant.& Size	0	0	0	0	0	0	0	0.000
1.1.5.2	Rounding	0.037	0	0	0	0	0	0	0.005
1.1.5.3	Estim. Comput.	0.029	0	0	0	0.007	0	0	0.005
1.1.5.4	Exponents&Mag.	0.034	0	0	0	0	0	0	0.005
1.2.1	Measurement Unit	0.036	0	0.041	0	0.038	0	0.000	0.017
1.2.2	Per.,Area,Volume	0.036	0	0.047	0	0.068	0	0.104	0.036
1.2.3	Estim. Errors	0	0	0	0	0	0	0	0.000
1.3.1	2D Geo:Coordinate	0.036	0	0.037	0	0.032	0	0.000	0.015
1.3.2	2D Geo:Basics	0.048	1.00	0.046	0	0.052	0.171	0.102	0.203
1.3.3	2D Geo: Polygons	0.043	0	0.058	0	0.094	0	0.153	0.050
1.3.4	3D Geo	0.042	0	0.063	0.500	0.065	0	0	0.096
1.3.5	Vectors	0	0	0	0	0	0	0	0.000
1.4.1	Geo. Transform.	0.041	0	0.052	0	0.053	0	0	0.021
1.4.2	Cong. & Sim.	0.039	0	0.049	0	0	0	0	0.012
1.4.3	Constructions	0.030	0	0.037	0	0	0	0	0.010
1.5.1	Proport. Concepts	0.037	0	0.046	0	0	0	0	0.012
1.5.2	Proport. Prob.	0.051	0	0.049	0	0.019	0	0	0.017
1.5.3	Slope & Trig.	0	0	0	0	0	0	0	0.000
1.5.4	Lin. Interp.	0	0	0	0	0	0	0	0.000
1.6.1	Pat., Rel., Func.	0.040	0	0.058	0	0.057	0.187	0.138	0.069
1.6.2	Equat. & Formulas	0.052	0	0.063	0.500	0.196	0.642	0.312	0.252
1.7.1	Data Rep. & Anal.	0.049	0	0.052	0	0.046	0	0.119	0.038
1.7.2	Uncer. & Prob.	0	0	0.038	0	0	0	0	0.005
1.8.1	Infinite Process.	0	0	0	0	0	0	0	0.000
1.8.2	Change	0	0	0	0	0	0	0	0.000
1.9.1	Val. & Just.	0	0	0	0	0	0	0	0.000
1.9.2	Struc. & Abs.	0	0	0	0	0	0	0	0.000
1.10.1	Other	0	0	0.047	0	0	0	0	0.007

guides ranged from around .03 to around .06. Topic weights on the intersection blueprints based on the textbooks had a larger range of .008 (1.1.5.3 *Estimating Computations*) to .196 (1.6.2 *Equations and Formulas*). Topic weight on the strict-intersection blueprints varied. Only one topic was included in the strict-intersection blueprint for the expert mapping; therefore, it received 100% of the weight. Two topics were included in the curriculum-guide strict-intersection blueprint, each receiving half of the weight. Three topics were included in the strict-intersection-test blueprint for the textbook. Two of the topics received around 20% of the weight, and one topic received around 60% of the weight. Disregarding topics not included in any of the intersection test blueprints, averages of the topic weights ranged from .005 to .251. Topics with the highest average weight were 1.3.2 *Basic 2D Geometry* (.203) and 1.6.2 *Equations & Formulas* (.251). Table 20 provides a summary of codes used throughout the remainder of this section.

Comparisons between the Field-Trial Instrument and Test Blueprints

I repeated the test-to-curriculum match analyses described earlier comparing the content of each of the inclusive-test blueprints (i.e., the union, 70%-intersection, and strict-intersection blueprints) to the content of the field-trial instrument. This resulted in comparisons of the actual test to other tests that could be developed based on a country's curriculum. Unique-test blueprints are identical to the data in each of the curriculum sources. Therefore, a comparison of these blueprints with the field-trial instrument would yield identical results to the initial sets of match analyses.

Table 20

Test-Blueprint Codes

Method of Specification	Curriculum Source			
	Expert Mapping	Curriculum Guides	Textbooks	Aggregate
Union of Topics across all Countries	EX-UN	CG-UN	TX-UN	AG-UN
Intersection of Topics in 70% of Countries	EX-7I	CG-7I	TX-7I	AG-7I
Intersection of Topics in all Countries	EX-SI	CG-SI	TX-SI	[a]
Unique Topic Coverage	EX-UQ	CG-UQ	TX-UQ	AG-UQ

[a] No topics existed in the test blueprint for the strict intersection of aggregate curriculum source.

Proportions of items/blueprints covered. Table 21 shows the proportions of items on the TIMSS field-trial instrument that tested topics included in each of the test blueprints, and the proportion of "items" in each of the test blueprints (i.e., sum of topic weights) for topics that were tested by items on the field-trial instrument. Proportions of items that tested topics on each of the blueprints ranged from .02 to 1.00 with an average of .61. All topics tested on the field-trial instrument were included in each of the union-test blueprints. The proportion of field-trial items measuring topics included on the 70%-intersection blueprints ranged from .29 to .84. The variability was quite substantial. The standard deviation was .38.

The proportions of "items" in each of the test blueprints that were allocated to topics tested on the field-trial instrument ranged from .78 for the curriculum-guide union-test blueprint to 1.00 for the three strict-intersection-test blueprints and the aggregate-70%-intersection-test blueprint. This meant that the field-trial instrument included all topics that were included on each of these test blueprints. The average proportion of emphasis was .91. As expected, proportions of items for topics included on the test blueprints that were also included on the field-trial instrument increased as the test blueprints became more restricted. The opposite was true when looking at the proportion of items on the field-trial instrument that tested topics in each of the test blueprints.

Differences in topic inclusion and emphasis. Differences in topic inclusion between each of the inclusive-test blueprints and the field-trial instrument are presented for all topics in Table 22. A check in the second column of the table indicates which items were included on the field-trial instrument. If a topic was included on the field-trial instrument but not in the blueprint, a value of -1 was entered in the corresponding cell of

Table 21

Proportions of Field-Trial Items in Each Test Blueprint and Proportions of Items in Each Test Blueprint Tested on Field-Trial Instrument

	Blueprints														
	EX-UN	EX-7I	EX-SI	CG-UN	CG-7I	CG-SI	TX-UN	TX-7I	TX-SI	AG-UN	AG-7I	AVE	SD	MIN	MAX
Proportion of Field-Trial Items in Test Blueprints	1.00	0.84	0.02	1.00	0.58	0.09	1.00	0.76	0.14	1.00	0.29	0.61	0.38	0.02	1.00
Proportion of Test Blueprints Tested on Field-Trial Inst.	0.82	0.90	1.00	0.78	0.83	1.00	0.87	0.97	1.00	0.88	1.00	0.91	0.08	0.78	1.00

Table 22

Differences in Topic Inclusion between the Field-Trial Instrument and Each Test Blueprint

CODE	Test	EX-UN	EX-TI	EX-SI	CG-UN	CG-TI	CG-SI	TX-UN	TX-TI	TX-SI	AG-UN	AG-TI	SUM	Prop. Match	# Match
1.1.1.1	✓	0	-1	-1	0	-1	-1	0	-1	-1	0	-1	-7	0.36	4
1.1.1.2	✓	0	-1	-1	0	0	-1	0	0	-1	0	-1	-5	0.55	6
1.1.1.3	✓	0	-1	-1	0	-1	-1	0	0	-1	0	-1	-6	0.45	5
1.1.2.1	✓	0	0	-1	0	-1	-1	0	0	-1	0	-1	-5	0.55	6
1.1.2.2	✓	0	0	-1	0	-1	-1	0	0	-1	0	-1	-5	0.55	6
1.1.2.3	✓	0	0	-1	0	-1	-1	0	0	-1	0	-1	-5	0.55	6
1.1.2.4	✓	0	0	-1	0	-1	-1	0	0	-1	0	-1	-5	0.55	6
1.1.2.5		1	1	0	1	0	0	1	0	0	1	0	5	0.55	6
1.1.3.1	✓	0	0	-1	0	0	-1	0	0	-1	0	0	-3	0.73	8
1.1.3.2		1	1	0	1	1	0	1	1	0	1	0	7	0.36	4
1.1.3.3		1	0	0	1	1	0	1	0	0	1	0	5	0.55	6
1.1.4.1		1	0	0	1	0	0	1	0	0	0	0	3	0.73	8
1.1.4.2	✓	0	0	-1	0	-1	-1	0	0	-1	0	-1	-5	0.55	7
1.1.4.3		0	0	0	1	0	0	1	0	0	0	0	2	0.82	9
1.1.4.4	✓	0	0	-1	0	0	-1	0	-1	-1	0	-1	-5	0.55	6
1.1.4.5		1	0	0	1	0	0	1	0	0	1	0	4	0.64	7
1.1.5.1	✓	0	-1	-1	0	-1	-1	0	-1	-1	0	-1	-7	0.36	4
1.1.5.2	✓	0	0	-1	0	-1	-1	0	-1	-1	0	-1	-6	0.45	5
1.1.5.3	✓	0	0	-1	0	-1	-1	0	0	-1	0	-1	-5	0.55	6
1.1.5.4	✓	0	0	-1	0	-1	-1	0	-1	-1	0	-1	-6	0.45	5
1.2.1	✓	0	0	-1	0	0	-1	0	0	-1	0	-1	-4	0.64	7
1.2.2	✓	0	0	-1	0	0	-1	0	0	-1	0	0	-3	0.73	8
1.2.3	✓	0	-1	-1	0	-1	-1	0	-1	-1	0	-1	-7	0.36	5
1.3.1	✓	0	0	-1	0	0	-1	0	0	-1	0	-1	-4	0.64	7
1.3.2	✓	0	0	0	0	0	-1	0	0	0	0	0	-1	0.91	10
1.3.3	✓	0	0	-1	0	0	-1	0	0	-1	0	0	-3	0.73	8
1.3.4	✓	0	0	-1	0	0	0	0	0	-1	0	-1	-3	0.73	8
1.3.5		1	0	0	1	0	0	1	0	0	1	0	4	0.64	4
1.4.1	✓	0	0	-1	0	0	-1	0	0	-1	0	-1	-4	0.64	7
1.4.2	✓	0	0	-1	0	0	-1	0	-1	-1	0	-1	-5	0.55	6
1.4.3		1	1	0	1	1	0	1	0	0	1	0	6	0.45	6
1.5.1	✓	0	0	-1	0	0	-1	0	-1	-1	0	-1	-5	0.55	6
1.5.2	✓	0	0	-1	0	0	-1	0	0	-1	0	-1	-4	0.64	7
1.5.3		1	0	0	1	0	0	1	0	0	1	0	4	0.64	7
1.5.4		1	0	0	1	0	0	1	0	0	0	0	3	0.73	8
1.6.1	✓	0	0	-1	0	0	-1	0	0	0	0	0	-2	0.82	9
1.6.2	✓	0	0	-1	0	0	0	0	0	0	0	0	-1	0.91	10
1.7.1	✓	0	0	-1	0	0	-1	0	0	-1	0	0	-3	0.73	8
1.7.2	✓	0	-1	-1	0	-1	-1	0	-1	-1	0	-1	-7	0.36	4
1.8.1		0	0	0	1	0	0	1	0	0	0	0	2	0.82	9
1.8.2		0	0	0	1	0	0	0	0	0	0	0	1	0.91	10
1.9.1		1	0	0	1	0	0	1	0	0	1	0	4	0.64	7
1.9.2		1	0	0	1	0	0	1	0	0	1	0	4	0.64	7
1.10.1		1	0	0	1	1	0	1	0	0	1	0	5	0.55	6
# Match		32	35	16	29	27	17	30	34	18	34	22			
Prop. Match		0.73	0.80	0.36	0.66	0.61	0.39	0.68	0.77	0.41	0.77	0.50			
In Blueprint		12	3	0	15	4	0	14	1	0	10	0			
Not in Blueprint		0	6	28	0	13	27	0	9	26	22	-22			

the test-blueprint vector. If a topic was not included on the field-trial instrument but was included in the blueprint, a 1 was entered in the corresponding cell of the blueprint vector. A 0 indicated correspondence between the field-trial instrument and the test blueprint (i.e., the topic was either on both the field-trial instrument and the test blueprint or the topic was off of both).

The proportion of the test blueprints that corresponded with the field-trial instrument in inclusion (or non-inclusion) of each topic (i.e., zeros) ranged from .36 to .91, The proportion of topics that either were included in a test blueprint and included on the field-trial instrument or not included on both ranged from .36 to .80. The topics with the lowest correspondence between the field-trial instrument and the test blueprints in topic inclusion were 1.1.1.1 *Whole Number Meanings*, 1.1.3.2 *Real Numbers*, 1.1.5.1 *Estimating Quantity & Size, 1.3.5 Vectors*, and 1.7.2 *Uncertainty and Probability.* Each of these topics was tested on the field-trial instrument but was not included in 7 of the 11 test blueprints. Those with the highest correspondence in topic inclusion were 1.3.2 *Basic 2D Geometry*, 1.6.2 *Equations and Formulas*, and 1.8.2 *Change*. Topics 1.3.2 and 1.6.2 were tested on the field-trial instrument and were included in all but one test blueprint each, and topic 1.8.2 was not included on the field-trial instrument and was only included in one test blueprint. The lowest correspondence of topic inclusion between the field-trial instrument and test blueprints was between the field-trial instrument and the expert-mapping strict-intersection blueprint, and the best correspondence of topic inclusion was between the field-trial instrument and the expert-mapping 70%-intersection blueprint.

Table 23

Differences in Topic Emphasis between the Field-Trial Instrument and Each Test Blueprint

CODE	EX-UN	EX-7I	EX-SI	CG-UN	CG-7I	CG-SI	TX-UN	TX-7I	TX-SI	AG-UN	AG-7I	AVE	*SD*	Ave. Pos. Dif.	Ave. Neg Dif.
1.1.1.1	0.00	-0.02	-0.02	0.01	-0.02	-0.02	-0.01	-0.02	-0.02	-0.01	-0.02	-0.01	0.01	0.007	-0.014
1.1.1.2	-0.04	-0.06	-0.06	-0.03	-0.02	-0.06	-0.03	-0.02	-0.06	-0.03	-0.06	-0.04	0.02	0	-0.042
1.1.1.3	0.01	-0.01	-0.01	0.01	-0.01	-0.01	0.01	0.01	-0.01	0.01	-0.01	0.00	0.01	0.010	-0.008
1.1.2.1	-0.11	-0.10	-0.14	-0.12	-0.14	-0.14	-0.11	-0.10	-0.14	-0.12	-0.14	-0.12	0.02	0	-0.124
1.1.2.2	-0.04	-0.04	-0.07	-0.05	-0.07	-0.07	-0.05	-0.05	-0.07	-0.05	-0.07	-0.06	0.01	0	-0.057
1.1.2.3	-0.02	-0.01	-0.05	-0.02	-0.05	-0.05	-0.04	-0.03	-0.05	-0.03	-0.05	-0.04	0.01	0	-0.035
1.1.2.4	0.00	0.00	-0.03	-0.01	-0.03	-0.03	0.00	0.00	-0.03	0.00	-0.03	-0.01	0.01	0.004	-0.018
1.1.2.5	0.03	0.03	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.01	0.015	0
1.1.3.1	0.02	0.03	-0.01	0.02	0.04	-0.01	0.02	0.03	-0.01	0.02	0.06	0.02	0.02	0.029	-0.012
1.1.3.2	0.03	0.04	0.00	0.03	0.04	0.00	0.02	0.03	0.00	0.03	0.00	0.02	0.02	0.031	0
1.1.3.3	0.02	0.00	0.00	0.03	0.04	0.00	0.02	0.00	0.00	0.02	0.00	0.01	0.01	0.025	0
1.1.4.1	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.004	0
1.1.4.2	0.02	0.03	-0.01	0.01	-0.01	-0.01	0.02	0.03	-0.01	0.01	-0.01	0.01	0.02	0.021	-0.012
1.1.4.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.002	0
1.1.4.4	0.02	0.03	0.00	0.02	0.03	0.00	0.01	0.00	0.00	0.01	0.00	0.01	0.01	0.021	-0.004
1.1.4.5	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.004	0
1.1.5.1	-0.02	-0.04	-0.04	-0.03	-0.04	-0.04	-0.04	-0.04	-0.04	-0.03	-0.04	-0.03	0.00	0	-0.034
1.1.5.2	0.00	0.00	-0.03	-0.01	-0.03	-0.03	-0.03	-0.03	-0.03	-0.02	-0.03	-0.02	0.01	0.004	-0.027
1.1.5.3	-0.01	0.00	-0.03	-0.01	-0.03	-0.03	-0.02	-0.02	-0.03	-0.02	-0.03	-0.02	0.01	0.000	-0.023
1.1.5.4	0.02	0.03	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.013	-0.004
1.2.1	-0.05	-0.04	-0.07	-0.05	-0.03	-0.07	-0.04	-0.04	-0.07	-0.04	-0.07	-0.05	0.02	0	-0.053
1.2.2	-0.04	-0.03	-0.07	-0.04	-0.02	-0.07	-0.01	0.00	-0.07	-0.02	0.04	-0.03	0.03	0.020	-0.040
1.2.3	0.01	-0.01	-0.01	0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	0.01	0.008	-0.012
1.3.1	0.00	0.01	-0.02	0.00	0.01	-0.02	0.00	0.01	-0.02	0.00	-0.02	-0.01	0.01	0.007	-0.017
1.3.2	0.01	0.02	0.97	0.00	0.02	-0.03	0.01	0.02	0.14	0.01	0.07	0.11	0.27	0.128	-0.029
1.3.3	0.00	0.01	-0.03	0.01	0.02	-0.03	0.04	0.06	-0.03	0.03	0.12	0.02	0.04	0.037	-0.033
1.3.4	0.02	0.03	-0.02	0.02	0.05	0.48	0.04	0.05	-0.02	0.03	-0.02	0.06	0.14	0.089	-0.017
1.3.5	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.006	0
1.4.1	-0.01	0.00	-0.04	-0.01	0.01	-0.04	0.00	0.01	-0.04	0.01	-0.04	-0.01	0.02	0.008	-0.026
1.4.2	-0.03	-0.02	-0.06	-0.03	-0.01	-0.06	-0.03	-0.06	-0.06	-0.02	-0.06	-0.04	0.02	0	-0.038
1.4.3	0.02	0.03	0.00	0.02	0.04	0.00	0.01	0.00	0.00	0.01	0.00	0.01	0.01	0.022	0
1.5.1	0.00	0.00	-0.03	0.00	0.01	-0.03	-0.03	-0.03	-0.03	-0.02	-0.03	-0.02	0.02	0.009	-0.024
1.5.2	-0.05	-0.04	-0.10	-0.06	-0.05	-0.10	-0.08	-0.08	-0.10	-0.07	-0.10	-0.07	0.02	0	-0.074
1.5.3	0.02	0.00	0.00	0.02	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.01	0.01	0.016	0
1.5.4	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.008	0
1.6.1	-0.02	-0.01	-0.05	-0.01	0.01	-0.05	0.00	0.01	0.14	0.01	0.09	0.01	0.05	0.050	-0.024
1.6.2	-0.10	-0.09	-0.14	-0.10	-0.07	0.36	0.02	0.06	0.50	0.00	0.17	0.06	0.20	0.225	-0.082
1.7.1	-0.07	-0.06	-0.11	-0.08	-0.06	-0.11	-0.07	-0.07	-0.11	-0.06	0.01	-0.07	0.03	0.007	-0.081
1.7.2	-0.03	-0.05	-0.05	-0.02	-0.01	-0.05	-0.04	-0.05	-0.05	-0.04	-0.05	-0.04	0.01	0	-0.038
1.8.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.002	0
1.8.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.005	0
1.9.1	0.01	0.00	0.00	0.01	0.00	0.00	0.02	0.00	0.00	0.01	0.00	0.00	0.01	0.011	0
1.9.2	0.01	0.00	0.00	0.02	0.00	0.00	0.02	0.00	0.00	0.01	0.00	0.01	0.01	0.014	0
1.10.1	0.02	0.00	0.00	0.03	0.05	0.00	0.03	0.00	0.00	0.02	0.00	0.01	0.02	0.029	0
SD	0.03	0.03	0.15	0.03	0.04	0.10	0.03	0.03	0.09	0.03	0.05				
Ave.Pos.Dif.	0.01	0.02	0.97	0.01	0.03	0.42	0.01	0.02	0.26	0.01	0.08				
Ave.Neg.Dif.	-0.03	-0.04	-0.05	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.03	-0.04				

Table 23 shows the differences between the topic weights in each test blueprint and the topic weights on the field-trial instrument. Positive differences occurred when topics received a higher weight in the blueprint than on the field-trial instrument, and negative differences occurred when topics received a higher weight on the field-trial instrument than in the blueprint. Tables show standard deviations of absolute weight differences for each topic and each test blueprint as well as the averages of the positive and negative weight differences for each topic and each test blueprint.

Looking within blueprints across topics, average positive weight differences (more weight in the test blueprint than on the field-trial instrument) ranged from .01 to .97. The largest average difference for topics emphasized more in the blueprint than in the field-trial instrument (positive) was for the expert-mapping strict-intersection blueprint. Strict-intersection blueprints had the largest positive weight differences (because they included few topics which received much weight), and they had the largest standard deviations of weight differences. All union and 70%-intersection blueprints had average positive weight differences of .08 or less, the aggregate70%-intersection blueprint having the largest difference. Average negative differences (topics emphasized more in the field-trial instrument) were between -.03 and -05.

On average across all test blueprints, nine topics had 0 as an average positive difference meaning that the topic did not receive more weight in any of the test blueprints than on the field-trial instrument. Topics with the largest average positive weight difference were 1.3.2 *Basic 2D Geometry* (.128) and 1.6.2 *Equations and Formulas* (.225). The 15 topics not included on the field-trial instrument had negative weight differences of 0. The topics with the highest average negative weight differences (more

weight on the field-trial instrument than on the test blueprint) were 1.1.2.1 *Common Fractions* (.124), 1.6.2 *Equations and Formulas* (.09), and 1.7.1 *Data Representation and Analysis* (.081). On average, topic 1.1.2.1 received over 10% more weight on the field-trial instrument than in the blueprints; topics 1.6.2 and 1.7.1 *Uncertainty and Probability* received around 8% to 9% more weight. However, topic 1.6.2 had a higher positive difference (.225) meaning that, overall, it received more weight on the test blueprints than on the field-trial instrument.

Correlations and Euclidean distance measures. Correlations between the topic weight patterns (profiles) on each test blueprint and the topic weight patterns (profiles) on the field-trial instrument are in Table 24.

Table 24

Correlations and Euclidean Distances between the Topic-Weight Profiles for Each Test Blueprint and the Topic-Weight Profiles for the Field-Trial Instrument

	Correlation	Euclidean Distance
EX-UN	0.590**	0.213
EX-7I	0.560**	0.208
EX-SI	-0.005	1.020
CG-UN	0.454**	0.226
CG-7I	0.374*	0.242
CG-SI	0.277	0.668
TX-UN	0.573**	0.220
TX-7I	0.603**	0.220
TX-SI	0.439**	0.609
AG-UN	0.635**	0.194
AG-7I	0.502**	0.349
Ave	0.45	0.38

*$p<.05$. **$p<.01$.

The correlations ranged from -.005 (expert-mapping strict-intersection blueprint) to .63 (aggregate-union blueprint). The overall average was .45. In most cases, higher correlations were between the field-trial instrument and the union blueprints. The exception was for the textbooks where the largest correlation was between the field-trial instrument topic-weight profiles and the 70%-intersection blueprint topic-weight profiles.

Euclidean distances between the topic weights on each test blueprint and the topic weights on the field-trial instrument are also shown in Table 24. The distances ranged from .194 (aggregate-union blueprint) to 1.02 (expert-mapping strict-intersection blueprint), with an overall average of .38. The smallest distances were between the field-trial instrument and the union blueprints, except for the expert mapping.

Re-Specification of Test Blueprints

I had intended to re-compute country scores on the field-trial instrument according to each test blueprint previously discussed. However, the field-trial instrument did not contain items for every topic of the framework, so I could not obtain country scores for all topics. I, therefore, had to re-write each test blueprint using only the topics included on the field-trial instrument. Again, I weighted topics according to averages in proportions of emphasis across countries. I also wrote blueprints for unweighted tests in which I gave each topic included in a blueprint equal weight. I then compared the correspondence in topic coverage between the new *weighted*-test blueprints and each corresponding curriculum source for each country.

Table 25 provides a summary of the union, 70%-intersection, and strict-intersection topic weights after removing topics not included on the field-trial instrument.

Table 25

Topic Weights on Specially-Constructed Test Blueprints

Topic Code	Test											
	EX1-UN	EX1-7I	EX1-SI	CG-UN	CG-7I	CG-SI	TX-UN	TX-7I	TX-SI	AG-UN	AG-7I	Average
1.1.1.1	0.015	0	0	0.030	0	0	0.013	0	0	0.007	0	0.006
1.1.1.2	0.021	0	0	0.035	0.042	0	0.036	0.0391	0	0.027	0	0.018
1.1.1.3	0.021	0	0	0.026	0	0	0.019	0.0205	0	0.018	0	0.009
1.1.2.1	0.039	0.044	0	0.022	0	0	0.036	0.0401	0	0.028	0	0.019
1.1.2.2	0.034	0.039	0	0.026	0	0	0.021	0.0232	0	0.029	0	0.016
1.1.2.3	0.030	0.034	0	0.027	0	0	0.012	0.013	0	0.019	0	0.012
1.1.2.4	0.031	0.035	0	0.028	0	0	0.032	0.0347	0	0.027	0	0.017
1.1.2.5	0	0	0	0	0	0	0	0	0	0	0	0.000
1.1.3.1	0.039	0.044	0	0.043	0.050	0	0.037	0.0407	0	0.036	0.073	0.033
1.1.3.2	0	0	0	0	0	0	0	0	0	0	0	0.004
1.1.3.3	0	0	0	0	0	0	0	0	0	0	0	0.004
1.1.4.1	0	0	0	0	0	0	0	0	0	0	0	0.000
1.1.4.2	0.044	0.050	0	0.029	0	0	0.037	0.0403	0	0.030	0	0.021
1.1.4.3	0	0	0	0	0	0	0	0	0	0	0	0.000
1.1.4.4	0.030	0.034	0	0.033	0.039	0	0.015	0	0	0.019	0	0.015
1.1.4.5	0	0	0	0	0	0	0	0	0	0	0	0.000
1.1.5.1	0.017	0	0	0.015	0	0	0.002	0	0	0.006	0	0.004
1.1.5.2	0.037	0.042	0	0.025	0	0	0.006	0	0	0.012	0	0.011
1.1.5.3	0.029	0.032	0	0.019	0	0	0.007	0.0077	0	0.013	0	0.010
1.1.5.4	0.033	0.038	0	0.020	0	0	0.006	0	0	0.006	0	0.009
1.2.1	0.035	0.040	0	0.035	0.041	0	0.036	0.0391	0	0.037	0	0.024
1.2.2	0.035	0.039	0	0.040	0.047	0	0.063	0.0696	0	0.051	0.104	0.041
1.2.3	0.022	0	0	0.030	0	0	0.002	0	0	0.006	0	0.005
1.3.1	0.036	0.040	0	0.031	0.037	0	0.030	0.033	0	0.024	0	0.021
1.3.2	0.047	0.054	1.00	0.039	0.046	0	0.049	0.0537	0.642	0.049	0.102	0.189
1.3.3	0.042	0.047	0	0.049	0.058	0	0.087	0.0961	0	0.074	0.153	0.055
1.3.4	0.041	0.046	0	0.053	0.063	0.500	0.061	0.0669	0	0.053	0	0.080
1.3.5	0	0	0	0	0	0	0	0	0	0	0	0.000
1.4.1	0.040	0.045	0	0.044	0.052	0	0.050	0.0549	0	0.056	0	0.031
1.4.2	0.038	0.043	0	0.041	0.049	0	0.036	0	0	0.042	0	0.023
1.4.3	0		0	0	0	0	0	0	0	0	0	0.004
1.5.1	0.037	0.041	0	0.039	0.046	0	0.008	0	0	0.019	0	0.017
1.5.2	0.051	0.057	0	0.042	0.049	0	0.018	0.020	0	0.031	0	0.024
1.5.3	0	0	0	0	0	0	0	0	0	0	0	0.000
1.5.4	0	0	0	0	0	0	0	0	0	0	0	0.000
1.6.1	0.039	0.044	0	0.049	0.058	0	0.053	0.0588	0.187	0.067	0.138	0.063
1.6.2	0.051	0.057	0	0.053	0.063	0.500	0.183	0.2015	0.171	0.151	0.312	0.158
1.7.1	0.048	0.054	0	0.044	0.052	0	0.043	0.0472	0	0.058	0.119	0.042
1.7.2	0.019	0	0	0.033	0.038	0	0.003	0	0	0.007	0	0.009
1.8.1	0	0	0	0	0	0	0	0	0	0	0	0.000
1.8.2	0	0	0	0	0	0	0	0	0	0	0	0.000
1.9.1	0	0	0	0	0	0	0	0	0	0	0	0.000
1.9.2	0	0	0	0	0	0	0	0	0	0	0	0.000
1.10.1	0	0	0	0	0	0	0	0	0	0	0	0.004

These proportions were scaled to sum to one across topics. Overall, the highest weights were given to topic 1.3.2 *Basic Geometry* and topic 1.6.2 *Equations and Formulas.* Table 26 provides an overview of the blueprints on which I compared country-level performance.

Comparisons of Curriculum to Unique Specially-Constructed-Test Blueprints

The first sets of comparisons I conducted were between the unique specially-constructed- (SC) test blueprints developed for each country and each corresponding curriculum-data source. That was, I compared each unique SC-test blueprint based on the expert mapping to the corresponding country's expert-mapping data, I compared each unique SC-test blueprint based on the curriculum-guide analyses to the corresponding country's curriculum-guide data, and so forth. This provided an indication of the best possible match that could occur between any test developed using the field-trial instrument topics and each country's corresponding data source.

I conducted much the same analyses as before, but adapted them as needed to fit the particular situation. I did not compute the proportion of items in each unique SC-test blueprint that were in each country's curricula since this would naturally be 100%. I likewise did not compute the proportion of topics in each country's curriculum that was included on the unique SC-test blueprints. No additional topics were included in the unique SC-test blueprints than were included on the field-trial instrument. Therefore, the proportions of curricula tested would be the same proportions as reported in Table 11.

Differences in topic inclusion and emphasis. Table 27 shows summaries of differences in topic inclusion between each country's curriculum data source and the

Table 26

Codes for Specially-Constructed-Test Blueprints

	Curriculum Source			
Method of Specification	Expert Mapping	Curriculum Guides	Textbooks	Aggregate Intersection
Union of Topics across all Countries	WEX1-UN[a]	WCG-Un[a]	WTX-UN[a]	WAG-UN[a]
Intersection of Topics in 70% of Countries	(W)EX1-7I	(W)CG-7I	(W)TX-7I	(W)AG-7I
Intersection of Topics in all Countries	EX1-Si[b]	CG-Si[c]	(W)TX-SI	[d]
Unique Topic Coverage	(W)EX1-UQ	CG-Uq[e]	(W)TX-UQ	(W)AG-UQ

NOTE. A"W" is used before a test to identify a weighted test.

[a] All union test blueprints consisted of the same topics. Therefore, only one unweighted union test was developed, simply labeled UNION [b] Only one topic existed on the test blueprint for EX-SI. Therefore, performance on the weighted test would be identical to performance on the unweighted test. [c] The two topics in the blueprint for CG-SI both received equal weight. Therefore, performance on the weighted test would be identical to performance on the unweighted test. [d] No topics existed in the test blueprint for the strict intersection of the aggregate data source. [e] Within a country, all topics in the curriculum guides receive equal weight. Therefore, performance on the weighted tests would be identical to performance on the unweighted tests.

corresponding unique SC-test blueprint only for topics not included on the field-trial instrument. These differences are identical to those in Table 12. All other differences were 0 since topics not in a country's curricula would not be included on its unique SC-test blueprint. What should be noted, however, are the numbers in the final row of the table. These can be compared to future test blueprints to determine if there is an improvement in test-curriculum match. An ideal match would result in all differences being 0 and proportions of match being 1.0. The topic inclusion on the test blueprints did not correspond exactly with the curricula because not all topics were included on the test blueprints. The inclusive test blueprints I develop, will not have lower differences or higher matches than these; the goal will be to come as close to these as possible.

Table 28 shows the summary information on the correspondence in topic inclusion between the test blueprints and the curriculum for each country. The numbers in the column "In Curr." are identical to those in Table 13. "Prop. Match" is the proportion of topics within a country's curriculum-data source that are included on the corresponding unique SC-test blueprints. Again, it is the best match expected given the topics on the field-trial instrument. Country agreement in topic inclusion with the field-trial instrument ranged from .66 (country Q, curriculum guide) to 1.00 (countries E and O, aggregate-data source). Averages of the proportions of topics both in the curriculum and the field-trial instrument or not in both were around .90 with the exception of country Q (.74). The average number of countries including topics in a particular data source that were not in the corresponding unique SC-test blueprint ranged from around three to six, with the lowest number being for the aggregate-data source and the highest number being for the textbooks.

Table 27

Numbers and Proportions of Countries Including Topics in Curriculum Sources that are not on Corresponding Unique-Test Blueprints

	Expert Mapping		Curriculum Guide		Textbook		Aggregate			
Topic Code	# Mis-Match	Prop. Match	# Mis-Match	Prop. Match	# Mis-Match	Prop. Match	# Mis-Match	Prop. Match	Ave # Mis-Match	Ave. Prop. Match
1.1.2.5	12	0.29	5	0.71	11	0.35	4	0.76	8	0.53
1.1.3.2	14	0.18	13	0.24	12	0.29	8	0.53	11.8	0.31
1.1.3.3	11	0.35	12	0.29	11	0.35	6	0.65	10	0.41
1.1.4.1	2	0.88	5	0.71	4	0.76	0	1	2.75	0.84
1.1.4.3	0	1	2	0.88	3	0.82	0	1	1.25	0.93
1.1.4.5	2	0.88	4	0.76	7	0.59	1	0.94	3.5	0.79
1.3.5	1	0.94	9	0.47	7	0.59	1	0.94	4.5	0.74
1.4.3	13	0.24	12	0.29	9	0.47	5	0.71	9.75	0.43
1.5.3	10	0.41	8	0.53	6	0.65	3	0.82	6.75	0.60
1.5.4	6	0.65	6	0.65	4	0.76	0	1	4	0.76
1.8.1	0	1	2	0.88	4	0.76	0	1	1.5	0.91
1.8.2	0	1	3	0.82	0	1	0	1	0.75	0.96
1.9.1	7	0.59	4	0.76	10	0.41	3	0.82	6	0.65
1.9.2	8	0.53	7	0.59	9	0.47	5	0.71	7.25	0.57
1.10.1	10	0.41	14	0.18	11	0.35	8	0.53	10.8	0.37
Average	4.64	0.73	5.52	0.68	5.16	0.64	6.20	0.36	5.38	0.60

Tables 29 and 30 show the differences between curriculum-data sources and each corresponding unique SC-test blueprint in topic emphasis. Table 29 highlights differences across topics, and Table 30 highlights differences across countries. Positive differences occurred when topics received a higher emphasis in the curriculum than on the corresponding test blueprint, and negative differences occurred when topics received a higher emphasis on the test blueprint than in the curriculum. Tables show standard deviations of absolute emphasis differences for each topic and country as well as the averages of the positive and negative differences for each topic (Table 29) and each country (Table 30). The table also shows averages of these numbers across data sources.

Table 28

Numbers and Proportions of Topics in Curriculum Sources that are Included on Corresponding Unique-Test Blueprints

	Expert Mapping		Curriculum Guide		Textbook		Aggregate			
Country	Prop. Match	In Curr.	Prop. Match	In Curr.	Prop. Match	In Curr.	Prop. Match	In Curr.	Ave. Prop. Match	Ave. In Curr.
A	0.86	6	0.91	4	0.84	7	0.98	1	0.90	4.5
B	0.86	6	0.89	5	0.91	4	0.95	2	0.90	4.3
C	0.86	6	0.73	12	0.73	12	0.86	6	0.80	9.0
D	0.82	8	0.86	6	0.82	8	0.91	4	0.85	6.5
E	0.84	7	0.95	2	0.86	6	1.00	0	0.91	3.8
F	0.95	2	0.75	11	0.89	5	0.98	1	0.89	4.8
G	0.93	3	0.95	2	0.86	6	0.98	1	0.93	3.0
H	0.89	5	0.95	2	0.84	7	0.95	2	0.91	4.0
I	0.89	5	0.84	7	0.82	8	0.91	4	0.86	6.0
J	0.86	6	0.84	7	0.95	2	0.98	1	0.91	4.0
K	0.93	3	0.89	5	0.89	5	0.98	1	0.92	3.5
L	0.84	7	0.91	4	0.86	6	0.95	2	0.89	4.8
M	0.84	7	0.82	8	0.84	7	0.91	4	0.85	6.5
N	0.77	10	0.86	6	0.80	9	0.86	6	0.82	7.8
O	0.91	4	0.89	5	0.98	1	1.00	0	0.94	2.5
P	0.98	1	0.89	5	0.93	3	0.98	1	0.94	2.5
Q	0.77	10	0.66	15	0.73	12	0.82	8	0.74	11.3
Average	0.87	5.6	0.86	6.24	0.86	6.35	0.94	2.6	0.88	5.21

Table 29

Differences in Topic Emphasis for Each Topic across Countries on Unique-Test Blueprints and Corresponding Curriculum Sources

	Expert Mapping			Curriculum Guide			Textbook		
Topic Code	*SD* of All Dif.	Ave. Pos. Dif.	Ave. Neg. Dif.	*SD* of All Dif.	Ave. Pos. Dif.	Ave. Neg. Dif.	*SD* of All Dif.	Ave. Pos. Dif.	Ave. Neg. Dif.
1.1.1.1	0.003	0	-0.005	0.006	0	-0.009	0.005	0.004	-0.001
1.1.1.2	0.003	0	-0.005	0.006	0	-0.010	0.005	0.005	-0.003
1.1.1.3	0.003	0	-0.006	0.006	0	-0.010	0.005	0.004	-0.004
1.1.2.1	0.008	0	-0.009	0.005	0	-0.008	0.005	0.005	-0.005
1.1.2.2	0.008	0	-0.008	0.004	0	-0.008	0.004	0.004	-0.001
1.1.2.3	0.008	0	-0.008	0.008	0	-0.012	0.002	0.002	-0.003
1.1.2.4	0.004	0	-0.007	0.005	0	-0.010	0.008	0.003	-0.017
1.1.2.5	0.026	0.035	0	0.013	0.028	0	0.010	0.009	0
1.1.3.1	0.004	0	-0.007	0.007	0	-0.011	0.006	0.004	-0.007
1.1.3.2	0.025	0.037	0	0.018	0.037	0	0.071	0.040	0
1.1.3.3	0.018	0.032	0	0.023	0.039	0	0.064	0.039	0
1.1.4.1	0.008	0.023	0	0.015	0.031	0	0.003	0.004	0
1.1.4.2	0.005	0	-0.009	0.008	0	0	0.013	0.011	-0.003
1.1.4.3	0.000	0	0	0.008	0.024	0	0.001	0.001	0
1.1.4.4	0.005	0	-0.008	0.008	0	-0.011	0.009	0.006	0.000
1.1.4.5	0.011	0.034	0	0.012	0.028	0	0.006	0.006	0
1.1.5.1	0.003	0	-0.005	0.004	0	0	0.001	0.001	0.000
1.1.5.2	0.006	0	-0.008	0.005	0	0	0.001	0.002	-0.001
1.1.5.3	0.004	0	-0.007	0.005	0	0	0.001	0.001	0.000
1.1.5.4	0.005	0	-0.008	0.005	0	0	0.001	0.002	0
1.2.1	0.004	0	-0.007	0.005	0	0	0.005	0.004	-0.008
1.2.2	0.004	0	-0.007	0.008	0	0	0.013	0.016	-0.004
1.2.3	0.005	0	-0.007	0.006	0	0	0.001	0.001	0
1.3.1	0.006	0	-0.008	0.005	0	-0.010	0.007	0.006	-0.012
1.3.2	0.005	0	-0.009	0.007	0	0	0.010	0.009	-0.003
1.3.3	0.006	0	-0.008	0.007	0	-0.012	0.022	0.023	-0.008
1.3.4	0.005	0	-0.009	0.007	0	-0.012	0.053	0.024	-0.040
1.3.5	0.007	0.028	0	0.017	0.033	0	0.013	0.013	0
1.4.1	0.006	0	-0.008	0.005	0	-0.011	0.011	0.014	-0.007
1.4.2	0.007	0	-0.010	0.008	0	-0.012	0.012	0.012	-0.002
1.4.3	0.018	0.031	0	0.017	0.035	0	0.012	0.016	0
1.5.1	0.004	0	-0.007	0.005	0	-0.010	0.001	0.002	-0.002
1.5.2	0.007	0	-0.010	0.005	0	-0.010	0.004	0.004	-0.001
1.5.3	0.022	0.035	0	0.027	0.045	0	0.025	0.041	0
1.5.4	0.017	0.031	0	0.018	0.036	0	0.004	0.007	0
1.6.1	0.007	0	-0.010	0.007	0	-0.012	0.024	0.015	-0.004
1.6.2	0.007	0	-0.011	0.007	0	-0.012	0.043	0.039	-0.013
1.7.1	0.007	0	-0.010	0.005	0	-0.011	0.005	0.006	-0.004
1.7.2	0.005	0	-0.007	0.005	0	-0.010	0.000	0.001	-0.001
1.8.1	0.000	0	0	0.012	0.034	0	0.001	0.003	0
1.8.2	0.000	0	0	0.011	0.028	0	0	0	0
1.9.1	0.018	0.031	0	0.016	0.034	0	0.072	0.037	0
1.9.2	0.014	0.027	0	0.020	0.036	0	0.034	0.040	0
1.10.1	0.015	0.028	0	0.017	0.038	0	0.062	0.056	0
Average	0.008	0.008	-0.005	0.010	0.011	-0.007	0.015	0.012	-0.004

Table 29 (Contd.)

	Aggregate							
Topic Code	*SD* of All Dif.	Ave. Pos. Dif.	Ave. Neg. Dif.	Ave. of Pos. Dif.	Ave.of Neg. Dif.	*SD* of Pos. Dif.	*SD* of Neg. Dif.	*SD* of All Dif.
1.1.1.1	0.002	0	-0.006	0.001	-0.005	0.002	0.003	0.004
1.1.1.2	0.003	0	-0.005	0.001	-0.006	0.002	0.003	0.004
1.1.1.3	0.002	0	-0.004	0.001	-0.006	0.002	0.002	0.004
1.1.2.1	0.007	0	-0.007	0	-0.007	0.002	0.001	0.005
1.1.2.2	0.014	0	-0.010	0.001	-0.007	0.002	0.004	0.005
1.1.2.3	0.003	0	-0.004	0.000	-0.007	0.001	0.003	0.004
1.1.2.4	0.004	0	-0.006	0.001	-0.010	0.001	0.004	0.006
1.1.2.5	0.014	0.031	0	0.026	0	0.010	0.000	0.015
1.1.3.1	0.004	0	-0.005	0.001	-0.008	0.002	0.002	0.005
1.1.3.2	0.076	0.068	0	0.045	0	0.013	0.000	0.025
1.1.3.3	0.028	0.048	0	0.040	0	0.006	0.000	0.020
1.1.4.1	0	0	0	0.014	0	0.013	0.000	0.012
1.1.4.2	0.006	0.000	-0.008	0.003	-0.008	0.005	0.003	0.007
1.1.4.3	0	0	0	0.006	0	0.010	0.000	0.008
1.1.4.4	0.005	0	-0.006	0.002	-0.006	0.003	0.004	0.005
1.1.4.5	0.007	0.031	0	0.025	0	0.011	0.000	0.015
1.1.5.1	0.002	0	-0.003	0.000	-0.004	0.000	0.003	0.003
1.1.5.2	0.002	0	-0.003	0.000	-0.005	0.001	0.004	0.004
1.1.5.3	0.002	0	-0.004	0.000	-0.005	0.001	0.003	0.003
1.1.5.4	0.002	0	-0.004	0.000	-0.005	0.001	0.004	0.004
1.2.1	0.003	0.000	-0.005	0.001	-0.007	0.002	0.002	0.005
1.2.2	0.006	0.000	-0.007	0.004	-0.008	0.007	0.003	0.008
1.2.3	0.001	0	-0.003	0.000	-0.005	0.000	0.004	0.004
1.3.1	0.004	0.000	-0.005	0.001	-0.009	0.002	0.002	0.006
1.3.2	0.009	0.000	-0.009	0.002	-0.008	0.004	0.003	0.006
1.3.3	0.016	0.000	-0.014	0.006	-0.011	0.010	0.003	0.011
1.3.4	0.018	0.000	-0.017	0.006	-0.020	0.011	0.012	0.017
1.3.5	0.009	0.037	0	0.028	0	0.009	0.000	0.015
1.4.1	0.030	0.000	-0.021	0.003	-0.012	0.006	0.005	0.009
1.4.2	0.017	0.000	-0.016	0.003	-0.010	0.005	0.005	0.008
1.4.3	0.013	0.027	0	0.027	0	0.007	0.000	0.014
1.5.1	0.002	0	-0.004	0.000	-0.006	0.001	0.003	0.004
1.5.2	0.003	0.000	-0.004	0	-0.006	0.002	0.004	0.005
1.5.3	0.028	0.054	0	0.044	0	0.007	0.000	0.022
1.5.4	0	0	0	0.018	0	0.015	0.000	0.014
1.6.1	0.021	0	-0.017	0.004	-0.010	0.007	0.005	0.009
1.6.2	0.043	0.000	-0.033	0.010	-0.017	0.017	0.009	0.019
1.7.1	0.014	0	-0.011	0	-0.009	0.003	0.003	0.006
1.7.2	0.002	0	-0.003	0.000	-0.005	0.000	0.003	0.004
1.8.1	0	0	0	0.009	0	0.014	0.000	0.011
1.8.2	0	0	0	0.007	0	0.012	0.000	0.009
1.9.1	0.024	0.047	0	0.037	0	0.006	0.000	0.019
1.9.2	0.019	0.037	0	0.035	0	0.005	0.000	0.018
1.10.1	0.032	0.046	0	0.042	0	0.010	0.000	0.022
Average	0.011	0.010	-0.006	0.010	-0.005	0.006	0.002	0.010

As was the case earlier, three topics (1.1.2.1 *Common Fractions*, 1.5.2 *Proportionality Problems*, 1.7.1 *Data Representation and Analysis*) had 0 as an average positive difference meaning that the topic was not more emphasized in the curriculum of any country than in the test blueprint (see Table 29). Topics with the highest positive differences were 1.1.3.2 *Rational Numbers* (.045); 1.1.3.3 *Real Numbers* (.04); 1.5.3 *Slope and Trigonometry* (.044); and 1.10 *Other Content* (.042).

The only topics with 0 as an average negative difference in emphasis were those averages for topics not included on the test blueprint. Topics with the highest negative differences were 1.3.4 *3-D Geometry* and 1.6.2 *Equations and Formulas*. Their average were -.02 and -.017 respectively. The average of the positive average differences was .01 while the average of the negative average differences was -.005. These numbers were much smaller than in Table 14. As was the case earlier, the largest differences were between topic emphasis in the textbook-data source and topic weight on the corresponding unique SC-test blueprints. Column sums and averages will be compared with those in future analyses.

Table 30 compares variability in topic emphasis differences within countries. Across data sources, standard deviations and means were similar. Averages across country differences in textbook emphasis versus emphasis in the unique SC-test blueprints based on the textbooks were smaller than the differences for other data sources and corresponding tests. Differences were largest between topic emphasis in the aggregate of the data sources and topic weight on the corresponding unique SC-test blueprint. Average positive differences in emphasis (more weight in curriculum-data source) within countries across topics ranged from 0 (countries E and O, aggregate) to

Table 30

Differences in Topic Emphasis for Each Country across Topics on Unique-Test Blueprintss and Corresponding Curriculum Sources

	Expert Mapping Age 13			Curriculum Guide			Textbook			Aggregate							
Country	*SD* of All Dif.	Ave. Pos. Dif.	Ave. Neg. Dif.	*SD* of All Dif.	Ave. Pos. Dif.	Ave. Neg. Dif.	*SD* of All Dif.	Ave. Pos. Dif.	Ave. Neg. Dif.	*SD* of All Dif.	Ave. Pos. Dif.	Ave. Neg. Dif.	Ave. Pos. Dif.	Ave. Neg. Dif.	*SD* of Pos. Dif.	*SD* of Neg. Dif.	*SD* of All Dif.
A	0.010	0.025	-0.006	0.011	0.034	-0.006	0.004	0.002	0	0.004	0.028	-0.002	0.022	-0.003	0.012	0.002	0.016
B	0.015	0.036	-0.009	0.016	0.042	-0.011	0.037	0.058	-0.010	0.022	0.083	-0.010	0.055	-0.010	0.018	0.001	0.035
C	0.016	0.036	-0.008	0.016	0.026	-0.011	0.011	0.012	0	0.012	0.028	-0.007	0.025	-0.007	0.009	0.004	0.017
D	0.010	0.020	-0.005	0.012	0.029	-0.006	0.013	0.006	0	0.009	0.026	-0.004	0.020	-0.004	0.009	0.002	0.014
E	0.022	0.043	-0.018	0.010	0.045	-0.005	0.006	0.016	-0.002	0.030	0	-0.074	0.026	-0.025	0.019	0.029	0.035
F	0.006	0.027	-0.002	0.017	0.028	-0.012	0.024	0.036	-0.013	0.005	0.030	-0.002	0.030	-0.007	0.004	0.005	0.019
G	0.027	0.088	-0.024	0.013	0.056	-0.007	0.045	0.016	0	0.057	0.327	-0.082	0.122	-0.028	0.121	0.032	0.116
H	0.013	0.034	-0.008	0.021	0.091	-0.020	0.009	0.005	0	0.024	0.095	-0.032	0.056	-0.015	0.038	0.012	0.045
I	0.009	0.024	-0.005	0.015	0.032	-0.009	0.019	0.014	0	0.012	0.034	-0.007	0.026	-0.005	0.008	0.003	0.017
J	0.020	0.043	-0.020	0.019	0.040	-0.016	0.001	0.005	-0.001	0.007	0.043	-0.005	0.033	-0.010	0.016	0.008	0.025
K	0.011	0.036	-0.005	0.019	0.050	-0.017	0.011	0.009	0	0.011	0.067	-0.007	0.040	-0.007	0.021	0.006	0.028
L	0.013	0.028	-0.007	0.011	0.033	-0.005	0.004	0.003	0	0.007	0.032	-0.003	0.024	-0.004	0.012	0.003	0.016
M	0.015	0.032	-0.011	0.016	0.032	-0.011	0.020	0.016	0	0.010	0.029	-0.006	0.027	-0.007	0.007	0.005	0.018
N	0.013	0.023	-0.008	0.027	0.059	-0.032	0.072	0.052	0	0.029	0.055	-0.030	0.047	-0.018	0.014	0.014	0.035
O	0.012	0.034	-0.008	0.018	0.048	-0.015	0.015	0.013	0	0	0	0	0.024	-0.006	0.018	0.006	0.020
P	0.011	0.071	-0.004	0.013	0.034	-0.007	0.002	0.003	-0.001	0.007	0.046	-0.003	0.039	-0.004	0.024	0.002	0.027
Q	0.013	0.022	-0.009	0.016	0.023	-0.012	0.014	0.007	0	0.014	0.026	-0.009	0.019	-0.007	0.008	0.004	0.015
Average	0.014	0.037	-0.009	0.016	0.041	-0.012	0.018	0.016	-0.002	0.015	0.056	-0.016	0.037	-0.010	0.021	0.008	0.029

.327 (country G, aggregate), with an average of around .04. Average negative differences in emphasis (more weight in unique SC-test blueprint) ranged from 0 (Country O, aggregate; most countries on textbooks) to -.08 (Country G, aggregate), with an average of -.01. Smaller overall positive and negative differences were noted for countries A and D, and larger differences in emphasis were noted for countries G and H.

Correlations and Euclidean distance measures. Correlations between the proportions-of-topic-emphasis profiles in each curriculum-data source and topic-weight profiles on each corresponding unique SC-test blueprint are in Table 31. The correlations ranged from 0 (Country Q, curriculum guide) to 1.00 (countries J and P, Textbooks; countries E and O, Aggregate) with an overall average of .84. The average correlation between topic emphasis profiles and field-trial topic profiles within data sources across countries was highest for the textbook- and aggregate-data-source topic-emphasis-profiles and the corresponding unique SC-test blueprints. The lowest was between the curriculum-guide-data-source topic-emphasis-profiles and the corresponding test-blueprint topic-weight profiles. However, this data source for each country consisted of either a 0 or a proportion which was always the same proportion across all topics included in a country's curriculum-guide-data source. Average correlations for countries across data sources varied. These ranged from a low of .655 (country Q) to a high of .925 (country O).

Euclidean distances between each country's topic emphasis profiles and topic weight profiles on the corresponding test blueprint are also shown in Table 31. The distances ranged from 0 (countries E and O, aggregate) to .54 (country N, textbooks), with an overall average of .10. The largest average distance was found between the

textbook-data-source topic-emphasis profiles and the corresponding unique SC-test blueprint topic weight profiles. The smallest was between the aggregate of the data sources and the corresponding unique SC-test blueprint. The smallest average distances for a country were for countries A (.049), L (.058), and P (.055). The largest were for countries N (.249) and G (.237). Average standard deviations of the Euclidean distances across countries (within curriculum-data sources) were between .24 and .34.

Table 31

Correlations and Euclidean Distances between The Proportions-of-Topic-Emphasis Profiles for Each Country in Each Curriculum-Data Source and the Topic-Weight Profiles for Each Corresponding Unique-Test Blueprint

	Correlations					Euclidean Distance				
Country	EX-UQ	CG-UQ	TX-UQ	AG-UQ	Average	EX-UQ	CG-UQ	TX-UQ	AG-UQ	Average
A	0.873	0.825	0.994	0.989	0.920	0.068	0.074	0.026	0.029	0.049
B	0.727	0.796	0.811	0.821	0.789	0.103	0.105	0.243	0.143	0.148
C	0.668	0.451	0.969	0.872	0.740	0.104	0.107	0.101	0.079	0.097
D	0.881	0.705	0.870	0.875	0.833	0.063	0.077	0.095	0.061	0.074
E	0.723	0.913	0.994	1.000	0.907	0.144	0.067	0.042	0.000	0.063
F	0.961	0.541	0.803	0.990	0.824	0.040	0.111	0.161	0.031	0.086
G	0.776	0.909	0.795	0.749	0.807	0.179	0.083	0.308	0.378	0.237
H	0.842	0.878	0.974	0.912	0.902	0.090	0.142	0.064	0.158	0.113
I	0.892	0.709	0.861	0.869	0.833	0.063	0.097	0.144	0.083	0.097
J	0.836	0.725	1.000	0.993	0.889	0.134	0.125	0.009	0.046	0.078
K	0.892	0.788	0.995	0.979	0.913	0.070	0.129	0.080	0.071	0.088
L	0.746	0.821	0.996	0.958	0.880	0.086	0.072	0.027	0.047	0.058
M	0.832	0.678	0.985	0.958	0.863	0.102	0.106	0.159	0.068	0.109
N	0.619	0.728	0.815	0.793	0.739	0.087	0.179	0.538	0.192	0.249
O	0.931	0.791	0.977	1.000	0.925	0.078	0.122	0.110	0.000	0.077
P	0.917	0.788	1.000	0.976	0.920	0.073	0.085	0.013	0.047	0.055
Q	0.820	0	0.932	0.868	0.655	0.088	0.108	0.101	0.090	0.097
Ave	0.820	0.709	0.928	0.918	0.843	0.092	0.105	0.131	0.090	0.104

Note. All correlations (but country Q, tx-uq) are significant; $p<.01$.

Comparisons of the Curriculum to Inclusive Specially-Constructed-Test Blueprints

I evaluated the correspondence in topic coverage between test blueprints and the curriculum one final time. For these final analyses, I evaluated the correspondence between topic coverage on each specially-constructed union- and 70%-intersection-test blueprint and topic coverage in each country's corresponding curriculum-data source. I did not use any of the strict-intersection-test blueprints in these analyses. These test blueprints were limited in scope, and I would not expect to find a high quantitative match between them and the data sources.

Proportions of items/curricula covered. Table 32 shows the proportions of "items" (i.e., sum of topic weights) on each inclusive SC-test blueprint that measured topics included in each of the corresponding curriculum-data sources for each country. These proportions ranged from .30 to 1.00 (30-100%) with an average of .82. Only two of the averages of the proportions of "items" on the inclusive SC-test blueprints measuring topics included in the corresponding curriculum-data sources were below .80 for any of the test blueprints. These exceptions were for the curriculum-guide union SC-test blueprint and aggregate union SC-test blueprint. For the countries, average proportions of "item" coverage ranged from .57 to .96. Country D had proportions of 1.0 for nearly all test blueprints. This meant that all topics included in most test blueprints also were included in the corresponding curriculum-data source.

The proportions of each country's curricula that were covered on the corresponding inclusive SC-test blueprints are shown in Table 33. Union-test blueprints had the same items as the field-trial instrument so proportions would be the same as in Table 11 and are not shown here. The proportions of coverage in Table 33 were more

Table 32

Proportions of Items on Inclusive-Test Blueprints in Each Corresponding Curriculum Source

	Country																				
Test	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	AVE	SD	MIN	MAX
EX-UN	0.95	0.90	0.90	1.00	0.66	0.92	0.43	0.84	0.92	0.52	0.87	0.99	0.79	1.00	0.66	0.71	0.90	0.82	0.16	0.43	1.00
EX-7I	0.96	0.93	0.91	1.00	0.72	0.93	0.46	0.93	0.95	0.56	0.92	1.00	0.89	1.00	0.73	0.73	0.96	0.86	0.15	0.46	1.00
CG-UN	0.87	0.70	0.97	1.00	0.75	0.92	0.63	0.38	0.86	0.72	0.61	0.94	0.83	0.46	0.62	0.84	1.00	0.77	0.18	0.38	1.00
CG-7I	0.83	0.83	0.95	1.00	0.74	1.00	0.69	0.43	0.91	0.96	0.78	1.00	0.96	0.60	0.69	0.88	1.00	0.84	0.16	0.43	1.00
TX-UN	0.90	0.89	0.99	1.00	0.59	0.76	0.67	0.91	0.96	0.52	0.91	0.91	1.00	0.93	0.88	0.88	0.99	0.86	0.14	0.52	1.00
TX-7I	0.94	0.93	1.00	1.00	0.61	0.83	0.70	0.97	0.98	0.51	0.92	0.95	1.00	0.95	0.90	0.95	1.00	0.89	0.14	0.51	1.00
AG-UN	0.77	0.76	0.90	0.99	0.47	0.73	0.30	0.46	0.82	0.51	0.52	0.87	0.83	0.59	0.60	0.54	0.91	0.68	0.19	0.30	0.99
AG-7I	0.90	0.88	1.00	0.72	1.00	0.55	0.72	1.00	0.82	0.75	1.00	1.00	0.88	0.67	0.55	1.00	0.48	0.82	0.17	0.48	1.00
Average	0.89	0.85	0.95	0.96	0.69	0.83	0.58	0.74	0.91	0.63	0.82	0.96	0.90	0.78	0.70	0.82	0.91	0.82	0.16	0.44	1.00
SD	0.06	0.08	0.04	0.09	0.14	0.14	0.15	0.25	0.06	0.15	0.16	0.05	0.08	0.20	0.12	0.14	0.16	0.06	0.02	0.07	0.00
Min	0.77	0.70	0.90	0.72	0.47	0.55	0.30	0.38	0.82	0.51	0.52	0.87	0.79	0.46	0.55	0.54	0.48	0.68	0.14	0.30	0.99
Max	0.96	0.93	1.00	1.00	1.00	1.00	0.72	1.00	0.98	0.96	1.00	1.00	1.00	1.00	0.90	1.00	1.00	0.89	0.19	0.52	1.00

Table 33

Proportions of Each Country's Curriculum Tested on Corresponding Inclusive-Test Blueprint

	Country																				
Test	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	AVE	SD	MIN	MAX
EX-7I	0.70	0.69	0.65	0.73	0.67	0.76	0.68	0.80	0.80	0.68	0.81	0.68	0.78	0.62	0.83	0.75	0.69	0.72	0.06	0.62	0.83
CG-7I	0.55	0.54	0.44	0.49	0.64	0.47	0.72	0.64	0.48	0.64	0.65	0.57	0.52	0.53	0.57	0.48	0.39	0.55	0.09	0.39	0.72
TX-7I	0.91	0.75	0.79	0.78	0.83	0.81	0.73	0.84	0.71	0.74	0.89	0.92	0.82	0.63	0.83	0.92	0.73	0.80	0.08	0.63	0.92
AG-71	0.47	0.30	0.23	0.29	0.57	0.42	0.55	0.51	0.32	0.69	0.63	0.37	0.47	0.37	0.44	0.29	0.34	0.43	0.13	0.23	0.69
Average	0.66	0.57	0.53	0.57	0.68	0.62	0.67	0.70	0.58	0.69	0.75	0.63	0.65	0.54	0.67	0.61	0.54	0.63	0.09	0.47	0.79
SD	0.17	0.17	0.21	0.20	0.10	0.17	0.07	0.13	0.19	0.04	0.11	0.20	0.15	0.10	0.17	0.24	0.18	0.15	0.02	0.17	0.09
Min	0.47	0.30	0.23	0.29	0.57	0.42	0.55	0.51	0.32	0.64	0.63	0.37	0.47	0.37	0.44	0.29	0.34	0.43	0.06	0.23	0.69
Max	0.91	0.75	0.79	0.78	0.83	0.81	0.73	0.84	0.80	0.74	0.89	0.92	0.82	0.63	0.83	0.92	0.73	0.80	0.13	0.63	0.92

variable than those in previous similar tables. They ranged from .23 to .92. The average was .63. Average proportions of coverage on SC test-blueprints for each data source were around .43 to .80. Average proportions of coverage on SC-test blueprints for each country ranged from .53 to .75. The highest proportions of coverage were for the textbook data sources. The lowest proportions were for the aggregate-data source.

Differences in topic inclusion and emphasis. Differences in topic inclusion between each curriculum-data source and the corresponding inclusive SC-test blueprints are presented for topics in Table 34. This table can be compared to Table 12 and Table 22. On average, the inclusive specially-constructed-test blueprints had 30% more of a mis-match to the curriculum in topic inclusion (topics on the test blueprint and not in the curriculum or in the curriculum and not in the test blueprint) than did the unique specially-constructed-test blueprints (see Table 22). The improvement over the correspondence in topic inclusion between the field-trial instrument and the curriculum was minimal (see Table 12). The most improvement was for the aggregate-test blueprint.

The average of the proportions of countries with a correspondence in topic inclusion between the inclusive SC-test blueprints and the corresponding data source ranged from .31 (1.3.2. *2-D Geometry*) to .97 (1.6.2 *Equations and Formulas*) for topics and from .64 to .72 for curriculum sources. The topics with the lowest and highest rates of match were the same as reported in Table 12. The lowest rate of correspondence in inclusion or non-inclusion for data sources was for the curriculum-guide-data source. The highest rate of correspondence was for the expert mapping.

Table 35 shows the summary information on correspondence in topic inclusion between the field-trial instrument and the curriculum-data sources for countries. This

Table 34

Differences in Topic Inclusion between Each Inclusive-Test Blueprint and Each Corresponding Curriculum Source for Each Topic

	EX 7I		CG 7I		TX 7I		AG 7I			
Topic Code	# Mis-Match	Prop. Match	# Mis-Match	Prop. Match	# Mis-Match	Prop. Match	# Mis-Match	Prop. Match	Ave. # Mis-Match	Prop. Match
1.1.1.1	8	0.53	10	0.41	11	0.35	3	0.82	8.0	0.53
1.1.1.2	9	0.47	5	0.71	2	0.88	8	0.53	6.0	0.65
1.1.1.3	8	0.53	9	0.47	2	0.88	6	0.65	6.3	0.63
1.1.2.1	3	0.82	9	0.47	1	0.94	8	0.53	5.3	0.69
1.1.2.2	3	0.82	10	0.41	2	0.88	9	0.47	6.0	0.65
1.1.2.3	5	0.71	10	0.41	2	0.88	9	0.47	6.5	0.62
1.1.2.4	5	0.71	11	0.35	3	0.82	8	0.53	6.8	0.60
1.1.2.5	12	0.29	5	0.71	11	0.35	4	0.76	8.0	0.53
1.1.3.1	3	0.82	2	0.88	2	0.88	5	0.71	3.0	0.82
1.1.3.2	14	0.18	13	0.24	12	0.29	8	0.53	11.8	0.31
1.1.3.3	11	0.35	12	0.29	11	0.35	6	0.65	10.0	0.41
1.1.4.1	2	0.88	5	0.71	4	0.76	0	1.00	2.8	0.84
1.1.4.2	2	0.88	6	0.65	3	0.82	10	0.41	5.3	0.69
1.1.4.3	0	1	2	0.88	3	0.82	0	1.00	1.3	0.93
1.1.4.4	4	0.76	5	0.71	11	0.35	8	0.53	7.0	0.59
1.1.4.5	2	0.88	4	0.76	7	0.59	1	0.94	3.5	0.79
1.1.5.1	9	0.47	6	0.65	9	0.47	4	0.76	7.0	0.59
1.1.5.2	3	0.82	10	0.41	10	0.41	7	0.59	7.5	0.56
1.1.5.3	5	0.71	8	0.53	5	0.71	7	0.59	6.3	0.63
1.1.5.4	5	0.71	8	0.53	7	0.59	3	0.82	5.8	0.66
1.2.1	2	0.88	4	0.76	2	0.88	11	0.35	4.8	0.72
1.2.2	2	0.88	3	0.82	4	0.76	5	0.71	3.5	0.79
1.2.3	11	0.35	11	0.35	7	0.59	4	0.76	8.3	0.51
1.3.1	3	0.82	5	0.71	3	0.82	8	0.53	4.8	0.72
1.3.2	0	1	3	0.82	0	1	3	0.82	1.5	0.91
1.3.3	1	0.94	1	0.94	1	0.94	2	0.88	1.3	0.93
1.3.4	2	0.88	0	1	4	0.76	11	0.35	4.3	0.75
1.3.5	1	0.94	9	0.47	7	0.59	1	0.94	4.5	0.74
1.4.1	2	0.88	2	0.88	4	0.76	11	0.35	4.8	0.72
1.4.2	3	0.82	3	0.82	11	0.35	9	0.47	6.5	0.62
1.4.3	13	0.24	12	0.29	9	0.47	5	0.71	9.8	0.43
1.5.1	2	0.88	3	0.82	10	0.41	9	0.47	6.0	0.65
1.5.2	1	0.94	2	0.88	5	0.71	10	0.41	4.5	0.74
1.5.3	10	0.41	8	0.53	6	0.65	3	0.82	6.8	0.60
1.5.4	6	0.65	6	0.65	4	0.76	0	1.00	4.0	0.76
1.6.1	2	0.88	1	0.94	0	1	2	0.88	1.3	0.93
1.6.2	1	0.94	0	1	0	1	1	0.94	0.5	0.97
1.7.1	1	0.94	2	0.88	3	0.82	4	0.76	2.5	0.85
1.7.2	9	0.47	12	0.29	6	0.65	4	0.76	7.8	0.54
1.8.1	0	1	2	0.88	4	0.76	0	1.00	1.5	0.91
1.8.2	0	1	3	0.82	0	1	0	1.00	0.8	0.96
1.9.1	7	0.59	4	0.76	10	0.41	3	0.82	6.0	0.65
1.9.2	8	0.53	7	0.59	9	0.47	5	0.71	7.3	0.57
1.10.1	10	0.41	14	0.18	11	0.35	8	0.53	10.8	0.37
Sum	210		267		238		233		237	
Average	4.77	0.72	6.07	0.64	5.41	0.68	5.30	0.69	5.39	0.68

Table 35

Differences in Topic Inclusion between Each Inclusive-Test Blueprint and Each Corresponding Curriculum Source for Each Country

	EX-7I			CG-7I			TX-7I			AG-7I					
Country	*SD* of All Dif.	Ave. Pos. Dif.	Ave. Neg. Dif.	*SD* of All	Ave. Pos. Dif.	Ave. Neg. Dif.	*SD* of All Dif.	Ave. Pos. Dif.	Ave. Neg. Dif.	*SD* of All Dif.	Ave. Pos. Dif.	Ave. Neg. Dif.	Ave. of All Dif.	Ave. of Pos. Dif.	Ave. of Neg. Dif.
A	0.73	11	-1	0.64	14	-2	0.68	12	-2	0.73	12	0	0.69	12.3	-1.3
B	0.73	10	-2	0.7	10	-3	0.82	7	-1	0.68	13	-1	0.73	10.0	-1.75
C	0.70	11	-2	0.5	22	0	0.57	19	0	0.43	24	-1	0.55	19.0	-0.75
D	0.68	14	0	0.59	18	0	0.64	16	0	0.43	25	0	0.59	18.3	0
E	0.66	8	-7	0.7	9	-4	0.59	7	-11	0.86	3	-3	0.70	6.8	-6.25
F	0.80	7	-2	0.57	19	0	0.75	6	-5	0.77	10	0	0.72	10.5	-1.75
G	0.61	4	-13	0.75	6	-5	0.66	10	-5	0.86	2	-4	0.72	5.5	-6.75
H	0.82	6	-2	0.68	4	-10	0.73	11	-1	0.84	4	-3	0.77	6.3	-4
I	0.77	9	-1	0.64	15	-1	0.66	14	-1	0.59	18	0	0.66	14.0	-0.75
J	0.59	7	-11	0.73	10	-2	0.57	5	-14	0.86	4	-2	0.69	6.5	-7.25
K	0.82	6	-2	0.7	8	-5	0.73	10	-2	0.82	6	-2	0.77	7.5	-2.75
L	0.73	12	0	0.7	13	0	0.68	13	-1	0.59	18	0	0.68	14.0	-0.25
M	0.77	7	-3	0.64	15	-1	0.66	15	0	0.64	16	0	0.68	13.3	-1
N	0.64	16	0	0.68	7	-7	0.68	13	-1	0.73	11	-1	0.68	11.8	-2.25
O	0.73	5	-7	0.73	8	-4	0.82	5	-3	0.77	7	-3	0.76	6.3	-4.25
P	0.77	4	-6	0.59	15	-3	0.82	7	-1	0.66	13	-2	0.71	9.8	-3
Q	0.68	13	-1	0.39	27	0	0.55	20	0	0.43	25	0	0.51	21.3	-0.25
Average	0.72	8.82	-3.53	0.64	12.9	-2.76	0.68	11.2	-2.8	0.69	12.4	-1.3	0.68	11.338	-2.603

table can be compared to Tables 13 and 23. Average proportions of correspondence in topic inclusion between the curricula and the corresponding inclusive SC-test blueprint ranged from .51 to .77 for countries. Correspondence in topic inclusion for the curriculum-data sources was the same as in Table 34. The average proportion of correspondence was .68, which again was a slight improvement over the field-trial instrument. However, it was .20 less than the correspondence in topic inclusion (or non-inclusion) between the curriculum and the unique SC-test blueprints. The average numbers of topics included in a country's curriculum but not on the corresponding inclusive-test blueprint was 11, and the average number of topics included on a test blueprint but not in the corresponding curriculum source was 3. The numbers of topics in curriculum sources not on the blueprints (positive differences) were about the same for all data sources. The lowest numbers of topics in blueprints not in a corresponding curriculum source (negative differences) were for topics included in the aggregate-data source. Average rates of non-tested topics ranged from 5 to 21 and rates for topics on the inclusive-test blueprints not in the curriculum ranged from 0 to 7.

Tables 36 and 37 show the differences between the curriculum-data sources and the corresponding inclusive SC-test blueprints in topic emphasis. Table 36 highlights differences for topics, and Table 37 highlights differences for countries. Positive differences occurred when topics received a higher emphasis in the curriculum than on the test blueprints, and negative differences occurred when topics receive a higher emphasis on the test blueprints than in the curriculum.

On average across all curriculum sources, topics with the lowest average positive difference in emphasis (more weight in the curriculum) were 1.8.1 *Infinite Processes*

Table 36

Difference in Topic Emphasis between Each Inclusive-Test blueprint and Each Corresponding Curriculum Source for Each Topic

	EX-UN			EX-7I			CG-UN			CG-7I			TX-UN		
Topic Code	*SD* of all Dif.	Ave. Pos. Dif.	Ave. Neg. Dif.	*SD* of all Dif.	Ave. Pos. Dif.	Ave. Neg. Dif.	*SD* of all Dif.	Ave. Pos. Dif.	Ave. Neg. Dif.	*SD* of all Dif.	Ave. Pos. Dif.	Ave. Neg. Dif.	*SD* of all Dif.	Ave. Pos. Dif.	Ave. Neg. Dif.
1.1.1.1	0.004	0.011	-0.015	0.013	0.026	0	0.016	0.019	-0.020	0.025	0.040	0	0.020	0.031	-0.011
1.1.1.2	0.013	0.015	-0.017	0.020	0.033	0	0.016	0.022	-0.017	0.023	0.020	-0.022	0.032	0.060	-0.026
1.1.1.3	0.012	0.021	-0.018	0.022	0.037	0	0.015	0.016	-0.021	0.024	0.038	0	0.013	0.031	-0.013
1.1.2.1	0.016	0.029	-0.018	0.015	0.032	-0.022	0.007	0.010	-0.022	0.016	0.032	0	0.019	0.035	-0.023
1.1.2.2	0.017	0.021	-0.015	0.017	0.023	-0.018	0.011	0.011	-0.020	0.018	0.034	0	0.014	0.031	-0.013
1.1.2.3	0.018	0.025	-0.015	0.018	0.021	-0.019	0.011	0.012	-0.022	0.020	0.036	0	0.005	0.013	-0.007
1.1.2.4	0.011	0.016	-0.017	0.012	0.012	-0.021	0.011	0.009	-0.020	0.017	0.034	0	0.020	0.030	-0.025
1.1.2.5	0.026	0.035	0	0.026	0.035	0	0.013	0.028	0	0.013	0.028	0	0.010	0.009	0
1.1.3.1	0.013	0.011	-0.019	0.014	0.012	-0.019	0.012	0.010	-0.015	0.016	0	-0.020	0.019	0.033	-0.028
1.1.3.2	0.025	0.037	0	0.025	0.037	0	0.018	0.037	0	0.018	0.012	0	0.071	0.040	0
1.1.3.3	0.018	0.032	0	0.018	0.032	0	0.023	0.039	0	0.023	0.033	0	0.064	0.039	0
1.1.4.1	0.008	0.023	0	0.008	0.023	0	0.015	0.031	0	0.015	0.031	0	0.003	0.004	0
1.1.4.2	0.012	0.010	-0.018	0.014	0.008	-0.020	0.012	0.011	-0.019	0.019	0.035	0	0.022	0.038	-0.025
1.1.4.3	0.000	0	0	0.000	0	0	0.008	0.024	0	0.008	0.024	0	0.001	0.001	0
1.1.4.4	0.011	0.011	-0.014	0.012	0.011	-0.016	0.013	0.012	-0.018	0.019	0.011	-0.021	0.014	0.026	-0.011
1.1.4.5	0.011	0.034	0	0.011	0.034	0	0.012	0.028	0	0.012	0.028	0	0.006	0.006	0
1.1.5.1	0.005	0.009	-0.017	0.013	0.026	0	0.004	0.018	-0.015	0.016	0.032	0	0.002	0.004	-0.002
1.1.5.2	0.011	0.013	-0.018	0.013	0.010	-0.021	0.009	0.009	-0.022	0.017	0.033	0	0.005	0.007	-0.006
1.1.5.3	0.011	0.017	-0.014	0.011	0.013	-0.018	0.005	0.013	-0.019	0.016	0.032	0	0.006	0.009	-0.005
1.1.5.4	0.013	0.014	-0.020	0.014	0.012	-0.023	0.005	0.013	-0.020	0.017	0.033	0	0.012	0.016	-0.006
1.2.1	0.010	0.011	-0.014	0.010	0.009	-0.017	0.013	0.013	-0.014	0.017	0.010	-0.019	0.028	0.044	-0.023
1.2.2	0.010	0.009	-0.013	0.011	0.006	-0.016	0.012	0.011	-0.017	0.017	0.006	-0.023	0.028	0.054	-0.045
1.2.3	0.012	0.009	-0.015	0.016	0.028	0	0.012	0.012	-0.020	0.019	0.036	0	0.002	0.006	-0.001
1.3.1	0.012	0.012	-0.017	0.012	0.013	-0.018	0.012	0.007	-0.019	0.017	0.007	-0.019	0.021	0.027	-0.023
1.3.2	0.007	0.011	-0.017	0.010	0.010	-0.020	0.012	0.010	-0.016	0.017	0.008	-0.021	0.023	0.043	-0.028
1.3.3	0.009	0.012	-0.016	0.011	0.007	-0.021	0.012	0.019	-0.017	0.019	0.017	-0.025	0.035	0.049	-0.034
1.3.4	0.011	0.013	-0.016	0.013	0.007	-0.021	0.010	0.015	-0.018	0.016	0.028	-0.024	0.095	0.177	-0.045
1.3.5	0.007	0.028	0	0.007	0.028	0	0.017	0.033	0	0.017	0.033	0	0.013	0.013	0
1.4.1	0.010	0.015	-0.019	0.011	0.012	-0.023	0.014	0.014	-0.017	0.020	0.039	-0.021	0.044	0.049	-0.043
1.4.2	0.011	0.017	-0.020	0.012	0.019	-0.021	0.013	0.009	-0.019	0.018	0.006	-0.021	0.042	0.081	-0.028
1.4.3	0.018	0.031	0	0.018	0.031	0	0.017	0.035	0	0.017	0.008	0	0.012	0.016	0
1.5.1	0.010	0.010	-0.014	0.011	0.009	-0.016	0.012	0.008	-0.018	0.017	0	-0.020	0.005	0.010	-0.007
1.5.2	0.012	0.012	-0.021	0.015	0.009	-0.023	0.012	0.008	-0.015	0.015	0	-0.019	0.016	0.020	-0.014
1.5.3	0.022	0.035	0	0.022	0.035	0	0.027	0.045	0	0.027	0.045	0	0.025	0.041	0
1.5.4	0.017	0.031	0	0.017	0.031	0	0.018	0.036	0	0.018	0.036	0	0.004	0.007	0
1.6.1	0.010	0.014	-0.018	0.011	0.011	-0.022	0.012	0.014	-0.019	0.019	0.017	-0.025	0.041	0.051	-0.025
1.6.2	0.012	0.009	-0.019	0.015	0.008	-0.021	0.010	0.015	-0.018	0.016	0.028	-0.024	0.060	0.134	-0.078
1.7.1	0.011	0.010	-0.022	0.014	0.008	-0.022	0.014	0.014	-0.017	0.020	0.039	-0.021	0.015	0.029	-0.028
1.7.2	0.008	0.012	-0.017	0.016	0.029	0	0.013	0.009	-0.019	0.018	0.009	0	0.007	0.015	-0.003
1.8.1	0.000	0	0	0.000	0	0	0.012	0.034	0	0.012	0.034	0	0.001	0.003	0
1.8.2	0.000	0	0	0.000	0	0	0.011	0.028	0	0.011	0.028	0	0.000	0	0
1.9.1	0.018	0.031	0	0.018	0.031	0	0.016	0.034	0	0.016	0.034	0	0.072	0.037	0
1.9.2	0.014	0.027	0	0.014	0.027	0	0.020	0.036	0	0.020	0.036	0	0.034	0.040	0
1.10.1	0.015	0.028	0	0.015	0.028	0	0.017	0.038	0	0.017	0.005	0	0.062	0.056	0
Average	0.012	0.018	-0.011	0.014	0.019	-0.010	0.013	0.020	-0.012	0.017	0.025	-0.010	0.024	0.033	-0.014

Table 36 (Contd.)

	TX-71			AG-UN			AG-71							
Topic Code	*SD* of all Dif.	Ave. Pos. Dif.	Ave. Neg. Dif.	*SD* of all Dif.	Ave. Pos. Dif.	Ave. Neg. Dif.	*SD* of all Dif.	Ave. Pos. Dif.	Ave. Neg. Dif.	Ave. of Pos. Dif.	Ave. of Neg. Dif.	*SD* of Pos. Dif.	*SD* of Neg. Dif.	*SD* of All Dif.
1.1.1.1	0.026	0.023	0	0.010	0.028	-0.007	0.014	0.035	0	0.026	-0.007	0.009	0.007	0.018
1.1.1.2	0.035	0.058	-0.021	0.019	0.024	-0.027	0.032	0.051	0	0.035	-0.016	0.017	0.010	0.029
1.1.1.3	0.015	0.034	-0.012	0.010	0.027	-0.018	0.023	0.044	0	0.031	-0.010	0.009	0.008	0.022
1.1.2.1	0.019	0.038	-0.025	0.008	0.025	-0.028	0.027	0.053	0	0.032	-0.017	0.011	0.010	0.027
1.1.2.2	0.015	0.029	-0.014	0.020	0.022	-0.026	0.031	0.048	0	0.027	-0.013	0.010	0.009	0.022
1.1.2.3	0.006	0.015	-0.006	0.007	0.015	-0.017	0.017	0.032	0	0.021	-0.011	0.009	0.008	0.018
1.1.2.4	0.020	0.027	-0.029	0.016	0.024	-0.027	0.030	0.051	0	0.025	-0.018	0.013	0.011	0.024
1.1.2.5	0.010	0.009	0	0.014	0.031	0	0.014	0.031	0	0.026	0.000	0.010	0.000	0.015
1.1.3.1	0.019	0.035	-0.030	0.015	0.019	-0.025	0.022	0	-0.049	0.017	-0.026	0.010	0.010	0.024
1.1.3.2	0.071	0.040	0	0.076	0.068	0	0.076	0.068	0	0.042	-0.003	0.017	0.007	0.026
1.1.3.3	0.064	0.039	0	0.028	0.048	0	0.028	0.048	0	0.039	-0.003	0.006	0.007	0.022
1.1.4.1	0.003	0.004	0	0.000	0	0	0.000	0	0	0.014	0.000	0.013	0.000	0.012
1.1.4.2	0.021	0.040	-0.025	0.016	0.019	-0.023	0.026	0.045	0	0.026	-0.016	0.014	0.010	0.024
1.1.4.3	0.001	0.001	0	0.000	0	0	0.000	0	0	0.006	0.000	0.010	0.000	0.008
1.1.4.4	0.022	0.025	0	0.007	0.016	-0.019	0.019	0.035	0	0.018	-0.012	0.009	0.008	0.017
1.1.4.5	0.006	0.006	0	0.007	0.031	0	0.007	0.031	0	0.025	0.000	0.011	0.000	0.015
1.1.5.1	0.003	0.005	0	0.006	0.017	-0.006	0.010	0.023	0	0.017	-0.005	0.010	0.007	0.014
1.1.5.2	0.008	0.012	0	0.005	0.014	-0.012	0.014	0.027	0	0.016	-0.010	0.009	0.009	0.016
1.1.5.3	0.006	0.010	-0.005	0.004	0.014	-0.013	0.014	0.027	0	0.017	-0.009	0.008	0.007	0.015
1.1.5.4	0.015	0.017	0	0.007	0.023	-0.006	0.011	0.029	0	0.020	-0.009	0.007	0.009	0.017
1.2.1	0.028	0.044	-0.025	0.016	0.022	-0.028	0.029	0.050	0	0.025	-0.018	0.016	0.008	0.025
1.2.2	0.030	0.053	-0.053	0.026	0.045	-0.027	0.031	0.024	-0.071	0.026	-0.033	0.020	0.020	0.036
1.2.3	0.003	0.005	0	0.005	0.016	-0.006	0.010	0.022	0	0.017	-0.005	0.011	0.007	0.014
1.3.1	0.021	0.026	-0.024	0.015	0.021	-0.024	0.027	0.046	0	0.020	-0.018	0.012	0.007	0.021
1.3.2	0.026	0.049	-0.025	0.019	0.020	-0.024	0.025	0	-0.063	0.020	-0.027	0.015	0.014	0.028
1.3.3	0.035	0.048	-0.035	0.021	0.032	-0.031	0.037	0	-0.087	0.023	-0.033	0.017	0.022	0.034
1.3.4	0.106	0.138	-0.030	0.031	0.063	-0.035	0.053	0.072	0	0.064	-0.024	0.059	0.013	0.061
1.3.5	0.013	0.013	0	0.009	0.037	0	0.009	0.037	0	0.028	0.000	0.009	0.000	0.015
1.4.1	0.043	0.058	-0.045	0.040	0.067	-0.038	0.061	0.076	0	0.041	-0.026	0.024	0.014	0.039
1.4.2	0.060	0.062	0	0.034	0.077	-0.030	0.053	0.069	0	0.043	-0.017	0.030	0.011	0.038
1.4.3	0.012	0.016	0	0.013	0.027	0	0.013	0.027	0	0.024	-0.002	0.009	0.006	0.015
1.5.1	0.010	0.014	0	0.008	0.013	-0.019	0.017	0.031	0	0.012	-0.012	0.008	0.008	0.014
1.5.2	0.016	0.023	-0.018	0.018	0.027	-0.021	0.029	0.046	0	0.019	-0.016	0.013	0.007	0.020
1.5.3	0.025	0.041	0	0.028	0.054	0	0.028	0.054	0	0.044	0.000	0.007	0.000	0.022
1.5.4	0.004	0.007	0	0.000	0	0	0.000	0	0	0.018	0.000	0.015	0.000	0.014
1.6.1	0.042	0.045	-0.025	0.020	0.032	-0.036	0.039	0.000	-0.079	0.023	-0.031	0.017	0.019	0.032
1.6.2	0.088	0.167	-0.043	0.045	0.099	-0.067	0.080	0.026	-0.192	0.061	-0.058	0.059	0.055	0.082
1.7.1	0.021	0.019	-0.033	0.024	0.029	-0.032	0.035	0.014	-0.073	0.020	-0.031	0.010	0.017	0.029
1.7.2	0.008	0.009	0	0.006	0.020	-0.007	0.012	0.027	0	0.016	-0.008	0.007	0.008	0.015
1.8.1	0.001	0.003	0	0.000	0	0	0.000	0	0	0.009	0.000	0.014	0.000	0.011
1.8.2	0.000	0	0	0.000	0	0	0.000	0	0	0.007	0.000	0.012	0.000	0.009
1.9.1	0.072	0.037	0	0.024	0.047	0	0.024	0.047	0	0.037	0.000	0.006	0.000	0.019
1.9.2	0.034	0.040	0	0.019	0.037	0	0.019	0.037	0	0.035	0.000	0.005	0.000	0.018
1.10.1	0.062	0.056	0	0.032	0.046	0	0.032	0.046	0	0.038	0.000	0.016	0.000	0.022
Average	0.026	0.033	-0.012	0.016	0.029	-0.015	0.025	0.033	-0.014	0.026	-0.012	0.014	0.008	0.023

(.009) and 1.8.2 *Change* (.007). The highest averages of positive differences were for 1.3.4 *3-D Geometry* (.064) and 1.6.2 *Equations and Formulas* (.06). Aside from the topics not in the blueprints, topics with the lowest negative difference were 1.1.5.1 *Estimating Quantity and Size* (-.005) and 1.2.3 *Estimation Errors* (-.005). The largest was for 1.6.2 *Equations and Formulas* - which also had a high positive difference. The average of the positive average differences was .026 while the average of the negative average differences was -.012. The averages of the average positive differences for the curriculum sources were all around .02; average negative differences were around -.01. In general, topics received more weight in the curriculum than on the test blueprints.

Table 37 shows the variability in topic emphasis differences for countries. Average positive differences in topic emphasis for countries ranged from .012 to .049, and negative differences ranged from -.013 to -.025. For data sources, the positive differences ranged from about .02 to .06. Negative differences ranged from .015 to .087. The poorest correspondence in topic emphasis was with the aggregate 70%-intersection-test blueprint, followed by the union aggregate-test blueprint. Lower numbers were for the expert-mapping and curriculum-guide blueprints.

Correlations and Euclidean distance measures. Correlations between the proportions-of-topic-emphasis profiles in each curriculum-data source and the topic-weight profiles in each inclusive SC-test blueprint are in Table 38. The correlations ranged from .00 to .90 with an overall average of .58. The average correlation within data sources across countries was highest between the text union- and 70%-intersection-test blueprint topic weight profiles and the topic emphasis profiles for the corresponding data source and lowest between the curriculum-guide union-test blueprint topic weight profiles

Table 37

Differences in Topic Emphasis between Each Inclusive-Test Blueprint and Each Corresponding Curriculum Source for Each Country

	EX-UN			EX-7I			CG-UN			CG-7I			TX-UN		
Country	*SD* of all Dif.	Ave. Pos. Dif.	Ave. Neg. Dif.	*SD* of all Dif.	Ave. Pos. Dif.	Ave. Neg. Dif.	*SD* of all Dif.	Ave. Pos. Dif.	Ave. Neg. Dif.	*SD* of all Dif.	Ave. Pos. Dif.	Ave. Neg. Dif.	*SD* of all Dif.	Ave. Pos. Dif.	Ave. Neg. Dif.
A	0.015	0.017	-0.014	0.020	0.021	-0.019	0.016	0.015	-0.014	0.024	0.034	-0.020	0.024	0.016	-0.019
B	0.019	0.018	-0.017	0.021	0.021	-0.020	0.022	0.022	-0.019	0.022	0.033	-0.016	0.052	0.059	-0.022
C	0.020	0.022	-0.015	0.023	0.028	-0.020	0.018	0.021	-0.012	0.024	0.026	-0.023	0.029	0.023	-0.019
D	0.012	0.011	-0.010	0.015	0.017	-0.014	0.014	0.015	-0.011	0.021	0.029	-0.019	0.028	0.019	-0.018
E	0.025	0.028	-0.020	0.026	0.028	-0.023	0.018	0.015	-0.021	0.025	0.031	-0.020	0.040	0.038	-0.021
F	0.015	0.017	-0.014	0.020	0.025	-0.017	0.018	0.021	-0.013	0.022	0.028	-0.020	0.038	0.037	-0.024
G	0.037	0.041	-0.032	0.038	0.049	-0.034	0.023	0.020	-0.028	0.025	0.023	-0.030	0.063	0.056	-0.026
H	0.017	0.017	-0.015	0.017	0.018	-0.017	0.036	0.056	-0.031	0.037	0.052	-0.044	0.031	0.026	-0.023
I	0.015	0.011	-0.015	0.017	0.014	-0.016	0.018	0.017	-0.015	0.022	0.032	-0.018	0.031	0.029	-0.018
J	0.028	0.029	-0.028	0.029	0.029	-0.033	0.021	0.021	-0.017	0.017	0.023	-0.012	0.049	0.059	-0.024
K	0.017	0.018	-0.016	0.018	0.020	-0.019	0.025	0.022	-0.025	0.022	0.019	-0.024	0.037	0.031	-0.017
L	0.016	0.016	-0.015	0.020	0.023	-0.019	0.014	0.013	-0.012	0.019	0.033	-0.014	0.019	0.016	-0.016
M	0.019	0.022	-0.016	0.019	0.022	-0.017	0.019	0.020	-0.015	0.021	0.032	-0.017	0.038	0.037	-0.014
N	0.015	0.018	-0.011	0.020	0.024	-0.016	0.031	0.032	-0.030	0.029	0.027	-0.037	0.092	0.091	-0.014
O	0.022	0.026	-0.019	0.022	0.024	-0.024	0.024	0.022	-0.024	0.027	0.026	-0.026	0.038	0.060	-0.020
P	0.024	0.020	-0.023	0.027	0.031	-0.021	0.018	0.018	-0.016	0.023	0.034	-0.018	0.043	0.032	-0.034
Q	0.016	-0.036	0.021	0.018	0.019	-0.016	0.018	0.019	-0.014	0.024	0.023	-0.025	0.022	0.017	-0.015
Average	0.020	0.017	-0.015	0.022	0.024	-0.020	0.021	0.022	-0.019	0.024	0.030	-0.023	0.040	0.038	-0.020

Table 37 (*Contd.*)

	TX-7I			AG-UN			AG-7I							
Country	*SD* of all Dif.	Ave. Pos. Dif.	Ave. Neg. Dif.	*SD* of all Dif.	Ave. Pos. Dif.	Ave. Neg. Dif.	*SD* of all Dif.	Ave. Pos. Dif.	Ave. Neg. Dif.	Ave. of Pos. Dif.	Ave. of Neg. Dif.	*SD* of Pos. Dif.	*SD* of Neg. Dif.	*SD* of All Dif.
A	0.019	0.016	-0.020	0.022	0.022	-0.022	0.047	0.045	-0.076	0.023	-0.025	0.010	0.019	0.029
B	0.054	0.050	-0.027	0.032	0.037	-0.020	0.058	0.054	-0.099	0.037	-0.030	0.015	0.026	0.040
C	0.029	0.024	-0.027	0.024	0.015	-0.026	0.057	0.032	-0.110	0.024	-0.032	0.005	0.030	0.035
D	0.022	0.019	-0.016	0.022	0.014	-0.024	0.055	0.028	-0.101	0.019	-0.027	0.006	0.029	0.031
E	0.041	0.046	-0.030	0.042	0.075	-0.024	0.052	0.144	-0.062	0.051	-0.028	0.039	0.013	0.049
F	0.034	0.033	-0.032	0.026	0.031	-0.021	0.052	0.053	-0.098	0.031	-0.030	0.010	0.026	0.036
G	0.067	0.057	-0.038	0.065	0.141	-0.028	0.063	0.158	-0.079	0.068	-0.037	0.049	0.016	0.064
H	0.027	0.026	-0.025	0.039	0.080	-0.023	0.053	0.101	-0.084	0.047	-0.033	0.029	0.021	0.047
I	0.030	0.031	-0.015	0.025	0.019	-0.026	0.053	0.038	-0.097	0.024	-0.028	0.009	0.026	0.032
J	0.057	0.061	-0.034	0.037	0.062	-0.023	0.037	0.065	-0.054	0.044	-0.028	0.018	0.012	0.039
K	0.041	0.035	-0.020	0.034	0.044	-0.025	0.039	0.052	-0.083	0.030	-0.029	0.012	0.021	0.034
L	0.021	0.016	-0.026	0.020	0.017	-0.020	0.049	0.035	-0.090	0.021	-0.027	0.008	0.024	0.030
M	0.042	0.034	-0.021	0.017	0.019	-0.016	0.041	0.033	-0.076	0.027	-0.024	0.007	0.020	0.030
N	0.098	0.092	-0.022	0.033	0.040	-0.021	0.053	0.057	-0.090	0.048	-0.030	0.028	0.024	0.047
O	0.042	0.055	-0.027	0.029	0.041	-0.021	0.050	0.071	-0.095	0.041	-0.032	0.018	0.024	0.042
P	0.040	0.035	-0.034	0.037	0.030	-0.035	0.064	0.055	-0.102	0.032	-0.035	0.011	0.026	0.039
Q	0.025	0.018	-0.016	0.019	0.016	-0.018	0.048	0.026	-0.095	0.013	-0.022	0.019	0.030	0.031
Average	0.041	0.038	-0.025	0.031	0.041	-0.023 #	0.051	0.062	-0.088	0.034	-0.029	0.017	0.023	0.038

Tab

A

B

D

E

F

H

K

M

N

P

Table 38

Correlations between the Proportions-of-Topic-Emphasis Profiles for Each Country in Each Curriculum-Data Source with the Topic-Weight Profiles for Each Corresponding Inclusive-Test Blueprint

Country	EX-UN	EX-7I	CG-UN	CG-7I	TX-UN	TX-7I	AG-UN	AG-7I	Ave.	SD
A	0.59**	0.51**	.56**	.34*	.76**	.77**	.73**	.66**	0.61	0.17
B	0.45**	0.47**	.38**	.54**	.42**	.43**	.53**	.34*	0.44	0.06
C	0.34*	0.31*	.31*	0.20	.65**	.64**	.58**	.30*	0.42	0.19
D	0.77**	0.74**	.63**	.47**	.59**	.60**	.67**	.46**	0.62	0.07
E	0.28	0.36*	.62**	.45**	.67**	.64**	.69**	.59**	0.54	0.08
F	0.69**	0.57**	.33*	.44**	.41**	.44**	.62**	.51**	0.50	0.10
G	0.18	0.22	.57**	.53**	.56**	.57**	.45**	.54**	0.45	0.05
H	0.60**	0.67**	.39**	.39**	.64**	.65**	.60**	.53**	0.56	0.12
I	0.66**	0.67**	.43**	.43**	.61**	.61**	.55**	.49**	0.56	0.08
J	0.33*	0.35*	.38*	.73**	.78**	.70**	.75**	.78**	0.60	0.15
K	0.62**	0.64**	.38*	.61**	.90**	.89**	.72**	.75**	0.69	0.19
L	0.51**	.48**	.69**	.64**	.85**	.87**	.76**	.63**	0.68	0.09
M	0.52**	.62**	.35*	.52**	.87**	.86**	.82**	.79**	0.67	0.21
N	0.56**	.49**	0.18	.42**	.56**	.56**	.50**	.48**	0.47	0.14
O	0.57**	.61**	.36*	.39**	.79**	.76**	.77**	.58**	0.60	0.19
P	0.49**	.40**	.47**	.43**	.34*	.36*	.29*	0.12	0.36	0.06
Q	0.55**	.58**	0.05	0.25	.81**	.82**	.73**	.70**	0.52	0.39
Ave	0.51	0.51	0.41	0.44	0.66	0.66	0.63	0.54	0.59	0.11
SD	0.15	0.14	0.17	0.16	0.16	0.15	0.13	0.17	0.09	0.08

*p <.05. *p <.01.

and the corresponding curriculum profiles. Average correlations for countries varied. These ranged from a low of .36 (country P) to a high of .68 (country K). Standard deviations of correlations for countries varied from .05 (country G) to .39 (country Q).

Euclidean distances between the proportion of topic-emphasis profiles in each curriculum-data source and topic-weight profiles for the corresponding test blueprints are shown in Table 39. The distances ranged from .08 to .66, with an overall average of .25. The largest average distance was found between the aggregate 70% intersection-test blueprint-topic profiles and the aggregate-data-source topic profiles. The smallest were between the expert-mapping- and curriculum-guide-test blueprint topic profiles and the corresponding data-source topic profiles. The smallest average distance for countries was country L (.14). The largest was for country N (.31). Average standard deviations of distances were generally less than .10. Table 40 shows differences between the Euclidean distances in Table 39 and those computed earlier using the unique-test blueprints. The largest difference was for the aggregate 70%-test blueprint. The smallest was for the curriculum-guide union-test blueprint.

Variations in Performance across Specially-Constructed Tests

Scores and Ranks

I computed country scores on SC tests using the following steps:

1. Identify topics included on each SC-test blueprint (i.e., either from the union, 70%-intersection, strict-intersection, or unique test blueprints).
2. Find average percent of students passing the items measuring each topic included on each test for each country by averaging across the percent of

Table 39

Euclidean Distances between the Proportions-of-Topic-Emphasis Profiles for Each Country in Each Curriculum-Data Source with the Topic-Weight Profiles for Each Corresponding Inclusive-Test Blueprint

Country	EX-UN	EX-7I	CG-UN	CG-7I	TX-UN	TX-7I	AG-UN	AG-7I	Ave.	SD
A	0.10	0.13	0.11	0.16	0.16	0.16	0.14	0.30	0.16	0.02
B	0.12	0.14	0.14	0.14	0.35	0.35	0.20	0.38	0.23	0.09
C	0.13	0.15	0.12	0.16	0.21	0.22	0.15	0.37	0.19	0.04
D	0.08	0.10	0.09	0.14	0.19	0.20	0.13	0.35	0.16	0.04
E	0.17	0.17	0.12	0.16	0.26	0.27	0.29	0.35	0.23	0.07
F	0.10	0.13	0.12	0.15	0.25	0.26	0.17	0.33	0.19	0.06
G	0.25	0.25	0.15	0.17	0.42	0.42	0.44	0.43	0.31	0.13
H	0.11	0.12	0.24	0.25	0.21	0.21	0.26	0.35	0.22	0.02
I	0.10	0.11	0.12	0.15	0.22	0.23	0.15	0.34	0.18	0.04
J	0.19	0.19	0.14	0.11	0.32	0.34	0.26	0.25	0.23	0.09
K	0.11	0.12	0.16	0.15	0.25	0.23	0.23	0.26	0.19	0.04
L	0.11	0.13	0.09	0.12	0.13	0.13	0.12	0.32	0.14	0.01
M	0.13	0.13	0.13	0.14	0.27	0.26	0.11	0.26	0.18	0.07
N	0.10	0.13	0.21	0.19	0.66	0.66	0.21	0.34	0.31	0.22
O	0.15	0.14	0.16	0.18	0.26	0.26	0.19	0.32	0.21	0.04
P	0.16	0.18	0.12	0.15	0.29	0.30	0.42	0.22	0.23	0.11
Q	0.11	0.12	0.12	0.16	0.15	0.16	0.31	0.22	0.17	0.07
Ave	0.13	0.14	0.14	0.16	0.27	0.27	0.22	0.32	0.25	0.06
SD	0.04	0.04	0.04	0.03	0.12	0.12	0.09	0.06	0.05	0.05

Table 40

Differences in Euclidean Distances between the Proportions-of-Topic-Emphasis Profiles for Each Country in Each Curriculum-Data Source with the Topic-Weight Profiles for Each Corresponding Inclusive-Test Blueprint

Country	EX-UN	EX-7I	CG-UN	CG-7I	TX-UN	TX-7I	AG-UN	AG-7I	Average	SD
A	0.03	0.06	0.03	0.09	0.13	0.14	0.11	0.27	0.11	0.04
B	0.02	0.04	0.04	0.04	0.10	0.11	0.06	0.23	0.08	0.03
C	0.03	0.05	0.01	0.05	0.11	0.12	0.07	0.29	0.09	0.04
D	0.01	0.04	0.02	0.07	0.09	0.11	0.07	0.29	0.09	0.03
E	0.02	0.03	0.05	0.10	0.22	0.23	0.29	0.35	0.16	0.09
F	0.06	0.09	0.01	0.03	0.09	0.10	0.14	0.30	0.10	0.05
G	0.07	0.07	0.07	0.08	0.11	0.11	0.06	0.05	0.08	0.02
H	0.03	0.03	0.10	0.11	0.14	0.15	0.10	0.19	0.11	0.02
I	0.03	0.05	0.02	0.05	0.07	0.08	0.07	0.26	0.08	0.02
J	0.05	0.06	0.02	-0.01	0.31	0.33	0.21	0.21	0.15	0.15
K	0.04	0.05	0.03	0.02	0.17	0.15	0.16	0.19	0.10	0.07
L	0.02	0.05	0.02	0.05	0.10	0.10	0.07	0.27	0.08	0.03
M	0.03	0.02	0.02	0.03	0.11	0.10	0.05	0.19	0.07	0.04
N	0.01	0.04	0.03	0.01	0.12	0.12	0.02	0.15	0.06	0.05
O	0.07	0.07	0.04	0.05	0.15	0.15	0.19	0.32	0.13	0.06
P	0.08	0.10	0.03	0.07	0.27	0.28	0.37	0.18	0.17	0.13
Q	0.02	0.03	0.01	0.05	0.05	0.06	0.22	0.13	0.07	0.07
Sum	0.64	0.88	0.55	0.89	2.36	2.42	2.26	3.87	2.36	0.80
Ave	0.04	0.05	0.03	0.05	0.14	0.14	0.13	0.23	0.14	0.05
Stdev	0.02	0.02	0.02	0.03	0.07	0.07	0.09	0.08	0.03	0.04

students within a country passing each item with codes corresponding to each topic on a given test blueprint.

3. If the test was an unweighted test, obtain an average of the topic averages for each topic on a given test blueprint. This was the country score on the unweighted test.
4. If the test was a weighted test, multiply the topic averages by the corresponding weight then sum over topics included on a given test blueprint. This was the country score on the weighted test.

Table 41 presents country scores on the field-trial instrument as well as a summary across scores on each specially-constructed test. Appendix D contains country scores on all tests. The field-trial instrument was scored by averaging over all items on the test. The unweighted union test represents an average of all topic scores on the test.

All country scores on the total field-trial instrument were higher than the average of scores on all other tests. Differences were around two to four points with the exception of country N (less than ½ a point) and country P (almost 6 points). Country M had the lowest scores on both the field-trial instrument and the average of all other test scores, and country J had the highest scores. The difference between the lowest and highest scores on both the field-trial instrument scores and the average of all other scores was nearly 30 percent. The difference between the average of each country's scores on the field-trial instrument and the grand average of all average country scores was three points. Standard deviations of the two sets of scores were almost identical. Standard deviations of each country's scores across all tests were around two to three percent. The lowest standard deviation was 1.5; the largest was 3.7.

Table 41

Summary of Country Scores on Field-Trial Instrument and across Specially-Constructed Tests

		Across Tests				
Country	Field Trial	AVE	*SD*	MIN	MAX	DIF
A	50.5	46.4	1.6	42.2	49.5	7.4
B	56.4	54.9	2.3	50.9	62.2	11.2
C	53.6	50.4	1.5	47.5	53.1	5.6
D	45.2	40.9	1.6	36.1	42.9	6.8
E	45.9	42.7	2.0	38.9	48.8	9.9
F	49.6	47.9	1.5	42.4	49.9	7.4
G	48.1	45.6	3.1	41.7	53.2	11.4
H	43.5	40.7	3.0	29.4	43.7	14.3
I	52.8	49.8	1.6	44.9	52.0	7.1
J	64.0	62.2	3.7	55.7	71.9	16.1
K	56.0	53.9	2.2	48.6	58.3	9.8
L	51.5	48.9	1.9	44.2	53.3	9.1
M	35.4	32.8	1.5	29.8	38.3	8.5
N	45.0	45.1	2.4	41.6	53.5	11.9
O	61.9	58.3	1.7	55.3	62.0	6.7
P	45.8	40.2	2.9	32.4	47.4	15.0
Q	56.4	52.6	1.9	45.5	55.9	10.4
AVE	50.7	47.8	2.1	42.8	52.7	9.9
SD	6.9	7.1	0.7	7.6	7.8	3.0

Test scores ranged from a low of 29.4 for country H to a high of 71.9 for country J. Two of the countries' minimum scores (countries J and O) were higher than the average maximum score (52.7), and one country's maximum score (country M) was lower than the average minimum score (42.8). Differences between minimum and maximum scores for each country ranged from 5.6 points (country C) to 16.1 points (country J). The average difference was almost 10 points.

Results of country ranks on tests are presented in Table 42. The second column shows each country's rank on the field-trial instrument, and the third column shows each country's average rank across all specially-constructed tests. On average, not much difference in ranks existed between the field-trial instrument and other tests. Most differences for countries ranged from less one than to slightly more than one rank. Standard deviations of ranks across tests for each country were around one to two ranks. One country (M) had a standard deviation of .20 ranks. Country G had the highest standard deviation (1.9 ranks).

Differences between minimum and maximum ranks showed much more variability than the averages did. No country received the same rank across all tests. However, two countries had a difference of only one rank across all tests. One country (country J) fluctuated between the first two ranks, while the other country (country M) fluctuated between the last two ranks. Three of the countries (E, G, and Q) had differences that were eight or more ranks out of 17. Six additional countries had differences of five or more ranks. The average difference between minimum and maximum ranks was 4.8.

Table 42

Summary of Country Ranks on Field-Trial Instrument and across Specially-Constructed Tests

		Across Tests				
Country	Field Trial	AVE	*SD*	MIN	MAX	DIF
A	9	10.6	0.9	9	13	4.0
B	3	3.3	0.6	2	4	2.0
C	6	6.5	0.6	5	8	3.0
D	14	14.8	0.8	13	16	3.0
E	12	13.2	1.6	8	16	8.0
F	10	9.2	1.1	7	12	5.0
G	11	10.8	1.9	5	13	8.0
H	16	14.4	1.3	11	17	6.0
I	7	7.0	1.4	6	12	6.0
J	1	1.3	0.4	1	2	1.0
K	5	4.0	0.9	3	7	4.0
L	8	8.3	1.1	4	10	6.0
M	17	17.0	0.2	16	17	1.0
N	15	11.0	1.9	6	13	7.0
O	2	1.8	0.6	1	3	2.0
P	13	15.0	1.5	10	16	6.0
Q	4	5.1	1.6	3	12	9.0
AVE	9.0	9.0	1.1	6.5	11.2	4.8
SD	4.9	4.8	0.5	4.1	4.7	2.4

Tables 43 and 44 show correlations between country performance on the field-trial instrument and on each of the specially-constructed tests. Table 43 shows score correlations using a Pearson product-moment correlation; Table 44 shows rank correlations using the Spearman rank-order correlation. Average correlations in both cases were quite high and significant ($p < .01$ in all cases). Only the correlation between field-trial scores and the expert-mapping strict-intersection-test scores was under .90, and only three of the rank correlations were below .90. These were for the expert mapping strict-intersection test (.85), the unweighted curriculum-guide strict-intersection test (.83), and the weighted textbook strict-intersection test (.85).

Performance Differences

I computed differences between each country score on the field-trial instrument and their score on each specially-constructed test, and I did the same with the ranks. Summaries are in Tables 45 to 48. Positive differences indicate higher performance on the field-trial instrument than on the specially-constructed test; negative differences indicate the opposite.

Tables 45 and 46 present the summary results for the score differences. Table 45 presents results for tests, and Table 46 presents results for countries. Most countries had positive score differences (i.e., higher performance on the filed-trial instrument). The main exception was for the strict-intersection test based on the curriculum guide. Differences were split in half for this test. The highest average *absolute* score differences were between the field-trial instrument and the weighted strict-intersection test based on the textbook data (6.32). This test also had high average positive differences and high

Table 43

Correlations between Country Scores on the the Field-Trial Instrument and Scores on Each Specially-Constructed Test

Test	Correlation
UNION	0.98
EX1-7I	0.96
EX1-SI	0.87
CG-71	0.98
CG-SI	0.90
TX-7I	0.98
TX-SI	0.94
AG1-UN	0.97
AG-71	0.97
WEX-UN	0.98
WEX-7I	0.97
WCG-UN	0.99
WCG-7I	0.98
WTX-UN	0.97
WTX-7I	0.97
WTX-SI	0.90
WAG-UN	0.98
WAG-7I	0.96
EX-UQ	0.97
WEX-UQ	0.97
CG-UQ	0.97
TX-UQ	0.99
WTX-UQ	0.93
AG-UQ	0.99
WAG-UQ	0.95
Average	0.96

Note. All corelations are significatn, $p < .01$.

Table 44

Correlations between Country Ranks on the the Field-Trial Instrument and Ranks on Each Specially-Constructed Test

Test	Correlation
UNION	0.97
EX-7I	0.94
EX-SI	0.85
CG-71	0.97
CG-SI	0.83
TX-7I	0.96
TX-SI	0.94
AG-71	0.96
WEX-UN	0.96
WEX-7I	0.94
WCG-UN	0.97
WCG-7I	0.96
WTX-UN	0.94
WTX-7I	0.94
WTX-SI	0.85
WAG-UN	0.96
WAG-7I	0.94
EX-UQ	0.95
WEX-UQ	0.94
CG-UQ	0.92
TX-UQ	0.96
WTX-UQ	0.90
AG-UQ	0.95
WAG-UQ	0.91
Average	0.93

Note. All correlations are significant, $p < .01$.

Table 45

Summary of Differences in Scores on the the Field-Trial Instrument and Scores on Each Specially-Constructed Test

Test	AVE ABS[a]	*SD*	MIN	MAX	Ave +[b]	Ave -[c]	Count +[b]	Count -[c]
UNION	2.17	1.46	0.19	6.65	2.29	-0.01	16	1
EX-7I	4.09	1.88	1.57	8.29	4.09	0	17	0
EX-SI	4.66	3.67	0.32	14.12	4.87	-0.79	14	3
CG-71	4.38	1.30	2.18	6.49	4.38	0	17	0
CG-SI	2.76	2.13	0.20	8.49	1.94	-3.92	8	9
TX-7I	1.37	1.13	0.08	4.42	1.66	-0.13	13	4
TX-SI	4.44	1.90	1.67	9.34	4.74	-0.28	15	2
AG-71	3.31	1.60	0.59	6.29	3.45	-0.07	16	1
WEX-UN	3.08	1.47	0.85	6.69	3.08	0	17	0
WEX-7I	3.82	1.80	1.44	7.67	3.82	0	17	0
WCG-UN	2.86	1.15	0.67	5.09	2.86	0	17	0
WCG-7I	3.99	1.35	1.36	6.03	3.99	0	17	0
WTX-UN	3.60	1.53	0.81	7.41	3.78	-0.05	16	1
WTX-7I	3.07	1.53	1.16	7.22	3.15	-0.11	16	1
WTX-SI	6.32	2.85	0.02	13.41	6.93	-0.24	15	2
WAG-UN	3.26	1.35	0.07	6.36	3.46	0.00	16	1
WAG-7I	4.42	1.88	0.01	7.34	4.96	-0.05	15	2
EX-UQ	2.41	1.39	0.35	4.92	2.58	-0.15	15	2
WEX-UQ	2.35	1.48	0.23	5.79	2.37	-0.12	16	1
CG-UQ	2.38	1.44	0.20	5.18	2.65	-0.05	15	2
TX-UQ	1.72	0.97	0.46	3.46	1.72	0	17	0
WTX-UQ	2.98	1.54	0.80	5.76	3.25	-0.65	13	4
AG-UQ	2.04	1.02	0.08	3.90	2.14	-0.02	16	1
WAG-UQ	2.23	1.28	0.54	5.02	2.28	-0.87	12	5
Average	3.24	1.63	0.66	6.89	3.35	-0.31	15.25	1.75
SD	1.1	0.6	0.6	2.5	1.2	0.8	2.0	2.0

[a]Average of the absolute value of the differences. [b]Average/Number of positive differences. [c]Average/Number of negative differences.

Table 46

Summary of Differences in Scores on the the Field-Trial Instrument and Scores on Each Specially-Constructed Test for Each Country

Country	AVE ABS[a]	*SD*	MIN	MAX	Ave +[b]	Ave -[c]	Count +[b]	Count -[c]
A	4.11	1.61	0.98	8.37	4.11	0	24	0
B	2.23	1.61	0.22	5.77	2.33	-1.86	19	5
C	3.16	1.47	0.51	6.06	3.16	0	24	0
D	4.37	1.58	2.33	9.12	4.37	0	24	0
E	3.57	1.25	0.84	7.02	3.70	-2.12	22	2
F	1.73	1.42	0.19	7.11	1.94	-0.27	21	3
G	3.61	1.84	0.08	6.42	3.70	-3.14	20	4
H	2.87	3.01	0.20	14.12	2.98	-0.20	23	1
I	2.95	1.60	0.79	7.91	2.95	0	24	0
J	3.36	2.39	0.02	8.29	3.85	-2.36	16	8
K	2.53	1.72	0.13	7.41	2.78	-1.29	20	4
L	2.88	1.44	1.21	7.34	2.98	-1.76	22	2
M	2.88	0.96	1.42	5.63	2.88	-2.86	23	1
N	1.71	1.70	0.01	8.49	1.38	-2.16	14	10
O	3.64	1.69	0.08	6.61	3.79	-0.08	23	1
P	5.72	2.62	1.61	13.41	5.90	-1.61	23	1
Q	3.74	1.88	0.41	10.83	3.74	0	24	0
AVE	3.24	1.75	0.65	8.23	3.33	-1.16	21.53	2.47
SD	0.95	0.49	0.65	2.39	1.00	1.11	2.85	2.85

[a]Average of the absolute value of the differences. [b]Average/Number of positive differences. [c]Average/Number of negative differences.

maximum differences. High maximum differences also existed for the strict-intersection test based on the expert-mapping data. The strict-intersection test for the curriculum guide had an average negative score difference of -3.9. This was higher than any of the other average negative score differences. Lower average *absolute* score differences were associated with the unweighted 70%-intersection test based on the textbook data (1.37) and the unweighted unique test based on the textbook data (1.72).

The overall average of average *absolute* score differences was around 3 points. The overall average of average positive score differences was about the same; the average of average negative score differences was only around -1. Across the 36 tests, most differences were positive, indicating higher scores on the field-trial instrument than other tests. Country N had the largest number of negative differences (10). Average *absolute* score differences ranged from around two to six points, and standard deviations were around two points. Score differences ranged from a minimum of less than one point to 14 points. High differences were found for countries H, P, and Q.

Tables 47 and 48 present summary information on the differences in ranks. Test information was in Table 47. All but two average *absolute* rank differences across countries within tests were around one rank or less. The two exceptions were the strict-intersection test based on the expert mapping (1.8) and the strict-intersection test based on the curriculum guides (1.9). Most differences were fairly evenly distributed among positive differences, negative differences, and no differences. The exceptions were for the unweighted 70%-intersection test based on the expert mapping (only two zero differences), the strict-intersection test based on the curriculum guides (only 3 zero differences), the unweighted strict-intersection test based on the textbook (2 positive and

Table 47

Summary of Differences in Ranks on the the Field-Trial Instrument and Ranks on Each Specially-Constructed Test for Each Test

Test	AVE ABS[a]	*SD*	MAX	Ave +[b]	Ave -[c]	Count +[b]	Count -[c]	Count 0[d]
UNION	0.9	0.9	3.0	1.3	-1	6	6	5
EX-7I	1.4	1.0	4.0	2.0	-2	6	9	2
EX-SI	1.8	2.0	8.0	2.5	-3	6	5	6
CG-71	0.9	0.7	2.0	1.6	-2	5	7	5
CG-SI	1.9	2.1	9.0	2.7	-3	6	8	3
TX-7I	1.1	0.9	3.0	1.8	-2	5	7	5
TX-SI	0.8	1.5	6.0	3.5	-4	2	4	11
AG-71	1.1	1.0	4.0	2.3	-2	4	8	5
WEX-UN	0.9	0.9	3.0	1.6	-2	5	6	6
WEX-7I	1.3	1.1	4.0	2.2	-2	5	8	4
WCG-UN	0.8	0.9	3.0	1.4	-1	5	5	7
WCG-7I	1.1	0.9	3.0	1.8	-2	5	7	5
WTX-UN	1.1	1.3	5.0	2.3	-2	4	6	7
WTX-7I	1.1	1.3	5.0	2.3	-2	4	6	7
WTX-SI	1.8	2.1	9.0	3.8	-4	4	9	4
WAG-UN	0.9	1.1	4.0	2.0	-2	4	6	7
WAG-7I	1.2	1.2	5.0	2.0	-2	5	7	5
EX-UQ	1.1	1.1	3.0	1.8	-2	5	5	7
WEX-UQ	1.2	1.3	3.0	2.5	-3	4	5	8
CG-UQ	1.4	1.4	5.0	2.4	-2	5	7	5
TX-UQ	0.8	1.0	3.0	1.8	-2	4	4	9
WTX-UQ	1.5	1.6	6.0	2.2	-2	6	6	5
AG-UQ	1.1	1.1	4.0	1.5	-2	6	5	6
WAG-UQ	1.4	1.5	6.0	3.0	-3	4	8	5
Average	1.2	1.3	4.6	2.2	-2.2	4.8	6.4	5.8
SD	0.3	0.4	1.9	0.6	0.6	1.0	1.4	1.9

[a]Average of the absolute value of the differences. [b]Average/Number of positive differences. [c]Average/Number of negative differences. [d]Number of ranks with no difference.

11 zero differences), and the unweighted unique test based on the textbook (9 zero differences). Some tests with higher maximum differences were the strict-intersection test for the expert mapping (8), the strict-intersection test based on the curriculum guides (9), and the weighted strict-intersection test based on the textbook (9).

The country information is in Table 48. It also shows minimal differences in ranks across tests. The average of average *absolute* rank differences was one rank. Most of the average absolute rank differences for each country were one rank or less. The exceptions were country H (1.7), country N (4.2), and country P (2.5). Large maximum rank differences were found for country N (9) and country Q (8). Countries B, C, J, L, M, and O had more zero differences than any other difference. Countries D, E, G, P, and Q had more negative rank differences, and countries F, H, K, and N had more positive rank differences. Country I had about as many positive rank differences as non-differences. Countries I and P had higher average negative rank difference than the other countries (-2.5 and -2.6 respectively), and country N had higher positive rank differences (4.0).

Variations in Topic Performance

Little variation existed within countries across total scores and ranks on the specially-constructed tests. However, significant variation did exist when looking at scores on individual topics. Table 49 presents the country scores on each topic. Within countries, standard deviation of topic scores ranged from 9 to 16 points, with an average of 9. Differences between minimum and maximum topic scores for a country were from around 30 to up to 70 points. Variations in scores for each topic across countries also

Table 48

Summary of Differences in Ranks on the the Field-Trial Instrument and Ranks on Each Specially-Constructed Test for Each Country

Country	AVE ABS[a]	*SD*	MAX	Ave +[b]	Ave -[c]	Count +[b]	Count -[c]	Count 0[d]
A	1.6	0.9	4.0	0.0	-1.7	0	23	1
B	0.4	0.5	1.0	1.0	-1.0	2	8	14
C	0.6	0.6	2.0	1.0	-1.1	1	11	12
D	0.8	0.7	2.0	1.0	-1.3	1	15	8
E	1.6	1.0	4.0	2.0	-1.8	3	20	1
F	1.2	0.8	3.0	1.5	-1.3	16	3	5
G	1.1	1.0	3.0	2.7	-1.3	7	10	7
H	1.7	1.2	5.0	2.1	-1.0	19	1	4
I	0.9	1.1	5.0	1.0	-2.5	10	4	10
J	0.3	0.4	1.0	0.0	-1.0	0	6	18
K	1.0	0.6	2.0	1.3	-2.0	19	1	4
L	0.7	0.9	4.0	2.5	-1.1	2	10	12
M	0.0	0.2	1.0	1.0	0.0	1	0	23
N	4.2	1.9	9.0	4.0	0.0	24	0	0
O	0.4	0.5	1.0	1.0	-1.0	6	2	16
P	2.5	0.8	3.0	2.0	-2.6	2	20	2
Q	1.2	1.5	8.0	1.0	-1.5	2	20	2
AVE	1.2	0.9	3.4	1.5	-1.3	6.8	9.1	8.2
SD	1.0	0.4	2.3	1.0	0.7	7.6	7.7	6.6

[a]Average of the absolute value of the differences. [b]Average/Number of positive differences. [c]Average/Number of negative differences. [d]Number of ranks with no difference.

Table 49

Topic Scores for Each Country

	Country																				
Topic	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	AVE	SD	MIN	MAX
Test	50.5	56.4	53.6	45.2	45.9	49.6	48.1	43.5	52.8	64.0	56.0	51.5	35.4	45.0	61.9	45.8	56.4	50.7	6.9	35.4	64.0
1.1.1.1	63.3	66.2	59.0	51.2	47.1	57.5	59.4	45.2	63.8	60.6	65.7	61.7	42.6	55.0	57.8	65.1	58.8	57.6	7.0	42.6	66.2
1.1.1.2	53.1	62.9	59.7	46.1	46.8	47.8	55.2	43.4	57.4	60.6	54.3	53.7	35.3	48.5	65.5	44.8	57.4	52.5	7.7	35.3	65.5
1.1.1.3	45.1	57.3	58.6	33.9	50.5	42.9	63.7	49.8	40.0	69.5	45.7	36.0	25.7	61.9	66.4	24.2	69.6	49.4	14.1	24.2	69.6
1.1.2.1	49.7	56.4	54.4	43.1	39.9	42.0	48.1	41.3	52.8	63.1	52.3	44.5	26.0	44.2	67.8	39.4	55.2	48.2	9.7	26.0	67.8
1.1.2.2	52.0	58.2	60.5	50.7	52.9	56.5	52.9	49.9	54.8	71.9	61.5	56.3	37.4	48.0	72.6	55.2	61.2	56.0	8.1	37.4	72.6
1.1.2.3	52.7	56.4	58.8	51.8	46.4	48.1	50.6	48.3	54.4	64.8	54.6	48.4	37.8	45.5	69.2	44.4	60.0	52.5	7.6	37.8	69.2
1.1.2.4	29.1	34.4	32.7	24.5	29.0	32.0	19.5	24.3	38.6	36.4	36.5	36.5	18.3	28.4	57.6	25.5	34.3	31.6	8.7	18.3	57.6
1.1.3.1	51.5	61.5	50.9	43.4	37.3	60.7	36.0	39.0	66.6	61.2	70.7	48.0	28.2	38.6	55.2	42.9	62.5	50.2	12.0	28.2	70.7
1.1.4.2	39.7	76.0	68.5	40.6	36.0	60.9	61.1	65.3	54.6	20.3	68.1	52.2	36.1	62.1	75.8	26.9	65.4	53.5	16.6	20.3	76.0
1.1.4.4	30.2	31.6	14.4	21.5	18.9	28.2	12.0	16.7	44.3	21.2	25.9	25.0	5.7	13.8	29.7	12.5	35.4	22.8	9.6	5.7	44.3
1.1.5.1	64.5	66.5	58.8	60.0	62.8	59.5	53.0	52.1	67.3	64.4	61.6	65.9	53.6	51.4	63.7	69.0	68.5	61.3	5.7	51.4	69.0
1.1.5.2	44.8	58.0	47.2	40.7	39.2	49.5	42.0	44.7	55.4	35.3	51.9	47.7	31.8	42.4	56.4	36.1	56.4	45.9	7.7	31.8	58.0
1.1.5.3	61.7	62.4	54.5	52.0	51.5	55.2	47.4	52.0	65.1	55.1	61.0	71.1	40.0	46.9	66.7	64.6	65.6	57.2	8.2	40.0	71.1
1.1.5.4	38.1	55.2	59.3	50.0	20.8	58.3	47.1	55.2	54.7	14.0	48.4	60.7	33.8	48.8	67.7	16.7	51.9	45.9	15.4	14.0	67.7
1.2.1	63.2	73.6	73.8	57.4	65.4	62.5	61.0	56.3	63.4	72.3	71.8	64.9	48.2	54.4	71.4	65.8	60.6	63.9	7.1	48.2	73.8
1.2.2	37.0	49.7	41.3	30.3	34.1	40.5	32.2	31.3	38.7	56.4	46.0	40.6	23.4	39.5	60.0	34.4	38.8	39.7	9.0	23.4	60.0
1.2.3	59.9	73.8	53.6	58.7	71.8	75.0	53.0	34.6	63.3	56.8	77.6	77.9	52.7	38.3	65.1	69.4	57.4	61.1	12.3	34.6	77.9
1.3.1	46.7	46.9	34.5	36.2	46.0	55.9	33.6	42.2	43.1	64.6	62.5	38.4	29.2	35.3	49.3	27.5	53.7	43.9	10.6	27.5	64.6
1.3.2	45.6	55.9	50.8	40.6	47.3	48.1	46.2	29.4	46.6	71.9	52.2	53.3	31.9	44.6	56.8	37.8	45.5	47.3	9.5	29.4	71.9
1.3.3	34.3	44.5	41.5	30.0	40.4	38.9	27.2	42.7	48.3	69.9	42.1	45.1	28.8	45.7	54.7	35.0	38.5	41.6	9.9	27.2	69.9
1.3.4	61.6	73.8	60.7	54.1	62.0	62.1	58.8	47.4	57.1	73.9	70.0	66.1	49.8	56.7	65.6	53.8	60.1	60.8	7.3	47.4	73.9
1.4.1	47.4	52.6	52.5	38.6	42.5	50.5	42.0	34.6	42.8	68.4	56.3	47.5	34.7	44.2	54.5	35.8	49.7	46.7	8.6	34.6	68.4
1.4.2	32.1	47.6	38.2	31.5	34.9	36.0	36.2	39.1	41.7	69.0	46.6	42.6	30.3	36.6	48.9	37.8	41.2	40.6	8.9	30.3	69.0
1.5.1	32.9	32.7	43.2	32.3	28.1	30.5	32.9	33.2	32.9	54.8	38.9	33.0	21.0	35.8	54.2	29.5	42.1	35.7	8.4	21.0	54.8
1.5.2	39.9	44.9	47.0	37.8	37.4	37.8	39.1	35.9	42.3	54.7	43.5	39.8	29.5	39.4	57.0	33.6	45.8	41.5	6.8	29.5	57.0
1.6.1	54.9	47.8	54.9	47.1	42.5	53.7	48.3	39.2	56.0	65.6	54.6	48.9	38.0	45.9	56.7	44.0	60.2	50.5	7.2	38.0	65.6
1.6.2	37.5	50.5	44.5	31.7	35.7	37.7	44.9	39.3	41.1	61.5	45.8	40.3	26.8	50.3	54.5	27.6	51.8	42.4	9.2	26.8	61.5
1.7.1	64.6	66.6	61.8	58.4	58.5	63.3	58.2	51.9	62.2	69.8	67.1	60.5	52.6	42.2	62.3	64.4	68.9	60.8	6.7	42.2	69.8
1.7.2	49.8	45.4	43.5	42.7	40.4	51.1	47.4	40.7	50.4	55.3	49.5	51.7	31.8	35.3	45.8	48.2	55.2	46.1	6.3	31.8	55.3
AVE	47.7	55.5	51.0	42.7	43.7	49.7	45.1	42.2	51.7	57.4	54.6	50.3	33.8	44.1	59.6	41.8	54.2	48.5	9.2	31.2	66.2
SD	11.0	11.9	11.8	10.4	12.4	11.3	12.5	10.0	9.6	16.0	11.8	12.1	10.7	9.7	9.3	15.7	10.1	9.4	2.6	10.2	7.3
MIN	29.1	31.6	14.4	21.5	18.9	28.2	12.0	16.7	32.9	14.0	25.9	25.0	5.7	13.8	29.7	12.5	34.3	22.8	5.7	5.7	44.3
MAX	64.6	76.0	73.8	60.0	71.8	75.0	63.7	65.3	67.3	73.9	77.6	77.9	53.6	62.1	75.8	69.4	69.6	63.9	16.6	51.4	77.9

existed. Standard deviations ranged from 6 to 17 points, with an average of 9. Differences between minimum and maximum scores within each topic were around 30 to 40 points.

Table 50 shows country ranks on each topic. Again, much variability existed. Nine countries had a rank of one on at least one topic. This meant that these nine countries had better performance than all other countries on at least one topic. Six countries ranked last on at least one topic. Two countries had ranks that placed them first on at least one topic and last on at least one. Standard deviations of ranks were larger across topics than they were across tests. They ranged from two to five places. Aside from country M, the lowest difference between minimum and maximum ranks was six places, with the next lowest being 10. All but four of the average ranks and five of the modal ranks differed from the same country's rank on the field-trial instrument.

Table 51 shows differences between the field-trial instrument total score and each topic score for each country. The average of *absolute* score differences was eight points, with a standard deviation of about four. The average minimum difference was 1.8 points; the average maximum was 20 points. Average positive differences (score higher on field-trial instrument) were larger than the average negative differences (score lower on field-trial instrument). The average of the positive differences was 7 points; the average of the negative differences was 5 points. The average number of countries with positive differences within a topic was nine; the average number of countries with negative differences was eight. Table 52 presents the same data for each country. This table reports differences that range from almost nothing to 50 points. Numbers and averages of positive and negative differences were about the same across countries.

Table 50

Country Ranks on Each Topic

	Country																
Topic	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
Test	9	3	6	14	12	10	11	16	7	1	5	8	17	15	2	13	4
1.1.1.1	5	1	9	14	15	12	8	16	4	7	2	6	17	13	11	3	10
1.1.1.2	10	2	4	14	13	12	7	16	5	3	8	9	17	11	1	15	6
1.1.1.3	11	7	6	15	8	12	4	9	13	2	10	14	16	5	3	17	1
1.1.2.1	8	3	5	12	15	13	9	14	6	2	7	10	17	11	1	16	4
1.1.2.2	13	6	5	14	11	7	12	15	10	2	3	8	17	16	1	9	4
1.1.2.3	8	5	4	9	14	13	10	12	7	2	6	11	17	15	1	16	3
1.1.2.4	10	6	8	14	11	9	16	15	2	5	3	4	17	12	1	13	7
1.1.3.1	8	4	9	11	15	6	16	13	2	5	1	10	17	14	7	12	3
1.1.4.2	13	1	3	12	15	9	8	6	10	17	4	11	14	7	2	16	5
1.1.4.4	4	3	13	9	11	6	16	12	1	10	7	8	17	14	5	15	2
1.1.5.1	6	4	13	11	9	12	15	16	3	7	10	5	14	17	8	1	2
1.1.5.2	9	1	8	13	14	6	12	10	4	16	5	7	17	11	2	15	2
1.1.5.3	7	6	11	12	14	9	15	13	4	10	8	1	17	16	2	5	3
1.1.5.4	13	5	3	9	15	4	12	5	7	17	11	2	14	10	1	16	8
1.2.1	10	2	1	14	7	11	12	15	9	3	4	8	17	16	5	6	13
1.2.2	11	3	5	16	13	7	14	15	10	2	4	6	17	8	1	12	9
1.2.3	9	4	13	10	5	3	14	17	8	12	2	1	15	16	7	6	11
1.3.1	7	6	14	12	8	3	15	10	9	1	2	11	16	13	5	17	4
1.3.2	11	3	6	14	8	7	10	17	9	1	5	4	16	13	2	15	12
1.3.3	14	6	9	15	10	11	17	7	3	1	8	5	16	4	2	13	12
1.3.4	8	2	9	14	7	6	11	17	12	1	3	4	16	13	5	15	10
1.4.1	9	4	5	14	12	6	13	17	11	1	2	8	16	10	3	15	7
1.4.2	15	3	9	16	14	13	12	8	6	1	4	5	17	11	2	10	7
1.5.1	9	12	3	13	16	14	9	7	11	1	5	8	17	6	2	15	4
1.5.2	8	5	3	12	14	13	11	15	7	2	6	9	17	10	1	16	4
1.6.1	5	11	6	12	15	8	10	16	4	1	7	9	17	13	3	14	2
1.6.2	13	4	8	15	14	12	7	11	9	1	6	10	17	5	2	16	3
1.7.1	5	4	10	13	12	7	14	16	9	1	3	11	15	17	8	6	2
1.7.2	6	11	12	13	15	4	9	14	5	1	7	3	17	16	10	8	2
AVE	9.1	4.6	7.4	12.8	12.1	8.8	11.7	12.9	6.9	4.7	5.3	7.2	16.3	11.8	3.6	12.2	5.6
SD	2.9	2.8	3.5	1.9	3.0	3.3	3.2	3.6	3.2	5.1	2.6	3.3	1.0	3.7	2.9	4.6	3.6
MIN	4	1	1	9	5	3	4	5	1	1	1	1	14	4	1	1	1
MAX	15	12	14	16	16	14	17	17	13	17	11	14	17	17	11	17	13
MODE	8	4	9	14	15	12	12	16	9	1	2	8	17	13	1	15	2

Table 51

Summary of Differences in Scores on the Field-Trial Instrument and Scores on Each Topic for Each Topic

Topic	AVE ABS[a]	*SD*	MIN	MAX	Ave +[b]	Ave -[c]	Count +[b]	Count -[c]
1.1.1.1	7.83	4.54	1.13	19.28	3.77	-8.37	2	15
1.1.1.2	2.76	2.16	0.09	7.02	1.33	-3.54	6	11
1.1.1.3	9.75	5.42	0.86	21.65	11.69	-8.03	8	9
1.1.2.1	3.25	3.04	0.04	9.42	3.72	-1.70	13	4
1.1.2.2	5.34	2.62	1.50	10.72	0	-5.34	0	17
1.1.2.3	2.67	2.10	0.04	7.30	1.47	-3.17	5	12
1.1.2.4	19.06	5.26	4.30	28.62	19.06	0	17	0
1.1.3.1	6.54	4.12	0.94	14.70	5.39	-8.63	11	6
1.1.4.2	13.19	9.85	0.65	43.71	17.60	-11.35	5	12
1.1.4.4	27.92	7.83	8.47	42.81	27.92	0	17	0
1.1.5.1	10.66	5.97	0.41	23.17	0	-10.66	0	17
1.1.5.2	5.46	6.31	0.02	28.70	6.72	-1.39	13	4
1.1.5.3	7.68	5.24	0.78	19.57	4.83	-8.06	2	15
1.1.5.4	10.84	12.43	1.04	50.01	14.73	-6.46	9	8
1.2.1	13.20	4.18	4.26	20.19	0	-13.20	0	17
1.2.2	11.02	3.84	1.93	17.52	11.02	0	17	0
1.2.3	13.10	8.77	0.03	26.37	7.57	-14.29	3	14
1.3.1	8.42	5.64	0.04	19.10	9.97	-3.37	13	4
1.3.2	4.66	3.67	0.32	14.12	4.87	-3.66	14	3
1.3.3	9.85	5.63	0.77	20.95	10.72	-3.34	15	2
1.3.4	10.11	4.35	3.65	17.43	0	-10.11	0	17
1.4.1	4.61	3.19	0.33	9.99	5.19	-1.89	14	3
1.4.2	10.68	3.83	4.39	18.40	11.03	-5.01	16	1
1.5.1	14.93	4.33	7.68	23.74	14.93	0	17	0
1.5.2	9.18	2.40	4.91	12.43	9.18	0	17	0
1.6.1	3.03	1.96	0.20	8.62	3.91	-2.41	7	10
1.6.2	8.87	4.11	2.52	18.24	9.09	-5.30	16	1
1.7.1	10.42	4.51	0.41	18.61	2.73	-10.90	1	16
1.7.2	5.04	4.44	0.25	16.12	5.82	-1.39	14	3
Average	9.31	4.89	1.79	20.29	7.73	-5.23	9.38	7.62
SD	5.2	2.3	2.3	10.4	6.5	4.2	6.3	6.3

[a]Average of the absolute value of the differences. [b]Average/Number of positive differences. [c]Average/Number of negative differences.

Table 52

Summary of Differences in Scores on the Field-Trial Instrument and Scores on Each Topic for Each Country

Country	AVE ABS[a]	*SD*	MIN	MAX	Ave +[b]	Ave -[c]	Count +[b]	Count -[c]
A	9.20	6.27	0.55	21.39	10.52	-8	17	12
B	9.14	7.28	0.04	24.79	9.99	-8.84	15	14
C	8.72	8.09	0.03	39.20	12.01	-6	14	15
D	8.83	5.82	0.52	23.73	9.97	-8	17	12
E	9.69	7.72	0.04	27.04	11.12	-8.60	16	13
F	9.13	6.30	0.02	25.48	9.56	-9.29	14	15
G	9.42	8.38	0.09	36.14	10.86	-8.13	17	12
H	7.63	6.37	0.09	26.81	7.82	-7.98	17	12
I	8.01	5.14	0.06	19.92	10.42	-7	13	16
J	10.72	13.29	0.41	50.01	16.05	-4.91	16	13
K	9.52	6.78	0.33	30.06	9.57	-10.20	17	12
L	9.63	7.06	0.25	26.50	10.11	-9.74	16	13
M	8.19	6.82	0.11	29.75	8.10	-8.99	18	11
N	6.93	6.65	0.32	31.17	7.72	-6.54	15	14
O	7.23	5.98	0.37	32.20	8.84	-5.79	16	13
P	13.44	8.57	0.34	33.30	12.98	-15.90	20	9
Q	7.91	6.34	0.06	22.10	10.71	-6	14	15
AVE	9.02	7.23	0.21	29.39	10.37	-8.20	16.00	13.00
SD	1.46	1.76	0.18	7.23	1.96	2.43	1.68	1.68

[a]Average of the absolute value of the differences. [b]Average/Number of positive differences. [c]Average/Number of negative differences.

Table 53 presents a summary of differences between each country's rank on the field-trial instrument and its rank on each topic. The average of absolute differences across topics was 2.7 ranks. The average maximum difference was 8 ranks and the average minimum difference was 3. An average of three countries for each topic had no difference in ranks. Table 54 shows the summary across countries. The average number of topics on which countries had no difference was five.

Table 55 reports topic ranks within countries. Seven of the topics had ranks of one for at least one country. No topic had an average rank of 1. The highest average ranks were for topics 1.1.5.1 *Estimating Quantity and Size* (5 - out of 29), 1.2.2 *Perimeter, Area, Volume* (3.6), 1.3.4 *3-D Geometry* (5.4), and 1.7.1 *Data Representation and Analysis* (5.9). The lowest average ranks were for 1.1.2.4 *Percentages* (26.8), 1.1.4.4 *Number Theory* (28.2), and 1.5.1 *Proportionality Concepts* (25.6).

Performance Expectations

The TIMSS mathematics framework code not only contained codes for topics, but also included codes for expected performance (See Appendix A). Textbook blocks were coded with topic and performance-expectation codes as was each test item. Therefore, country performance can also be evaluated in light of the performance expectations and the combination of topic by performance expectation.

Table 56 presents the proportion of textbook blocks devoted to each performance expectation. The highest of the average proportions was devoted to 2.1.3 *Recalling Mathematical Objects and Property* (.313). This was followed by 2.2.2 *Performing Routine Procedures* (.294), 2.3.3 *Solving Problems* (.114), and 2.1.1 *Representing* (.112).

Table 53

Summary of Differences in Ranks on the Field-Trial Instrument and Ranks on Each Topic for Each Topic

Test	AVE ABS[a]	*SD*	MAX	Ave +[b]	Ave -[c]	Count +[b]	Count -[c]	Count 0[d]
1.1.1.1	3.4	2.8	10.0	3.6	-4	8	6	3
1.1.1.2	1.6	1.2	4.0	2.3	-2	6	8	3
1.1.1.3	3.8	2.7	10.0	5.3	-5	6	10	1
1.1.2.1	1.6	1.1	4.0	1.8	-2	8	6	3
1.1.2.2	1.5	1.3	4.0	1.9	-2	7	6	4
1.1.2.3	1.8	1.4	5.0	2.1	-2	7	7	3
1.1.2.4	2.1	1.6	5.0	2.3	-2	8	6	3
1.1.3.1	2.7	1.6	5.0	2.6	-3	9	7	1
1.1.4.2	3.9	3.9	16.0	3.7	-4	9	7	1
1.1.4.4	3.3	2.6	9.0	3.5	-4	8	6	3
1.1.5.1	3.9	2.7	12.0	4.1	-4	8	8	1
1.1.5.2	2.6	3.5	15.0	2.9	-3	8	5	4
1.1.5.3	3.2	2.6	9.0	3.4	-3	8	7	2
1.1.5.4	4.6	3.8	16.0	5.0	-5	8	8	1
1.2.1	2.4	2.5	9.0	3.3	-3	6	8	3
1.2.2	2.0	1.7	7.0	2.1	-2	8	7	2
1.2.3	4.4	3.1	11.0	5.3	-5	7	9	1
1.3.1	3.2	2.1	8.0	3.4	-3	8	7	2
1.3.2	1.8	2.0	8.0	2.5	-3	6	5	6
1.3.3	3.5	3.2	11.0	5.0	-5	6	8	3
1.3.4	2.4	1.8	6.0	2.5	-3	8	6	3
1.4.1	1.6	1.6	5.0	2.8	-3	5	7	5
1.4.2	2.4	2.1	8.0	3.3	-3	6	7	4
1.5.1	2.8	3.2	9.0	4.8	-5	5	5	7
1.5.2	1.5	1.3	5.0	2.2	-2	6	7	4
1.6.1	1.9	1.9	8.0	2.3	-2	7	6	4
1.6.2	2.4	2.4	10.0	5.0	-5	4	10	3
1.7.1	2.5	1.9	7.0	3.0	-3	7	7	3
1.7.2	3.3	2.5	8.0	3.1	-3	9	6	2
Average	2.7	2.3	8.4	3.3	-3.3	7.1	7.0	2.9
SD	0.9	0.8	3.3	1.1	1.1	1.3	1.3	1.5

[a]Average of the absolute value of the differences. [b]Average/Number of positive differences. [c]Average/Number of negative differences. [d]Number of ranks with no difference.

Table 54

Summary of Differences in Ranks on the Field-Trial Instrument and Ranks on Each Topic for Each Country

Country	AVE ABS[a]	*SD*	MAX	Ave +[b]	Ave -[c]	Count +[b]	Count -[c]	Count 0[d]
A	2.3	1.7	6.0	0.0	-3.0	13	12	4
B	2.2	2.3	9.0	1.5	-3.1	6	18	5
C	3.0	2.2	8.0	2.3	-4.3	11	15	3
D	1.6	1.5	5.0	2.6	-1.4	16	5	8
E	2.6	1.5	7.0	3.4	-2.4	11	16	2
F	3.0	1.7	7.0	3.7	-2.3	17	12	0
G	2.6	1.8	7.0	2.7	-3.1	11	16	2
H	3.3	3.4	11.0	4.7	-1.0	20	4	5
I	2.7	1.7	6.0	3.2	-3.0	13	13	3
J	3.5	5.0	16.0	0.0	-5.9	0	18	11
K	2.1	1.5	6.0	2.2	-2.8	13	13	3
L	2.6	2.1	7.0	3.9	-2.5	13	11	5
M	0.6	1.0	3.0	1.7	0.0	11	0	18
N	3.7	3.1	11.0	4.8	0.0	21	7	1
O	2.1	2.5	9.0	1.0	-4.2	8	13	8
P	3.6	2.8	12.0	5.9	-2.6	11	16	2
Q	2.8	2.6	9.0	1.7	-5.0	11	13	5
AVE	2.6	2.3	8.2	2.7	-2.7	12.1	11.9	5.0
SD	0.7	0.9	3.0	1.6	1.5	4.8	4.9	4.2

[a]Average of the absolute value of the differences. [b]Average/Number of positive differences. [c]Average/Number of negative differences. [d]Number of ranks with no difference.

Table 55

Within-Country Ranks of Topic Scores

	Country																				
Topic	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	AVE	SD	MIN	MAX
1.1.1.1	3	7	8	8	10	9	4	11	4	18	7	6	6	4	15	4	12	8.0	4	3	18
1.1.1.2	9	8	6	12	11	19	7	13	8	17	15	10	12	9	10	10	14	11.2	3	6	19
1.1.1.3	18	13	11	22	8	20	1	8	26	7	24	27	25	2	8	27	1	14.6	9	1	27
1.1.2.1	14	14	14	14	18	21	13	16	16	14	16	20	24	17	5	14	17	15.7	4	5	24
1.1.2.2	11	11	5	9	6	10	10	7	12	3	10	9	10	10	2	7	7	8.2	3	2	12
1.1.2.3	10	15	10	7	12	17	11	9	15	11	13	15	9	14	4	11	11	11.4	3	4	17
1.1.2.4	29	27	28	28	26	27	28	28	28	25	28	26	28	28	16	26	29	26.8	3	16	29
1.1.3.1	12	10	17	13	21	6	23	21	2	16	3	16	22	22	21	13	6	14.4	7	2	23
1.1.4.2	21	1	2	18	22	5	2	1	14	28	5	12	11	1	1	25	5	10.2	9	1	28
1.1.4.4	28	29	29	29	29	29	29	29	20	27	29	29	29	29	29	29	28	28.2	2	20	29
1.1.5.1	2	6	9	1	3	7	8	4	1	13	9	4	1	6	12	2	3	5.4	4	1	13
1.1.5.2	19	12	19	16	19	16	19	12	11	26	18	17	17	18	20	17	15	17.1	3	11	26
1.1.5.3	5	9	13	6	7	12	15	5	3	22	11	2	7	11	7	5	4	8.5	5	2	22
1.1.5.4	22	17	7	10	28	8	16	3	13	29	20	7	14	8	6	28	19	15.0	8	3	29
1.2.1	4	4	1	4	2	3	3	2	5	2	2	5	5	5	3	3	8	3.6	2	1	8
1.2.2	24	20	25	26	25	22	26	26	27	20	22	22	26	20	14	20	26	23.0	3	14	27
1.2.3	7	3	15	2	1	1	8	24	6	19	1	1	2	23	11	1	13	8.1	8	1	24
1.3.1	16	23	27	21	13	11	24	15	21	12	8	25	20	27	26	24	18	19.5	6	8	27
1.3.2	17	16	18	17	9	18	17	27	19	4	17	11	15	15	18	15	23	16.2	5	4	27
1.3.3	25	26	24	27	17	23	27	14	18	5	26	19	21	13	22	19	27	20.8	6	5	27
1.3.4	6	2	4	5	4	4	5	10	9	1	4	3	4	3	9	8	10	5.4	3	1	10
1.4.1	15	18	16	19	14	15	20	23	22	9	12	18	13	16	24	18	21	17.2	4	9	24
1.4.2	27	22	26	25	24	26	22	20	24	8	21	21	18	24	27	16	25	22.1	5	8	27
1.5.1	26	28	23	23	27	28	25	25	29	23	27	28	27	25	25	22	24	25.6	2	22	29
1.5.2	20	25	20	20	20	24	21	22	23	24	25	24	19	21	17	21	22	21.6	2	17	25
1.6.1	8	21	12	11	15	13	12	19	10	10	14	14	8	12	19	12	9	12.9	4	8	21
1.6.2	23	19	21	24	23	25	18	18	25	15	23	23	23	7	23	23	20	20.8	4	7	25
1.7.1	1	5	3	3	5	2	6	6	7	6	6	8	3	19	13	6	2	5.9	4	1	19
1.7.2	13	24	22	15	16	14	14	17	17	21	19	13	16	26	28	9	16	17.6	5	9	28

Table 56

Proportions of Textbook Blocks Allocated to Each Performance Expectation by Each Country

	Country																					
CODE	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	AVE	SD	MDN	MAX	Count
2.1.1	0.034	0.005	0.083	0.119	0.543	0.053	0.301	0.052	0.136	0.002	0.015	0.021	0.187	0.299	0.002	0.024	0.036	0.11	0.14	0.05	0.54	17
2.1.2	0.028	0.009	0.066	0.019	0.003	0.014	0.010	0.031	0.043	0.001	0.022	0.010	0.049	0.295	0	0.026	0.032	0.04	0.07	0.02	0.30	16
2.1.3	0.549	0.440	0.028	0.122	0.006	0.397	0.126	0	0.480	0.524	0.247	0.443	0.325	0.552	0.408	0.345	0.323	0.31	0.19	0.35	0.55	16
2.2.1	0.039	0.130	0.036	0.091	0	0.033	0.001	0.017	0.019	0.007	0.026	0.094	0.055	0.008	0.044	0.019	0.033	0.04	0.03	0.03	0.13	16
2.2.2	0.312	0.276	0.263	0.412	0	0.321	0.093	0.288	0.146	0.229	0.365	0.237	0.544	0.416	0.533	0.295	0.264	0.29	0.14	0.29	0.54	16
2.2.3	0.068	0.023	0.059	0.129	0	0.029	0.001	0.282	0.035	0.028	0.068	0.068	0.096	0.182	0.031	0.101	0.213	0.08	0.08	0.07	0.28	16
2.3.1	0.001	0.006	0.031	0.007	0.069	0	0.251	0.001	0.018	0.043	0.098	0.006	0.134	0.028	0.008	0.037	0.009	0.04	0.06	0.02	0.25	16
2.3.2	0.001	0.002	0.024	0.013	0.011	0.005	0.108	0.002	0.023	0.002	0.075	0	0.021	0.035	0.001	0	0.013	0.02	0.03	0.01	0.11	15
2.3.3	0.029	0.036	0.034	0.072	0.354	0.057	0.077	0.296	0.182	0.014	0.199	0.083	0.185	0.075	0.022	0.131	0.092	0.11	0.10	0.08	0.35	17
2.3.4	0.003	0.001	0.008	0.015	0	0.019	0.002	0.001	0.002	0.005	0.007	0.003	0.019	0.016	0.001	0.003	0.031	0.01	0.01	0.00	0.03	16
2.3.5	0.001	0	0.022	0.004	0	0	0.055	0.006	0.008	0.009	0.058	0	0.006	0.088	0	0.002	0.016	0.02	0.03	0.01	0.09	12
2.4.1	0.035	0	0.079	0.039	0.003	0	0.064	0.146	0	0.010	0.107	0	0.051	0.042	0	0.002	0.001	0.03	0.04	0.01	0.15	12
2.4.2	0.011	0.006	0.018	0.020	0	0	0.188	0.002	0.045	0.002	0.055	0.005	0.002	0.245	0	0	0.005	0.04	0.07	0.01	0.25	13
2.4.3	0.001	0.014	0.013	0.012	0.010	0	0.101	0.021	0	0	0.056	0.001	0.015	0.094	0	0	0.026	0.02	0.03	0.01	0.10	12
2.4.4	0.001	0.002	0.019	0.002	0	0	0.010	0.001	0	0.004	0.035	0.010	0.070	0.205	0	0.001	0.012	0.02	0.05	0.00	0.21	13
2.4.5	0.006	0.009	0.053	0.001	0	0	0.017	0.082	0.011	0.113	0.022	0	0.045	0.182	0.011	0	0.015	0.03	0.05	0.01	0.18	13
2.4.6	0.001	0.002	0.012	0	0	0	0.012	0	0	0	0.002	0	0	0	0	0	0	0.00	0.00	0.00	0.01	5
2.5.1	0.017	0.028	0.035	0.012	0	0.005	0.015	0	0.145	0.001	0.051	0.003	0.138	0.014	0.008	0	0.070	0.03	0.04	0.01	0.15	14
2.5.2	0.003	0.001	0.008	0.005	0	0	0.122	0.000	0.049	0.007	0.012	0.004	0.017	0.147	0	0	0.071	0.03	0.04	0.01	0.15	14
2.5.3	0.007	0.010	0.037	0.017	0	0	0.022	0.088	0.003	0.003	0.029	0.011	0.057	0.068	0	0.012	0.151	0.03	0.04	0.01	0.15	14
2.5.4	0.000	0.002	0.026	0.002	0	0	0.013	0.008	0	0	0.006	0	0.002	0.083	0	0.001	0.012	0.01	0.02	0.00	0.08	11
Average	0.063	0.055	0.049	0.061	0.056	0.052	0.080	0.068	0.072	0.055	0.084	0.055	0.108	0.154	0.059	0.055	0.066	0.07	0.06	0.05	0.23	14.2
SD	0.137	0.114	0.056	0.096	0.143	0.111	0.087	0.105	0.115	0.126	0.093	0.11	0.134	0.153	0.147	0.101	0.094	0.09	0.05	0.10	0.16	2.8
Max	0.549	0.440	0.263	0.412	0.543	0.397	0.301	0.296	0.480	0.524	0.365	0.443	0.544	0.552	0.533	0.345	0.323	0.31	0.19	0.35	0.55	17

The average of the proportions of textbook blocks for 2.4.6 *Axiomatizing*, 2.3.4 *Predicting*, and 2.5.4 *Critiquing* were small (.002, .008, .009 respectively). Aside from the most frequently reported performance expectations, countries clearly had different levels of expectations. Performance Expectation 2.1.1 *Representing* and 2.3.3 *Solving* were the only performance expectations included in the textbooks of all countries

Tables 57 and 58 show country performance on items with the same performance-expectation codes, regardless of content. As with topic scores, much variation in performance clearly existed. For most countries, differences in minimum and maximum scores on performance expectations were 50 to 70 percentage points. Differences in minimum and maximum ranks were around 10 ranks. Six countries had a rank of 1 on at least one performance expectation. Average ranks for each country differed from average total score ranks. Also, ranks on the two performance expectations included in the intersection of all countries (2.1.1, 2.3.3) differed from the ranks on the field-trial instrument for many countries.

Tables 59 and 60 present the same information as in Tables 57 and 58, except that performance expectations have been grouped into six categories: knowing (2.1 and 2.5.1), performing routine procedures (2.2.1 and 2.2.2), performing complex procedures (2.2.3), problem solving (2.3), reasoning (2.4), communicating (2.5 - except 2.5.1). Again, some striking variation exists. Country J ranks first in all categories except category 5 *communication.* Country K ranks first on this category. Country C had lower ranks on category 3 *performing complex procedures* and category 5 *reasoning.* Country O had its lowest rank on category 1 *knowing*, while country Q had its highest rank in this category.

Table 57

Performance-Expectation Scores

	Performance Expectations																
Country	2.1.1	2.1.2	2.1.3	2.2.1	2.2.2	2.2.3	2.3.1	2.3.2	2.3.3	2.3.4	2.4.3	2.4.5	2.5.1	2.5.3	AVE	MIN	MAX
# Items	27	15	32	5	70	61	10	1	62	7	3	2	13	7			
A	57.8	59.8	51.7	65.5	47.9	57.8	33.8	34.9	42.9	51.6	24.5	18.7	48.1	28.5	44.5	18.7	65.5
B	62.9	68.8	58.4	73.5	53.9	62.4	36.7	38.0	49.2	61.6	45.0	23.5	59.9	29.3	51.6	23.5	73.5
C	60.1	67.8	55.7	71.1	55.3	56.9	32.5	18.6	45.8	53.4	27.9	15.3	52.3	21.6	45.3	15.3	71.1
D	51.9	56.0	48.7	57.1	42.1	51.0	28.5	44.9	36.9	48.9	18.6	12.5	43.2	20.4	40.0	12.5	57.1
E	51.3	58.4	52.7	68.7	41.1	53.9	31.7	20.8	37.6	51.6	17.5	16.6	44.9	20.8	40.5	16.6	68.7
F	53.6	64.6	48.6	69.0	46.3	58.5	30.9	8.6	42.7	55.4	26.0	19.6	50.3	18.4	42.3	8.6	69.0
G	53.4	59.1	54.0	58.5	46.7	53.1	35.3	7.9	39.8	50.9	19.9	14.8	50.6	15.4	40.0	7.9	59.1
H	49.5	55.8	47.0	58.8	41.1	48.3	31.1	17.5	37.0	47.0	24.4	17.0	43.2	21.6	38.5	17.0	58.8
I	58.0	60.0	54.1	59.7	51.2	58.3	33.8	55.3	48.6	54.9	26.2	27.1	49.2	34.1	47.9	26.2	60.0
J	66.7	66.0	67.2	70.8	65.2	70.8	48.7	17.9	57.9	60.2	54.5	34.8	62.4	23.0	54.7	17.9	70.8
K	62.1	69.9	58.4	74.4	52.7	62.4	36.9	40.8	48.7	60.8	30.6	20.0	60.2	34.8	50.9	20.0	74.4
L	56.8	64.3	53.3	74.1	47.2	59.7	40.6	65.5	44.1	53.9	29.8	16.4	50.6	34.1	49.3	16.4	74.1
M	41.0	44.1	40.0	55.4	32.6	42.2	25.6	4.8	28.2	42.0	14.9	5.0	38.0	13.2	30.5	4.8	55.4
N	50.9	55.1	49.3	49.7	47.1	43.2	34.4	0.0	40.1	47.2	33.8	23.6	46.5	12.8	38.1	0.0	55.1
O	65.1	71.2	58.0	58.3	63.6	65.3	49.5	58.9	59.4	58.0	32.9	24.9	53.2	34.1	53.7	24.9	71.2
P	52.5	56.9	49.7	74.8	37.9	55.9	31.6	31.3	40.0	57.2	26.2	6.3	50.0	23.1	42.4	6.3	74.8
Q	63.5	66.4	58.9	61.7	55.1	60.3	38.5	51.3	48.0	58.5	31.4	31.7	51.9	33.0	50.7	31.4	66.4
Ave	56.3	61.4	53.3	64.8	48.6	56.5	35.3	30.4	43.9	53.7	28.5	19.3	50.3	24.6	44.8	15.8	66.2
SD	6.5	6.7	5.9	7.6	8.3	7.2	6.1	19.7	7.5	5.3	9.5	7.7	6.2	7.4	6.4	8.1	6.9
Min	41.0	44.1	40.0	49.7	32.6	42.2	25.6	0.0	28.2	42.0	14.9	5.0	38.0	12.8	30.5	0.0	55.1
Max	66.7	71.2	67.2	74.8	65.2	70.8	49.5	65.5	59.4	61.6	54.5	34.8	62.4	34.8	54.7	31.4	74.8

Table 58

Performance-Expectation Ranks

	Performance Expectations																
Country	2.1.1	2.1.2	2.1.3	2.2.1	2.2.2	2.2.3	2.3.1	2.3.2	2.3.3	2.3.4	2.4.3	2.4.5	2.5.1	2.5.3	AVE	MIN	MAX
A	8	10	11	9	8	9	9	8	9	12	12	9	12	7	9.5	7	12
B	4	3	3	4	5	4	6	7	3	1	2	6	3	6	4.1	1	7
C	6	4	6	5	3	10	11	11	7	10	8	13	5	10	7.8	3	13
D	13	14	14	15	13	14	16	5	16	14	15	15	15	13	13.7	5	16
E	14	12	10	8	14	12	12	10	14	11	16	11	14	12	12.1	8	16
F	10	7	15	7	12	7	15	14	10	7	11	8	9	14	10.4	7	15
G	11	11	8	13	11	13	7	15	13	13	14	14	8	15	11.9	7	15
H	16	15	16	12	15	15	14	13	15	16	13	10	15	11	14.0	10	16
I	7	9	7	11	7	8	10	3	5	8	9	3	11	2	7.1	2	11
J	1	6	1	6	1	1	2	12	2	3	1	1	1	9	3.4	1	12
K	5	2	4	2	6	3	5	6	4	2	6	7	2	1	3.9	1	7
L	9	8	9	3	9	6	3	1	8	9	7	12	7	2	6.6	1	12
M	17	17	17	16	17	17	17	16	17	17	17	17	17	16	16.8	16	17
N	15	16	13	17	10	16	8	17	11	15	3	5	13	17	12.6	3	17
O	2	1	5	14	2	2	1	2	1	5	4	4	4	2	3.5	1	14
P	12	13	12	1	16	11	13	9	12	6	10	16	10	8	10.6	1	16
Q	3	5	2	10	4	5	4	4	6	4	5	2	6	5	4.6	2	10

Table 59

Performance-Expectation Category Scores

	Performance Expectation Categories									
Country	Knowing	Routine Procedures	Complex Procedures	Problem Solving	Reasoning	Comm-unication	AVE	SD	MIN	MAX
# Items	87	75	61	80	5	7				
A	54.4	49.1	57.8	42.4	22.2	28.5	42.4	12.1	22.2	57.8
B	61.8	55.2	62.4	48.6	36.4	29.3	48.9	11.5	29.3	62.4
C	58.6	56.3	56.9	44.5	22.9	21.6	43.5	14.5	21.6	58.6
D	50.1	43.1	51.0	37.0	16.1	20.4	36.3	12.6	16.1	51.0
E	52.1	43.0	53.9	37.9	17.1	20.8	37.4	13.1	17.1	53.9
F	53.2	47.8	58.5	41.9	23.4	18.4	40.5	13.7	18.4	58.5
G	54.2	47.5	53.1	39.8	17.8	15.4	38.0	14.6	15.4	54.2
H	48.7	42.3	48.3	36.9	21.4	21.6	36.5	10.5	21.4	48.7
I	55.6	51.8	58.3	47.4	26.6	34.1	45.6	10.7	26.6	58.3
J	66.1	65.6	70.8	56.5	46.6	23.0	54.7	15.0	23.0	70.8
K	61.8	54.1	62.4	48.2	26.4	34.8	47.9	12.4	26.4	62.4
L	55.9	49.0	59.7	44.8	24.4	34.1	44.6	11.3	24.4	59.7
M	40.7	34.2	42.2	28.8	11.0	13.2	28.4	11.4	11.0	42.2
N	50.4	47.3	43.2	39.5	29.7	12.8	37.1	11.8	12.8	50.4
O	61.8	63.2	65.3	58.0	29.7	34.1	52.0	13.4	29.7	65.3
P	51.9	40.4	55.9	40.3	18.2	23.1	38.3	12.8	18.2	55.9
Q	60.6	55.5	60.3	47.7	31.5	33.0	48.1	11.1	31.5	60.6
Ave	55.2	49.7	56.5	43.5	24.8	24.6	42.4	12.5	21.5	57.1
SD	6.0	7.9	7.2	7.0	8.2	7.4	6.5	1.4	5.8	6.5
Min	40.7	34.2	42.2	28.8	11.0	12.8	28.4	10.5	11.0	42.2
Max	66.1	65.6	70.8	58.0	46.6	34.8	54.7	15.0	31.5	70.8

Table 60

Performance-Expectation Category Ranks

	Performance Expectation Categories							
Country	Knowing	Routine Procedures	Complex Procedures	Problem Solving	Reasoning	Commun-ication	AVE	SD
A	9	8	9	9	11	7	8.8	1.1
B	2	5	4	3	2	6	3.7	1.4
C	6	3	10	8	10	10	7.8	2.4
D	15	13	14	15	16	13	14.3	1.0
E	12	14	12	14	15	12	13.2	1.1
F	11	10	7	10	9	14	10.2	2.0
G	10	11	13	12	14	15	12.5	1.6
H	16	15	15	16	12	11	14.2	1.8
I	8	7	8	6	6	2	6.2	1.9
J	1	1	1	2	1	9	2.5	2.7
K	3	6	3	4	7	1	4.0	1.9
L	7	9	6	7	8	2	6.5	2.1
M	17	17	17	17	17	16	16.8	0.3
N	14	12	16	13	5	17	12.8	3.6
O	4	2	2	1	4	2	2.5	1.0
P	13	16	11	11	13	8	12.0	2.3
Q	5	4	5	5	3	5	4.5	0.7

I used the textbook information on performance expectations as well as the combination of performance expectation by topic and constructed unique tests for each country. The results were in Table 61. Again, variation clearly existed when tests were developed that closely match the topics students were taught and the performance that was expected of them. Scores and ranks on these tests for some countries are different from those on the field-trial instrument. Some of the scores are 10 or more points different, and some of the ranks are up to 7 places different.

Table 61

Country Performance on Unique Tests based on Performance Expectaions and Topics Crossed with Performance Expectations

	Field Trial	Scores		Field Trial	Ranks	
	Scores	PE Unique	CxPE Unique	Ranks	PE Unique	CxPE Unique
A	50.5	48.9	49.6	9	5	9
B	56.4	55.3	57.1	3	3	3
C	53.6	39.8	53.0	6	13	5
D	45.2	43.2	44.1	14	12	15
E	45.9	44.3	47.0	12	9	12
F	49.6	46.1	48.1	10	6	10
G	48.1	33.9	46.9	11	16	13
H	43.5	34.0	41.2	16	15	16
I	52.8	44.9	53.0	7	7	6
J	64.0	56.8	65.3	1	2	1
K	56.0	43.9	51.5	5	10	7
L	51.5	52.1	50.8	8	4	8
M	35.4	28.7	33.5	17	17	17
N	45.0	35.1	47.4	15	14	11
O	61.9	57.0	58.9	2	1	2
P	45.8	44.8	46.2	13	8	14
Q	56.4	43.3	54.1	4	11	4

CHAPTER V

Discussion, Summary, and Recommendations

I set out in this study to answer the following questions:

1. How much variation in content exists across the 17 nations in the mathematics curricula for 13-year-old students? How well does the content of the TIMSS field-trial instrument match these curricula?
2. What test specifications provide a curricular match across countries? How well does the content of the TIMSS field-trial instrument match these test specifications?
3. What test specifications would improve the content match between the TIMSS field-trial instrument and the countries' math curricula? How well do these specifications match the curricula?
4. How stable are country scores and ranks across tests that increase the correspondence between the TIMSS field-trial instrument and the curricula of the 17 countries? How stable are country results across topics and performance expectations?

Each of these questions is discussed below. Following the discussion is a summary of conclusions and recommendations for future work.

How Much Variation Exists in Curricular Content?

A surprising amount of variation in curricular content is present both within and across countries as well as within and across data sources. However, some commonalties do exist.

Variation in Coverage for Topics within Each Data Source

In the expert-topic-mapping data source, for example, just over half of the 44 topics in the mathematics framework are intended for inclusion in instruction at age 13 in at least 70% of the countries; however, only one topic is intended for inclusion in all countries. Likewise, only 3 topics are intended to be *excluded* from instruction at age 13 in all countries; two topics are intended to be excluded in all but one or two countries. The average number of countries intending instruction on a topic is 11 (65%); on the other hand, the average number of countries intending that a topic receive special focus is only 4.

The patterns in the other data sources are similar. The average number of countries including a topic in their corresponding curriculum-guide and textbook sample is 11. No topic is excluded from the curriculum-guide samples of all countries, and only one is excluded from all country textbook samples. Only two topics appear in the curriculum-guide samples of all countries, and only three appear in the textbook samples of all countries. In both these data sources, less than half of the topics are included in the document samples of at least 70% of the countries. Variation in topic emphasis is about twice as large in the textbook-data source as in the curriculum-guide- or expert-mapping-data sources.

Several topics were rarely included in the curriculum. These topics are the *Number* topics of *Binary Arithmetic* (1.1.4.1), *Complex Numbers* (1.1.4.3), and *Systematic Counting* (1.1.4.5) as well as the *Calculus* topics of *Infinite Processes* (1.8.1) and *Change* (1.8.2). These are generally included in the data sources of only four or less countries each. On the other hand, *Basic 2D Geometry* (1.3.4) is included in the expert-

topic-mapping- and textbook-data sources of all countries, and *Equations and Formulas* (1.6.2) is included in the curriculum-guide and textbook samples of all countries. Other topics with high inclusion rates across data sources are *Polygons and Circles* (1.3.3), *3D Geometry* (1.3.2), *Proportionality Problems* (1.5.2), *Patterns, Relations, and Functions* (1.6.1), and *Data Representation and Analysis* (1.7.1). These appear in the data sources of around 15 to all 17 of the countries each. In each of the data sources, at least 10 countries include *Other Content* (1.10). Overall, the topic that seems to have the highest inclusion rate and most emphasis across data sources is the algebra topic 1.6.2 *Equations and Formulas*.

Variation in Topic Coverage for Countries within Each Data Source

Topic inclusion and emphasis also vary across countries within each data source. The difference between the largest number of topics included by a country in a data source and the minimum number of topics in the same data source is around 25 to 30 topics. However, the average number of topics included across the country-data sources is 28. The average of the average proportion of emphasis countries devote to topics included in the expert-topic-mapping and curriculum-guide-data sources is .04, or about 7 class periods. The average proportions of emphasis for countries across all data sources range from .02 (4 class periods) to .09 (16 class periods). The average proportion of textbook blocks devoted to topics included in textbooks is .05, with a range of .03 to .09.

The countries include different numbers of topics in each of the three data sources. At the most extreme, country N has an inclusion difference of 22 topics across data sources (39 topics included in the expert-topic-mapping and 17 topics in the

curriculum-guide-data sources). On average, countries included about 18 topics in all three data sources, and 7 in none of them. Thus, countries had an average rate of agreement of topic inclusion across data sources of just over half the topics.

Potential Explanations of Variation

Variations in topic coverage across countries is to be expected. Countries approach schooling in different ways. Some cover many topics over a period of many years, and others focus on select topics for shorter periods of time. Some countries begin with the "basics," adding topics only as students grasp necessary concepts; others want to continually challenge their students. Schooling at age 13 is just a slice in the pattern of schooling that for students in most countries began eight years earlier. However, the commonalties in topic inclusion and exclusion (e.g., the focus on geometry and algebra, the exclusion of certain complex numbers and calculus topics) show that there may be an underlying pattern of mathematics sequencing followed by most countries.

What is surprising to see is the great variation within countries. Some of this may be explained by the differing roles of curriculum guides and textbooks across countries as well as the structure of the educational systems. It is useful to think of educational systems as falling along a continuum of centralization. More centralized systems have common curricula that are often mandated. In these cases, curriculum guides, textbooks, and often lesson plans, may be written from the same "blueprint." One would expect to see higher agreement among documents in countries with highly centralized educational systems than in other less centralized systems. Less centralized systems often leave curriculum development completely up to local authorities, resulting in collections of

curriculum guides and textbooks within the country. As a result, it may be more difficult to find agreement of topic coverage across data sources in less centralized countries.

Additionally, the experts who completed the expert topic mapping in some countries may have had better knowledge of their country's curricula than experts in other countries. It is also possible that expert-mapping data in a particular country may reflect recent reform movements; movements that may not have found their way yet into the curricular documents. On the other hand, new curriculum guides may have been written in a country that has not gone through the process of writing new textbooks. Furthermore, in some classrooms, textbooks may be used only as a resource, while in others they may provide a daily map for instruction.

The variations in curricula, and potential reasons behind these variations, demonstrate the need to consider multiple sources of information to obtain a complete picture of curricular intentions across countries. These variations in patterns both within and across countries need to be considered in test development. Additionally, the variations stress the need for test developers and researchers to be specific about the inferences they intend to make from the tests they develop or use. Tests that are used to demonstrate student "achievement of mathematics curriculum" may face criticism concerning their validity because "achievement of mathematics curriculum" is open to many interpretations.

How Well Does the Content of the Field-Trial Instrument Match the Content of the Curriculum-Data Sources?

Topics

Definite gaps exist between country-level treatment of topics in each data source and topic inclusion on the field-trial instrument. However, three of the topics with the highest coverage across data sources (1.6.1 *Patterns, Relations, Functions*, 1.6.2 *Equations & Formulas*, 1.7.1 *Data Representation & Analysis*) also have high numbers of items on the field-trial instrument. Topic 1.6.2, which is heavily emphasized in all three curriculum sources, is one of the topics included in the field-trial instrument that is also included in most data sources for most countries (Table 12). On the other hand, the three geometry topics that are prevalent in the curriculum sources (1.3.2 *Coordinate Geometry*, 1.3.3 *Polygons Circles*, 1.3.4 *3D Geometry*) have few items on the field-trial instrument. Topic 1.3.4 is covered by only four items even though over 70% of the countries include it in each data source. Other topics that deserve more items are 1.1.3.1 *Negative Numbers*, 1.1.3.3 *Real Numbers*, 1.1.4.2 *Exponents*, and 1.1.4.4 *Number Theory*. Topics deserving less items are 1.1.5.1 *Estimating Quantity and Size* and 1.7.2 *Uncertainty and Probability*. The highest proportion of test items on the field-trial instrument is for topic 1.1.2.1 *Common Fractions* which appears in the curriculum-guide samples of only nine countries. This topic also has an average of 11% more emphasis on the field-trial instrument than across the curriculum sources.

On the other hand, none of the topics previously mentioned as having low coverage across data sources (1.1.4.1, 1.1.4.3, 1.1.4.5, 1.8.1, and 1.8.2) are included on

the field-trial instrument. However, topics 1.1.3.1 *Negative Numbers* and 1.1.3.3 *Real Numbers* also have no items on the field-trial instrument but are prevalent in the curricula.

Data Sources

Determining to which data source the content field-trial instrument is most similar depends upon how one chooses to evaluate the similarity. On average, 87% of the items on the field-trial instrument test topics included in the textbook-data source; only 55% of the items test topics in the aggregate-data source. The data source with the lowest coverage on the field-trial instrument is the curriculum-guide-data source. The expert-mapping data source has the best correspondence of topic inclusion (or non-inclusion) with the field-trial instrument. The worst correspondence of topic inclusion with the field-trial instrument is for the aggregate-data source. The difference between topic emphasis in each data source and topic weight on the field-trial instrument is fairly even across all data sources, except that the textbooks and aggregate-data sources tend to have higher emphasis on certain topics than did the field-trial instrument. The largest correlations between topic-weight profiles on the field-trial instrument and topic-emphasis profiles in the data sources are between the field-trial instrument and the textbook- and aggregate-data sources; however, this is likely due to the fact that more variation exists in the proportions in the textbook-data source as opposed to the other data sources. The lowest correlations are between the field-trial instrument and the curriculum-guide-data source.

The largest Euclidean distance between topic-emphasis profiles for each data source and topic-weight profiles for the filed-trial instrument is between the topic weights on the field-trial instrument and topic emphasis in the textbooks. Other Euclidean distances are fairly similar to one another.

The Euclidean distances, which take into take into account topic means, standard deviations, and rank ordering of topic emphasis within countries, indicate that the textbook would seem to have the overall poorest match of topic coverage to the field-trial instrument. However, not all "mis-matches" are the same. One type of topic-coverage mis-match occurs when students receive instruction on more than is tested; the other type of mis-match occurs when students are tested on topics they have not been taught. These have different consequences for validity. Mis-matches between the field-trial instrument and the textbooks generally result when more emphasis is placed on certain topics in the textbooks than topics on the field-trial instrument. This represents the first type of mis-match (instruction on topics not tested). However, the higher correlation between topic-emphasis profiles in the textbooks and topic-weight profiles on the field-trial instrument suggests that the relative ranking of topics on the field-trial instrument is more similar to the textbook topic rankings than the rankings from the expert-mapping and curriculum-guide-data sources. The same is true of the ranking of topics in the aggregate-data source for each country, but, on average, the aggregate-data sources contain fewer topics than are contained on the field-trial instrument. The curriculum-guide-topic profiles have a poor correlation with the field-trial-instrument-topic profiles because their profiles are basically flat – all topics included in a country's curriculum-guide sample have the same proportion. Correlations probably are not the best measure for the similarity between the

content of the curriculum guides and the content of the field-trial instrument. The expert mapping seems to fare well across all analyses.

Countries

The correspondence between topic coverage on the field-trial instrument and in the data sources varies also across countries. The field-trial instrument covers about 70% to 90% of what most countries include across the three data sources. For one country (D) almost 100% of the items on the field-trial instrument test topics that are included in the curriculum; however, for another country (J), less than 50% of the items test topics that are in the curriculum. On average, 20% of the items on the field-trial instrument test topics not included in the curriculum of a country. On the other hand, an average of 10% to 30% of the curriculum of the countries is untested by the field-trial instrument. Overall negative and positive differences between each country and the field-trial instrument in emphasis are fairly even, although variation exists across countries and data sources. A difference of .40 exists across countries between the highest and lowest average correlations between topic-weight profiles on the field-trial instrument and topic-emphasis profiles in each of the data sources (SD of .07). The average correlation is only .36. Differences in Euclidean distances between topic-weight profiles on the field-trial instrument and topic-emphasis profiles in each data source are .20 (SD .04).

Conclusions about Test-to-Curriculum Match

The summaries of most indices show reasonable correspondence between topic coverage on the field-trial instrument and in the data sources. On average, 75% of tested items are in the curriculum-data sources; an average of 84% of the curricula are tested on

the field-trial instrument; an average of a 68% of the topics are included on the field-trial instrument and included in the curriculum or not included in both; an average of only an 8% absolute difference exists in topic emphasis (4% positive and 4% negative). However, some data sources or countries do not show as much consistency. The field-trial instrument is more similar to some data sources in topic coverage than others. Therefore, a final evaluation of content validity depends upon the purpose of the test and which curriculum source represents the most appropriate comparison. A more serious problem is the differential match in topic coverage of the field-trial instrument to each country. Such differences raise serious questions about validity. If the test content is more similar to the curriculum of some countries than it is to the curriculum of others, it would constitute a better measure of the intended or potentially implemented curriculum in those countries. Comparisons referenced to such a test would be difficult to interpret across countries.

How Does the Content of the Test Blueprints Compare with the Content of the Field-Trial Instrument?

Focus of the Test Blueprints

The test blueprints described above were based on different testing purposes and differed in two respects. First, they differed in the curricular domain they represent. Test blueprints based on the expert mapping and curriculum-guide-data sources represent the intended curriculum of the nations. Test blueprints based on the textbooks and aggregate-data sources represent the potentially implemented curriculum of the nations.

Second, the blueprints differed in specificity of the intended inferences. Intended inferences from tests developed using the union-test blueprints relate to student achievement of the topics covering the full range of math topics. They represent all of what is possible in the mathematics curriculum. Tests based on the 70% intersections yield inferences related to achievement of "prevalent" math topics for 13-year-old students. Tests based on the strict intersections yield inferences related to those topics that all countries find important. Finally, inferences based on the "unique" tests relate to student achievement of the topics they were intended to learn. Each combination of specific domain and type of inference is meaningful. Validity needs to be evaluated in light of the purposes behind each particular combination.

The amount of variation in the content of the test blueprints showed that they were not equivalent. The blueprints differed on topic inclusion as well as emphasis. The 70%-intersection-test blueprints included about half of the 44 framework topics for all test blueprints except those based on the aggregate of the data sources. The strict-intersection-test blueprints included only a few topics. No strict-intersection-test blueprints could be developed for the aggregate-data source because no topic appeared in all three data sources of all countries.

Variation in Correlations between the Test Blueprints and the Field-Trial Instrument

Although I compared the content of the field-trial test instrument to the content of all three types of test blueprints (union, 70% intersection, and strict intersection), I did not expect a good quantitative match with the strict-intersection tests. Overall, however, approximately 60% of the topics on average appeared in both the field-trial instrument

and the blueprints. An average of 61% of the field-trial items were covered by topics in the test blueprints, and an average of 91% of the "items" on the test blueprints were covered by topics on the field-trial instrument. Differences in topic emphasis on the field-trial instrument and topic emphasis on the test blueprints is about 4% across topics, but 21% across curriculum sources. However, disregarding the strict intersection blueprints would lower the difference in emphasis across curriculum sources to approximately 3%.

On average, the new test blueprints place more emphasis on topics 1.3.2 *Basic Geometry*, 1.3.4 *3D Geometry*, and 1.6.2 *Equations and Formulas* than is on the field-trial instrument. However, when looking at data sources, the expert-topic-mapping and curriculum-guide blueprints place less emphasis on topic 1.6.2 than it receives on the field-trial instrument. The field-trial instrument also places more emphasis on topics 1.1.2.1 *Common Fractions* and 1.7.1 *Data Representation and Analysis* than is in the new blueprints. Correlations between topic-weight profiles on the field-trial instrument and topic-weight profiles on the test blueprints average around .45 and Euclidean distances between topic-weight profiles for the test blueprints and the topic-weight profile for the field-trial instrument average around .38 (larger than the distances seen earlier). Topic coverage on the field-trial instrument seems most similar to topic coverage on the union-test blueprint based on the aggregate of the data sources and the 70%-intersection-test blueprint based on the textbook-data source; topic coverage on the field-trial instrument seems least similar to topic coverage on the 70%-intersection test blueprints based on the curriculum-guide-data source and the aggregate of the data sources (disregarding the strict-intersection-test blueprints).

How Well Does the Content of the Specially Constructed Test Blueprints Match the Content of the Curriculum-data sources?

The content of the unique specially-constructed-test blueprints written using only topics included on the field-trial instrument is more similar to the content of the curriculum than was the content of the field-trial instrument. An average of 80%-90% of the topics were either included in the unique-test blueprints and in curriculum-data sources or not included in both. This is a 10% to 15% increase in the correspondence in topic inclusion between the field-trial instrument and the curriculum-data sources across topics and a 15%-30% increase across countries. Less than a 1% difference between topic emphasis on the unique-test blueprints and in the curriculum-data sources exists across topics (2% less difference than between the field-trial instrument and the curriculum), and across countries a 4% difference in topic emphasis between specially-constructed tests and the curriculum sources results in 2%-3% less difference than between the field-trial instrument and the curriculum. Correlations between topic-weight profiles on the test blueprints and topic-emphasis profiles in the curriculum sources improve by .50 (to an average of .84) and distances between the topic-weight profiles of the test blueprints and the topic-emphasis profiles of the curriculum sources shrink by .17 (to an average of .11). Unique-test blueprints based on the aggregate-data source were most similar to the curriculum on topic inclusion and topic-pattern profiles but was most dissimilar from the field-trial instrument on overall topic emphasis. The unique-test blueprints based on the textbook data have the highest of the correlations of topic emphasis with the corresponding curriculum and the most similarity in topic emphasis,

and the unique-test blueprints based on the curriculum-guide data has the lowest of the topic-emphasis correlations with the corresponding data source and the least similarity with the data source in topic inclusion.

The content of the inclusive test blueprints was more similar to the content of each corresponding curriculum source than was the content of the field-trial instrument. In some cases the improvements are small, but overall improvement in content similarity looks good; however, not as good as was seen with the unique tests. The proportion of tested items included in each country's curricula increases by 5% with the unique-test blueprints, to a new overall proportion of 82%. However, the tests based on 70% intersections are only testing an overall average of 63% of the curricula - a decline of 21%. This coverage ranges, however, from 43% for the aggregate-data source to 80% for the textbook-data source. No improvement over the field-trial instrument is seen in the similarity of topic inclusion between the and only a slight improvement is seen in the similarity of topic emphasis averaged across topics. However, the similarity between topic inclusion in the aggregate-data source and the corresponding test blueprints was better than between the data source and the field-trial instrument; topic emphasis on the tests based on the expert-mapping and curriculum-guide-data sources was more similar to these data sources than was topic emphasis on the field-trial instrument. Differences between the curriculum and the tests in topic emphasis decreased by 2.5% over countries with most improvement for the expert mapping and textbook-data sources. A decline was seen for the aggregate-data source.

The largest improvement for specially-constructed test blueprints over the field-trial instrument is seen in the correlations between the topic-weight profiles on the test

blueprints and the topic-emphasis profiles in the curricula. This means that the "profiles" of the specially-constructed tests had a more similar shape (i.e., relative weighting of topics) to the curriculum-topic-emphasis profiles than did the field-trial-instrument topic-weight profile. The average correlation between topic emphasis on a specially-constructed-test blueprint and topic-emphasis profiles in the corresponding data source rises to .58, an increase of .20. However, this is still .30 less than the average correlation between topic-weight profiles on the unique specially-constructed-test blueprints and topic-emphasis profiles in the data sources.

Euclidean distances between topic-weight profiles on the test blueprints and topic-emphasis profiles in the curricula improved slightly over those between the field-trial instrument and the curriculum-topic-emphasis profiles. As stated earlier, Euclidean distances are best interpreted as relative measures. However, I provided several benchmarks earlier. Two topic profiles with no difference between them would have a distance of 0 and topic profiles of tests and curricula emphasizing completely different topics would have a distance of 1.4. The original distances between the topic profiles of the field-trial instrument and the topic profiles of the curricula ranged from .30 to .04. The distances between the test blueprints and the curricula ranged from .10 to .20. If all topics differed in emphasis by .01, the distance would be .07; if all topics differed in emphasis by .025, the distance would be .165; if all topics differed by .05, the distance would be .33. Additionally, if ¼ of the topics differed by .1 and the others had no difference, the distance would be .33. One percent of a 180 period mathematics school year is around 2 class periods; 2 ½ percent of the school year is almost 5 class periods; five percent if almost 10 class periods, and ten percent is 18 class periods.

The distances between test-blueprint topic-weight profiles and curriculum-topic-emphasis profiles for the expert mapping and curriculum guides are very close to those between the unique test-blueprint topic-weight profiles and the curriculum-topic-emphasis profiles. The textbook- and aggregate-union-test blueprint Euclidean distances are about .10 higher than those for the unique-test blueprints. The distance between the aggregate 70%-intersection-test-blueprint topic-weight profiles and the aggregate-data-source topic-emphasis profiles is larger than the distance between the field-trial-instrument topic-weight profiles and that data source's topic-emphasis profile.

The new test blueprints based on field-trial topics did improve the similarity in topic coverage between the tests and the curriculum for all but the tests based on the strict intersections. However, data sets and countries did not all show the same degree of similarity with the field-trial instrument or the test blueprints. This was due in part to the missing topics on the field-trial instrument. The missing topics, though, can be treated as topics for which "good" items do not exist or that may have been negotiated out of the test. Thus, the test blueprints may represent the best overall match possible between the test and country-curriculum sources. Even under the best circumstances, mis-match will exist and needs to be considered in test interpretation.

How Does Country Performance Vary?

I evaluated variation in test performance across the specially-constructed tests to obtain a sense of the impact that test-curriculum mis-matches in content coverage might have on test interpretation. I used data on the proportions of students passing each item to calculate potential country scores on the new "tests."

Differences in Total Scores and Ranks

When scores and ranks are averaged over all the specially-constructed tests, little difference is seen from the original field-trial scores and ranks. Average differences in passing rates on the field-trial instrument and passing rates on other tests are only about 3%, and all country ranks are nearly identical. Correlations of all specially-constructed-test scores and ranks with the field-trial scores and ranks are near .90 and above. However, as much as a 16% difference in passing rate is seen across all specially-constructed tests, with an average of a 10% difference. Differences in highest and lowest ranks within a country are as high as 9 places with an average of 5. This means that, on average, countries would rank in different quartiles (of this distribution of 17 countries) based on their highest and lowest ranks, with some countries ranking in the top half of the distribution for some tests and in the bottom half for others. Largest differences in scores are between the field-trial instrument and the strict-intersection tests and the smallest are between the field-trial instrument and the unique tests. The lowest correlations between the field-trial instrument and specially-constructed tests are also with the strict-intersection tests. On average little difference could be found between performance on the field-trial instrument and performance on the specially-constructed tests. Although, some countries (H, J, P) do display score fluctuations, and others (E, G, Q) display rank fluctuations.

Differences in Topic Scores and Ranks

More variation is evident within countries, however, when looking at individual topic scores. Within countries, standard deviations across topic scores are 10%, and

differences between minimum and maximum passing rates within a country are as high as 50%. Countries J and P have larger score fluctuations than other countries. Additionally, nine different countries rank first on at least one topic, and six countries rank last on at least one topic. This shows clear patterns of strength and weakness across countries. Larger differences in performance are seen on topics 1.1.4.4 *Number Theory* and 1.1.2.4 *Percentages* than other topics. Differences in scores for these two topics are around 10%. Both these topics are included by over half the countries in the three primary data sources. However, 1.1.4.4 had only one item on the field-trial instrument.

Topic ranks within countries varied across countries. Countries, on average, performed their best on topic 1.2.1 *Measurement Units* and their worst on topic 1.1.4.4 *Number Theory.* Reviews of prior coverage of these two topics in each country's curricula many explain these results. Countries N and J have different patterns of performance than other countries.

Differences in Performance-Expectation Results

Even more significant are performance results when looking at performance expectation codes and topic by performance-expectation intersections. Overall, students perform better on basic-understanding and routine-computation items (which had the highest proportions of textbook blocks) and more poorly on reasoning and communication items. More of these items also happened to be extended response. Six countries rank first on at least one performance expectation and two rank last on at least one. Average ranks across performance-expectation categories were within one to two places of field-trial ranks. However, when developing unique tests for countries that only

include items with performance-expectation codes included in their curriculum, slightly more variation is seen.

Within-Country Variation

Within-country variation was noticeably larger in performance at the item and topic level than at the total score level. Figure 2 shows box plots of country scores across all specially-constructed tests, across all TIMSS reporting sub-scales, across all topics, across all performance expectations, and across all items within two of the topics. Topic 1.6.2 *Equations and Formulas* is a topic emphasized in the curriculum across all countries and data sources. Topic 1.7.2 *Uncertainty and Probability* is a topic that is not highly emphasized.

The reduction in variation as scores get further away from individual items is striking. Some of this variation may be explained by measurement error - especially when looking at the item level. However, measurement error most likely does not account for all the variation. Table 62 provides estimated reliabilities and standard errors for the scores in each set of box plots. These are estimates treating each country as a case and each total, scale, topic, performance-expectation, or item score as an item. Standard errors for each individual country were estimated under the assumption that the reliabilities were the same for each country as it was for the group. Reliabilities were all .90 and above. Standard errors for the scales ranged from .31 to almost 2, producing error bands of ±.6 to ±4 points). Across countries though, standard errors could be close to 9 points. These would need to be investigated further to determine the effect of measurement error on score variation.

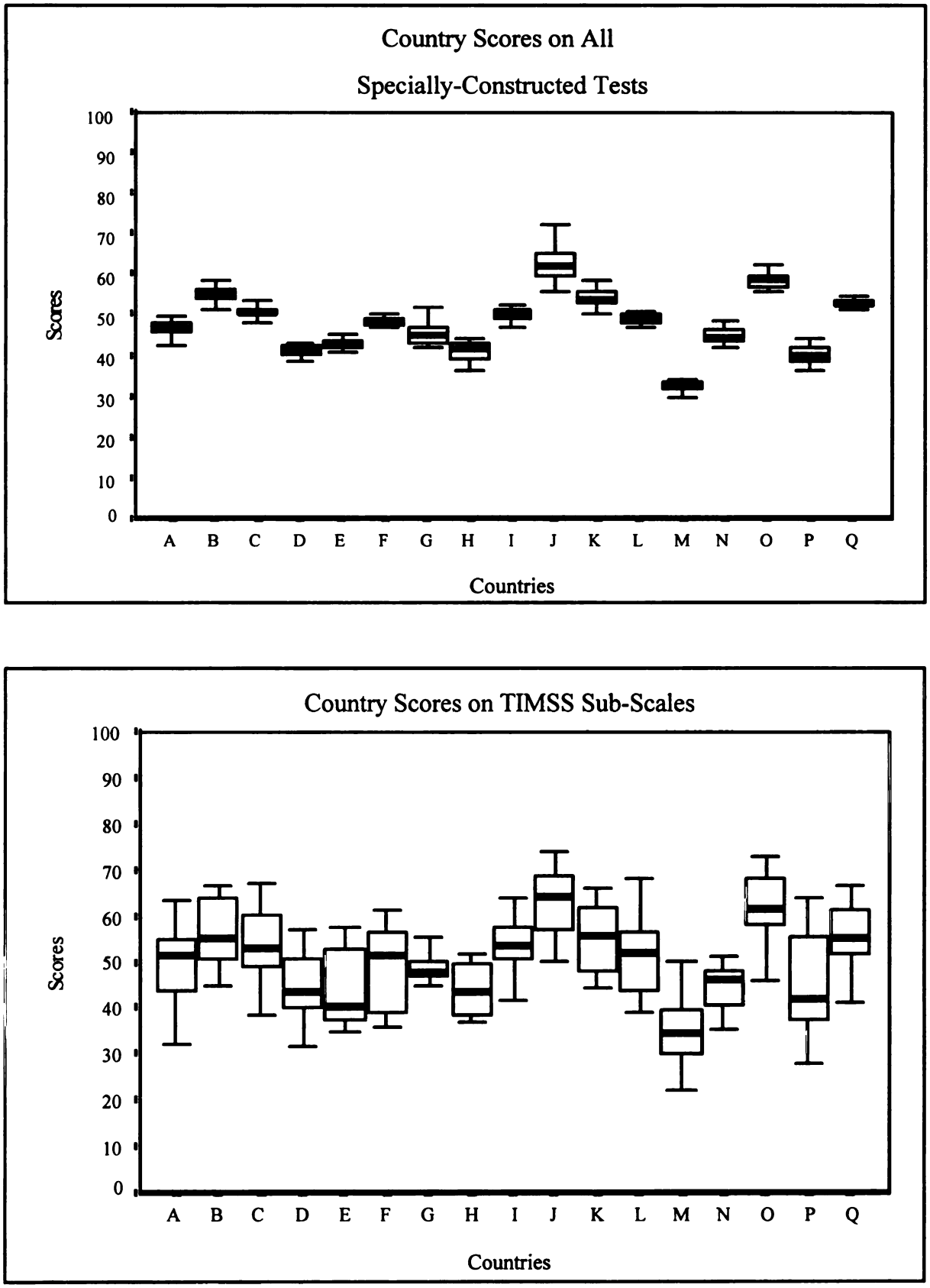
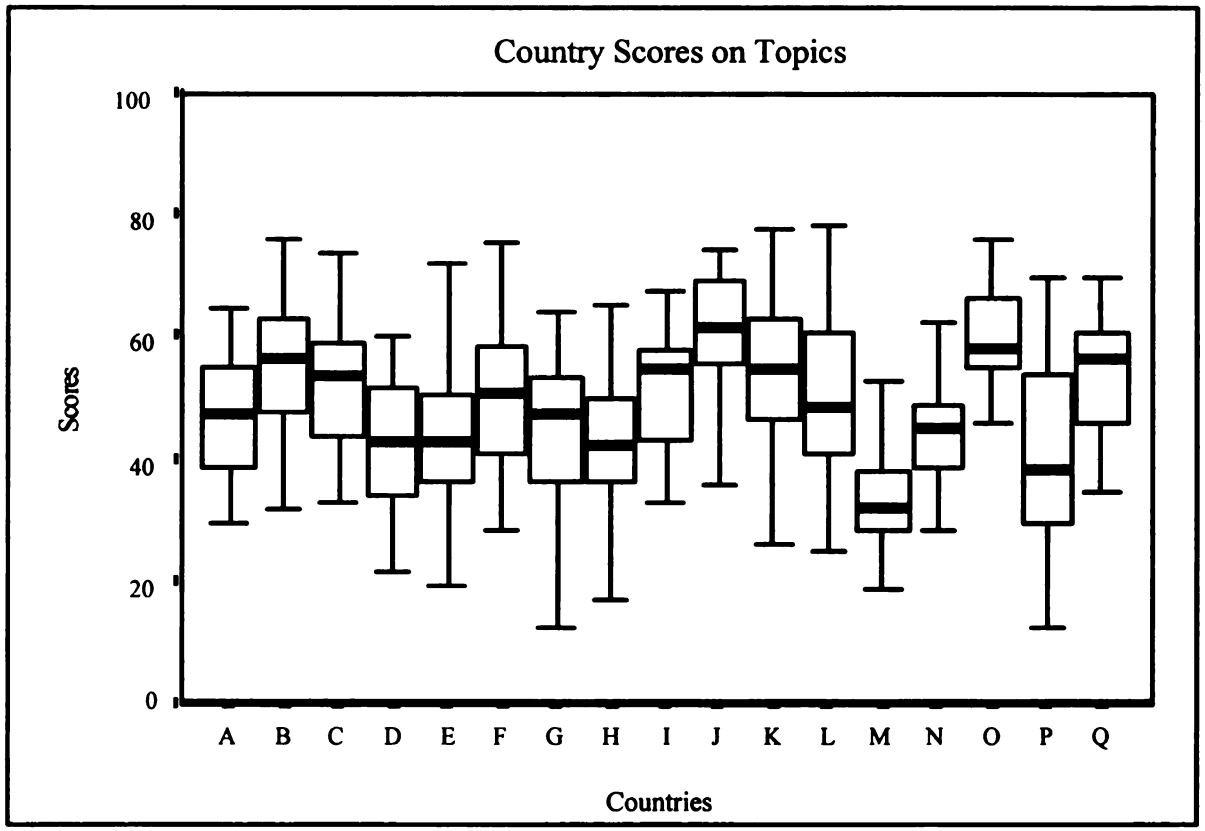


Figure 2. Boxplots of scores on all specially-constructed tests, TIMSS sub-scales, topics, performance expectations, items for topic 1.6.2, and items for topic 1.7.2.



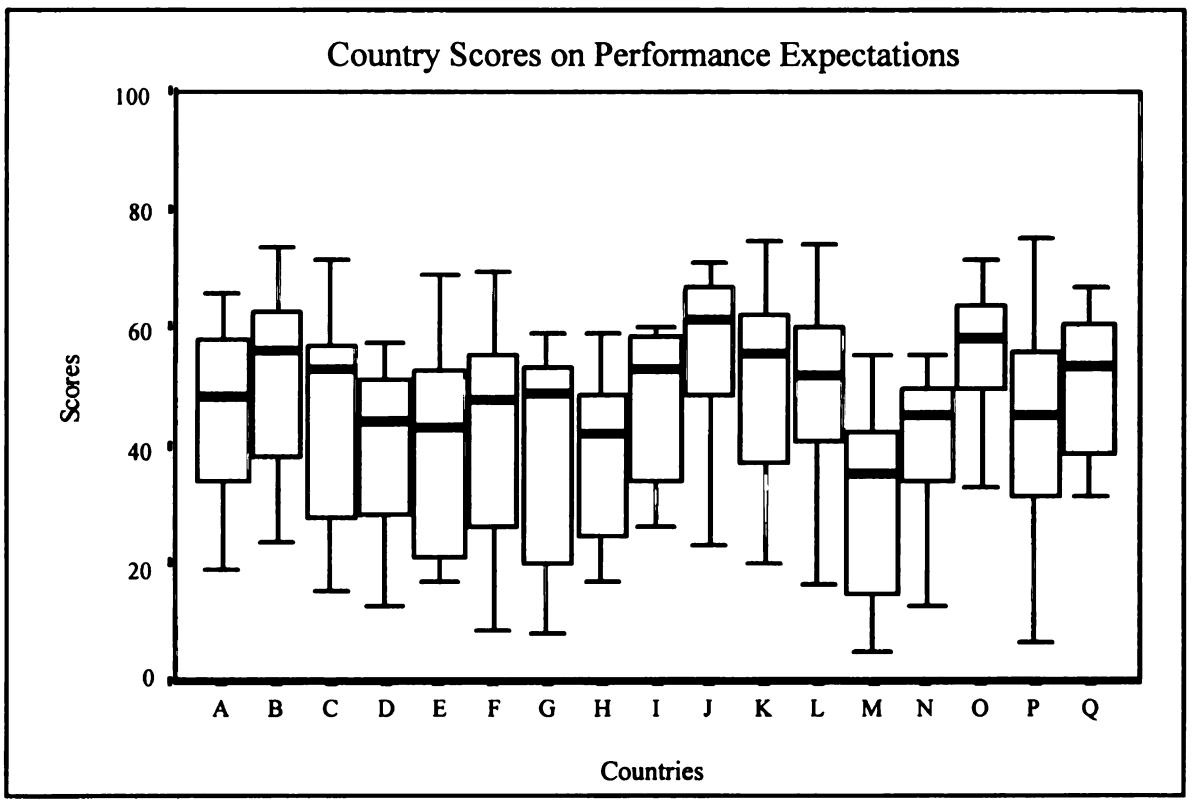
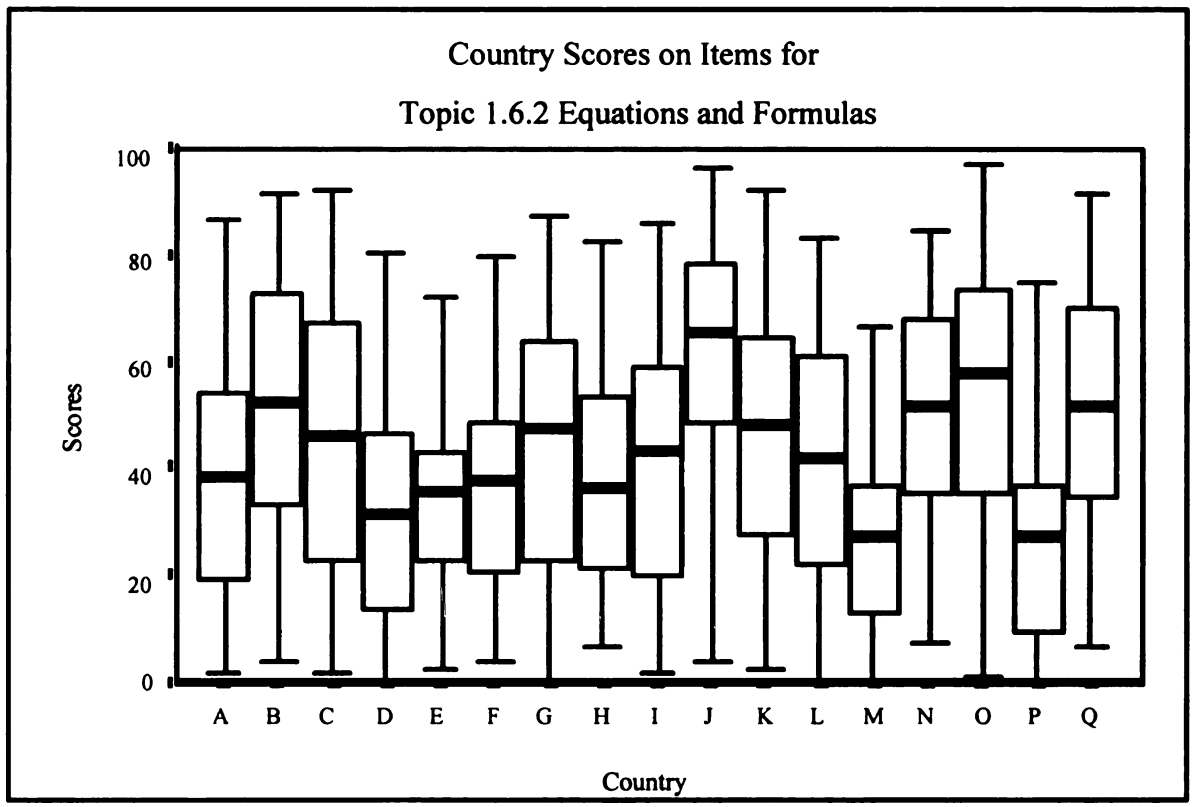


Figure 2. (Cont'd.)



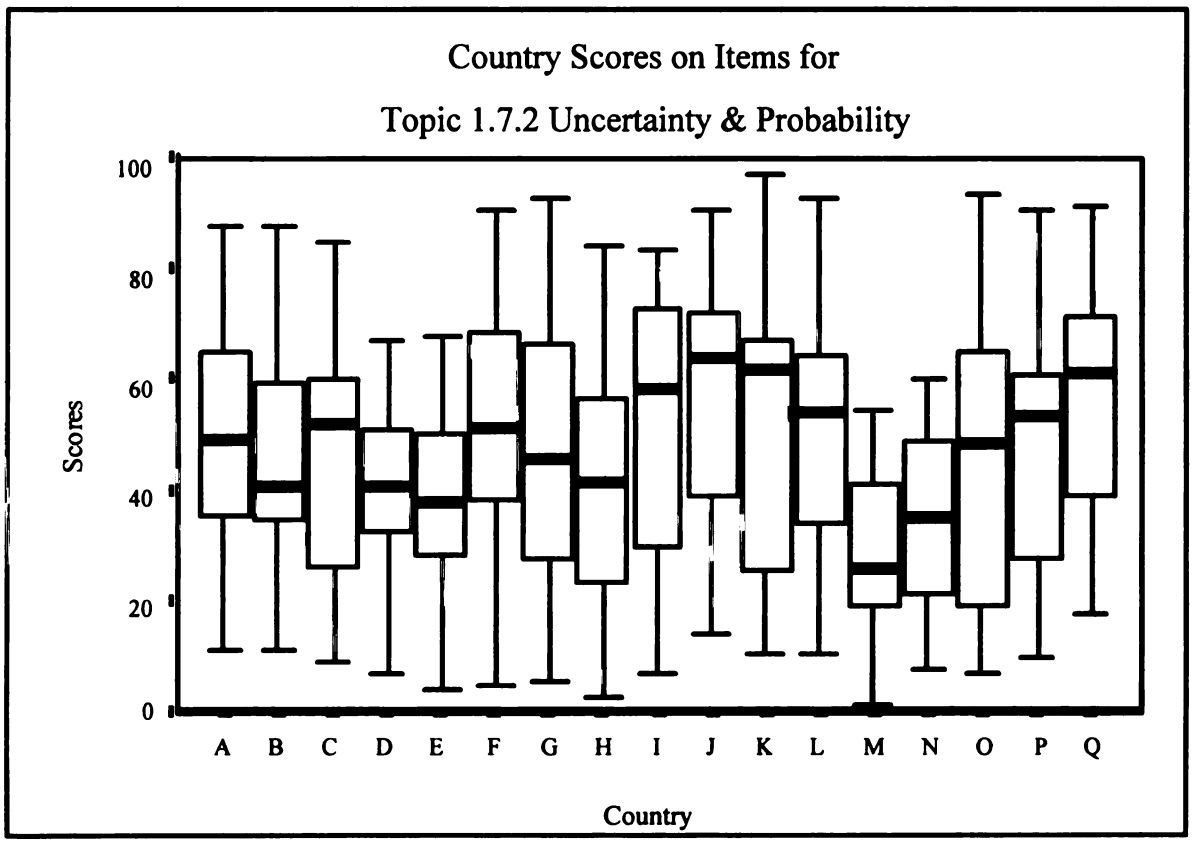


Figure 2. (*Cont'd.*)

Table 62

Estimated Reliabilities and Standard Errors

	Reliability (*a*)	SD	Std.Err.	SD Range across Countries	Estimated Std.Err. Range across Countries
All Tests	0.99	7.1	0.31	1.7- 3.7	.07-.16
Scales	0.95	6.8	1.38	4.9-11.9	.99-2.4
Topics	0.96	9.2	1.84	9.3-16.0	1.9-2.4
PEs	0.90	6.4	1.99	12-18.0	3.7-5.6
1.6.2	0.96	9.2	1.73	17-25.0	3.2-4.7
1.7.2	0.90	6.3	1.95	17-28.0	5.3-8.7

Note. Reliabilities and standard errors were calculated treating each country as a case and each total, scale, topic, performance expectation, or item score as an item. Because individual student results were not available, reliabilities and standard errors could not be calculated for each country. However, reliabilities and standard errors were estimated for each country under the assumption that the test was as reliable for each country as it was for the group of countries as a whole.

The variation in performance on items, topics, and performance expectations highlights the complexity of curriculum-to-test matching, and, hence, evaluations of test validity. The items on the field-trial instrument each had their own "signature" which identifies the unique nature of each item. Few items had the exact same signature. This signature was created by a particular combination of topic and performance-expectation codes and was further enriched by item format. No one country have consistent performance or consistent rankings across items. Evaluations of the signatures of the items on which countries demonstrate strength and weakness provide more information than knowing that students were "low in geometry." The item signatures allow one to evaluate the specific nature of a problem on which students are excelling or with which

they are having difficulty. Additionally, items represent the ways in which the curriculum is delivered in the classroom. Teachers do not teach "math" or "geometry." They teach complex interrelated topics and performances. Thus, the "implemented" curriculum for geometry and the expectations for performance will not be the same from class to class. Aggregations of items and topics, however, begin to mask the multi-dimensional nature of mathematics. What is left over is general-math achievement.

> Schools do not directly teach global constructs, but instead try to develop specific skills, introducing one skill at a time for the student to integrate with previously acquired skills...Hence, there are more apt to be differences between schools and programs at the specific objective level than at the total score level. (Airasian & Madaus, 1983, p. 105-106)

While rankings of countries on general math achievement may provide information of interest to some stakeholders, such rankings certainly do not provide very descriptive information on student achievement. They also may complicate evaluations of test-curriculum match. Certainly each country's mathematics curriculum is more than a collection of isolated topics. Therefore, variation in performance across countries cannot be explained by variation in topic coverage alone. Variations in expectations for performance and the complex blending of topics and performance expectations also must be considered.

Summary

I set out in this study to develop test blueprints for cross-national assessments that validly measure student achievement of topics in the mathematics curriculum for 13-year-old students. However, the variation within and across nations in curriculum and lack of an adequate item pool complicated this goal. Through my analyses, I found that

1. The intended and potentially implemented mathematics curriculum varies across nations and also varies within nations. Some countries include few topics in their curricula (as indicated by the data sources), and others include many. Some countries focus on particular topics; others spread their focus across many topics. However, some commonalties do exist, with a handful of topics either missing from most countries' curriculum sources or being highly emphasized in most countries. Variations within each country's data sources point out the need for multiple representations of math curricula.
2. The content of the field-trial instrument is more similar to the content of the curriculum of some countries than others and is more similar to the content of some of the data sources than others. This "differential match" has implications for the validity of inferences made from the test, but final conclusions about test validity will depend on the purpose for which the test will be used.
3. Test blueprints varied according to test purpose. Topic coverage and emphasis were inconsistent across the blueprints due to the variability in the curriculum sources. Some blueprints, though, were very similar to one another (e.g., all the union blueprints), while others were very different (e.g., the strict intersections). Each blueprint provides a different look at student achievement.
4. The content of the test blueprints for specially-constructed tests were more similar to the curriculum sources than was the content of the field-trial instrument, especially in the relative emphasis of topics. Thus, an increase in validity and less bias would be expected. However, variations in the similarity of the content of these tests to the content of the curriculum of particular countries still existed - primarily due to the

missing topics. Impact of the mis-match needs to be balanced with other information provided by the tests. For example, because the strict-intersection tests do not represent the entire curriculum of any country, the weighting of the topics on the test relative to other topics in a country's curriculum is lost. However, the strict-intersection tests do provide information on how students perform on topics included in the curriculum of all countries. Furthermore, the unique tests provide an indication on how students performed on their unique curricula, and these tests have a good fit to the curriculum of each country. However, comparisons of student performance when all students do not take the same test are complicated.

5. Variation in county scores and ranks on specially-constructed tests was minimal; however, some isolated differences did exist. Patterns suggest that tests covering a comprehensive range of math topics are unlikely to produce striking variations in performance, suggesting that, at the total-score level, the impact of test-curriculum mis-match is likely to be minimal. However, variation in performance across topics and performance expectations indicate that country ranks of total scores may be reflecting a general-math achievement, rather than achievement of a particular curriculum. Performance across countries did vary when unique tests were developed based on topics and performance expectations. The concept of test-curriculum match is more complex than merely matching on topic coverage. The content of the curriculum is made up not only of topics but also of expectations for performance on each topic. Both vary separately and together across countries. Thus, all countries may include "algebra" topics in their curriculum, but may have widely different intentions for student achievement. These differing expectations,

undoubtedly, would result in subtle differences in goals, textbooks, and instruction. Such differences will need to be considered in evaluating test-content validity.

Limitations

Some limitations of the results should be pointed out. First, the study used pilot data. As such, the student samples were not random within countries and items were being tested for inclusion on the final TIMSS assessment. However, population sample sizes were within or close to the IEA guidelines in all countries. Furthermore, the item pool on the field-trial instrument was much larger than on the final TIMSS assessment. In fact, one of the main purposes of the field trial was to try out the extended-response items, and many of these items were dropped from the final assessment due to testing-time constraints.

Another limitation was the lack of student-level data for the field trial. Only country-level *p*-values were provided. Thus, it was impossible to construct scores for individual students or evaluate with any certainty variation on items. For this reason, significance tests were not conducted for most analyses. It is difficult to determine, therefore, the statistical significance of these results.

The lack of items for some topics was another limitation. Country-level performance was not available on all topics, some of which factored highly in the curriculum of some countries. Results may have changed had these data been available. The lack of depth of coverage in all topic areas was also a problem. A full range of items covering all topics crossed with all performance expectations would be ideal. However, the item sample from the field-trial instrument likely reflects the reality of test

development. The item pool of rich items covering all topics and performances is not yet available (Garden & Orpwood, 1996).

Differences in the level of specificity of topic codes presents another issue to consider. Some select topics on the framework are coded in detail (e.g., the fractions topics). Thus, a richer picture of curriculum is available for these topics. Other topics are reported at the more global level. For example, algebra is covered by only two codes. However, algebra (1.6.2) is prevalent in the curriculum of the countries. A finer distinction of algebra topics may have pointed out more variation in curriculum coverage and differences in the similarity of the test items to these curricula.

Finally, my analyses are based on the assumption that all items are good measures of the behaviors they represent. As discussed earlier, content validity is just one aspect of validity, and should never be taken as a final indication of test validity. It is important that items not only represent the content of the domain, but also do it well.

Recommendations and Conclusion

My first recommendation is that a higher quality item pool be developed for cross-national work. Several topics important to many countries were missing from the test, and items measuring complex applications of topic knowledge and understanding were not available for all topics. The items were not a comprehensive representation of the performance expectation aspect of the framework. It is difficult to determine how country performance might vary if more items measuring higher order skills were included on the test. Many countries expect their students to demonstrate complex use of subject matter. If researchers want to adequately measure such skills, better items will

need to be developed. Fortunately, within the U.S. research is being conducted on content standards as well as performance standards (Linn & Baker, 1995). Cross-national researchers should look to these studies to guide their research. Until better item pools are developed, results of cross-national achievement testing should be interpreted with caution.

Second, researchers developing cross-national achievement tests should clearly state the purpose and the domain of the instrument. Without such information, it is not possible to evaluate how well a test represents a domain. This is one of the first rules of thumb taught in any measurement course. Unfortunately, it is not often followed, and consumers are left to guess at the domain, or researchers imply that the test represents more than it actually does. Secondary analysts may also be guilty of applying test results to too broad a domain. These situations can be avoided by clearly describing the item domain.

Third, this study has shown variation in curriculum, in test-curriculum match, and in performance on topics and performance expectations. These variations should be reported with the results of cross-national achievement tests so stakeholders can better interpret findings. The study has demonstrated the importance of the first rule in test development - identify the purpose of the testing. Simply starting with collections of items and piecing them together to fit a content map is not adequate. Test developers need to clearly articulate what they are attempting to measure and what types of inferences are appropriate and inappropriate.

Finally, my recommendation is that researchers take into account the complexity of the curriculum and items when evaluating test-curriculum match. A clear match with

curriculum is unlikely to emerge by focusing only on topics. Two countries may demonstrate the same level of coverage on a topic, but have different expectations for performance. Likewise, two items may measure the same topic, but be very different in the type of performance or application expected. Replications of the analyses in this study may produce different results if performance expectations were included in the analyses.

In the current period of educational reform, cross-national studies are receiving renewed attention as educational systems across the world strive for "world class" standards and fight to maintain or gain competitive economic footing (Linn & Baker, 1995; Porter, 1990; Schmidt & Valverde, 1995). The results of such studies are useful for both accountability and school improvement. However, researchers and policy-makers cannot allow themselves to be lured into the international horse race and to be swayed by public demands for simplistic results and explanations. The international educational system is varied and complex, and analyses of this system should reflect this complexity.

> My answer to people who want comparative standings is to give them comparative standings - lots of them: in different topics, at different ages, with different kinds of tasks, both unadjusted and adjusted for factors such as national curricula and proportion of students in school. Recognizing that no single index of achievement can tell the full story and that each has its own limitations, we increase our understanding of how nations compare by increasing the breadth of our vision. Even so, however, simply ascertaining nations' relative standing tells us little about how to set educational policy or improve instructional practice.(Mislevy, 1995, p.419)

APPENDICES

APPENDIX A

Appendix A

Mathematics Curriculum-Framework Categories*

Content

1.1 Numbers
- 1.1.1 Whole numbers
  - 1.1.1.1 Meaning
  - 1.1.1.2 Operations
  - 1.1.1.3 Properties of operations
- 1.1.2 Fractions and decimals
  - 1.1.2.1 Common fractions
  - 1.1.2.2 Decimal fractions
  - 1.1.2.3 Relationships of common and decimal fractions
  - 1.1.2.4 Percentages
  - 1.1.2.5 Properties of common and decimal fractions
- 1.1.3 Integer, rational, and real numbers
  - 1.1.3.1 Negative numbers, integers, and their properties
  - 1.1.3.2 Rational numbers and their properties
  - 1.1.3.3 Real numbers, their subsets, and their properties
- 1.1.4 Other numbers and number concepts
  - 1.1.4.1 Binary arithmetic and/or other number bases
  - 1.1.4.2 Exponents, roots, and radicals (integer, rational, and real exponents)
  - 1.1.4.3 Complex numbers and their properties
  - 1.1.4.4 Number theory
  - 1.1.4.5 Counting
- 1.1.5 Estimation and number sense
  - 1.1.5.1 Estimating quantity and size
  - 1.1.5.2 Rounding and significant figures
  - 1.1.5.3 Estimating computations
  - 1.1.5.4 Exponents and orders of magnitude

---

* from Robitaille, D.F., McKnight, C., Schmidt, W.H., Britton, E., Raizen, S., & Nicol, C. (1993). *Curriculum frameworks for mathematics and science.* Vancouver, Canada: Pacific Educational Press.

1.2 Measurement
- 1.2.1 Units
- 1.2.2 Perimeter, area, and volume
- 1.2.3 Estimation and errors

1.3 Geometry: position, visualization, and shape
- 1.3.1 Two-dimensional geometry: coordinate geometry
- 1.3.2 Two-dimensional geometry: basics
- 1.3.3 Two-dimensional geometry: polygons and circles
- 1.3.4 Three-dimensional geometry
- 1.3.5 Vectors

1.4 Geometry: symmetry, congruence, and similarity
- 1.4.1 Transformations
- 1.4.2 Congruence and similarity
- 1.4.3 Constructions using straight-edge and compass

1.5 Proportionality
- 1.5.1 Proportionality concepts
- 1.5.2 Proportionality problems
- 1.5.3 Slope and Trigonometry
- 1.5.4 Linear Interpolation and Extrapolation

1.6 Functions, relations, and equations
- 1.6.1 Patterns, relations, and functions
- 1.6.2 Equations and formulas

1.7 Data representation, probability, and statistics
- 1.7.1 Data representation and analysis
- 1.7.2 Uncertainty and probability

1.8 Elementary analysis
- 1.8.1 Infinite processes
- 1.8.2 Change

1.9 Validation and structure
- 1.9.1 Validation and justification
- 1.9.2 Structuring and abstracting

1.10 Other Content
- 1.10.1 Informatics

Performance Expectations

2.1 Knowing
- 2.1.1 Representing
- 2.1.2 Recognizing equivalents
- 2.1.3 Recalling mathematical objects and properties

2.2 Using routine procedures
- 2.2.1 Using equipment
- 2.2.2 Performing routine procedures
- 2.2.3 Using more complex procedures

2.3 Investigating and problem solving
- 2.3.1 Formulating and clarifying problems and situations
- 2.3.2 Developing strategy
- 2.3.3 Solving
- 2.3.4 Predicting
- 2.3.5 Verifying

2.4 Mathematical reasoning
- 2.4.1 Developing notation and vocabulary
- 2.4.2 Developing algorithms
- 2.4.3 Generalizing
- 2.4.4 Conjecturing
- 2.4.5 Justifying and proving
- 2.4.6 Axiomatizing

2.5 Communicating
- 2.5.1 Using vocabulary and notation
- 2.5.2 Relating representations
- 2.5.3 Describing/discussing
- 2.5.4 Critiquing

Perspectives

3.1 Attitudes towards science, mathematics, and technology

3.2 Careers involving science, mathematics, and technology
- 3.2.1 Promoting careers in science, mathematics, and technology
- 3.2.2 Promoting the importance of science, mathematics, and technology in non-technical careers

3.3 Participation in science and mathematics by underrepresented groups
3.4 Science, mathematics, and technology to increase interest
3.5 Scientific and mathematical habits of mind

APPENDIX B

Table B1

Topic Coverage on the TIMSS Mathematics Item Field-Trial Instrument for Population 2

	Book 3				Book 5				Book 6				Book 8				MC		SA		ER		Total	
code	mc	sa	er	T	mc	sa	er	T	mc	sa	er	T	mc	sa	er	T	u	l	u	l	u	l	u	l
1.1.1.1	0	0	0	0	3	0	0	3	1	0	0	1	0	0	0	0	4	0	0	0	0	0	4	0
1.1.1.2	0	2	0	2	2	0	1	3	**3**	0	1	4	**3**	2	1	6	7	1	4	0	3	0	14	1
1.1.1.3	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	2	0	0	0	0	0	2	0
1.1.2.1	**9**	**4**	0	13	**5**	0	0	5	**9**	2	0	11	**7**	1	0	8	27	3	7	0	0	0	34	3
1.1.2.2	**3**	1	1	5	**3**	0	0	3	5	0	0	5	5	0	0	5	15	1	1	0	1	0	17	1
1.1.2.3	2	0	0	2	**4**	0	0	4	2	2	0	4	**2**	0	0	2	9	1	2	0	0	0	11	1
1.1.2.4	1	0	1	2	2	0	0	2	1	0	0	1	1	0	1	2	5	0	0	0	2	0	7	0
1.1.2.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.1.3.1	1	0	0	1	0	0	0	0	1	0	0	1	1	0	0	1	3	0	0	0	0	0	3	0
1.1.3.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.1.3.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.1.4.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.1.4.2	1	0	0	1	1	0	0	1	1	0	0	1	0	0	0	0	3	0	0	0	0	0	3	0
1.1.4.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.1.4.4	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0
1.1.4.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.1.5.1	**1**	0	0	1	**7**	0	0	7	**2**	0	0	2	**2**	0	0	2	9	3	0	0	0	0	9	3
1.1.5.2	0	0	0	0	3	1	0	4	1	0	0	1	3	0	0	3	7	0	1	0	0	0	8	0
1.1.5.3	1	0	0	1	5	0	1	6	0	0	0	0	0	0	0	0	6	0	0	0	1	0	7	0
1.1.5.4	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0
1.2.1	4	0	0	4	3	0	0	3	**3**	0	0	3	**8**	1	0	9	17	1	1	0	0	0	18	1
1.2.2	1	0	0	1	**3**	0	0	3	**3**	0	1	4	5	2	2	9	11	1	2	0	3	0	16	1
1.2.3	0	0	0	0	2	0	0	2	0	0	0	0	1	0	0	1	3	0	0	0	0	0	3	0

mc=multiple choice; sa=short answer; er=extended response;
u=unique items; l=linked items

Bolded items indicate where linked items occur.
Shading highlights cells with no items.

Table B1 (*Cont'd.*)

	Book 3				Book 5				Book 6				Book 8				MC		SA		ER		Total	
code	mc	sa	er	T	mc	sa	er	T	mc	sa	er	T	mc	sa	er	T	u	l	u	l	u	l	u	l
1.3.1	**1**	0	0	1	**4**	0	0	4	**0**	0	0	0	1	0	1	2	5	1	0	0	1	0	6	1
1.3.2	**0**	0	0	0	**2**	1	0	3	**3**	0	1	4	1	0	0	1	5	1	1	0	1	0	7	1
1.3.3	**5**	0	0	5	**2**	0	0	2	1	0	0	1	1	0	0	1	8	1	0	0	0	0	8	1
1.3.4	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	4	0	0	0	0	0	4	0
1.3.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.4.1	**1**	0	1	2	**3**	0	1	4	**2**	0	1	3	2	0	0	2	7	1	0	0	3	0	10	1
1.4.2	**9**	0	2	11	**5**	0	1	6	2	0	0	2	0	0	0	0	12	4	0	0	2	1	14	5
1.4.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.5.1	0	0	**1**	1	**1**	0	**1**	2	**5**	1	0	6	0	0	1	1	5	1	1	0	2	1	8	2
1.5.2	2	2	0	4	2	0	0	2	**11**	2	2	15	**3**	0	0	3	17	1	4	0	2	0	23	1
1.5.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.5.4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.6.1	**2**	0	0	2	**3**	0	0	3	**6**	0	0	6	**4**	0	1	5	11	4	0	0	1	0	12	4
1.6.2	**8**	1	0	9	**6**	0	0	6	**15**	**1**	0	16	**8**	**1**	4	13	27	10	2	1	4	0	33	11
1.7.1	4	1	3	8	5	1	0	6	5	0	0	5	7	0	1	8	21	0	2	0	4	0	27	0
1.7.2	**1**	**1**	0	2	**10**	**1**	0	11	**1**	0	0	1	0	0	0	0	10	2	1	1	0	0	11	3
1.8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Total	48	9	6	63	65	5	4	75	64	7	3	74	50	6	7	63	197	30	25	2	19	1	241	33

mc=multiple choice; sa=short answer; er=extended response;
u=unique items; l=linked items

Bolded items indicate where linked items occur. Shading indicates cells with no items.
The last row of totals indicate the number of items within each book or on the test. Because items may have more than one topic code, columns may not add up to the totals.

APPENDIX C

Appendix C

Curriculum Data for Each Country and Each Data Source

Table C1

Expert Topic Mapping Proportions for 13 Year Old Students

Topic Code	Country A	B	C	D	E	F	G	H	I	J
1.1.1.1	0	0.024	0.027	0.020	0	0.027	0	0	0.020	0
1.1.1.2	0.025	0.024	0.027	0.020	0	0.054	0	0	0.020	0
1.1.1.3	0.050	0	0.027	0.020	0	0.054	0	0	0.020	0
1.1.2.1	0.025	0.024	0.027	0.039	0.030	0.054	0.105	0.024	0.041	0
1.1.2.2	0.025	0.024	0.027	0.039	0.030	0.027	0.105	0.024	0.041	0
1.1.2.3	0.025	0	0.027	0.039	0	0.027	0.105	0.024	0.041	0
1.1.2.4	0.050	0.048	0.027	0.020	0.030	0.054	0	0	0.041	0
1.1.2.5	0.025	0.024	0.027	0.020	0.030	0.027	0.105	0.024	0	0
1.1.3.1	0.025	0.024	0.027	0.039	0	0.027	0	0.049	0.041	0
1.1.3.2	0.025	0.024	0.054	0.020	0.030	0	0.105	0.049	0.020	0
1.1.3.3	0	0.048	0.054	0.020	0.030	0	0.053	0.024	0.020	0
1.1.4.1	0	0	0	0	0	0	0	0	0	0.032
1.1.4.2	0.025	0.048	0.027	0.039	0	0.027	0.053	0.049	0.041	0
1.1.4.3	0	0	0	0	0	0	0	0	0	0
1.1.4.4	0.025	0	0.027	0.020	0.030	0.027	0	0.024	0.041	0.032
1.1.4.5	0	0	0	0	0	0	0	0	0	0
1.1.5.1	0.025	0.024	0.027	0.020	0	0	0	0	0	0
1.1.5.2	0.050	0.048	0.027	0.020	0	0.027	0	0.049	0.041	0.065
1.1.5.3	0.025	0.024	0.027	0.020	0	0	0	0.049	0.041	0
1.1.5.4	0	0.024	0.054	0.039	0	0.027	0	0	0.041	0
1.2.1	0.025	0.024	0.054	0.020	0.030	0.054	0	0.049	0.041	0
1.2.2	0.025	0.048	0.027	0.020	0.030	0.027	0	0.024	0.041	0
1.2.3	0.025	0.024	0.027	0.020	0	0.027	0	0.024	0.020	0.065
1.3.1	0.025	0.024	0.000	0.039	0	0.000	0.053	0.024	0.020	0.065
1.3.2	0.025	0.024	0.027	0.039	0.030	0.027	0.053	0.024	0.041	0.065
1.3.3	0.025	0.048	0.027	0.039	0.030	0.054	0	0.049	0.020	0.065
1.3.4	0.050	0.048	0.027	0.039	0.030	0.027	0.053	0.049	0	0.032
1.3.5	0	0	0	0	0	0	0	0	0	0
1.4.1	0.025	0.024	0.027	0.020	0.061	0.054	0	0.049	0.020	0.032
1.4.2	0.025	0.024	0.027	0.020	0.061	0.027	0	0.024	0.041	0.065
1.4.3	0.025	0.024	0.027	0.020	0.061	0.027	0	0	0.020	0.032
1.5.1	0.050	0.024	0.054	0.039	0.030	0.027	0	0.049	0.041	0
1.5.2	0.050	0.024	0.027	0.039	0.061	0.027	0	0.049	0.020	0.065
1.5.3	0.025	0.048	0	0.020	0.061	0	0	0.049	0	0.065
1.5.4	0	0	0	0.020	0.061	0	0	0	0	0.032
1.6.1	0.050	0.048	0.027	0.020	0.061	0.027	0.053	0.024	0.041	0.065
1.6.2	0.025	0.048	0.027	0.039	0.061	0.054	0.053	0.049	0.041	0.065
1.7.1	0.025	0.024	0	0.039	0.061	0.054	0.053	0.049	0.020	0.065
1.7.2	0.025	0	0	0.020	0.030	0.027	0.053	0	0	0
1.8.1	0	0	0	0	0	0	0	0	0	0
1.8.2	0	0	0	0	0	0	0	0	0	0
1.9.1	0	0	0.027	0	0.030	0	0	0	0.020	0.065
1.9.2	0.025	0	0.027	0.020	0	0	0	0	0.041	0
1.10.1	0.025	0.048	0	0.020	0	0	0	0.024	0	0.032
Average[b]	0.030	0.032	0.031	0.032	0.042	0.036	0.071	0.037	0.032	0.053
Deviation	0.016	0.018	0.016	0.013	0.023	0.02	0.036	0.02	0.017	0.028
Count[a]	33/7	31/11	32/5	37/14	24/9	28/9	14/5	27/14	31/18	19/12

[a] Number of non-zero topics: number of emphasized topics. [b] Average of non-zero numbers.

Table C1(*Cont'd.*)

Topic Code	K	L	M	N	O	P	Q	*X*	*SD*	Median	Max. Prop.	Count[a]
1.1.1.1	0.027	0	0	0.027	0.034	0	0	0.012	0.013	0	0.034	8/1
1.1.1.2	0	0.024	0	0.027	0	0.071	0	0.017	0.020	0.020	0.071	9/3
1.1.1.3	0	0.024	0	0.027	0	0.071	0	0.017	0.022	0	0.071	8/4
1.1.2.1	0.027	0.024	0.028	0.027	0	0.071	0	0.032	0.025	0.027	0.105	14/6
1.1.2.2	0.027	0.024	0	0.027	0	0.036	0.020	0.028	0.023	0.027	0.105	14/4
1.1.2.3	0.027	0.024	0	0.027	0	0.036	0.020	0.025	0.024	0.025	0.105	12/4
1.1.2.4	0.054	0.024	0	0.027	0	0.036	0.020	0.025	0.019	0.027	0.054	12/6
1.1.2.5	0	0.024	0	0.027	0	0.071	0.020	0.025	0.026	0.024	0.105	12/3
1.1.3.1	0.054	0.049	0.028	0.027	0.034	0.071	0.041	0.032	0.019	0.028	0.071	14/8
1.1.3.2	0.054	0.024	0.028	0.027	0.034	0	0.020	0.030	0.025	0.025	0.105	14/5
1.1.3.3	0	0.024	0	0.027	0.034	0	0.020	0.021	0.018	0.020	0.054	11/3
1.1.4.1	0	0	0	0.014	0	0	0	0.003	0.008	0	0.032	2/0
1.1.4.2	0.054	0.024	0.056	0.027	0.069	0.036	0.041	0.036	0.018	0.039	0.069	15/9
1.1.4.3	0	0	0	0	0	0	0	0	0	0	0	0/0
1.1.4.4	0	0.049	0.056	0.027	0.034	0	0.020	0.024	0.016	0.027	0.056	13/4
1.1.4.5	0	0	0.028	0	0	0	0.041	0.004	0.011	0	0.041	2/1
1.1.5.1	0.027	0.024	0	0.027	0	0.036	0.020	0.014	0.013	0.020	0.036	9/1
1.1.5.2	0.027	0.049	0.028	0.027	0	0.036	0.020	0.030	0.018	0.027	0.065	14/7
1.1.5.3	0.054	0.049	0.028	0.027	0	0.036	0.020	0.023	0.018	0.025	0.054	12/5
1.1.5.4	0.027	0.024	0.056	0.027	0.069	0.036	0.041	0.027	0.021	0.027	0.069	12/7
1.2.1	0.027	0.024	0.028	0.027	0.034	0.036	0.020	0.029	0.015	0.027	0.054	15/5
1.2.2	0.054	0.024	0.028	0.027	0.034	0.036	0.041	0.029	0.014	0.027	0.054	15/5
1.2.3	0	0.024	0	0.027	0	0	0.020	0.018	0.016	0.020	0.065	11/2
1.3.1	0.054	0.024	0.028	0.027	0.034	0.036	0.041	0.029	0.018	0.027	0.065	14/5
1.3.2	0.054	0.049	0.028	0.027	0.034	0.071	0.041	0.039	0.014	0.034	0.071	17/8
1.3.3	0.027	0.024	0.056	0.027	0.034	0.036	0.020	0.034	0.016	0.030	0.065	16/7
1.3.4	0.027	0.024	0.028	0.027	0.069	0	0.041	0.034	0.017	0.030	0.069	15/7
1.3.5	0	0	0.028	0	0	0	0	0.002	0.007	0	0.028	1/0
1.4.1	0.054	0.024	0.028	0.027	0.069	0	0.041	0.033	0.019	0.027	0.069	15/7
1.4.2	0	0.024	0.056	0.027	0.069	0	0.041	0.031	0.021	0.027	0.069	14/7
1.4.3	0	0.049	0.056	0.014	0.034	0	0.020	0.024	0.018	0.024	0.061	13/3
1.5.1	0.027	0.024	0.028	0.027	0.034	0.036	0.020	0.030	0.015	0.028	0.054	15/6
1.5.2	0.054	0.024	0.056	0.027	0.069	0.071	0.041	0.041	0.019	0.041	0.071	16/11
1.5.3	0.027	0	0.028	0.014	0	0	0.020	0.021	0.022	0.020	0.065	10/4
1.5.4	0.027	0	0	0.027	0	0	0.020	0.011	0.017	0	0.061	6/2
1.6.1	0.027	0.024	0.028	0.027	0	0	0.020	0.032	0.018	0.027	0.065	15/6
1.6.2	0.027	0.024	0.056	0.027	0.069	0	0.041	0.041	0.017	0.041	0.069	16/11
1.7.1	0.027	0.024	0.056	0.027	0.069	0.036	0.041	0.039	0.018	0.039	0.069	19/9
1.7.2	0.027	0.024	0	0.014	0	0	0.041	0.015	0.016	0.014	0.053	9/1
1.8.1	0	0	0	0	0	0	0	0	0	0	0	0/0
1.8.2	0	0	0	0	0	0	0	0	0	0	0	0/0
1.9.1	0	0.024	0	0.027	0	0	0.020	0.013	0.018	0	0.065	7/2
1.9.2	0	0.024	0.028	0.027	0	0	0.020	0.012	0.014	0	0.041	8/2
1.10.1	0	0.024	0.028	0.027	0.034	0	0.020	0.017	0.015	0.020	0.048	10/2
Average[b]	0.037	0.029	0.037	0.026	0.048	0.048	0.029	0.024	0.018	0.027	0.064	12/5
Deviation	0.021	0.014	0.021	0.009	0.026	0.026	0.015	0.012	0.006	0.013	0.025	5/3
Count[a]	27/10	35/6	27/9	39/35	21/8	21/7	35/14	41	41	33	41	41/39

[a] *Number of non-zero topics/number of emphasized topics.* [b] *Average of non-zero numbers.*

Table C2

Curriculum Guide Topic Coverage Data

Topic Code	Country A	B	C	D	E	F	G	H	I	J
1.1.1.1	0.034	0	0.026	0.029	0.045	0.028	0.056	0.091	0.032	0
1.1.1.2	0.034	0.042	0.026	0.029	0.045	0.028	0	0.091	0.032	0
1.1.1.3	0.034	0	0.026	0.029	0.045	0.028	0	0.091	0.032	0
1.1.2.1	0.034	0	0.026	0.029	0.045	0.028	0	0	0.032	0
1.1.2.2	0.034	0	0.026	0.029	0.045	0.028	0.056	0	0.032	0
1.1.2.3	0	0	0.026	0.029	0.045	0.028	0	0	0.032	0
1.1.2.4	0	0.042	0.026	0.029	0.045	0.028	0	0	0.032	0
1.1.2.5	0	0	0.026	0.029	0.000	0.028	0	0	0	0
1.1.3.1	0.034	0.042	0.026	0.029	0.045	0.028	0.056	0	0.032	0.04
1.1.3.2	0	0.042	0.026	0.029	0.000	0.028	0.056	0	0.032	0.04
1.1.3.3	0	0.042	0.026	0.029	0.000	0.028	0	0.091	0.032	0.04
1.1.4.1	0.034	0	0.026	0	0	0	0	0	0	0.04
1.1.4.2	0	0.042	0.026	0.029	0.000	0.028	0	0	0.032	0
1.1.4.3	0	0	0.026	0	0	0	0	0	0	0
1.1.4.4	0.034	0	0.026	0.029	0.045	0.028	0.056	0	0.032	0.04
1.1.4.5	0.034	0	0.026	0	0	0.028	0	0	0	0
1.1.5.1	0.034	0.042	0	0.029	0.000	0	0	0	0	0
1.1.5.2	0.034	0.042	0.026	0.029	0.000	0	0	0	0.032	0
1.1.5.3	0.034	0.042	0.026	0.029	0.000	0	0	0	0	0
1.1.5.4	0.034	0.042	0	0.029	0.000	0	0	0	0.032	0.04
1.2.1	0.034	0	0.026	0.029	0.000	0.028	0.056	0	0.032	0.04
1.2.2	0.034	0	0.026	0.029	0.000	0.028	0.056	0	0.032	0.04
1.2.3	0.034	0	0.026	0.029	0.000	0.028	0.056	0	0	0.04
1.3.1	0.034	0.042	0.026	0.029	0.045	0.028	0	0	0.032	0.04
1.3.2	0.034	0.042	0.026	0.029	0.000	0.028	0.056	0	0.032	0.04
1.3.3	0.034	0.042	0.026	0.029	0.045	0.028	0.056	0.091	0.032	0.04
1.3.4	0.034	0.042	0.026	0.029	0.045	0.028	0.056	0.091	0.032	0.04
1.3.5	0.000	0.042	0.026	0.029	0.000	0.028	0	0	0.032	0.04
1.4.1	0.034	0.042	0.026	0.029	0.045	0.028	0	0.091	0.032	0.04
1.4.2	0	0.042	0.026	0.029	0.045	0.028	0.056	0	0	0.04
1.4.3	0	0.042	0.026	0.029	0.000	0.028	0	0	0.032	0.04
1.5.1	0.034	0.042	0.026	0.029	0.045	0.028	0.056	0	0.032	0.04
1.5.2	0.034	0.042	0.026	0.029	0.045	0.028	0.056	0	0.032	0.04
1.5.3	0	0	0.026	0	0	0	0	0.091	0.032	0
1.5.4	0	0	0.026	0	0	0.028	0	0	0	0.04
1.6.1	0.034	0.042	0.026	0.029	0.045	0.028	0.056	0.091	0.032	0.04
1.6.2	0.034	0.042	0.026	0.029	0.045	0.028	0.056	0.091	0.032	0.04
1.7.1	0.034	0.042	0.026	0.029	0.045	0.028	0	0.091	0.032	0.04
1.7.2	0.034	0	0.026	0.029	0.045	0.028	0.056	0	0	0.04
1.8.1	0	0	0	0	0.045	0	0	0	0	0
1.8.2	0.034	0	0	0	0	0.028	0	0	0	0
1.9.1	0	0	0.026	0	0	0.028	0	0	0	0
1.9.2	0	0	0.026	0	0	0.028	0.056	0	0.032	0
1.10.1	0.034	0.042	0	0.029	0.045	0.028	0	0	0.032	0.04
Average[a]	0.034	0.042	0.026	0.029	0.045	0.028	0.056	0.091	0.032	0.040
Standard Deviation	0.016	0.021	0.008	0.012	0.023	0.011	0.027	0.039	0.015	0.020
Count	29	24	39	35	22	36	18	11	31	25

[a]Average of non-zero numbers.

Table C2 (*Cont'd.*)

Topic Code	Country K	L	M	N	O	P	Q	*X*	*SD*	Median	Max. Prop.	Count
1.1.1.1	0	0.033	0	0	0	0	0.023	0.023	0.025	0.026	0.091	10
1.1.1.2	0.05	0.033	0	0	0	0.034	0.023	0.027	0.023	0.029	0.091	12
1.1.1.3	0	0	0	0	0	0.034	0.023	0.020	0.024	0.023	0.091	9
1.1.2.1	0	0.033	0	0	0	0.034	0.023	0.017	0.016	0.023	0.045	9
1.1.2.2	0	0.033	0	0	0	0.034	0.023	0.020	0.018	0.026	0.056	10
1.1.2.3	0.05	0.033	0	0.059	0	0.034	0.023	0.021	0.020	0.026	0.059	10
1.1.2.4	0	0.033	0.032	0	0.048	0.034	0.023	0.022	0.017	0.028	0.048	11
1.1.2.5	0	0	0	0	0	0.034	0.023	0.008	0.013	0.000	0.034	5
1.1.3.1	0.05	0.033	0.032	0.059	0	0.034	0.023	0.033	0.016	0.033	0.059	15
1.1.3.2	0.05	0.033	0.032	0.059	0	0.034	0.023	0.028	0.018	0.032	0.059	13
1.1.3.3	0	0.033	0.032	0.059	0	0.034	0.023	0.028	0.023	0.029	0.091	12
1.1.4.1	0	0	0.032	0	0	0	0.023	0.009	0.015	0	0.040	5
1.1.4.2	0	0.033	0.032	0.059	0.048	0.034	0.023	0.023	0.019	0.028	0.059	11
1.1.4.3	0	0	0	0	0	0	0.023	0	0	0	0	2
1.1.4.4	0	0.033	0.032	0.059	0	0	0.023	0.026	0.019	0.029	0.059	12
1.1.4.5	0	0	0	0	0	0	0.023	0.007	0.012	0.000	0.034	4
1.1.5.1	0	0	0.032	0	0	0.034	0.023	0.011	0.016	0.000	0.042	6
1.1.5.2	0	0.033	0.032	0	0.048	0.034	0.023	0.020	0.017	0.026	0.048	10
1.1.5.3	0	0.033	0.032	0	0	0.034	0.023	0.015	0.016	0.000	0.042	8
1.1.5.4	0	0	0.032	0	0	0.034	0.023	0.016	0.017	0.000	0.042	8
1.2.1	0.05	0.033	0.032	0	0.048	0.034	0.023	0.027	0.017	0.032	0.056	13
1.2.2	0.05	0.033	0.032	0.059	0.048	0.034	0.023	0.031	0.017	0.032	0.059	14
1.2.3	0.05	0.033	0.032	0	0.048	0	0.023	0.023	0.019	0.028	0.056	11
1.3.1	0	0.033	0.032	0	0.048	0	0.023	0.024	0.017	0.029	0.048	12
1.3.2	0	0.033	0.032	0.059	0.048	0.034	0.023	0.030	0.017	0.032	0.059	14
1.3.3	0	0.033	0.032	0.059	0.048	0.034	0.023	0.038	0.019	0.034	0.091	16
1.3.4	0.05	0.033	0.032	0.059	0.048	0.034	0.023	0.041	0.016	0.034	0.091	17
1.3.5	0	0	0.032	0	0.048	0	0.023	0.018	0.017	0	0.048	9
1.4.1	0.05	0.033	0.032	0	0.048	0.034	0.023	0.035	0.020	0.033	0.091	15
1.4.2	0.05	0.033	0.032	0.059	0.048	0.034	0.023	0.032	0.018	0.033	0.059	14
1.4.3	0.05	0.033	0.032	0	0.048	0.034	0.023	0.024	0.017	0.029	0.050	12
1.5.1	0.05	0.033	0.032	0	0.048	0	0.023	0.030	0.017	0.032	0.056	14
1.5.2	0.05	0.033	0.032	0	0.048	0.034	0.023	0.032	0.015	0.033	0.056	15
1.5.3	0.05	0	0.032	0.059	0.048	0	0.023	0.021	0.027	0.000	0.091	8
1.5.4	0.05	0	0	0	0.048	0	0.023	0.013	0.018	0	0.050	6
1.6.1	0.05	0.033	0.032	0.059	0	0.034	0.023	0.038	0.019	0.034	0.091	16
1.6.2	0.05	0.033	0.032	0.059	0.048	0.034	0.023	0.041	0.016	0.034	0.091	17
1.7.1	0.05	0.033	0.032	0	0.048	0.034	0.023	0.035	0.020	0.033	0.091	15
1.7.2	0.05	0.033	0.032	0	0	0.034	0.023	0.025	0.018	0.029	0.056	12
1.8.1	0	0	0	0	0	0	0.023	0	0	0	0	2
1.8.2	0	0	0	0	0	0	0.023	0	0	0	0	3
1.9.1	0	0	0	0.059	0	0	0.023	0.008	0.016	0	0.059	4
1.9.2	0	0	0.032	0.059	0	0	0.023	0.015	0.020	0	0.059	7
1.10.1	0.05	0.033	0.032	0.059	0.048	0.034	0.023	0.031	0.017	0.033	0.059	14
Average[a]	0.050	0.033	0.032	0.059	0.048	0.034	0.023	0.023	0.017	0.030	0.060	11
Standard Deviation	0.025	0.016	0.015	0.029	0.024	0.016	0.000	0.010	0.003	0.014	0.019	4
Count	20	30	31	17	21	29	44	44	44	31	44	44

[a]Average of non-zero numbers.

Table C3

Proportion of Blocks Devoted to Topics in Each Country's Textbook(s)

Topic Code	Country A	B	C	D	E	F	G	H	I	J
1.1.1.1	0.003	0	0.040	0.018	0	0	0.004	0	0.106	0
1.1.1.2	0.043	0.005	0.072	0.035	0	0.010	0.106	0.008	0.093	0
1.1.1.3	0.031	0.001	0.069	0.016	0.009	0.057	0.049	0.004	0.061	0
1.1.2.1	0.038	0.016	0.070	0.076	0.002	0.057	0.126	0.007	0.074	0
1.1.2.2	0.019	0.001	0.061	0.064	0	0.005	0.014	0.004	0.062	0
1.1.2.3	0.006	0.004	0.030	0.021	0	0.029	0.010	0.019	0.010	0
1.1.2.4	0.035	0.052	0.037	0.072	0	0.129	0.002	0	0.046	0
1.1.2.5	0.010	0.001	0.042	0.001	0.014	0.005	0.002	0	0.001	0
1.1.3.1	0.050	0.001	0.010	0.081	0.076	0.110	0.098	0.083	0.035	0
1.1.3.2	0.001	0	0.031	0.008	0.012	0.010	0.306	0.013	0.010	0
1.1.3.3	0.001	0.001	0.022	0.000	0.024	0	0.003	0.011	0.020	0
1.1.4.1	0	0	0.001	0	0.012	0	0	0.002	0	0
1.1.4.2	0.038	0.024	0.041	0.034	0	0	0.100	0.002	0.024	0
1.1.4.3	0	0	0.001	0	0	0	0	0.002	0	0
1.1.4.4	0.029	0	0.015	0.026	0	0	0.025	0.002	0	0.010
1.1.4.5	0.001	0	0.001	0.001	0	0.005	0.000	0.002	0	0
1.1.5.1	0	0.001	0	0.011	0	0	0.002	0	0.004	0.007
1.1.5.2	0	0	0.001	0.018	0	0	0	0.008	0.028	0
1.1.5.3	0	0.001	0.010	0.032	0	0	0.005	0.017	0.006	0
1.1.5.4	0.002	0	0	0.004	0	0	0.012	0.062	0.023	0
1.2.1	0.060	0.002	0.077	0.031	0	0.067	0.012	0.035	0.062	0
1.2.2	0.145	0	0.041	0.083	0	0.024	0	0.141	0.075	0
1.2.3	0.001	0	0.009	0	0	0	0	0	0	0
1.3.1	0.092	0.081	0.038	0.022	0	0	0.036	0.032	0.043	0
1.3.2	0.076	0.012	0.130	0.043	0.142	0.024	0.004	0.004	0.094	0.104
1.3.3	0.067	0.093	0.086	0.065	0.099	0.043	0	0.084	0.111	0.110
1.3.4	0.001	0.282	0.064	0.025	0	0	0	0.136	0.008	0
1.3.5	0	0	0.020	0.002	0	0	0	0	0.001	0
1.4.1	0	0.021	0.163	0.060	0.243	0.067	0.007	0.098	0.052	0.007
1.4.2	0	0.008	0.046	0.012	0.090	0	0	0	0.007	0.231
1.4.3	0.010	0	0.028	0.031	0	0	0.015	0.003	0.035	0
1.5.1	0.011	0.008	0.020	0.025	0	0	0	0.028	0.019	0
1.5.2	0.024	0.003	0.032	0.020	0	0	0	0.040	0	0
1.5.3	0	0	0.009	0	0	0	0	0.058	0.046	0
1.5.4	0	0	0.006	0	0	0.014	0	0	0	0
1.6.1	0.031	0.012	0.014	0.049	0.076	0.053	0.084	0.075	0.038	0.061
1.6.2	0.134	0.120	0.110	0.064	0.194	0.086	0.374	0.091	0.123	0.388
1.7.1	0.070	0.030	0.012	0.059	0.054	0.033	0	0.087	0.026	0.071
1.7.2	0	0	0.001	0.015	0	0.005	0	0	0	0
1.8.1	0	0	0	0.001	0	0	0	0	0.002	0
1.8.2	0	0	0	0	0	0	0	0	0	0
1.9.1	0	0.008	0.003	0.002	0.007	0	0.001	0	0	0.003
1.9.2	0.022	0	0.016	0	0.024	0	0.014	0	0.117	0
1.10.1	0.001	0.223	0.000	0.087	0	0.148	0	0	0	0.006
Average	0.024	0.023	0.033	0.028	0.025	0.022	0.032	0.026	0.033	0.023
Standard Deviation	0.035	0.057	0.037	0.027	0.053	0.037	0.075	0.039	0.037	0.070
Max	0.145	0.282	0.163	0.087	0.243	0.148	0.374	0.141	0.123	0.388
Count	30	26	39	36	16	21	25	30	33	11

Table C3 (*Cont'd.*)

Topic Code	Country K	L	M	N	O	P	Q	*X*	*SD*	Median	Max. Prop.	Count
1.1.1.1	0	0.014	0.004	0.009	0.003	0.042	0.011	0.015	0.026	0.004	0.106	11
1.1.1.2	0.010	0.073	0.002	0.009	0.005	0.184	0.022	0.040	0.049	0.010	0.184	15
1.1.1.3	0	0.002	0.004	0.013	0.031	0.001	0.007	0.021	0.023	0.009	0.069	15
1.1.2.1	0.007	0.067	0.030	0.009	0.013	0.065	0.036	0.041	0.034	0.036	0.126	16
1.1.2.2	0.007	0.065	0.021	0.003	0.013	0.040	0.022	0.024	0.024	0.014	0.065	15
1.1.2.3	0.005	0.031	0.019	0.006	0.002	0.024	0.008	0.013	0.010	0.010	0.031	15
1.1.2.4	0.013	0.047	0.015	0.003	0.017	0.073	0.059	0.035	0.034	0.035	0.129	14
1.1.2.5	0	0.002	0.015	0	0	0.003	0	0.006	0.010	0.001	0.042	11
1.1.3.1	0.040	0.047	0.002	0.014	0	0.013	0.043	0.041	0.036	0.040	0.110	15
1.1.3.2	0.011	0	0.002	0.046	0	0	0.025	0.028	0.071	0.010	0.306	12
1.1.3.3	0.040	0	0.002	0.278	0	0	0.032	0.026	0.064	0.002	0.278	11
1.1.4.1	0	0	0	0	0	0	0.001	0.001	0.003	0	0.012	4
1.1.4.2	0.062	0.018	0.117	0.101	0	0.001	0.055	0.041	0.038	0.034	0.117	14
1.1.4.3	0	0	0	0	0	0	0.001	0	0	0	0	3
1.1.4.4	0.003	0.007	0.064	0.072	0	0	0.026	0.016	0.022	0.007	0.072	11
1.1.4.5	0	0.006	0	0	0	0	0.025	0.002	0.006	0.000	0.025	7
1.1.5.1	0.003	0	0.008	0	0	0.004	0.001	0.002	0.003	0.001	0.011	9
1.1.5.2	0.011	0.014	0.008	0	0.010	0.018	0.007	0.007	0.008	0.007	0.028	10
1.1.5.3	0.001	0.023	0.008	0.006	0	0.007	0.017	0.008	0.009	0.006	0.032	12
1.1.5.4	0	0.007	0.008	0	0	0	0	0.007	0.015	0.000	0.062	7
1.2.1	0.010	0.083	0.013	0.014	0.005	0.167	0.038	0.040	0.042	0.031	0.167	15
1.2.2	0.148	0.023	0.127	0.092	0.164	0.081	0.059	0.071	0.057	0.075	0.164	13
1.2.3	0	0.001	0.006	0.006	0.001	0	0.008	0.002	0.003	0.000	0.009	7
1.3.1	0.010	0.010	0.015	0.039	0.112	0.002	0.038	0.034	0.032	0.032	0.112	14
1.3.2	0.057	0.063	0.025	0.073	0.022	0.033	0.023	0.055	0.042	0.043	0.142	17
1.3.3	0.199	0.118	0.202	0.174	0.125	0.037	0.049	0.098	0.054	0.093	0.202	16
1.3.4	0.037	0.051	0.047	0.469	0.014	0.004	0.019	0.068	0.121	0.019	0.469	13
1.3.5	0	0	0.053	0.010	0	0.001	0.001	0.005	0.013	0	0.053	7
1.4.1	0	0	0.062	0.069	0.079	0	0.021	0.056	0.064	0.052	0.243	13
1.4.2	0.025	0	0.034	0.101	0.118	0	0.015	0.040	0.060	0.012	0.231	11
1.4.3	0	0.009	0	0.002	0	0	0.008	0.008	0.012	0.002	0.035	9
1.5.1	0.011	0.001	0.004	0	0	0	0.016	0.008	0.010	0.004	0.028	10
1.5.2	0.040	0.017	0.011	0.006	0.095	0.024	0.030	0.020	0.023	0.017	0.095	12
1.5.3	0	0	0	0.032	0.083	0	0.016	0.014	0.025	0.000	0.083	6
1.5.4	0	0.001	0	0.006	0	0	0	0.002	0.004	0	0.014	4
1.6.1	0.055	0.037	0.174	0.208	0.001	0.003	0.046	0.060	0.054	0.049	0.208	17
1.6.2	0.356	0.174	0.323	0.371	0.296	0.043	0.236	0.205	0.118	0.174	0.388	17
1.7.1	0.058	0.057	0.066	0	0	0.099	0.094	0.048	0.032	0.057	0.099	14
1.7.2	0	0.001	0	0	0	0.001	0.034	0.003	0.008	0.000	0.034	6
1.8.1	0	0	0	0	0	0	0.003	0	0	0	0	4
1.8.2	0	0	0	0	0	0	0	0	0	0	0	0
1.9.1	0.011	0	0.013	0.309	0	0	0.016	0.022	0.072	0	0.309	10
1.9.2	0.015	0	0.051	0.098	0	0	0.007	0.021	0.034	0	0.117	9
1.10.1	0.001	0.020	0.006	0.029	0	0.006	0.085	0.036	0.062	0.006	0.223	11
Average	0.028	0.025	0.035	0.061	0.029	0.022	0.029	0.029	0.032	0.020	0.119	10.95
Standard Deviation	0.063	0.036	0.063	0.107	0.058	0.041	0.039	0.035	0.029	0.032	0.109	4
Max	0.356	0.174	0.323	0.469	0.296	0.184	0.236	0.205	0.121	0.174	0.469	17
Count	28	32	35	32	22	26	40	43	43	33	43	43

Table C4

Number of Data Sources in which Topics Appear within a Country

Topic Code	Country									
	A	B	C	D	E	F	G	H	I	J
1.1.1.1	2	1	3	3	1	2	2	1	3	0
1.1.1.2	3	3	3	3	1	3	1	2	3	0
1.1.1.3	3	1	3	3	2	3	1	2	3	0
1.1.2.1	3	2	3	3	3	3	2	2	3	0
1.1.2.2	3	2	3	3	2	3	3	2	3	0
1.1.2.3	2	1	3	3	1	3	2	2	3	0
1.1.2.4	2	3	3	3	2	3	1	0	3	0
1.1.2.5	2	2	3	3	2	3	2	1	1	0
1.1.3.1	3	3	3	3	2	3	2	2	3	1
1.1.3.2	2	2	3	3	2	2	3	2	3	1
1.1.3.3	1	3	3	2	2	1	2	3	3	1
1.1.4.1	1	0	2	0	1	0	0	1	0	2
1.1.4.2	2	3	3	3	0	2	2	2	3	0
1.1.4.3	0	0	2	0	0	0	0	1	0	0
1.1.4.4	3	0	3	3	2	2	2	2	2	3
1.1.4.5	2	0	2	1	0	2	0	1	0	0
1.1.5.1	2	3	1	3	0	0	1	0	1	1
1.1.5.2	2	2	3	3	0	1	0	2	3	1
1.1.5.3	2	3	3	3	0	0	1	2	2	0
1.1.5.4	2	2	1	3	0	1	1	1	3	1
1.2.1	3	2	3	3	1	3	2	2	3	1
1.2.2	3	1	3	3	1	3	1	2	3	1
1.2.3	3	1	3	2	0	2	1	1	1	2
1.3.1	3	3	2	3	1	1	2	2	3	2
1.3.2	3	3	3	3	2	3	3	2	3	3
1.3.3	3	3	3	3	3	3	1	3	3	3
1.3.4	3	3	3	3	2	2	2	3	2	2
1.3.5	0	1	2	2	0	1	0	0	2	1
1.4.1	2	3	3	3	3	3	1	3	3	3
1.4.2	1	3	3	3	3	2	1	1	2	3
1.4.3	2	2	3	3	1	2	1	1	3	2
1.5.1	3	3	3	3	2	2	1	2	3	1
1.5.2	3	3	3	3	2	2	1	2	2	2
1.5.3	1	1	2	1	1	0	0	3	2	1
1.5.4	0	0	2	1	1	2	0	0	0	2
1.6.1	3	3	3	3	3	3	3	3	3	3
1.6.2	3	3	3	3	3	3	3	3	3	3
1.7.1	3	3	2	3	3	3	1	3	3	3
1.7.2	2	0	2	3	2	3	2	0	0	1
1.8.1	0	0	0	1	1	0	0	0	1	0
1.8.2	1	0	0	0	0	1	0	0	0	0
1.9.1	0	1	3	1	2	1	1	0	1	2
1.9.2	2	0	3	1	1	1	2	0	3	0
1.10.1	3	3	0	3	1	2	0	1	1	3
Ave	2.1	1.8	2.5	2.5	1.4	1.9	1.3	1.5	2.2	1.3
Med	2	2	3	3	1	2	1	2	3	1
# 3s	19	19	30	32	7	17	5	8	25	9
# 0s	5	9	3	3	10	6	10	9	6	15
agrmt.	0.55	0.64	0.75	0.80	0.39	0.52	0.34	0.39	0.70	0.55

Table C4 (*Cont'd.*)

Topic Code	Country K	L	M	N	O	P	Q	AVE	MED	# 3s	# 0s	Agrmt.
1.1.1.1	1	2	1	2	2	1	2	2	2	3	1	0.24
1.1.1.2	2	3	1	2	1	3	2	2	2	8	1	0.53
1.1.1.3	0	2	1	2	1	3	2	2	2	6	2	0.47
1.1.2.1	2	3	2	2	1	3	2	2	2	8	1	0.53
1.1.2.2	2	3	1	2	1	3	3	2	3	9	1	0.59
1.1.2.3	3	3	1	3	1	3	3	2	3	9	1	0.59
1.1.2.4	2	3	2	2	2	3	3	2	2	8	2	0.59
1.1.2.5	0	2	1	1	0	3	2	2	2	4	3	0.41
1.1.3.1	3	3	3	3	1	3	3	3	3	12	0	0.71
1.1.3.2	3	2	3	3	1	1	3	2	2	8	0	0.47
1.1.3.3	1	2	2	3	1	1	3	2	2	6	0	0.35
1.1.4.1	0	0	1	1	0	0	2	1	0	0	9	0.53
1.1.4.2	2	3	3	3	3	3	3	2	3	10	2	0.71
1.1.4.3	0	0	0	0	0	0	2	0	0	0	14	0.82
1.1.4.4	1	3	3	3	1	0	3	2	2	8	2	0.59
1.1.4.5	0	1	1	0	0	0	3	1	0	1	9	0.59
1.1.5.1	2	1	2	1	0	3	3	1	1	4	4	0.47
1.1.5.2	2	3	3	1	2	3	3	2	2	7	2	0.53
1.1.5.3	2	3	3	2	0	3	3	2	2	7	4	0.65
1.1.5.4	1	2	3	1	1	2	2	2	1	3	1	0.24
1.2.1	3	3	3	2	3	3	3	3	3	11	0	0.65
1.2.2	3	3	3	3	3	3	3	2	3	12	0	0.71
1.2.3	1	3	2	2	2	0	3	2	2	4	2	0.35
1.3.1	2	3	3	2	3	2	3	2	2	8	0	0.47
1.3.2	2	3	3	3	3	3	3	3	3	14	0	0.82
1.3.3	2	3	3	3	3	3	3	3	3	15	0	0.88
1.3.4	3	3	3	3	3	2	3	3	3	11	0	0.65
1.3.5	0	0	3	1	1	1	2	1	1	1	6	0.41
1.4.1	2	2	3	2	3	1	3	3	3	11	0	0.65
1.4.2	2	2	3	3	3	1	3	2	3	9	0	0.53
1.4.3	1	3	2	2	2	1	3	2	2	5	0	0.29
1.5.1	3	3	3	1	2	1	3	2	3	9	0	0.53
1.5.2	3	3	3	2	3	3	3	3	3	10	0	0.59
1.5.3	2	0	2	3	2	0	3	1	1	3	4	0.41
1.5.4	2	1	0	2	1	0	2	1	1	0	7	0.41
1.6.1	3	3	3	3	1	2	3	3	3	15	0	0.88
1.6.2	3	3	3	3	3	2	3	3	3	16	0	0.94
1.7.1	3	3	3	1	2	3	3	3	3	13	0	0.76
1.7.2	2	3	1	1	0	2	3	2	2	4	4	0.47
1.8.1	0	1	0	0	0	0	2	0	0	0	12	0.71
1.8.2	0	0	0	0	0	0	1	0	0	0	14	0.82
1.9.1	1	1	1	3	0	0	3	1	1	3	4	0.41
1.9.2	1	1	3	3	0	0	3	1	1	5	5	0.59
1.10.1	2	3	3	3	2	2	3	2	2	8	2	0.59
Ave	1.7	2.2	2.1	2	1.5	1.7	2.7	1.905	1.98	7	2.7	0.57
Med	2	3	3	2	1	2	3	2.088	2	8	1	0.59
# 3s	11	25	23	17	11	18	32	0	16	4	1	0
# 0s	8	5	4	4	11	11	0	0	5	5	17	0
agrmt.	0.43	0.68	0.61	0.48	0.50	0.66	0.73	0.00	0.48	0.20	0.41	0.00

Table C5

Average Emphasis Devoted to Topics across Expert Topic Mapping, Curriculum Guides, and Textbooks

Topic Code	Country A	B	C	D	E	F	G	H	I	J
1.1.1.1	0.000	0.000	0.030	0.022	0.000	0.000	0.000	0.000	0.053	0.000
1.1.1.2	0.047	0.029	0.040	0.028	0.000	0.046	0.000	0.000	0.048	0.000
1.1.1.3	0.053	0.000	0.040	0.021	0.000	0.070	0.000	0.000	0.038	0.000
1.1.2.1	0.045	0.000	0.040	0.048	0.054	0.070	0.000	0.000	0.049	0.000
1.1.2.2	0.036	0.000	0.037	0.044	0.000	0.030	0.122	0.000	0.045	0.000
1.1.2.3	0.000	0.000	0.027	0.029	0.000	0.042	0.000	0.000	0.028	0.000
1.1.2.4	0.000	0.059	0.029	0.040	0.000	0.107	0.000	0.000	0.040	0.000
1.1.2.5	0.000	0.000	0.031	0.016	0.000	0.030	0.000	0.000	0.000	0.000
1.1.3.1	0.050	0.028	0.020	0.049	0.000	0.083	0.000	0.000	0.036	0.000
1.1.3.2	0.000	0.000	0.036	0.019	0.000	0.000	0.327	0.000	0.021	0.000
1.1.3.3	0.000	0.037	0.033	0.000	0.000	0.000	0.000	0.074	0.024	0.000
1.1.4.1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1.1.4.2	0.000	0.047	0.030	0.034	0.000	0.000	0.000	0.000	0.032	0.000
1.1.4.3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1.1.4.4	0.041	0.000	0.022	0.025	0.000	0.000	0.000	0.000	0.000	0.045
1.1.4.5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1.1.5.1	0.000	0.028	0.000	0.020	0.000	0.000	0.000	0.000	0.000	0.000
1.1.5.2	0.000	0.000	0.017	0.022	0.000	0.000	0.000	0.000	0.034	0.000
1.1.5.3	0.000	0.028	0.020	0.027	0.000	0.000	0.000	0.000	0.000	0.000
1.1.5.4	0.000	0.000	0.000	0.024	0.000	0.000	0.000	0.000	0.032	0.000
1.2.1	0.055	0.000	0.051	0.026	0.000	0.075	0.000	0.000	0.045	0.000
1.2.2	0.094	0.000	0.030	0.043	0.000	0.040	0.000	0.000	0.049	0.000
1.2.3	0.028	0.000	0.020	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1.3.1	0.070	0.061	0.000	0.030	0.000	0.000	0.000	0.000	0.032	0.000
1.3.2	0.062	0.032	0.059	0.037	0.000	0.040	0.079	0.000	0.056	0.114
1.3.3	0.058	0.076	0.045	0.044	0.121	0.063	0.000	0.131	0.054	0.117
1.3.4	0.039	0.154	0.038	0.031	0.000	0.000	0.000	0.161	0.000	0.000
1.3.5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1.4.1	0.000	0.036	0.070	0.036	0.242	0.075	0.000	0.139	0.035	0.043
1.4.2	0.000	0.030	0.032	0.020	0.136	0.000	0.000	0.000	0.000	0.183
1.4.3	0.000	0.000	0.026	0.026	0.000	0.000	0.000	0.000	0.029	0.000
1.5.1	0.044	0.030	0.032	0.031	0.000	0.000	0.000	0.000	0.031	0.000
1.5.2	0.050	0.028	0.028	0.029	0.000	0.000	0.000	0.000	0.000	0.000
1.5.3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.116	0.000	0.000
1.5.4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1.6.1	0.053	0.042	0.022	0.032	0.126	0.054	0.135	0.111	0.037	0.090
1.6.2	0.089	0.087	0.053	0.044	0.208	0.085	0.338	0.135	0.065	0.269
1.7.1	0.059	0.040	0.000	0.042	0.111	0.058	0.000	0.133	0.026	0.096
1.7.2	0.000	0.000	0.000	0.021	0.000	0.030	0.000	0.000	0.000	0.000
1.8.1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1.8.2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1.9.1	0.000	0.000	0.018	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1.9.2	0.000	0.000	0.022	0.000	0.000	0.000	0.000	0.000	0.063	0.000
1.10.1	0.028	0.129	0.000	0.045	0.000	0.000	0.000	0.000	0.000	0.043
Standard Deviation	0.028	0.035	0.019	0.016	0.057	0.032	0.073	0.049	0.022	0.055
Count	19	19	30	32	7	17	5	8	25	9

Table C5 (*Cont'd.*)

Topic Code	Country K	L	M	N	O	P	Q	*X*	*SD*	Median	Max. Prop.	Count
1.1.1.1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.006	0.014	0.000	0.053	3
1.1.1.2	0.000	0.049	0.000	0.000	0.000	0.121	0.000	0.024	0.032	0.000	0.121	8
1.1.1.3	0.000	0.000	0.000	0.000	0.000	0.045	0.000	0.016	0.023	0.000	0.070	6
1.1.2.1	0.000	0.047	0.000	0.000	0.000	0.072	0.000	0.025	0.027	0.000	0.072	8
1.1.2.2	0.000	0.046	0.000	0.000	0.000	0.046	0.023	0.025	0.031	0.023	0.122	9
1.1.2.3	0.048	0.034	0.000	0.023	0.000	0.039	0.018	0.017	0.017	0.018	0.048	9
1.1.2.4	0.000	0.040	0.000	0.000	0.000	0.060	0.036	0.024	0.030	0.000	0.107	8
1.1.2.5	0.000	0.000	0.000	0.000	0.000	0.046	0.000	0.007	0.014	0.000	0.046	4
1.1.3.1	0.083	0.049	0.021	0.025	0.000	0.050	0.038	0.031	0.027	0.028	0.083	12
1.1.3.2	0.067	0.000	0.021	0.034	0.000	0.000	0.024	0.032	0.076	0.000	0.327	8
1.1.3.3	0.000	0.000	0.000	0.093	0.000	0.000	0.027	0.017	0.028	0.000	0.093	6
1.1.4.1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0	0.000	0
1.1.4.2	0.000	0.029	0.068	0.048	0.089	0.030	0.042	0.026	0.026	0.030	0.089	10
1.1.4.3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0	0	0	0	0
1.1.4.4	0.000	0.034	0.050	0.040	0.000	0.000	0.024	0.017	0.019	0.000	0.050	8
1.1.4.5	0.000	0.000	0.000	0.000	0.000	0.000	0.031	0.002	0.007	0.000	0.031	1
1.1.5.1	0.000	0.000	0.000	0.000	0.000	0.031	0.016	0.006	0.010	0.000	0.031	4
1.1.5.2	0.000	0.036	0.022	0.000	0.000	0.037	0.018	0.011	0.014	0.000	0.037	7
1.1.5.3	0.000	0.040	0.022	0.000	0.000	0.032	0.021	0.011	0.014	0.000	0.040	7
1.1.5.4	0.000	0.000	0.032	0.000	0.000	0.000	0.000	0.005	0.011	0.000	0.032	3
1.2.1	0.050	0.053	0.024	0.000	0.039	0.099	0.029	0.032	0.029	0.029	0.099	11
1.2.2	0.146	0.031	0.062	0.045	0.111	0.063	0.043	0.045	0.041	0.043	0.146	12
1.2.3	0.000	0.022	0.000	0.000	0.000	0.000	0.018	0.005	0.010	0.000	0.028	4
1.3.1	0.000	0.026	0.025	0.000	0.087	0.000	0.036	0.022	0.027	0.000	0.087	8
1.3.2	0.000	0.055	0.028	0.041	0.047	0.058	0.031	0.043	0.028	0.041	0.114	14
1.3.3	0.000	0.066	0.096	0.066	0.093	0.045	0.033	0.065	0.037	0.063	0.131	15
1.3.4	0.066	0.041	0.035	0.142	0.059	0.000	0.029	0.047	0.053	0.035	0.161	11
1.3.5	0.000	0.000	0.037	0.000	0.000	0.000	0.000	0.002	0.009	0	0.037	1
1.4.1	0.000	0.000	0.040	0.000	0.088	0.000	0.030	0.049	0.061	0.036	0.242	11
1.4.2	0.000	0.000	0.040	0.048	0.106	0.000	0.028	0.037	0.053	0.020	0.183	9
1.4.3	0.000	0.034	0.000	0.000	0.000	0.000	0.018	0.008	0.013	0.000	0.034	5
1.5.1	0.051	0.022	0.021	0.000	0.000	0.000	0.021	0.017	0.017	0.021	0.051	9
1.5.2	0.083	0.028	0.033	0.000	0.095	0.054	0.033	0.027	0.029	0.028	0.095	10
1.5.3	0.000	0.000	0.000	0.027	0.000	0.000	0.021	0.010	0.028	0.000	0.116	3
1.5.4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0	0.000	0
1.6.1	0.076	0.036	0.077	0.075	0.000	0.000	0.032	0.059	0.039	0.053	0.135	15
1.6.2	0.251	0.088	0.136	0.117	0.186	0.000	0.106	0.133	0.087	0.106	0.338	16
1.7.1	0.078	0.043	0.051	0.000	0.000	0.071	0.056	0.051	0.038	0.051	0.133	13
1.7.2	0.000	0.022	0.000	0.000	0.000	0.000	0.035	0.006	0.012	0.000	0.035	4
1.8.1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0	0	0	0	0
1.8.2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0	0	0	0	0
1.9.1	0.000	0.000	0.000	0.101	0.000	0.000	0.021	0.008	0.024	0	0.101	3
1.9.2	0.000	0.000	0.037	0.047	0.000	0.000	0.018	0.011	0.019	0	0.063	5
1.10.1	0.000	0.029	0.022	0.029	0.000	0.000	0.045	0.022	0.032	0.000	0.129	8
Standard Deviation	0.049	0.023	0.030	0.036	0.044	0.031	0.020	0.024	0.019	0.023	0.075	4
Count	11	25	23	17	11	18	32	39	39	16	39	39

APPENDIX D

Table D1

Unweighted Specially-Constructed Test Scores

Country	UNION	EX-7I	EX-SI	CG-71	CG-SI	TX-7I	TX-SI	AG-71	AVE	SD	MIN	MAX
A	47.69	45.11	45.59	46.02	49.55	48.37	46.01	46.49	46.85	1.43	45.11	49.55
B	55.49	53.60	55.91	52.26	62.17	56.61	51.40	53.78	55.15	3.14	51.40	62.17
C	51.00	49.57	50.80	47.84	52.62	53.09	50.08	49.40	50.55	1.62	47.84	53.09
D	42.65	40.63	40.59	39.98	42.90	42.41	39.80	40.20	41.15	1.20	39.80	42.90
E	43.66	40.64	47.29	42.25	48.84	45.10	41.80	42.24	43.98	2.69	40.64	48.84
F	49.75	47.99	48.06	47.36	49.88	49.84	46.47	48.97	48.54	1.19	46.47	49.88
G	45.13	42.25	46.19	41.83	51.82	46.30	46.46	41.85	45.23	3.14	41.83	51.82
H	42.23	41.53	29.39	38.94	43.31	43.16	35.94	38.96	39.18	4.39	29.39	43.31
I	51.71	50.07	46.57	49.11	49.13	51.30	47.92	51.36	49.65	1.70	46.57	51.71
J	57.36	55.72	71.86	61.83	67.68	61.60	66.30	65.18	63.44	5.03	55.72	71.86
K	54.58	53.20	52.17	52.81	57.92	55.83	50.86	54.06	53.93	2.07	50.86	57.92
L	50.29	47.93	53.30	47.03	53.23	49.61	47.51	48.11	49.63	2.33	47.03	53.30
M	33.83	31.67	31.94	32.08	38.31	33.89	32.28	32.82	33.35	2.03	31.67	38.31
N	44.13	42.81	44.64	41.58	53.46	46.11	46.93	43.84	45.44	3.42	41.58	53.46
O	59.62	59.17	56.79	55.41	60.01	61.98	55.99	57.17	58.27	2.13	55.41	61.98
P	41.78	38.03	37.84	39.73	40.66	41.37	36.46	40.87	39.59	1.81	36.46	41.78
Q	54.19	52.12	45.53	51.02	55.95	55.23	52.49	52.32	52.36	3.02	45.53	55.95
AVE	48.53	46.59	47.32	46.30	51.61	49.52	46.75	47.51	48.02	1.69	46.30	51.61
SD	6.55	6.93	9.54	7.01	7.55	7.21	7.89	7.64	7.54	0.86	6.55	9.54

Table D2

Weighted Specially-Constructed Test Scores

Country	WEX1-UN	WEX1-7I	WCG-UN	WCG-7I	WTX-UN	WTX-7I	WTX-SI	WAG2-UN	WAG2-7I	AVE	SD	MIN	MAX
A	46.30	45.19	46.90	46.18	45.12	45.75	42.16	45.76	44.43	45.31	1.31	42.16	46.90
B	54.55	54.01	54.56	52.71	54.54	55.23	50.92	54.22	52.66	53.71	1.28	50.92	55.23
C	50.42	50.03	50.19	48.43	49.89	50.89	47.54	50.12	48.23	49.53	1.09	47.54	50.89
D	41.56	40.70	41.62	40.21	39.36	39.74	36.11	40.19	38.33	39.76	1.62	36.11	41.62
E	42.51	41.00	43.22	42.56	42.34	43.21	38.92	42.77	41.15	41.97	1.31	38.92	43.22
F	48.70	48.12	48.88	47.51	47.04	47.66	42.45	47.46	46.53	47.15	1.81	42.45	48.88
G	44.13	42.90	44.29	42.29	44.01	44.72	45.75	44.11	41.71	43.77	1.18	41.71	45.75
H	41.66	41.68	41.06	39.24	41.10	41.52	37.56	41.04	39.06	40.44	1.38	37.56	41.68
I	50.39	49.94	50.57	49.25	48.96	49.18	44.86	49.32	49.29	49.09	1.59	44.86	50.57
J	57.32	56.58	58.92	62.65	61.60	62.63	64.03	61.60	64.82	61.13	2.73	56.58	64.82
K	53.81	53.49	54.05	53.08	53.11	53.82	48.55	53.26	51.91	52.79	1.61	48.55	54.05
L	48.96	47.98	49.21	47.45	47.58	47.94	44.16	47.95	46.72	47.55	1.39	44.16	49.21
M	33.09	32.09	33.44	32.65	32.35	32.76	29.82	32.93	31.72	32.32	1.01	29.82	33.44
N	43.69	43.37	43.70	42.42	45.78	46.72	48.49	45.03	44.97	44.91	1.78	42.42	48.49
O	59.06	59.23	58.38	55.87	58.72	59.70	55.29	58.54	56.69	57.94	1.49	55.29	59.70
P	40.03	38.13	41.05	40.02	38.39	38.58	32.39	39.43	38.45	38.50	2.34	32.39	41.05
Q	53.01	52.31	52.98	51.27	52.12	52.83	52.28	52.53	51.35	52.30	0.60	51.27	53.01
AVE	47.60	46.87	47.82	46.70	47.18	47.82	44.78	47.43	46.35	46.95	0.90	44.78	47.82
SD	6.67	6.94	6.64	7.04	7.28	7.44	8.27	7.07	7.69	7.23	0.49	6.64	8.27

Table D3

Unique Specially-Constructed Test Scores

Country	EX1-UQ	WEX1-UQ	CG-UQ	TX-UQ	WTX-UQ	AG2-UQ	WAG2-UQ	AVE	SD	MIN	MAX
A	47.47	46.76	49.18	47.07	44.77	47.76	48.42	47.35	1.30	44.77	49.18
B	56.74	56.16	54.77	55.87	58.41	54.55	57.35	56.27	1.27	54.55	58.41
C	51.51	52.26	50.41	50.41	52.11	50.88	52.78	51.48	0.87	50.41	52.78
D	42.65	42.73	42.65	42.08	40.68	42.08	40.86	41.96	0.80	40.68	42.73
E	41.76	41.80	41.86	42.94	42.20	42.04	43.44	42.29	0.60	41.76	43.44
F	48.92	48.27	48.80	48.51	47.05	48.51	46.65	48.10	0.82	46.65	48.92
G	50.01	50.12	42.96	46.42	47.34	48.06	53.16	48.29	3.00	42.96	53.16
H	41.55	42.83	43.70	42.99	41.87	42.50	42.96	42.63	0.68	41.55	43.70
I	50.96	50.85	50.49	51.98	50.49	50.86	50.94	50.94	0.46	50.49	51.98
J	60.20	60.95	58.96	62.41	65.79	62.16	65.34	62.26	2.35	58.96	65.79
K	55.53	55.44	56.49	53.89	50.23	56.31	58.32	55.17	2.37	50.23	58.32
L	49.88	49.75	49.84	50.09	50.13	49.74	49.54	49.85	0.19	49.54	50.13
M	32.30	31.77	33.75	33.90	31.34	32.63	34.03	32.82	1.01	31.34	34.03
N	44.13	44.28	43.57	44.51	48.04	43.57	45.77	44.84	1.48	43.57	48.04
O	57.37	58.46	59.00	61.12	56.87	58.94	59.05	58.69	1.27	56.87	61.12
P	40.88	40.01	41.99	43.91	47.40	43.76	43.98	43.13	2.27	40.01	47.40
Q	53.23	53.19	54.19	54.27	53.10	53.28	52.10	53.34	0.68	52.10	54.27
AVE	48.53	48.57	48.39	48.96	48.70	48.69	49.69	48.79	0.40	48.39	49.69
SD	7.04	7.24	6.78	7.04	7.55	7.06	7.52	7.17	0.26	6.78	7.55

Table D4

Ranks on Unweighted Specially-Constructed Tests

Country	UNION	EX1-7I	EX1-SI	CG-71	CG-SI	TX-7I	TX-SI	AG2-71	AVE	SD	MIN	MAX
A	10	10	11	10	11	10	12	10	11	1	10	12
B	3	3	3	4	2	3	4	4	3	1	2	4
C	7	7	6	7	8	6	6	7	7	1	6	8
D	14	15	14	14	15	15	14	15	15	1	14	15
E	13	14	8	11	13	13	13	12	12	2	8	14
F	9	8	7	8	10	8	10	8	9	1	7	10
G	11	12	10	12	9	11	11	13	11	1	9	13
H	15	13	17	16	14	14	16	16	15	1	13	17
I	6	6	9	6	12	7	7	6	7	2	6	12
J	2	2	1	1	1	2	1	1	1	0	1	2
K	4	4	5	3	4	4	5	3	4	1	3	5
L	8	9	4	9	7	9	8	9	8	2	4	9
M	17	17	16	17	17	17	17	17	17	0	16	17
N	12	11	13	13	6	12	9	11	11	2	6	13
O	1	1	2	2	3	1	2	2	2	1	1	3
P	16	16	15	15	16	16	15	14	15	1	14	16
Q	5	5	12	5	5	5	3	5	6	2	3	12

Table D5

Ranks on Weighted Specially-Constructed Tests

Country	WEX1-UN	WEX1-7I	WCG-UN	WCG-7I	WTX-UN	WTX-7I	WTX-SI	WAG2-UN	WAG2-7I	AVE	SD	MIN	MAX
A	10	10	10	10	11	11	12	10	11	11	1	10	12
B	3	3	3	4	3	3	4	3	3	3	0	3	4
C	6	6	7	7	6	6	7	6	7	6	0	6	7
D	15	15	14	14	15	15	15	15	16	15	1	14	16
E	13	14	13	11	13	13	13	13	13	13	1	11	14
F	9	8	9	8	9	9	11	9	9	9	1	8	11
G	11	12	11	13	12	12	8	12	12	11	1	8	13
H	14	13	15	16	14	14	14	14	14	14	1	13	16
I	7	7	6	6	7	7	9	7	6	7	1	6	9
J	2	2	1	1	1	1	1	1	1	1	0	1	2
K	4	4	4	3	4	4	5	4	4	4	0	3	5
L	8	9	8	9	8	8	10	8	8	8	1	8	10
M	17	17	17	17	17	17	17	17	17	17	0	17	17
N	12	11	12	12	10	10	6	11	10	10	2	6	12
O	1	1	2	2	2	2	2	2	2	2	0	1	2
P	16	16	16	15	16	16	16	16	15	16	0	15	16
Q	5	5	5	5	5	5	3	5	5	5	1	3	5

Table D6

Ranks on Unique Specially-Constructed Tests

Country	EX1-UQ	WEX1-UQ	CG-UQ	TX-UQ	WTX-UQ	AG2-UQ	WAG2-UQ	AVE	SD	MIN	MAX
A	11	11	9	10	13	11	10	11	1	9	13
B	3	3	4	3	2	4	4	3	1	2	4
C	6	6	7	7	5	6	6	6	1	5	7
D	13	14	14	16	16	15	16	15	1	13	16
E	14	15	16	15	14	16	14	15	1	14	16
F	10	10	10	9	12	9	11	10	1	9	12
G	8	8	13	11	11	10	5	9	2	5	13
H	15	13	11	14	15	14	15	14	1	11	15
I	7	7	6	6	6	7	8	7	1	6	8
J	1	1	2	1	1	1	1	1	0	1	2
K	4	4	3	5	7	3	3	4	1	3	7
L	9	9	8	8	8	8	9	8	0	8	9
M	17	17	17	17	17	17	17	17	0	17	17
N	12	12	12	12	9	13	12	12	1	9	13
O	2	2	1	2	3	2	2	2	1	1	3
P	16	16	15	13	10	12	13	14	2	10	16
Q	5	5	5	4	4	5	7	5	1	4	7

LIST OF REFERENCES

LIST OF REFERENCES

Anastasi, A. (1982). *Psychological testing.* New York: MacMillan Publishing Co., Inc.

Airasian, P.W., & Madaus, G.F. (1983). Linking testing and instruction: Policy issues. *Journal of Educational Measurement, 20*(2), 103-118.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: Author.

Baker, D.P. (1993, April). Compared to Japan the U.S. is a low achiever...really: New evidence and comments on Westbury. *Educational Researcher, 22*(3), 18-20.

Berliner, D.C. (1993, Fall). International comparisons of student achievement: A false guide for reform. *National Forum*, 25-29.

Bracey, G.W. (1995, Oct.). The fifth Bracey report on the condition of public education. *Phi Delta Kappan*, 149-160.

Burstein, L. (1991). Conceptual considerations in instructionally sensitive assessment. Los Angeles: Center for Research in Evaluation, Standards, and Student Testing. (ERIC Document Reproduction Service No. ED 335367)

Burstein, L. (1993). Studying learning, growth, and instruction cross-nationally: Lessons learned about why and why not engage in cross-national studies. In Burstein, L. (Ed.) The IEA Study of Mathematics III: Student Growth and Classroom Processes. New York: Pergamon Press. (p.xxvi-lii).

Burstein, L., Aschbacher, P., Chen, Z., Lin, L., & Sen, Q. (1990). *Establishing the content validity of tests designed to serve multiple purposes: Bridging secondary-postsecondary mathematics.* Los Angeles, CA: UCLA Graduate School of Education. CSE Dissemination Office.

Cattell, R.B. (1949). r and other coefficients of pattern similarity. *Psychometrica, 14*(4), 279-298.

Cronbach, L.J. & Gleser, G.C. (1953). Assessing similarity between profiles. *The Psychological Bulletin, 50*(6), 456-473.

Cronbach, L.J. (1971). Test validation. In Thorndike, R.L. (Ed.), *Educational Measurement*. Washington, DC: American Council on Education.

Cohen, M. (1988, April). Designing state assessment systems. *Phi Delta Kappan*, 583-588.

Crocker, L.M., Miller, M.D., & Franks, E.A. (1989). Quantitative methods for assessing the fit between test and curriculum. *Applied Measurement in Education, 2(*2), 179-194.

Crocker, L., Llabre, M., & Miller, M.D. (1988). The generalizability of content validity ratings. *Journal of Educational Measurement, 25*, 287-299.

Fitzpatrick, A.R. (1983). The meaning of content validity. *Applied Psychological Measurement, 7*(1), 3-13.

Freeman, D.J., Belli, G.M., Porter, A.C., Floden, R.E., Schmidt, W.H., Schwille, J.R. (1983, Fall). The influence of different styles of textbook use on the instructional validity of standardized tests. *Journal of Educational Measurement, 20*(3), 259-270.

Frederikson, J.R. & Collins, A. (1989). A systems approach to.educational testing. *Educational Researcher, 18*(9), 27-32.

Gamoran, A., Porter, A.C., Smithson, J., & White, P.A. (1996, March). *Upgrading high school math instruction: Improving opportunities for low-achieving, low income youth.* A paper presented at the annual meeting of the American Education Research Association, New York, NY.

Garden, R.A., & Orpwood, G. (1996). Development of the TIMSS achievement tests. In IEA (Eds.), *Third International Mathematics and Science Study Technical Report Volume I: Design and Development*. Boston: Bocton College.

Guion, R.M. (1978). Scoring of content domain samples: The problem of fairness. *Journal of Applied Psychology, 63*(4), 499-506.

Guiton, G., & Oakes, J. (1995). Opportunity to learn and conceptions of educational equality. *Educational Evaluation and Policy Analysis, 17*(3), 323-336.

Guskey, T.R., & Kifer, E.W. (1990). Ranking school districts on the basis of statewide test results: Is it meaningful or misleading? *Educational Measurement: Issues and Practice, 9*(1), 11-16.

Guthrie, J.T. (1986). Roles of the National Assessment of International Progress in international studies (Publication NO. TM 870 049). *The Nation's Report Card.* (ERIC Document Report Service NO. ED279678).

Haertel, E. & Calfee, R. (1983, Sum.). School achievement: Thinking about what to test. *Journal of Educational Measurement, 20*(2), 119-132.

Husen, T. (1982). *A cross-national perspective on assessing the quality of learning.* Washington, DC: National Commission on Excellence in Education. (ERIC Document Reproduction Service NO. ED225992).

Husen, T. (1983). Are standards in US schools really lagging behind those in other countries? *Phi Delta Kappan, 64,* 455-461.

Husen, T. (1987). Policy impact of IEA research. *Comparative Education Review, 20,* 81-92.

International Association for the Evaluation of Educational Achievement. (1994a). *Teacher questionnaire population 2 math.* (Doc. Ref. ICC 880/NRC417). The Hague: Author.

International Association for the Evaluation of Educational Achievement. (1994b). *TIMSS field trial data analysis plan.* The Hague: Author.

International Association for the Evaluation of Educational Achievement. (1994c). *TIMSS field trial manual for national research coordinators.* (Doc. Ref. ICC 714/NRC277). The Hague: Author.

Kaestle, C. (1985, Feb.). Education reform and the swinging pendulum. *Phi Delta Kappan,* 422-423.

Kupermintz, H., Ennis, M.M., Hamilton, L.S., Talbert, J.E., Snow, R.E. (1995, Fall). Enhancing the validity and usefulness of large-scale educational assessments: I. NELS: 88 mathematics achievement. *American Educational Research Journal, 32*(3), 523-554.

LaPointe, A.E. (1991). NAEP: A national report card for education and the public. *The Assessment of National Goals: Proceedings of the 1990 ETS Invitational Conference,* 47-62.

Leinhardt, G., & Seewald, A.M. (1981). Overlap: What's tested, what's taught? *Journal of Educational Measurement, 18*(2), 85-96.

Leinhardt, G. (1983). Overlap: Testing whether it is taught. In Madaus, G.F. (Ed.), *The Courts, Validity, and Minimum Competency Testing.* Boston: Kluwer-Nijhoff Publishing.

Linn, R.L. (1987). *State-by-state comparisons of student achievement: The definition of the content domain for assessment.* (Technical report #275). Los Angeles, CA: University of California Los Angeles, Center for Research on Evaluation, Standards, and Student Testing.

Linn, R.L. (1988). Accountability: The comparison of educational systems and the quality of test results. *Educational Policy, 1,* 181-198.

Linn, R.L., & Baker, E.L. (1995). What do international assessments imply for world-class standards? *Educational Evaluation and Policy Analysis, 17*(4), 405-418.

Linn, R.L. Baker, E.L., & Dunbar, S.D (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 5-21.

Maeroff, G. (1991). The public's expectations for assessment of National Educational Goals. *The Assessment of National Goals: Proceedings of the 1990 ETS Invitational Conference,* 87-95.

McDonnell, L. M. (1995). Opportunity to learn as a research concept and a policy instrument. *Educational Evaluation and Policy Analysis,17*(3), 305-322.

McKnight, C.C., Crosswhite, F.J., Dossey, J.A., Kifer, E., Swafford, J.O., Travers, K.J., Cooney, T.J. (1987). *The underachieving curriculum: Assessing U.S. school mathematics from an international perspective.* Champaign, IL: Stipes Publishing Company.

Mehrens, W.A. (1984, Fall). National tests and local curriculum: Match or mismatch? *Educational Measurement: Issues and Practice.* 9-15.

Mehrens, W.A., & Lehmann, I.J. (1991). *Measurement and evaluation in education and psychology.* Fort Worth: Holt, Rinehart and Winston, Inc.

Mehrens, W.A., & Phillips, S.E. (1987). Sensitivity of item difficulties to curricular validity. *Journal of Educational Measurement, 24*(4), 357-370.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement* (pp. 13-103). New York, New York: American Council on Education, Macmillan Publishing Company.

Millman, J. & Greene, J. (1989). The specification and development of tests of achievement and ability. In R.L. Linn (Ed.), *Educational Measurement* (pp. 335-366). New York, New York: American Council on Education, Macmillan Publishing Company.

Mislevy, R.J. (1995). What can we learn from international assessments? *Educational Evaluation and Policy Analysis, 17*(4), 419-437.

Moss, P.A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62*(3), 229-258.

Muthen, B., Huang, L., Jo, B., Khoo, S., Goff, G., Novak, J., & Shih, J. (1995). Opportunity-to-learn effects on achievement: Analytical aspects. *Educational Evaluation and Policy Analysis,17*(3), 371-403.

National Commission on Excellence in Education. (1983). *A Nation at Risk.* Washington, DC: U.S. Dept. of Ed.

Nitko, A.J. (1989). Designing tests that are integrated with instruction. In R.L. Linn (Ed.), *Educational Measurement* (pp. 453-474). New York, New York: American Council on Education, Macmillan Publishing Company.

Passow, A.H. (1984). The IEA national case study. *Educational Forum, 48*, 469-487.

Pelgrum, W.J. (1989). *Educational assessment: Monitoring, evaluation and the curriculum.* Enschende, The Netherlands: University of Twente Department of Education.

Phillips, S.E., & Mehrens, W.A. (1988). Effects of curricular differences on achievement test data at item and objective levels. *Applied Measurement in Education, 1*(1), 33-51.

Porter, A.C. (1990). Assessing national goals: Some measurement dilemmas. *The Assessment of National Goals: Proceedings of the 1990 ETS Invitational Conference,* 21-42.

Postlethwaite, T.N. (1987). Comparative educational achievement research: Can it be improved? *Comparative Education Review, 31,* 150-163.

Purves, A.C. (1987). The evolution of the IEA: A memoire. *Comparative Education Review, 31*, 10-28.

Raizen, S.A. & Jones, L.V. (1985). *Indicators of precollege education in science and mathematics: A preliminary review.* Washington, D.C.: National Academy Press.

Resnick, L.B., Nolan, K.J., & Resnick, D,P. (1995). Benchmarking education standards. *Educational Evaluation and Policy Analysis, 17,* 4, 438-461.

Robitaille. D.F. & Gardden. R.A. (1996). *Research questions & study design.* The Third International Mathematics and Science Study Monograph No. 2. Vancouver, Canada: Pacific Educational Press.

Robitaille, D.F., McKnight, C., Schmidt, W.H., Britton, E., Raizen, S., & Nicol, C. (1993). *Curriculum frameworks for mathematics and science.* Vancouver, Canada: Pacific Educational Press.

Romberg, T.A., & Wilson, L.D. (1992). Alignment of tests with the Standards. *The Arithmetic Teacher, 40*(1), 18-22.

Schmidt, W.H. (1983). Content biases in achievement tests. *Journal of Educational Measurement,20*(2), 165-178.

Schmidt, W.H., & McKnight, C.C. (1995). Surveying educational opportunity in mathematics and science: An international perspective. *Educational Evaluation and Policy Analysis,17*(3), 337-353.

Schmidt, W.H., McKnight, C.C., Valverde, G.A., Houang, R.T., & Wiley, D.E. (in press). *Many visions, many aims: A cross-national investigation of curricular intentions.* Third International Mathematics and Science Study, Michigan State University, MI.

Schmidt, W.H., Porter, W.H., Schwille, J.R., Floden, R.E., & Freeman, D.J. (1983). Overlap: Testing whether it is taught. In Madaus, G.F. (Ed.), *The Courts, Validity, and Minimum Competency Testing.* Boston: Kluwer-Nijhoff Publishing.

Schmidt, W.H., & Valverde, G.A. (1995). *National policy and cross-national research: United States participation in the Third International Mathematics and Science* Study. Manuscript in preparation, East Lansing, MI: Michigan state University, Third International Mathematics and Science Study.

Sireci, S.G. (1990). *Applying empirical analyses to the evaluation of test content.* Paper presented at the annual meeting of the Northeastern Educational Research Association, Ellenville, NY, November, 1990.

Skinner, H.A. (1978). Differentiating the contribution of elevation, scatter, and shape in profile similarity. *Educational and Psychological Measurement, 38,* 297-308.

Stedman, L.C. (1994, Oct.). Incomplete explanations: The case of U.S. performance in the international assessments of education. *Educational Researcher, 23*(7), 24-32.

Survey of Mathematics and Science Opportunities (SMSO). (1993, Nov.). *A description of the TIMSS' achievement test content design test blueprints.* East Lansing, MI: Michigan State University.

Walker, D. & Schaffarzick, T. (1974). Comparing curricula. *Review of Educational Research, 44*(1), 83-11.

Westbury, I. (1992, June-July). Comparing American and Japanese achievement: Is the United States really a low achiever? *Educational Researcher, 21*(5), 18-24.

Westbury, I. (1993, April). American and Japanese achievement...again. *Educational Researcher, 22*(3), 21-25.

Wolf, R.M. (1988, April). The NAEP and international comparisons. *Phi Delta Kappan, 69,* 580-582.