Patterns of Diagnosed Mathematical Content and Process Skills in TIMSS-R Across a Sample of 20 Countries

Kikumi K. Tatsuoka, James E. Corter and Curtis Tatsuoka

Am Educ Res J 2004 41: 901
DOI: 10.3102/00028312041004901

American Educational Research Journal
Winter 2004, Vol. 41, No. 4, pp. 901–926

Patterns of Diagnosed Mathematical Content and Process Skills in TIMSS-R Across a Sample of 20 Countries

Kikumi K. Tatsuoka and James E. Corter
Teachers College, Columbia University
Curtis Tatsuoka
George Washington University

This study used a diagnostic testing approach to compare the mathematics achievement of eighth-grade students across a sample of 20 countries, analyzing data from the Third International Math and Science Study–Revised (TIMSS-R, 1999). Using the rule-space method, student mastery was measured on 23 specific content knowledge and processing subskills ("attributes") underlying students' item scores, using 23 attributes previously defined and validated. Mean mastery levels for each attribute were compared for the 20 selected countries. Clear differences among the countries were found in patterns of subskill achievement. U.S. students were strong in some content and quantitative reading skills, but weak in others, notably geometry. Interestingly, success in geometry was found to be highly associated with logical reasoning and other important mathematical thinking skills across the sampled countries.

KEYWORDS: international comparisons, mathematics achievement, mathematics problem-solving, TIMSS.

The Third International Math and Science Study–Revised (TIMSS-R), conducted in 1999, is the successor to the 1995 Third International Mathematics and Science Study. The TIMSS-R (1999) contained revised versions of the 1995 TIMSS items, using the 1999 TIMSS assessment frameworks and specifications, and collected data solely from eighth-grade students. Thirty-eight countries participated (Mullis et al., 2001). The data collected are multifaceted, including not only achievement data from the actual math and science test, but also background questionnaires aimed at measuring various aspects of students, teachers, and schools, and even including videotaped observations of actual math and science lessons in the various countries. The TIMSS-R (1999) study also included benchmarking data, providing to states, school districts, and educational consortia an unprecedented opportunity to

evaluate the comparative international standing of their students' achievement (Mullis et al., 2001).

The TIMSS studies have had political and educational impact. According to Macnab (2000), participating countries have reacted in a variety of ways to the comparative performance of their students. Other studies have examined how these reactions have played out in terms of changes to mathematics curricula and teaching methods (Macnab, 1999; Robitaille, 1997). These investigations document that some countries, including Canada, England, Germany, Japan, Sweden, and the United States, have used TIMSS results to help plan or implement changes designed to improve their educational systems.

The first TIMSS study and the TIMSS-R also have had considerable scientific impact, generating much research aimed at understanding differences between countries in math and science teaching practices and in student achievement. But because of the huge volume of data collected by the studies, there is much work still to be done in understanding student performance in math and science in these countries. In particular, few studies have attempted analyses of students' underlying performance on the mathematics portion of the TIMSS studies by using cognitive "diagnostic" approaches, nor have any studies tried to compare differences among countries at this microlevel. The present article aims to accomplish these two goals, using a statistical approach called the rule-space method (RSM) developed by K. Tatsuoka and her associates (e.g., K. Tatsuoka, 1985, 1995; K. Tatsuoka & M. Tatsuoka, 1987). The RSM is an example of a diagnostic testing approach to analyzing large-scale tests, one aimed at discovering and measuring important subskills involved in domain competence.

Recently there has been increased interest in such diagnostic testing models (e.g., Haertel & Wiley, 1993; Stout, 2002; C. Tatsuoka & Ferguson, 1999; C. Tatsuoka, 2002; Yan, Mislevy, & Almond, 2003), but in mathemat-

---

KIKUMI K. TATSUOKA is a Senior Research Professor in the Department of Human Development at Teachers College, Columbia University, Box 118, 525 W. 120th Street, New York, NY 10027; e-mail *kt2005@columbia.edu*. Her area of specialization is rule-space methodology.

JAMES E. CORTER is an Associate Professor of Statistics and Education and Chair of the Department of Human Development at Teachers College, Columbia University, Box 41, 525 W. 120th Street, New York, NY 10027; e-mail *jec34@columbia.edu*. His research specializations include the learning and teaching of problem solving (especially in mathematics), human categorization, judgment, and decision making, as well as psychometric methods, in particular, multidimensional scaling and clustering methods.

CURTIS TATSUOKA is an Assistant Professor in the Department of Statistics at George Washington University, 2140 Pennsylvania Avenue, NW, Washington, DC 20052; e-mail *tatsuoka@gwu.edu*. His research has focused on theoretical and methodological aspects of cognitively diagnostic adaptive testing.

---

ics education there is a long tradition of attempting to analyze mathematics ability and achievement into component skills. Fifty years ago, Polya (1954) described how highly effective mathematicians think. His description of creative and effective patterns of mathematical thinking is wholly consistent with recent descriptions of effective information processing in mathematics problem solving (Pressley, 1995). Polya emphasized that students should always be attempting to understand when and how particular problem solutions can be applied, and to determine if they might already know how to solve the problem at hand. Polya's perspectives are still relevant today, and substantial data have accrued to support his theories about instruction and mathematical thinking (Pressley, 1995). Other studies have described a variety of thinking skills required in mathematics (Jones & Idol, 1990; Marzano et al., 1988) and the effects of education on cognitive competencies (Pascarella & Terenzini, 1991). Different cognitive models of learning and teaching mathematics have been developed (Carpenter & Moser, 1982; Davis, 1992; Greeno, 1991; Schneider & Graham, 1992), and validated by empirical evidence.

The RSM has been developed for analyzing latent variables, such as whether or not a student possesses particular pieces of knowledge or cognitive processing and thinking skills required in solving a particular problem. The RSM has been applied successfully to generate scoring reports prescribing individual weaknesses and strengths for large-scale assessments. It has also been applied to reading comprehension tests (Buck, Tatsuoka, & Kostin, 1997) listening (Buck & Tatsuoka, 1998), hands-on tasks in science (Yepes-Baraya & Allen, 2002), and other mathematics tests (K. Tatsuoka, Linn, M. Tatsuoka, & Yamamoto, 1988; K. Tatsuoka, 1990, 1995; K. Tatsuoka & Boodoo, 2000). Because the theoretical foundation of RSM is still relatively new to the field of educational and psychological measurement, some background and a brief introduction to basic concepts of the RSM are presented.

Introduction to the Rule-Space Method

In psychometrics and educational measurement, one of the best known examples of a statistical modeling approach is item response theory (IRT). Many models of test performance assume some kind of algebraic relationship on a latent variable (or a few latent variables) to explain observed responses. In IRT models, logistic functions are used on the latent variable θ (ability) to explain students' item responses. The latent variable θ is viewed as ability, or a trait to perform well on test items. In a statistical modeling approach, it is crucial to test how well the model fits the observed responses. Various fit statistics have been developed for this purpose (Glas & Meijer, 2003).

The value of a diagnostic profile that enumerates particular strengths and weaknesses in individual performance has been recognized by various investigators (e.g., K. Tatsuoka & M. Tatsuoka, 1997; VanLehn, 1982). However, to be most valuable, diagnostic profiles should provide information about how well test takers performed on the underlying knowledge and

cognitive processing skills required in answering problems. In the rule-space approach, posited knowledge and thinking skills are termed *attributes,* and binary attribute patterns that express mastery/nonmastery of attributes define what are called *knowledge states* or *latent knowledge states.* While physical objects or events in science applications are usually observable, attributes and knowledge states are not observable. Measurement of such unobservable latent variables can be performed only indirectly from observable item scores by making inferences about what misconceptions, leading to what incorrect responses, did a tested individual most likely have. Given a response pattern, the goal in RSM is to determine the closest knowledge state to that pattern as well as the probability that a test taker's observed responses came from that state.

Factor analysis, cluster analysis, and traditional latent class models produce factors, clusters, and classes, but they are exploratory methods that merely group observed responses into similar classes or patterns. For this reason, they may produce solutions with no clear interpretation of the resulting groups of items or respondents. Ideally, diagnostic analyses of test results should be descriptive and objective, uniquely expressing an individual's state of knowledge, which must be free from ambiguous interpretations.

To achieve these goals, we need a new methodology that transforms unobservable knowledge and subskill variables into observable variables. The RSM transforms unobservable latent variables (attributes) into observable attribute mastery probabilities. Once the RSM results (the attribute mastery probabilities) are estimated, further statistical methods, such as factor analysis, cluster analysis, hierarchical multi-level analysis, and other multivariate analyses, can be applied to these outputs of the RSM analysis.

To explain how RSM works, it is helpful to relate it to statistical pattern recognition and classification problems, in which an observed pattern will be classified into one of the predetermined classification groups (Fukunaga, 1990; Ripley, 1996). Typical examples are to enable computers to recognize handwritten letters or to scan to X-ray images to diagnose whether or not a tumor is cancerous. For example, the letter recognition problem has 52 predetermined groups representing lower and upper cases of 26 alphabetic characters. These letters are expressed uniquely by 52 binary patterns of features. The set of features is predesigned by examining shapes, strokes, and geometric characteristics of the 52 letters. After this design stage is done, statistical classifiers are usually estimated. The classifiers classify an observed input pattern into one of 52 predetermined groups, and compute error probabilities. The group with the smallest error probability is usually taken as the letter to which the observed input pattern would belong. This general outline of a pattern recognition system applies to cognitive diagnosis, as follows: in a diagnostic analysis of a test, feature variables become attributes and the 52 letters represented by the patterns of the feature variables are analogous to knowledge states. However, the predetermined groups in the letter recognition example are expressed by observable feature variables, and hence they are directly measurable. Attributes are feature variables that are impos-

sible to measure directly, and knowledge states defined by patterns of attributes are also impossible to measure directly.

Therefore, RSM has to be extended to include an additional step to deal with latent feature variables. K. Tatsuoka (1990, 1991) solved this difficulty by introducing an item by attribute matrix Q where the cell q_{jk} in a Q matrix is coded by 1 if item j involves attribute k for answering item j correctly, and 0 if not. Thus, the Q matrix is a cognitive model for test item performance hypothesized by cognitive researchers, teachers, or other domain experts. It explains performance on the n observable test items in terms of competencies on k latent attributes. A knowledge state KS_m is defined by a (latent) attribute pattern of 1s and 0s. If a student can use attribute k correctly, then the kth element of KS_m is 1, and 0 if not. It is assumed that the right answer for item j is obtained if and only if all attributes involved in item j are successfully applied. Furthermore, the probability of answering item j correctly is assumed to be calculated by multiplying the probabilities of correct use of the involved attributes for item j.

K. Tatsuoka (1991) described an algorithm that generates all possible knowledge states from a given Q matrix, incorporated in a program called BUGSHELL (C. Tatsuoka, Varadi, & K. Tatsuoka, 1992). These possible knowledge states generate a set of expected or "ideal" item score patterns, so called in order to differentiate them from students' observed item response patterns. The knowledge states, or equivalently the ideal item score patterns, form a set of predetermined classification groups in RSM. The unobservable attribute patterns correspond to ideal item score patterns, which are directly observable.

One of the unique characteristics of RSM is that it entails developing a one-to-one correspondence between a subject's observed item response pattern and the corresponding ideal item score pattern(s). By so doing, we can make an inference about how well an individual has performed on latent attributes from his or her performance on observable item responses. Given a student's observed item response pattern, statistical classifiers estimate his or her knowledge state by estimating a mastery probability for each attribute. In this way a student's observed item responses are transformed into estimated attribute mastery probabilities. In other words, RSM transforms a data set of students by item scores into a data set of students by attribute mastery probabilities. The benefit of this approach is that it allows diagnosis of students in terms of very detailed content knowledge and processing skills.

In the present study we apply the RSM to investigate student mastery of critical knowledge and cognitive processing skills underlying performance on the TIMSS-R Mathematics achievement test. These skills, including domain knowledge, cognitive processing skills, and mathematical thinking skills, are referred to as *attributes* in applications of the RSM. The attributes used in the present analyses were developed specifically for the TIMSS-R eighth-grade mathematics test (Corter & Tatsuoka, 2002).

The present article includes analyses of how well students in different countries, with different cultures and educational environments, perform on

the TIMSS math items, and recast this performance in terms of mastery levels on the knowledge, skill, and process attributes. The organization of the article is as follows. First, the present rule-space analysis is briefly described. Then, a list of 20 countries to be studied is introduced, and descriptive statistics on mastery of specific attributes in these countries are presented. Some of the results are discussed and explored in more detail, for example, evidence concerning particular weaknesses in the U.S. students in geometry and certain process attributes. Third, a number of composite variable subscales are defined to summarize patterns of mastery on conceptually related specific attributes, and used to explore the performance characteristics of several top-ranked countries. Finally, we discuss how our results relate to findings of previous researchers, for example, international comparisons of classroom practices derived from TIMSS video studies.

Method

The present analyses use the rule-space methodology (K. Tatsuoka, 1983, 1985, 1990, 1995, 1997, in press; K. Tatsuoka & M. Tatsuoka, 1987; M. Tatsuoka & K. Tatsuoka, 1989) to diagnose each student in terms of inferred mastery of specific "attributes" (knowledge and subskill components) assumed to underlie test performance. The present work followed the general outline of any rule-space analysis, as follows. First, a set of underlying cognitive processing skills and knowledge believed to be involved in solution of the TIMSS-R "population 2" (eighth-grade) mathematics test items was identified. Then, a team of experts coded the test items in terms of which attributes are required for successful solution of each item, a process that defines the Q matrix. After preparation of the data set, the rule-space analysis was performed using special purpose software developed for this purpose. Results of the RSM include diagnosis of each student in terms of a vector of attribute mastery probabilities, as well as classification of each student into a closest knowledge state. In the present article, we used these results of the RSM to compare mathematics achievement across a sample of 20 countries participating in TIMSS-R. These steps are described in more detail below.

Identifying Attributes

In order to identify the specific knowledge and subskill attributes required for successful performance on the TIMSS-R eighth-grade math test items, we gathered and analyzed written student protocols. In addition, all items were solved by a team of domain experts. These experts were professors or graduate students in measurement, all of whom had experience teaching secondary-school mathematics or college-level introductory statistics courses. Finally, we interviewed two secondary-school mathematics teachers about the attributes in an attempt to gauge their validity and usefulness to educators, and used their feedback to revise the coding scheme. Further detail on the development and

validation of these attributes is given in Corter and K. Tatsuoka (2002) and K. Tatsuoka, Corter, & Guerrero (2004).

The set of attributes we developed fall into three general categories: content knowledge variables, cognitive process variables, and what we term "skill" or "item-type" variables (Table 1). The content attributes we used are not unlike the content categories used in the TIMSS-R test framework. The skill attributes include certain context-specific and format-specific process skills. This category of skills was deemed necessary because many specific skills in arithmetic and mathematics are associated closely with certain types of items. For example, skill attribute S3 involves reading data or understanding

Table 1

Knowledge, Skill, and Process Attributes Derived to Explain Performance on Mathematics Items From the TIMSS-R (1999) for Population 2 (Eighth Graders)

Content attributes	
C1	Basic concepts and operations in whole numbers and integers
C2	Basic concepts and operations in fractions and decimals
C3	Basic concepts and operations in elementary algebra
C4	Basic concepts and operations in two-dimensional geometry
C5	Data, probability, and basic statistics
C6	Measuring or estimating: length, time, angle, temperature, etc.
Process attributes	
P1	Translate/formulate equations and expressions to solve a problem
P2	Computational applications of knowledge in arithmetic and geometry
P3	Judgmental applications of knowledge in arithmetic and geometry
P4	Applying rules in algebra
P5	Logical reasoning—includes case reasoning, deductive thinking skills, if-then, necessary and sufficient, generalization skills
P6	Problem search; analytic thinking, problem restructuring; inductive thinking
P7	Generating, visualizing, and reading figures and graphs
P8	Applying and evaluating mathematical correctness
P9	Management of data and procedures
P10	Quantitative and logical reading
Skill (item type) attributes	
S1	Unit conversion
S2	Apply number properties and relationships; number sense/number line
S3	Using figures, tables, charts, and graphs
S4	Approximation/estimation
S5	Evaluate/verify/check options
S6	Patterns and relationships (inductive thinking skills)
S7	Using proportional reasoning
S8	Solving novel or unfamiliar problems
S9	Comparison of two/or more entities
S10	Open-ended items, in which an answer is not given
S11	Understanding verbally posed questions

relationships from graphs and figures provided in the actual math item. As another example, proportional reasoning (skill S7) is a process variable, but this type of reasoning is often associated with a particular format of item. In contrast, the process attributes are more general skills that may be used across a wider variety of item types. For example, logical reasoning (process attribute P5) is used in many types of items, and encompasses several varieties of logical inference.

Developing the Q Matrix

Once the set of attributes thought to be involved in successful solution of the items was identified, each actual math item used in the test had to be coded for the attributes needed to solve it. Here, 163 different test items were used across the eight forms of the mathematics test booklet. These 163 test items were each coded by three raters, to specify the attributes involved in solution of each item. The three coders consisted of two researchers with doctoral degrees and college teaching experience in statistics and measurement and a graduate student with a background in measurement. Each of the three raters coded items independently first, then they met to discuss any discrepancies (including reviewing student protocols), until a consensus was reached.

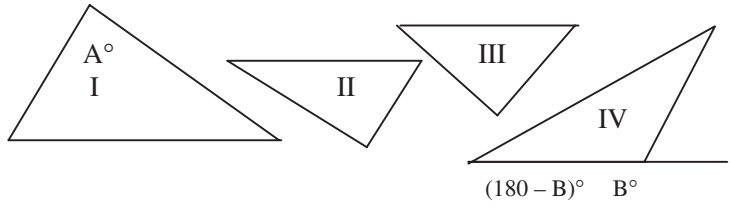
One complication that can arise in item coding arises if the raters (or a single rater) identify several possible strategies for solving the item (e.g., an algebraic strategy versus plug-in). For the present study, in these cases we took the approach of coding both strategies, then comparing each solution to student work and student performance, in an attempt to identify the dominant strategy used by this population (cf. Tatsuoka, 1990).

After consensus was reached on an initial Q matrix, some iterative refinement of the coding scheme was performed. As a start, a linear multiple regression analysis was performed predicting item difficulties from the coded binary entries in the Q matrix (that is, using the coded attributes as predictors). Also, a preliminary RSM analysis was performed using this Q matrix to derive estimated attribute mastery probabilities for each student. Then, descriptive statistics (including a correlation matrix of the attribute mastery probabilities) were computed and used to eliminate statistically weak attributes for the final stage of RSM analysis. Some attributes that were highly correlated were merged into single attributes. In this way, the final Q matrix was developed.

An example item involving attributes C4, S3, S5, P3, P5, P7, and P9 with the proportion correct of .1025 is given in Figure 1. This problem is very difficult because it involves P3, P5, P9, and P7.

Selection of the Sample of Countries

For the comparative rule-space analyses, we chose to focus on the following countries: the United States, Australia, Belgium-Flemish, Canada, Chile, Czech Republic, England, Finland, Hong Kong, Indonesia, Israel, Italy, Japan, Jordan, Korea, Netherlands, Philippine, Russia, and Turkey. Our selection was based on several criteria, intended to achieve a diverse sample both cultur-



Which two triangles are similar?

A. I and IV
B. I and II
C. II and III
D. II and IV
E. III and IV

This is a geometry problem…….……..…….. C4
Use figures……..…….…….…….…….…….. S4
Must evaluate options to get answer. ……... S5
Apply some property to judge "similar or not"...…….. P3
One angle in IV is (180 – B)° > 90°; A° in I is 90°.
Since I and II have 3 parallel sides, I and II have the
same angles. Therefore, two triangles are similar…….. P5
Any other pairs do not have this property…….……. P9
Comprehend the relationships of figures, such as
which sides are parallel. Add a line to get (180 – B)°... P7

Figure 1. **A sample TIMSS math item, with attribute coding.**

ally and in terms of achievement levels. Specifically, the following countries were selected because they were the six top-achieving countries based on mean "plausible-value" total achievement scale score: Singapore, Korea, Hong Kong, Japan, the Czech Republic, and Belgium-Flemish (Mullis et al., 2001). The United States, Canada, and Australia were all included because Canada and Australia are similar to the United States in being English-speaking countries with heterogeneous populations, yet their math ranking is considerably higher than that of the United States. Chile was selected as a representative Spanish-speaking country. Several European countries were selected, including Russia, Italy, Finland, Netherlands, and England. Finally, Israel and several other Middle Eastern or Islamic countries were selected, namely, Turkey, Jordan, Indonesia, and the Philippines.

RSM Diagnosis and Classification of Students

Rule-space analyses were run separately for each of the 20 countries. For each analysis, we used data from only Booklets (i.e., forms) 1, 3, 5, and 7, because the other booklets showed a "patchy" distribution of attributes (that is, few or no items measuring certain attributes). Still, we found an insufficient number of items representing some of the attributes (C6, S1, S9, and P8) described in Table 1. Because this data sparseness means we cannot estimate these attributes reliably, they were not included in the RSM analysis.

Results

Before comparing the results of estimated attribute mastery probabilities across countries, we will report some overall descriptive statistics. For the combined sample ($N = 51{,}435$) of students from all 20 countries, the mean Mahalonobis distance (D^2) value of students' attribute mastery vectors from the closest knowledge state was less than .5. For the second closest knowledge state, mean D^2 was approximately 1.0. These mean Mahalonobis distances are relatively small. Based on the following logic, we adopted a cutoff criterion of 4.5, judging as a successful classification any Mahalonobis distance of less than 4.5 from a student's attribute mastery vector to the nearest knowledge state. The logic behind this criterion is as follows. In general, D^2 may be modeled by the gamma density function with expectation $(\beta+1)/\alpha = p$ and variance $(\beta+1)/\alpha^2 = 2p$, where p is the dimensionality of the classification space (Fukunaga, 1990; Hogg & Craig, 1978). Because a RSM analysis defines a three-dimensional space with orthonormal coordinates, D^2 follows a special case of gamma, namely the chi-square distribution with 3 degrees of freedom, hence an expected value of 3 and a variance of 6. Thus, a cutoff of $D^2 < 4.5$ corresponds to accepting approximately 80% of the expected population values as satisfactory or not unusual.

Using this criterion for successful classification resulted in an average 99.5% classification rate in these 20 countries. In other words, almost all observed item response patterns were classified into one of the logically derived knowledge states from our Q matrix. This finding can be seen as providing one form of validation for the set of proposed attributes used here, because it demonstrates that the attributes perform well in explaining eighth-graders' performance on the TIMSS-R math items.

Another check on the validity of the proposed attributes was performed using ordinary multiple regression analysis. To do this, we tested whether the variance in item difficulties (measured by mean proportion correct values across all respondents) can be predicted using only information about which attributes are involved in that item (as specified by the Q matrix). This can be set up as a regression problem in which the observations are the $N = 163$ items, and the predictors are the columns of the Q matrix, namely 27 vectors (attribute variables) with values equal to 1 if the corresponding attribute is involved in the item being considered, and 0 otherwise. In this regression, we obtained an adjusted R^2 value of .869 using the present set of attributes. This result shows that the coded attribute composition of a math item does a good job of predicting its difficulty.

Comparison of Performance on Single Attributes Across Countries

The descriptive statistics on attribute performance across countries showed that the standard deviations of attribute mastery probabilities, calculated within each of the 20 countries, vary greatly. For the most part these discrepancies in variability seem explainable. For example, some countries showing low standard deviations, such as Hong Kong and Korea, are very

homogeneous in terms of students, teacher training, and schools, while other countries showing much larger standard deviations for some attributes have diversified cultures and distribution of resources.

In order to compare more easily attribute mastery probabilities across countries, all attribute mastery probabilities were standardized by subtracting their values from the grand means of each attribute (across the 20 countries) and dividing by the pooled standard deviation. Since sample sizes range from a minimum of 1,700 (for the Czech Republic) to a maximum of 4,500 (for the United States), even small differences between two standardized attribute scores tend to be significant.

The most difficult attribute (i.e., with the lowest mastery probability) across the 20 countries is S6, inductive thinking skills. The easiest attributes are S5 (evaluate and verify options), S3 (using figures and tables), and P1 (translate word expressions into equation or algebraic expressions). By looking into which attributes the students in each of the 20 countries have successfully mastered, one can form hypotheses about the skills and values that countries emphasize in their teaching, curriculum, and culture.

The top five countries excel in the cognitive process variables. Japanese students in particular have the highest mastery probabilities for the most difficult cognitive processing variables, P5 (deductive thinking), P6 (analytical thinking), P3 (application of concepts, theories), P9 (management of processes and data), and S6 (inductive thinking skills). This may be related to the problem-solving approach used in math teaching in Japanese classrooms (Kawanaka & Stigler, 1999). Other examples of very high attribute mastery probabilities can be seen by listing countries that had the maximum mean performance on specific attributes (Table 2).

Another way to compare attribute performance across countries is to compare how well a country does on a specific attribute to how well it performs overall as measured by the mean item percent-correct (PC) score. Discrepancies between the specific attribute performance and the general PC measure may give clues to problems or strengths in a country's typical approach to teaching a specific skill or topic. For example, Figure 2 (p. 913) shows the profile plot across countries for performance on the specific attribute C4 (geometry) as well as on the mean of standardized item proportion correct (PC) score. Geometry seems to be a relatively weak content skill area for U.S. students at this grade level, as shown by the fact that the profile trend for this attribute dips sharply below the relative mean achievement (PC) line for the United States. In contrast, Russia and Italy are doing very well on geometry at the eighth-grade level because the profile trend rises sharply there above the PC curve there.

Figure 3 (p. 914) shows the profile curves of three other content variables: C1, C2, and C3. The United States performed much better on these than in geometry. Hong Kong, Netherlands, and Russia performed especially well on C3 (algebra). The top 11 countries performed nearly equally well on C1 (whole numbers/integers) and C2 (fractions) but not on C3 (algebra). In contrast, students in Indonesia, Chile, and the Philippines are doing relatively poorly

Table 2

Countries With Maximal Mean Achievement on Specific Attributes

Country	Maximal on attribute(s)	
Aus	S4	Approximation/estimation
Bfl	C1	Basic concepts, properties, and operations in whole numbers and integers
	C2	Basic concepts, properties, and operations in fractions and decimals
	S11	Understanding verbally posed questions
Czk	S8	Solving novel or unfamiliar problems
Fin	S2	Apply number properties/relationships; number sense/number line
Hkg	C3	Basic concepts and operations in elementary algebra
	C4	Basic concepts and operations in two-dimensional geometry
Jpn	C5	Data, probability, and basic statistics
	S6	Patterns and relationships (inductive thinking skills)
	P3	Judgmental applications of knowledge in arithmetic and geometry
	P5	Logical reasoning—includes case reasoning, deductive thinking skills, if-then, necessary and sufficient, generalization skills
	P6	Problem search: Analytic thinking, problem restructuring, and inductive thinking
Kor	C4	Basic concepts and properties of two-dimensional geometry
	S3	Using figures, tables, charts, and graphs
	P9	Management of data and procedures
	P10	Quantitative and logical reading
Nld	S5	Evaluate/verify/check options
	P7	Generating, visualizing, and reading figures and graphs
Sgr	S7	Using proportional reasoning
	P1	Translate/formulate equations and expressions to solve a problem
	P2	Computational applications of knowledge in arithmetic and geometry
	P4	Applying rules in algebra
	S10	Open-ended item, in which an answer is not given

Note. Aus = Australia, Bfl = Belgium-Flemish, Czk = Czech Republic, Fin = Finland, Hkg = Hong Kong, Jpn = Japan, Kor = Korea, Nld = Netherlands, Sgr = Singapore.

on C2. The students in the four Asian countries (especially Hong Kong) performed well on C3 (algebra). Russia and the United States also had relatively strong scores in algebra, but Finland, Italy, and England are doing less well.

Defining Composite Attribute Variables to Compare Achievement Across Countries

The diagnosis of students in terms of these very detailed content and process attributes is one of the primary benefits of a rule-space analysis. However, in order to see larger patterns in the data across a large number of countries, it can be useful to create composite variables comprised of several specific attributes that are related either conceptually or statistically. Countries may then be compared on these composite or summary variables.

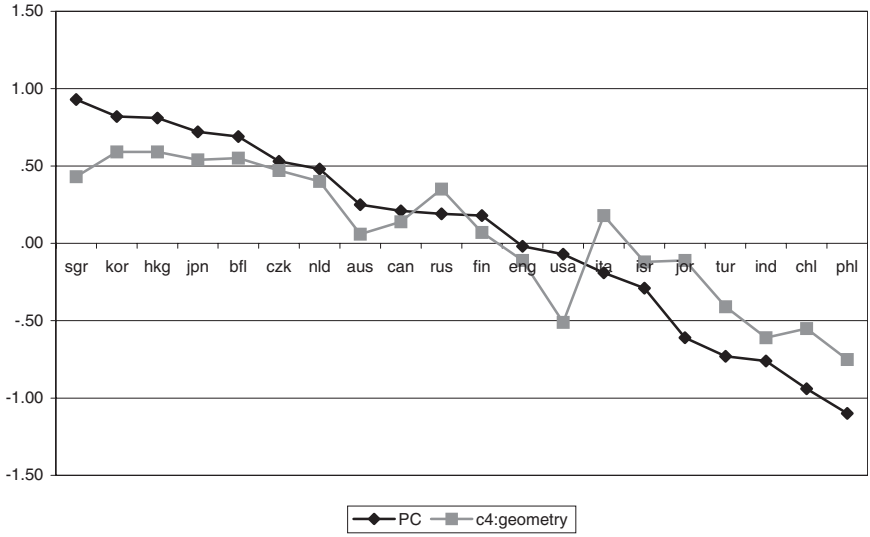


Figure 2. **Profile across 20 countries of geometry knowledge (C4) achievement, with the standardized item proportion correct (PC). Country codes (alphabetical): Aus = Australia, Bfl = Belgium-Flemish, Can = Canada, Chl = Chile, Czk = Czech Republic, Eng = England, Fin = Finland, Hkg = Hong Kong, Ind = Indonesia, Isr = Israel, Ita = Italy, Jor = Jordan, Jpn = Japan, Kor = Korea, Nld = Netherlands, Phl = Philippines, Rus = Russia, Sgr = Singapore, Tur = Turkey, USA = United States.**

We created three composite variables based on theoretical grounds and three based on statistical evidence (i.e., correlations of attributes). The three conceptually based composite variables consisted of the sum of all nine process skills ("Process"), the sum of three spatial skills ("Spatial"), and the sum of three reading skills ("Reading"). These composite variables have obvious interpretations. We investigated if the Process attribute was too broadly defined by computing correlations among the nine process skills and examining them via a principal components analysis with Varimax rotation. This analysis resulted in the extraction of three components: p1 = (P1, P2, P6, P7, P10), p2 = (P3, P5), and p3 = (P4, P9). They are summarized as follows:

Process:	P1 + P2 + P3 + P4 + P5 + P6 + P7 + P9 + P10
Spatial:	C4 + S3 + P7
Reading:	S11 + P1 + P10
p1:	P1 + P2 + P6 + P7 + P10 (application skills P1, P2; verbal skills P1, P10; problem search P6; spatial skill P7)
p2:	P3 + P5 (logical and abstract reasoning skills)
p3:	P4 + P9 (algebraic skills; plus data and complex process management)

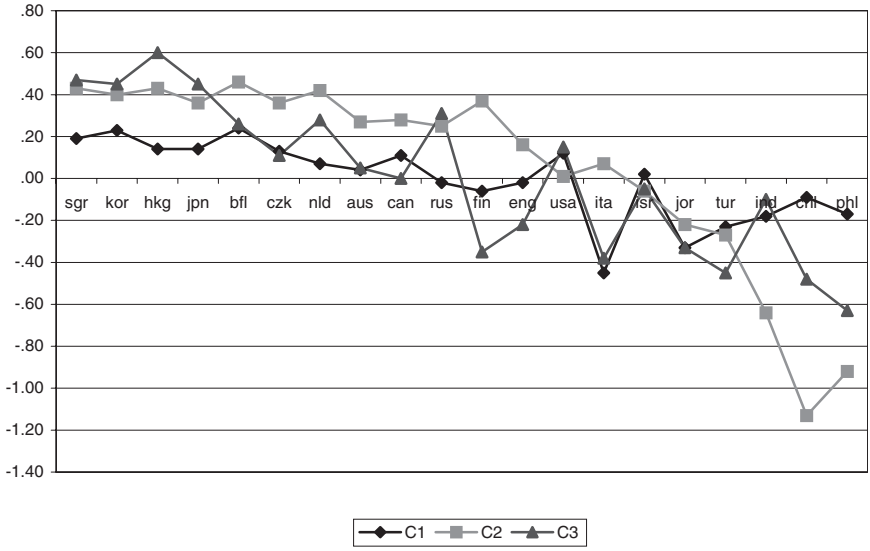


Figure 3. **The profiles across 20 countries of three content variables: C1 (whole numbers), C2 (fractions), and C3 (algebra). For country codes, see Figure 2.**

Table 3 reports the means for these summary variables by country. Profiles of the summary variables across countries are shown in the following figures.

The first summary variable, Process, represents the sum of the attribute mastery probabilities for the nine process-skill attributes. Interestingly, this summary score carries much of the variance of the original total mean item scores of the test, as seen by the fact that the ordering of the highest countries is the same by both types of summary score. Specifically, by either mean item PC score or by the process variables, summary variable, the ordering of the top six countries is: Singapore, followed by Hong Kong, Korea, Japan, Belgium-Flemish, and the Czech Republic. This ordering can be seen in Figure 4 (p. 916), which plots mean item PC across the countries, ordered by this score. The trend line for the Process composite across the 20 countries is monotonic with the line for PC.

Figure 5 (p. 916) shows the spatial skills composite variable, Spatial, which consists of C4 (geometry), S3 (using figures/graphs), and P7 (generating figures/graphs). Hong Kong shows the highest mean performance on this composite. The seven top-ranking countries performed this skill equally well, followed by Australia, Canada, Russia, Finland, and England. The United States dips sharply on this composite variable, like Israel and Jordan. The profile curve for this composite variable is almost identical to that of geometry alone. On the composite skill, Reading, Japan dips sharply and Russia dips moderately.

Table 3

Standardized Means, by Country, of Defined Composite Achievement Variables

Country	Process	Spatial	Reading	F1	F2	F3
Sgr	.48	.43	.38	.45	.37	.70
Kor	.45	.49	.32	.40	.40	.64
Hkg	.44	.51	.30	.39	.37	.66
Jpn	.39	.45	.12	.31	.51	.49
Bfl	.38	.50	.34	.38	.37	.39
Czk	.28	.45	.25	.34	.20	.23
Nld	.26	.45	.29	.32	.23	.13
Aus	.17	.25	.20	.28	.09	−.03
Can	.16	.24	.21	.26	.03	.04
Rus	.12	.24	.13	.11	.09	.18
Fin	.10	.28	.32	.32	−.02	−.32
Eng	.00	.19	.11	.22	−.10	−.46
USA	−.03	−.16	.15	.11	−.33	−.07
Ita	−.06	.08	−.11	−.06	−.02	−.13
Isr	−.14	−.15	.01	−.09	−.15	−.24
Jor	−.30	−.15	−.38	−.42	−.10	−.22
Tur	−.36	−.60	−.26	−.50	−.11	−.27
Ind	−.48	−.67	−.48	−.51	−.70	−.21
Chl	−.50	−.62	−.48	−.45	−.39	−.71
Phl	−.74	−.94	−.83	−.98	−.18	−.67
Top country	Sgr	Hkg	Sgr	Sgr	Jpn	Sgr

Note. Process = sum of all nine cognitive process variables; Spatial = C4 + S3 + P7; Reading = S11 + P1 + P10; F1 = P1 + P2 + P6 + P7 + P10; F2 = P3 + P5; F3 = P4 + P9. Country codes (alphabetical): Aus = Australia, Bfl = Belgium-Flemish, Can = Canada, Chl = Chile, Czk = Czech Republic, Eng = England, Fin = Finland, Hkg = Hong Kong, Ind = Indonesia, Isr = Israel, Ita = Italy, Jor = Jordan, Jpn = Japan, Kor = Korea, Nld = Netherlands, Phl = Philippines, Rus = Russia, Sgr = Singapore, Tur = Turkey, USA = United States.

Figure 6 (p. 917) shows the profiles of the three process subcomponents: p1 = (P1, P2, P6, P7, P10), p2 = (P3, P5), and p3 = (P4, P9). The curve for the overall Process composite is also given for reference. It is clear from the figure that students in Singapore, Korea, and Hong Kong performed at a high level on p3, consisting of algebraic skills and complex data/procedure management skills. However, they performed relatively less strongly on p1 (consisting of P1, P2, P6, P7, P10) and p2 (consisting of P3, P5). We can conclude that Singapore, Korea, and Hong Kong students achieve their high mean performance mainly by mastering algebraic skills and complex management skills. England, Finland, Australia, Canada, and Netherlands were weaker on p3 (algebra/complex management) than on p1. On p2, abstract judgment and logical reasoning skills, Japanese students achieved the highest. The U.S. students did not do well on p2, in fact ranking 17th in the sample of countries. It is important to note that U.S. students did relatively well on C3 (algebra),

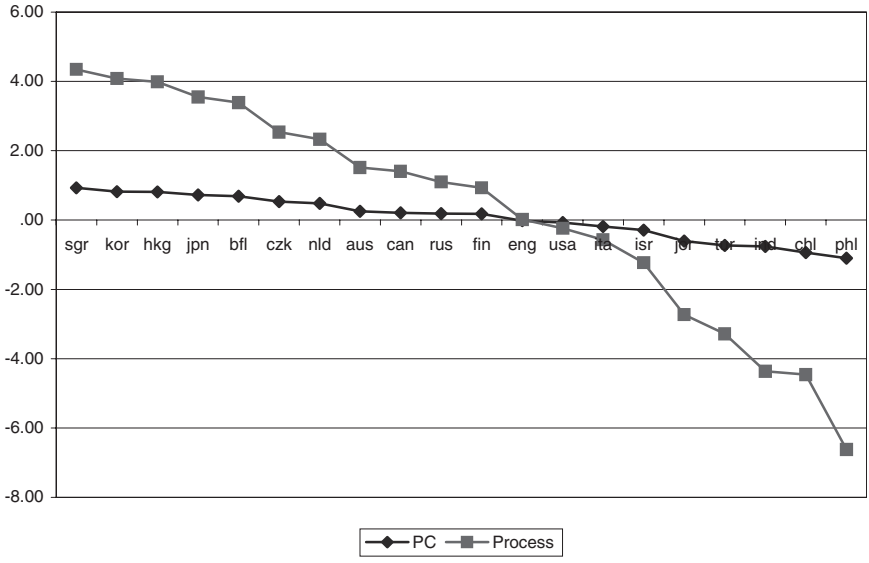


Figure 4. **The profiles across 20 countries of the standardized mean item proportion correct (PC) scores, along with the composite variable (*Process*) representing the (standardized) sum of all process skills. For country codes, see Figure 2.**

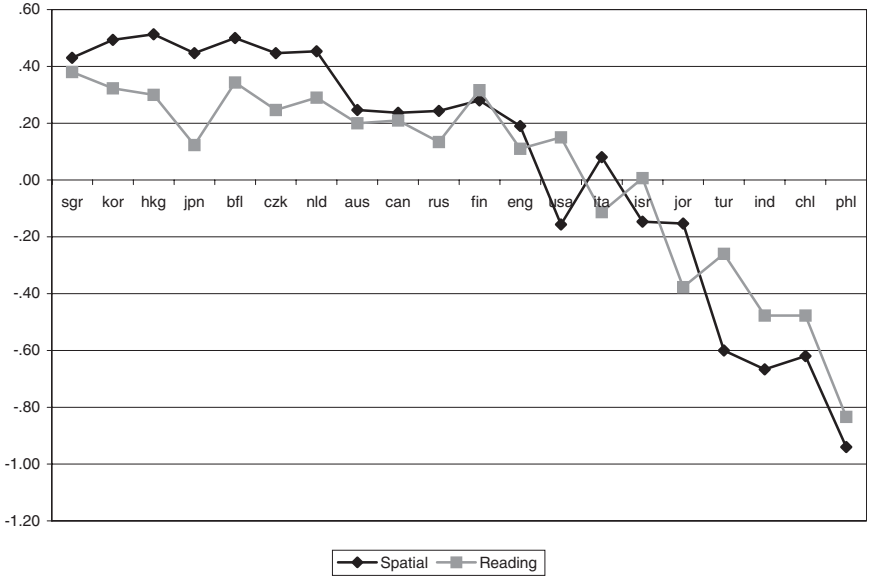


Figure 5. **Profiles across 20 countries of the composite variables *Spatial*, representing the sum of skills C4, S3, and P7, and *Reading*, representing the sum of S11, P1, and P10. For country codes, see Figure 2.**

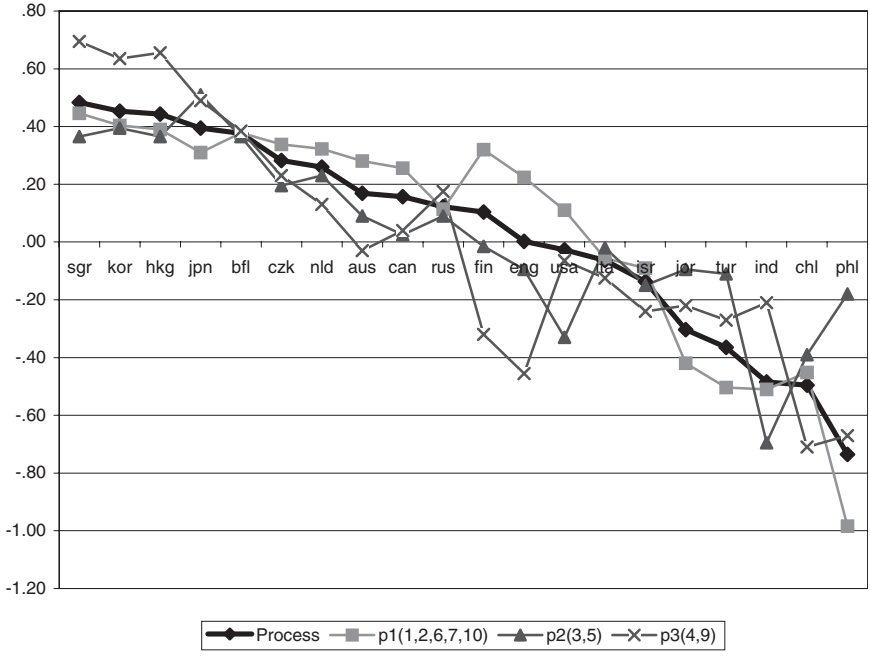


Figure 6. **Profiles across 20 countries of the composite variable *Process* representing the sum of all process skills, with the mean standardized scores of three components obtained from a principal components analysis with Varimax rotation performed on the nine process variables. For country codes, see Figure 2.**

C1 (numbers), and p1, but not well on C4 (geometry) and p2 (the logical reasoning skills). A natural question arises—is there a connection between the relative weakness of U.S. students on geometry and on higher order thinking skills, or are these two areas that just happen to be less effectively taught at or before this grade level in the United States? We will return to this issue later.

Principal Components Analysis of All Attribute Means

The dimensionality of the set of attributes was explored via a principal components analysis of the matrix of mean standardized attribute mastery probability scores across the 20 countries. Because the results are based on the correlations between mean attribute vectors across the 20 countries, the results are affected by profile shape but not by overall difficulty level. Evidence was found for four dimensions using the criterion of number of eigenvalues greater than 1. A Varimax rotation with Kaiser normalization converged in 12 iterations. The resulting components matrix is given in Table 4. Using a cutoff of .55 for interpretation purposes, the first component consists of attributes C2,

Table 4

Rotated Component Matrix From the Principal Components Analysis of All Attributes, Performed on Mean Attribute Mastery Probability Profiles Across a Sample of 20 Countries

Attribute description	F1	F2	F3	F4
C2: Fractions	0.66			
S3: Figures, tables, and graphs	0.68			
P6: Problem search, analytical thinking skills	0.73			
P1: Translate words to equations and expressions	0.76			
S11: Understanding verbally posed problems	0.54			
S10: Open-ended items	0.57	0.56		
C1: Whole numbers and integers		0.75		
C3: Elementary algebra		0.71		
C5: Data, probability, and statistics		0.75		
S6: Recognize patterns and their relations		0.62		
P2: Computational applications		0.74		
P4: Apply rules to solve equations, derive algebraic expressions		0.61		
C4: Geometry	0.57		0.65	
S7: Proportional reasoning			0.78	
P3: Judgmental applications of knowledge and concepts			0.73	
P9: Executive control, manage data, process			0.71	
P5: Logical reasoning skills			0.65	
S2: Number properties and relations				0.84
S3: Figures, tables, and graphs	0.59			0.68
S4: Approximation and estimation				0.91
S5: Evaluate and verify options				0.75
P7: Generating, visualizing and reading graphs, figures, tables	0.59			0.62
P10: Quantitative and logical reading				0.80

Note. Only component loadings above a cutoff of .50 are shown.

S3, P1, P6, and S10, with large coefficients for P6 (problem search) and P1 (translating words to algebraic equations). The second component consists of attributes C1, C3, C5, S6, P2, and P4. It can be characterized by C3 (algebra), P4 (algebraic skills), and C5 (data, probability, statistics). The third component consists of C4, S7, P3, P5, and P9, with larger coefficients for P3 (judgmental applications of knowledge) and S7 (proportional reasoning). The fourth and last component consists of S2, S4, S5, and P10. Attribute S11 did not show up in any component with a coefficient larger than .55, but it appeared with a coefficient of .54, loading on the first component. The first component includes P1 (translate equations and expressions), and the third component includes C4 (geometry) and P5 (logical reasoning). Geometry (C4) also appears in the first component (with a weaker coefficient than in component 3), which is

characterized by S3 (using figures, tables, and graphs). S3 also appeared in the fourth component with P7 (spatial skills). The fourth component includes number sense and properties (S2) and the approximation/estimation attribute (S4), evaluation skills (S5), and quantitative and logical reading skills (P10).

It is interesting that geometry skill loads on the same component as some of the important mathematical thinking skills, namely P5, P3, and P9. That suggests that the observation that U.S. students are relatively weak on both these skills is not coincidence, but may be due to some meaningful correlation between these skills. Geometry also relates closely to proportional reasoning skills, which can be argued to be an important thinking skill in everyday life, and also may be aided by spatial reasoning. On the other hand, algebra (C3) loads on the same components as the important algebraic and arithmetic computational skills, P2 and P4.

Figure 7 underscores that U.S. students did not perform well on geometry (C4) and the higher level mathematical thinking skills attributes P5, P3, P9, and S7, although they had relatively high achievement in algebra (C3), approximation and estimation (S4), and quantitative and logical reading (P10).

It was unexpected to find that algebra did not correlate with these important mathematical thinking skills. However, it did correlate highly with C1, C5, S6, P2, and P4, which involve the computational application of content knowledge and algebraic computational skills. Along with the finding, described

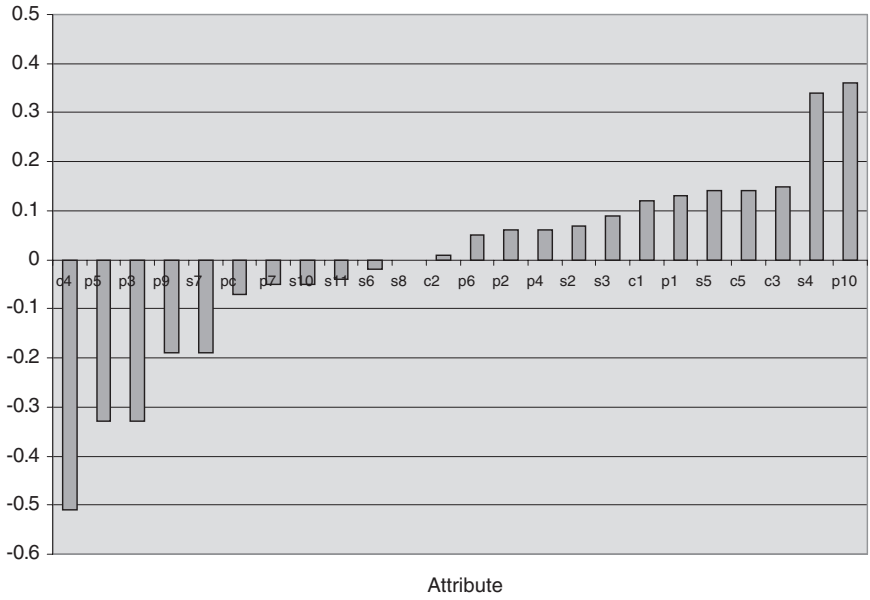


Figure 7. **Profile showing mean standardized performance of U.S. students for the 23 content, skill, and process attributes, ordered by relative performance level.**

above, that geometry does correlate highly with the attributes measuring higher order mathematical thinking, this raises the question of whether geometry might be a gateway skill to certain higher order mathematical and logical reasoning skills, just as algebra has long been considered a gateway skill in applied mathematics and technical fields.

In order to explore this unanticipated result, we examined the items requiring C3, C4, and P5, and found that the items involving geometry were slightly more difficult than the items involving algebra, with mean proportions correct of .44 (31 items) and .41 (38 items), respectively. However, the geometry items that involve P5 (logical reasoning) were easier than the algebra items that involve P5, .32 (14 items) versus .28 (11 items), respectively. These statistics suggest that for middle-school students, geometry may be a better topic than algebra with which to teach important mathematical thinking skills. While it is obvious that algebra is important as a gateway to mathematics, the content of elementary algebra suitable to this age group may not include opportunities to teach challenging mathematical thinking skills. Since the Asian and some European countries in this sample are teaching mathematical thinking skills very well (as can be seen in Figure 6), we may be able to teach U.S. children these higher level mathematics thinking skills better than we are doing now.

Cluster Analysis of Countries

As another way of finding general patterns of attribute mastery across countries, a type of cluster analysis was performed on the standardized attribute mean vectors for the 20 countries. Specifically, the correlation between each pair of countries was computed across the 23 attributes, plus the standardized percent correct. Then, an additive tree was fit to the matrix of correlations among countries, using the GTREE program (Corter, 1998). The resulting tree accounts for 72% of the variance in the correlations and is shown in Figure 8. The root for the tree in Figure 8 is selected so as to minimize the variance of the distances from the root to the leaves.

Four distinct branches or clusters are apparent in the tree. At the top is a cluster consisting at its core of Singapore, Korea, and Hong Kong. Japan and Belgium-Flemish then join this cluster, followed by the Czech Republic and then the Netherlands. These are the highest performing countries. The next cluster consists of Italy and Jordan, joined by Russia, a cluster that has no obvious interpretation. The third cluster consists of England and Finland, joined by Australia, Canada, and then finally the United States. Note that all of these except Finland have English as their national language and share some historical and cultural ties. Finally, the fourth cluster consists of two pairs: Turkey joins Indonesia (both predominately Muslim countries), and Chile joins with the Philippines (both Catholic countries that are former Spanish colonies). These two pairs are then joined by Israel. Examination of these patterns of similarity among countries' achievement profiles suggest that patterns of student achievement are influenced at least in part by shared culture

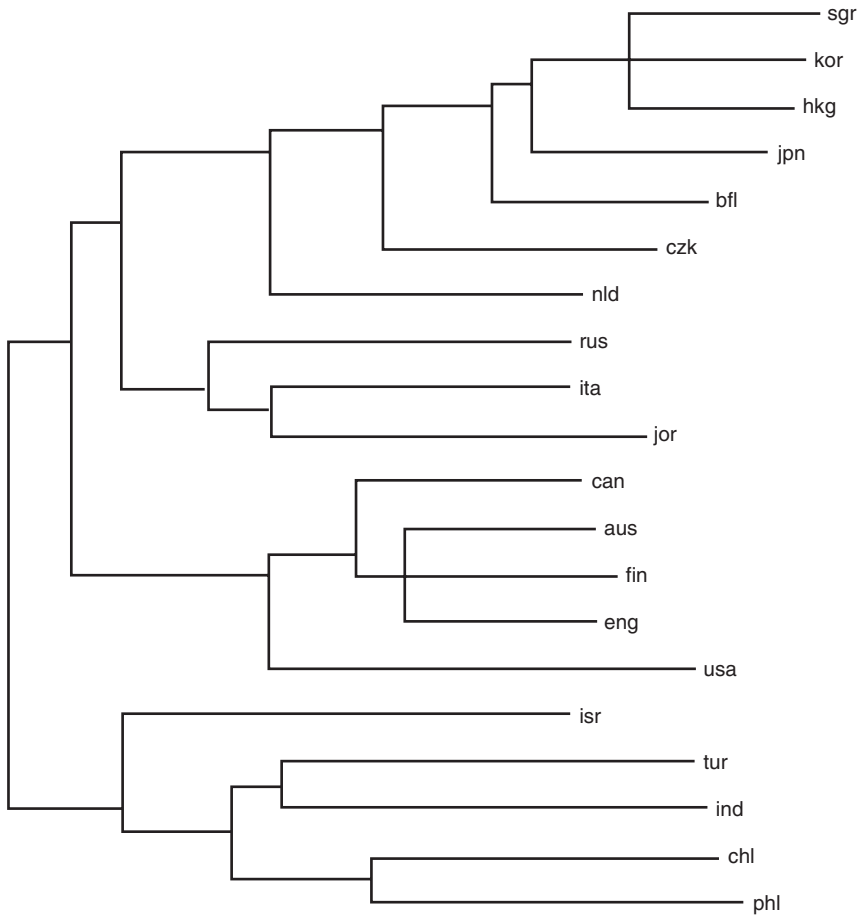


Figure 8. **Additive tree of the correlations of pairs of countries across 23 attribute means plus the standardized proportion correct (PC). Proportion of variance accounted for (PVAF) = .72. For country codes, see Figure 2.**

or language. Whether these cultural commonalities are also reflected in school and curriculum practices may be a fruitful avenue for future investigations.

Summary and Discussion

The present results demonstrate that the investigation of microlevel cognitive processing, mathematical thinking skills, and knowledge can lead us to new findings and provide us with insights into problems in educational practice. RSM is a method that researchers can use to implement and test their cognitive models of educational skills, thereby providing information that can be

useful for diagnosis of students and the improvement of educational practice. Since RSM, unlike other methods using statistical modeling approaches, is a pattern analysis technique requiring only weak assumptions, the model can be applied widely. RSM does incorporate the assumption that a student obtains the correct answer to an item if and only if all attributes involved in the item are applied correctly. RSM also assumes that the density function of the knowledge states follows a normal distribution.

Note that a RSM analysis converts a data set of performance on specific test items for each student into an output vector of attribute mastery probabilities. One advantage to this transformation is that it allows merging the performance data for several different tests to a single data set of students by attributes, as long as tests share the same set of attributes. This property of the RSM is particularly useful for the sampling design of the TIMSS studies, because the RSM results from several booklets can be merged into a single data set of students by attribute mastery probabilities.

Our results show that high-achieving countries in the eighth-grade TIMSS-99 mathematics assessment attained their level of performance in different ways. For example, Singapore students obtained the top performance on TIMSS mainly by showing excellence in reading and computational skills. Japanese students demonstrated excellent higher level thinking skills, while Belgian students achieved high scores through strength in fractions and proportional reasoning skills. Hong Kong and Korean students show relatively balanced knowledge and processing skills. In contrast, students in the industrialized countries that were not grouped into the highest achieving cluster (see Figure 8) tend to show weaker scores in these higher level mathematical thinking skills. These industrialized countries in lower-achieving subgroups include Russia, Italy (in the second cluster), Canada, Australia, Finland, England, United States (in the third cluster), and Israel (in the fourth cluster). These higher level skills are extremely important for students to master in order to succeed at study or employment in science and technology fields.

Unlike most other industrialized countries, U.S. students also did not perform well on the content area of geometry. Since geometry here correlated highly with the important mathematical skills S7 (proportional reasoning), P3 (judgmental application of knowledge, concepts, properties), and P9 (managing data and processing skills), geometry may be something of a gateway skill to the teaching of higher order mathematics thinking skills. This may be because geometry is a domain that serves as an effective context in which to teach logical reasoning and higher level judgmental skills.

These findings suggest that the curriculum in the United States should put more emphasis on teaching geometry, because geometry may enable teaching of important mathematical thinking skills needed in physical science and engineering. These fields are of course extremely important in maintaining a technological edge in global industries. Surprisingly, algebra did not correlate with these mathematical thinking skills but was related to computational skills. Thus, educators concerned with designing effective mathematics curricula might ask: Is the emphasis on algebra in current U.S.

mathematics curricula sufficient to effectively teach logical reasoning and higher level judgmental skills to this age group (eighth graders)? Also, can aspects of U.S. curricula, textbooks, or teaching practices be found that seem to be related to these weaknesses in U.S. students?

Analyses of the TIMSS video studies of international classroom practices (e.g., Hiebert & Stigler, 2000; Kawanaka & Stigler, 1999; Schumer, 1999; Stigler, Gonzales, Kawanaka, Knoll, & Serrano, 1999) offer some evidence that these differences among countries in higher order mathematics thinking skills may be due to differences in teaching practices. In addition, it has been shown that learning which takes place outside of regular instruction hours can have a dramatic influence on the learning processes and learning success (Chen & Stevenson, 1995) of students.

Using the TIMSS video data, Kawanaka and Stigler (1999) investigated teachers' use of questions in eighth-grade mathematics classrooms in their video study. They concluded that teachers dominated conversation in the classroom in countries such as Japan, Germany, and the United States, and students' conversation was mostly in the form of responses to teacher questions. Interestingly, teachers in these three countries asked different kinds of higher order questions, apparently reflecting differing pedagogical goals. German teachers often asked students to explain what they know and what they thought, but the information tended to be elicited and evaluated by teachers. Japanese teachers emphasized divergent thinking in problem solving and the solution of open-ended problems. For example, students were asked to solve nonroutine problems entirely on their own, using any methods they want to use. The students were encouraged to think about how to solve the problem rather than actually solving the problem. "(Japanese) teachers let students go through the process of identifying a problem, investigating solution methods, shared individual thoughts and correctively arriving at a conclusion" (Kawanaka & Stigler, 1999, p. 277). Kawanaka and Stigler also concluded that mathematics instruction in the United States emphasizes mastery of principles and procedures and the production of correct answers, despite reform ideas that encourage teachers to shift their instruction toward nonroutine problem solving.

It is also true that curriculum studies have established that U.S. textbooks include far more topics than is typical internationally and are considerably less focused than comparable textbooks in other countries (Schmidt et al., 1999). The video studies by Stigler and his associates also concluded that U.S. teachers tend to teach less advanced topics than Japan and Germany, tend to splinter their lessons into many small activity components, and spend more class time on homework (Jakwerth, 2004). Teaching geometry may require more focus on a few carefully selected topics than current textbooks do, and may need to go conceptually deeper, in order to teach mathematical proof skills. Thus, it may be that geometry is an area in which such higher order skills can easily be learned or applied. If this is true, then it seems desirable that math curricula in junior high schools in the United States be modified to increase the time devoted to teaching geometry.

Broad surveys and investigations of math achievement across a large sample of countries, like the present one, can be interesting to illuminate patterns of attribute achievement across countries. However, useful insights concerning curricula, schools, and teaching practices may emerge more easily via in-depth investigation of only a few countries at a time. By so doing, one can go deeper into the effects of culture and educational systems, making it easier to extract useful information to improve education internationally.

Notes

This study was supported by a grant from the National Science Foundation (REC award No. 0126064) to Kikumi Tatsuoka and James Corter. Portions of this research were presented at the annual meeting of the National Council for Measurement in Education, April 2004, San Diego, CA. We also wish to acknowledge the contributions of our research team: Annabelle Guerrero, Michael Dean, Enis Dogan, Jennifer Grossman, Seongah Im, Toshihiko Matsuka, Eunkyung Um, Tao Xin, and Tomoko Yamada.

References

Buck, G., & Tatsuoka, K. K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing, 15*(2), 119–157.

Buck, G., Tatsuoka, K. K., & Kostin I. (1997). The skills of reading: Rule-space analysis of a multiple-choice test of second language reading comprehension. *Language Learning, 47,* 423–466.

Carpenter, T. P., & Moser, J. M. (1982). The development of addition and subtraction problem solving skills. In T. P Carpenter, J. M. Moser, & T. A. Romberg (Eds.), *Addition and subtraction: A cognitive perspective* (pp. 9–24). Hillsdale, NJ: Erlbaum.

Chen, C., & Stevenson, H. W. (1995). Motivation and mathematics achievement: A comparative study of Asian-American, Caucasian-American, and East Asian high school students. *Child Development, 66,* 1215–1234.

Corter, J. E. (1998). An efficient metric combinatorial algorithm for fitting additive trees. *Multivariate Behavioral Research, 33,* 249–272.

Corter, J. E., & Tatsuoka, K. K. (2002). *Diagnostic Assessments for Mathematics Tests Grades 6–12.* Technical Report, College Board.

Davis, R. B. (1992). Understanding "understanding." *Journal for Research in Mathematics Education, 22,* 362–365.

Fukunaga, K. (1990). *Introduction to statistical pattern recognition* (2nd ed.). Boston: Academic Press.

Glas, C. A. W., & Meijer, R. R. (2003). A Bayesian approach to person-fit analysis in item response theory models. *Applied Psychological Measurement, 27,* 217–233.

Greeno, J. G. (1991). Number sense as situated knowing in a conceptual domain. *Journal for Research in Mathematics Education, 22,* 170–218.

Haertel, E. H., & Wiley, D. E. (1993). Representations of ability structures: Implications for testing. In N. Frederiksen, R. J. Mislevy, & I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 359–384). Hillsdale, NJ: Erlbaum.

Hiebert, J., & Stigler, J. W. (2000). A proposal for improving classroom teaching: Lessons from the TIMSS video study. *The Elementary School Journal, 101,* 3–20.

Hogg, R. V., & Craig, A. T. (1978). *Introduction to mathematical statistics.* New York: Macmillan.

Jakwerth, P. (2004). *Splintered vision: An investigation of US science and mathematics education.* Boston: Kluwer.

Jones, B. F., & Idol, L. (1990). *Dimensions of thinking: Review of research*. Hillsdale, NJ: Erlbaum.

Kawanaka, T., & Stigler, J. W. (1999). Teachers' use of questions in eighth-grade mathematics classrooms in Germany, Japan, and the United States. *Mathematical Thinking and Learning, 1,* 255–278.

Macnab, D. S. (1999). *Implementing change in mathematics education*. Paper presented at the 1999 Annual Conference of the Scottish Educational Research Association, Northern College, Aberdeen.

Macnab, D. S. (2000). Raising standards in mathematics education: Values, vision, and TIMSS. *Educational Studies in Mathematics, 42,* 61–80.

Marzano, R. J., Brandt, R. S., Highs, C. S., Jones, B. F., Presseissen B. Z., Rankin S. C., & Suhor, C. (1988). *Dimensions of thinking: A framework for curriculum and instruction*. Alexandria, VA: Association of Supervision and Curriculum Development.

Mullis, I. V. S., Martin, M. O., Gonzales, E. J., Gregory, K. D., Garden, R. A., O'Connor, K. M., Chrostowski, S. J., & Smith, T. A. (2000). *TIMSS 1999 International Mathematics Report*. Chestnut Hill, MA: International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Gonzales, E. J., O'Connor, K. M., Chrostowski, S. J., Gregory, K. D., Garden, R. A., & Smith, T. A. (2001). *Mathematics Benchmarking Report—TIMSS 1999 Eighth Grade*. Chestnut Hill, MA: International Study Center, Boston College.

Pascarella, E. T., & Terenzini, P. T. (1991). *How college affects students*. San Francisco, CA: Jossey-Bass.

Polya, G. (1954). *Mathematics and plausible reasononing I. Induction and analogy in mathematics II. Patterns of plausible inference*. Princeton, NJ: Princeton University Press.

Pressley, M. (1995). *Cognition, teaching and assessment*. New York: HarperCollins.

Robitaille, D. F. (1997). *National contexts for mathematics and science education: An encyclopedia of the educational systems participating in TIMSS*. Vancouver, Canada: Pacific Educational Press.

Schmidt, W. H., McKnight, C. C., Cogan, L. S., Jakwerth, P. M., Houang, R. T., Wiley, D. E., Wolfe, R. G., Bianchi, L. J., Vaverde, G. A., Raizen, S. A., & Demars, C. E. (1999). *Facing the consequences: Using TIMSS for a closer look at U.S. mathematics and science education*. Boston: Kluwer.

Schneider, W. A., & Graham, D. J. (1992). Introduction to connectionist modeling in education. *Educational Psychologist, 27,* 513–530.

Schumer, G. (1999). Mathematics education in Japan. *Journal of Curriculum Studies, 31,* 399–427.

Stigler, J. W., Gonzales, P., Kawanaka, T., Knoll, & Serrano, A. (1999). *The TIMSS videotape classroom study: Methods and findings from an exploratory research project on eighth grade mathematics instruction in Germany, Japan, and the United States* (NCES 1999-074). Washington, DC: National Center for Education Statistics.

Stout, W. (2002). Psychometrics: From practice to theory and back: 15 years of nonparametric multidimensional IRT, DIF/test equity, and skills diagnostic assessment. *Psychometrika, 67,* 485–518.

Tatsuoka, C. (2002). Data analytic methods for latent partially ordered classification models: Applied statistics. *Journal of the Royal Statistical Society Series C, 51,* 337–350.

Tatsuoka, C., & Ferguson, T. (1999). *Sequential classification on partially ordered sets* (Tech. Rep. No. 99-05). Washington, DC: Department of Statistics, George Washington University.

Tatsuoka, C., Varadi, F., & Tatsuoka, K. K. (1992). BUGSHELL [computer software]. Ewing, NJ: Tanar Software.

Tatsuoka, K. K. (1983). Rule-space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20,* 34–38.

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. *Journal of Educational Statistics, 1,* 55–73.

Tatsuoka, K. K. (1990). Toward an integration of item response theory and cognitive analysis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. C. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 543–588). Hillsdale, NJ: Erlbaum.

Tatsuoka, K. K. (1991). *Boolean algebra applied to determination of the universal set of knowledge states* (Research Rep. No. 91-2-ONR). Princeton, NJ: Educational Testing Service.

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern classification approach. In P. Nichol, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327–360). Hillsdale, NJ: Erlbaum.

Tatsuoka, K. K. (1997). Use of generalized person-fit indices for statistical pattern classification. An invited paper for a special issue of person-fit statistics. *Journal of Applied Educational Measurement, 9,* 65–75.

Tatsuoka, K. K. (in press). *Statistical pattern recognition and classification of latent knowledge states: Cognitively diagnostic assessment.* Mahwah, NJ: Erlbaum.

Tatsuoka, K. K., & Boodoo, G. M. (2000). Subgroup differences on the GRE quantitative test based on the underlying cognitive processes and knowledge. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 821–857). Mahwah, NJ: Erlbaum.

Tatsuoka, K. K., Corter, J. E., & Guerrero, A. (2004). *Coding manual for identifying involvement of content, skill, and process subskills for the TIMSS-R 8th grade and 12th grade general mathematics test items* (Tech. Rep.). New York: Department of Human Development, Teachers College, Columbia University.

Tatsuoka, K. K., Linn, R. L., Tatsuoka, M. M., & Yamamoto, K. (1988). Differential item functioning resulting from the use of different solution strategies. *Journal of Educational Measurement, 25,* 301–319.

Tatsuoka, M. M., & Tatsuoka, K. K. (1989). Rule space. In S. Kotz and N. L. Johnson (Eds.), *Encyclopedia of statistical sciences.* New York: Wiley.

Tatsuoka, K. K., & Tatsuoka, M. M. (1987). Bug distribution and pattern classification. *Psychometrika, 52,* 193–206.

Tatsuoka, K. K., & Tatsuoka, M. M. (1997). Computerized adaptive diagnostic testing. *Journal of Educational Measurement, 34,* 3–20.

VanLehn, K. (1982). Bugs are not enough: Empirical studies of bugs, impasses and repairs in procedural skills. *Journal of Mathematical Behavior, 3,* 3–71.

Yan, D., Mislevy, R. J., & Almond, R. G. (2003). *Design and analysis in cognitive assessment* (ETS Research Rep. No. ETS RR-03-32). Princeton, NJ: Educational Testing Service.

Yepes-Baraya, M., & Allen, N. (2002). *An attribute-based study to obtain diagnostic assessment information on the attainment of cognitive dimensions relevant in science education.* Paper presented at the 2003 AERA meeting, Chicago, IL.

Manuscript received July 31, 2004
Revision received August 1, 2004
Accepted August 16, 2004