ELSEVIER

Studies in Educational Evaluation 30 (2004) 151-173

Studies in Educational Evaluation

www.elsevier.com/stueduc

DIAGNOSTIC ASSESSMENT IN TIMSS-R: BETWEEN-COUNTRIES AND WITHIN-COUNTRY COMPARISONS OF EIGHTH GRADERS' MATHEMATICS PERFORMANCE[1]

Menucha Birenbaum*, Curtis Tatsuoka and Tomoko Yamada*****

**Tel Aviv University, Israel*
***George Washington University, USA*
****Columbia University, USA*

The importance and relevance of mathematics knowledge to every day life in the 21st century has been widely acknowledged. Moreover, from an economical standpoint it is universally recognized that a country whose students excel in mathematics is more likely to establish its competitive advantage in an increasingly global economy. Education systems in many countries are therefore concerned about their students' performance in international tests and are seeking ways to improve their mathematics instruction (Wagemaker, 2002).

International studies of mathematics achievement such as the Third International Mathematics and Science Study (TIMSS-R) and its predecessors, FIMS, SIMS, and TIMSS, have reported internationally comparable achievement profiles for over four decades (Wagemaker, 2002). Information regarding the attained curricula from an international perspective provides a "calibrated yardstick" which, if carefully and thoughtfully used, could support a country's education system in its effort to prepare an internationally competitive workforce.

Attempts to explain differences in mathematics achievement from an international perspective address the effectiveness of education systems as well as the confounding effects of the cultural contexts in which they operate (Martin, Mullis, Gregory, Hoyle, & Shen, 2000; Ramakrishnan, 2000). Yet, there is a portion of score variance in international tests that cannot be accounted for by cultural effects, school effectiveness, and their interactions. That portion is due, at least partially, to the validity of the international test itself; that is, to the correspondence between the test and the intended curricula in the

0191-491X/04/$ - see front matter © 2004 Published by Elsevier Science Ltd.
doi: 10.1016/j.stueduc.2004.06.004

participating countries. Indeed, in view of existing differences in intended curricula among countries, the validity of international comparisons such as TIMSS has frequently been questioned (Bechger, van-Schooten, De Glopper, & Hox-Joop, 1998; Jaeger, 1994; Jaeger & Hattie, 1996). Less susceptible to this threat to validity are within-country comparisons of subpopulations of national importance, due to the fact that they share the same intended curriculum.

The current study combines between-countries and within-country comparisons. First, a comparison of mathematics achievement is made among three countries – the U.S, Japan, and Israel – which differ in culture and education system. Second, a comparison is made within one of these countries, Israel, between two culturally diverse subpopulations – Jews and Arabs. The reasons for selecting the three particular countries are twofold: the U.S. and Japan are the major world economic superpowers that are in constant competition, and therefore it is of great interest to compare their students' mathematics achievement. Israel provides an interesting context for within-country comparisons, as it comprises two culturally diverse populations with little inter-group contact; Jews, who maintain a Western lifestyle, and Arabs, who maintain a traditional patriarchal way of life. The other reason for selecting these particular countries is the familiarity of the researchers with the cultures and the education systems they investigate. In the case of the current study, each of the authors has a first-hand acquaintance with the culture and the education system in one of these countries, which are their respective home countries.

Diagnosing individual performance in large-scale assessment is not a common practice. Results of national and international tests provide interpretations to test scores (i.e., scale scores), where all test takers who get the same scale score or are within a pre-specified range of the total score distribution receive the same interpretation. For instance, the diagnostic approach reported in TIMSS allows for diagnostic feedback at four benchmarks, set at the 90th, 75th, 50th, and 25th percentiles of the international score distribution. Performance of students whose scores were around these percentiles was examined in terms of the educational requirements for solving anchored items that at least 65% of the students in a given such group successfully answered and more than 50% of the lower benchmark group failed to answer correctly. The mastery profile for that benchmark is specified in terms of skills judged by experts to be necessary for successfully solving those particular items (Mullis et al., 2001). However, students in the same quartile can widely vary in their response pattern to the test items, and therefore a diagnosis on an individual level is much to be desired. Tatsuoka (1983, in press) has developed a methodology for such assessment, which she termed "Rule Space". Following is a brief account of this methodology.

Rule-Space Methodology

Rule-space (RS) is a probabilistic model for cognitive diagnosis that employs pattern analysis to classify students' item responses on a test according to their profile of strengths and weaknesses on the underlying constructs termed *attributes,* and which are measured by the test. An attribute is a description of a procedure, skill, or content knowledge that a student must possess in order to successfully complete the target task. Performing an RS analysis involves five phases:

1. *Defining attributes*: domain experts define the attributes of the target task that are of interest and write/select a set of items that tap this set of attributes.
2. *Assigning attributes to items*: an item-by-attribute incidence matrix (referred to as *Q matrix* in RS) is created where every column represents an attribute and every row an item. For every item, 1s are assigned to attributes whose mastery is required for answering that item correctly and 0s otherwise. These item-by-attribute involvement relationships are essential for the success of the classification process, as they specify the hypothesized underlying constructs being measured by the test.
3. *Determining identifiable knowledge states*: Actual mastery or non-mastery of a set of attributes cannot be measured directly and therefore must be inferred from the student's pattern of responses to the set of items. In an ideal case, a student who has mastered some, but not other, attributes would answer correctly those items that require only the attributes that s/he has mastered and answer incorrectly those items that require at least one attribute that s/he has not mastered. Such a student would produce an *ideal item-score pattern*. This ideal item-score pattern can be expressed in terms of an attribute pattern so that every item that is correctly answered (denoted as 1 in the ideal item-score pattern) is expressed in terms of the attributes required for its successful completion (denoted as 1s in the Q matrix with respect to that item.) There are thus two representations of a *knowledge state*, one in the item space and the other in the attribute space. If only one attribute is involved with each item, then the number of knowledge states is equal to the number of attributes; however, this is rarely the case. In most cases, several attributes are involved with each item. Consequently, the number of possible knowledge states can become very large (the maximum being 2^k, where k is the number of attributes). However, not all of the knowledge states are relevant, given the involvement relationships in the Q matrix. In order to reduce the number of possible knowledge states to the relevant ones and to map them into ideal item-score patterns, RS uses the degenerative properties of Boolean algebra (Tatsuoka, 1991).
4. *Formulating the classification space*: a multidimensional classification space is formulated with respect to various dimensions: θ (theta), ζ (zeta), and generalized ζs (Tatsuoka, 1997). *Theta* is the ability continuum derived from item-response theory (IRT) (Lord & Novick, 1968). *Zeta* is a measure of "unusualness of response". The higher the absolute value on this dimension, the less common the respective item-response pattern (Tatsuoka, 1984; Tatsuoka & Linn, 1983). *Generalized ζs* have been introduced in order to have orthogonal coordinates in a multidimensional RS; while _ measures the unusualness of *n* item-score patterns, generalized ζs measure the unusualness of item-score patterns in subsets of *n* items (Tatsuoka, 1997). In this multidimensional space certain points represent the predetermined knowledge states. However, students' performance on test items is often subject to fluctuations; therefore, an observed item-response pattern that corresponds to a knowledge state is likely to be rare. Students' item-response patterns that deviate from a knowledge state are considered as "fuzzy" response patterns. Points corresponding to the *fuzzy response patterns* swarm around their respective knowledge state and generate regions within probability ellipses with the ideal item-score pattern that corresponds to a knowledge state as their center. A 90% probability ellipse encloses 90% of the

fuzzy-response-pattern points; a 95% probability ellipse encloses 95% of them, and so forth (Tatsuoka & Tatsuoka, 1987).

5. *Classifying examinees' responses:* In this phase RS classifies students' fuzzy response patterns into the closest ellipse by measuring how far from the centroid the student's point is, in terms of squared Mahalanobis distance (D^2). Bayes' decision rules for minimizing errors are used to classify a student into one of the predetermined knowledge states. The probability of misclassification and the posterior probability of the student's response pattern coming from the group (knowledge state) under which it was classified are computed. Once the most likely knowledge state for a particular student is identified, the most conservative attribute-mastery pattern for that ideal item-score pattern is assigned by RS to that student and his/her probabilities of mastering each attribute are listed. This diagnosis is expected to spur a remedial strategy that would be most likely to target the student's weaknesses in the domain tested.

RS has been shown to perform quite well in various areas such as subtraction of fractions (Tatsuoka & Tatsuoka, 1992), signed numbers operations (Tatsuoka, 1990), algebra (Birenbaum, Kelly, & Tatsuoka, 1993), the quantitative parts of the Scholastic Aptitude Test (SAT-M; Tatsuoka, Birenbaum, Lewis, & Sheehan, 1993), and the Graduate Record Examination (GRE; Tatsuoka & Boodoo, 2000), as well as in listening comprehension (Buck & Tatsuoka, 1998). Although the RS methodology has already been successfully applied in quite a few studies of mathematics performance, comparisons of group performances using this methodology are sparse (Tatsuoka & Boodoo, 2000).

Applying the RS methodology for analyzing international test data sets seems to hold great potential for a significant contribution to the teaching and learning of the tested subject. The current study applied this methodology to examine between-countries and within-country differences in 8th grade mathematics knowledge.

Method

Participants

The study comprised three samples of 8th graders who participated in the 1999 TIMSS-R study, consisting of 4411 students from the U.S., 2371 from Japan, and 2092 from Israel (1684 Jews and 408 Arabs).[2]

Instruments

The 1999 TIMMS-R mathematics test included a pool of 162 items and was assembled in eight booklets, each requiring 90 minutes to complete. Classified by content, 38% of the items addressed fractions and number sense; 15% measurement; 13% data presentation, analysis, and probability; 13% geometry; and 22% algebra. Classified by format, about 25% of the items were open-ended and the rest were of the choice-response format (Gonzalez & Miles, 2001). Only four booklets (1, 3, 5, and 7) were used in the current study; the other four booklets (2, 4, 6, and 8) had few or no items measuring certain attributes and were therefore eliminated from the analyses.

Analysis

The set of attributes used in this study was developed by Tatsuoka and her associates for analyzing TIMSS-R-1999 mathematics items for 8th graders (Tatsuoka, Corter, & Guerrero, 2003). They grouped the attributes into three clusters of content (5 attributes), processes (9 attributes), and skill/item-type (9 attributes). *Content attributes* refer to basic concepts and properties in whole numbers and integers; fractions and decimals; elementary algebra; two-dimensional geometry, data and basic statistics. *Process attributes* include attributes such as: judgmental applications of knowledge in arithmetic and geometry; rule application in algebra; logical reasoning; problem search; generating, visualizing and reading figures and graphs; managing of data and procedures. *Skill (item-type) attribute* include attributes such as: applying number properties and relationships (number sense); approximation/estimation; recognizing patterns and sequences; solving open-ended items. The full list of the attributes used in the current study appears in the Appendix.

The test items for each test booklet were coded according to the set of 23 attributes. For data analysis, the BILOG-MG program (Zimowski, Muraki, Mislevy, & Bock, 1996) was used to estimate the IRT *a* and *b* parameters for the items and the BUGLIB program (Tatsuoka, Varadi, & Tatsuoka, 1992) was used for the RS analysis.

Following the computations of students' attribute mastery probabilities, the mean probabilities for the three countries – the U.S., Japan, and Israel – were compared. Next, clusters of hierarchically related latent knowledge states were identified and the relative proportions of students from the three countries in each cluster were computed. Similarly, within the Israeli sample, comparisons were made between the Jewish and Arab subpopulations.

Results

The adequacy of the Q matrix was assessed by predicting the total test score by the attribute probabilities. The squared multiple correlations (R^2 and adjusted R^2) for the entire sample and for each of the three countries are presented in Table 1. All values in the table are considered satisfactory (Tatsuoka, in press).

Table 1: R^2 (R^2 adjusted) for Predicting the Average Test Score from the Attribute Probabilities

Sample	Booklet 1	Booklet 3	Booklet 5	Booklet 7	4 Booklets
Israel (*n* = 2092)	.83 (.82)	.94 (.94)	.97 (.96)	.98 (.98)	.90 (.90)
Japan (*n* = 2371)	.97 (.97)	.97 (.97)	.97 (.97)	.98 (.98)	.95 (.95)
USA (*n* = 4411)	.82 (.81)	.95 (.95)	.96 (.96)	.98 (.98)	.89 (.89)
Entire Sample (*N* = 8874)	.86 (.86)	.95 (.95)	.96 (.96)	.97 (.97)	.90 (.90)

Another measure of the adequacy of the Q matrix is the overall rate of classification by the RS. This value indicates the percentage of students' response patterns that were located within the 95% probability ellipse of a latent knowledge state. The rates of classification for the American, Japanese, and Israeli samples in the current study were 99.7%, 99.1%, and 99.6%, respectively.

Comparisons Between the U.S., Japan, and Israel

A. Test Score Distributions

Before presenting the results of the RS analyses, statistics regarding the test score distribution are summarized. Figure 1 displays box plots of percentage correct responses on the test for the three countries. As can be seen in the figure, Japan has the highest median (77.78) and the smallest dispersion of scores (Q3 - Q1 = 26.20). The U.S. has a higher median than Israel (52.38 compared to 46.15) and similar dispersion of scores (Q3 - Q1 = 35.72 and 35.14, respectively). One-way ANOVA yielded a significant effect of country on the test scores ($F_{2,\ 8821}$ = 914.41; p < .0001). The respective means (and standard deviations) for the U.S., Japan, and Israel are 53.78 (22.00); 73.18 (18.88); and 48.34 (21.80). All means are significantly different (p < .0001) from each other according to Scheffé's procedure.

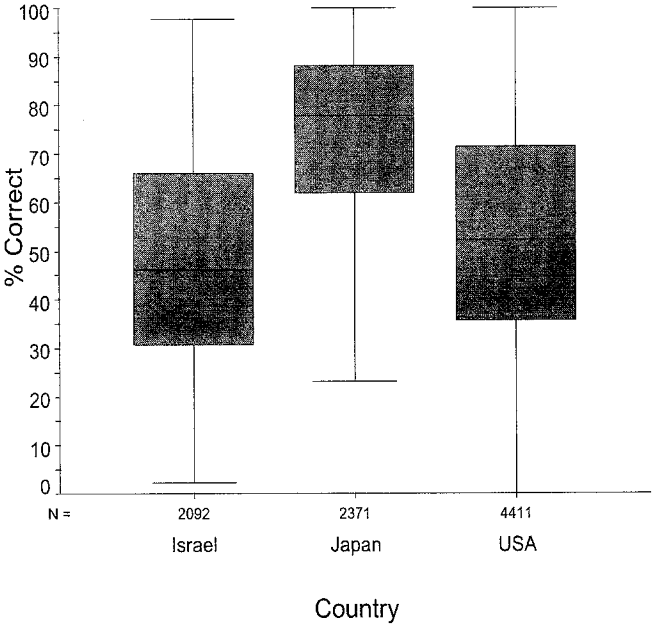

Figure 1: Box Plots of Percentage Correct Responses on the Test for Israel, Japan, and the USA

B. *Attribute Mastery Probabilities*

The results of one-way ANOVAs for the effect of country on the attribute mastery probabilities are presented in Table 2.

Table 2: One-Way ANOVA Results for the Effect of Country on the Attribute Mastery Probabilities

Attribute	Country	Mean	SD	F	Multiple Comparisons(1)
C1: whole numbers and integers	Israel	.91	.19	10.58*	J > I; U > I
	Japan	.93	.17		
	USA	.93	.18		
C2: Fractions and decimals	Israel	.84	.32	10.58*	J > U > I
	Japan	.97	.13		
	USA	.86	.32		
C3: Algebra	Israel	.71	.27	143.83*	J > U >I
	Japan	.84	.20		
	USA	.76	.25		
C4: Geometry	Israel	.78	.25	190.33*	J > I > U
	Japan	.96	.14		
	USA	.68	.30		
C5: Data & statistics	Israel	.67	.27	946.14*	J > U > I
	Japan	.81	.21		
	USA	.73	.26		
P1: Translate	Israel	.94	.14	12.09*	U > J; U > I
	Japan	.94	.14		
	USA	.96	.13		
P2: Computation application	Israel	.88	.21	149.37*	J > U > I
	Japan	.96	.12		
	USA	.92	.16		
P3: Judgmental applications	Israel	.90	.16	632.31*	J > I > U
	Japan	.98	.07		
	USA	.85	.17		
P4: Rule application in algebra	Israel	.53	.31	163.72*	J > U > I
	Japan	.69	.28		
	USA	.59	.29		
P5: Logical reasoning	Israel	.66	.25	644.89*	J > I ; J > U
	Japan	.88	.19		
	USA	.65	.30		
P6: Problem search	Israel	.80	.25	342.35*	J > U > I
	Japan	.95	.12		
	USA	.83	.22		
P7: Visualize/ Fig. & Graph	Israel	.72	.25	431.30*	J > U > I
	Japan	.90	.18		
	USA	.77	.23		
P9: Data management	Israel	.64	.30	840.57*	J > U > I
	Japan	.93	.16		
	USA	.68	.30		
P10: Quantitative reading	Israel	.80	.24	80.43*	U > J > I
	Japan	.84	.20		
	USA	.87	.20		

/Cont.

Table 2/cont.

S2: Number sense	Israel	.74	.28	155.07*	J > U > I
	Japan	.81	.18		
	USA	.78	.24		
S3: Figures tables & graphs	Israel	.93	.17	58.45*	J > U > I
	Japan	.99	.03		
	USA	.95	.14		
S4: Approximation & estimation	Israel	.76	.25	178.75*	U > J > I
	Japan	.86	.18		
	USA	.88	.21		
S5: Evaluate / verify options	Israel	.95	.15	243.08*	J > U > I
	Japan	.99	.04		
	USA	.97	.11		
S6: Recognize patterns	Israel	.46	.38	107.38*	J > I ; J > U
	Japan	.72	.22		
	USA	.46	.33		
S7: Proportional reasoning	Israel	.94	.15	415.45*	J > U ; I > U
	Japan	.94	.16		
	USA	.91	.20		
S8: Unfamiliar problems	Israel	.85	.23	58.43*	J > U > I
	Japan	.91	.17		
	USA	.87	.22		
S10: Open-ended items	Israel	.54	.40	521.47*	J > U > I
	Japan	.85	.23		
	USA	.60	.38		
S11: Wordy problems	Israel	.87	.23	12.19*	J > U ; J > I
	Japan	.90	.24		
	USA	.88	.24		

$p < .0001$
$df = 2, 8871$
(1)Scheffé test
J=Japan n= 2371) ; I= Israel (n= 2090) ; U=USA (n=4411)

The table includes mean probabilities and standard deviations for each country on the 23 attributes along with *F* values and results of multiple comparisons using the Scheffé procedure. As can be seen in the table, all attributes yielded significant effects. Japan had the highest probabilities on 20 of the 23 attributes and the U.S. had the highest mean on three attributes – Translation (P1), Quantitative reading (P10), and Approximation and estimation; (S4). The U.S. also had the lowest mean probabilities on three attributes: Geometry (C4), Judgmental applications (P3), and Proportional reasoning (S7). On three attributes, the U.S. and Israel had similar means – Logical thinking (P5), Pattern recognition (S6), and Word problems (S11). On two attributes Israel and Japan had similar means – Translate (P1) and Proportional reasoning (P7) – and on the remaining 16 attributes Israel had the lowest mean probabilities. Setting the cut-off point for mastery probability at 0.70 indicates that the average student in the Japanese sample has mastered 22 attributes and has failed to master one attribute – Rule application in Algebra (P4); the average American student and his/her Israeli counterpart have each mastered 17 attributes and has failed to reach mastery on the six remaining attributes (the one missed by their Japanese counterpart – Rule application in algebra [P4] – as well as Logical reasoning [P5], Data management [P9], Pattern recognition [S6], and Open-ended questions [S10]). The sixth attribute the average American student failed to master was Geometry (C4), whereas the sixth attribute missed by the average Israeli student was Data and statistics (C5).

C. *Latent Knowledge States*

Clusters of hierarchically related latent knowledge states were derived from a cluster analysis of students' attribute mastery probability patterns in the combined sample. An eight-cluster solution is presented in Table 3, and a map of the transitional relations among the clusters is presented in Figure 2.

Table 3: Cluster Analysis Results (Combined Sample *N*=8874)

Attribute	Cluster Centers							
	1	2	3	4	5	6	7	8
C1: whole numbers and integers	.58	.98	.56	.93	.94	.95	.96	1.00
C2: Fractions and decimals	.46	1.00	.91	.99	.24	.98	.97	.07
C3: Algebra	.45	.89	.53	.72	.72	.87	.54	.70
C4: Geometry	.61	.93	.52	.89	.64	.71	.33	.57
C5: Data & statistics	.52	.83	.66	.73	.85	.83	.43	.37
P1: Translate	.62	.99	.82	.97	.92	.98	.98	.90
P2: Computation application	.63	.99	.66	.88	.85	.95	.97	.99
P3: Judgmental applications	.65	.94	.98	.94	.68	.88	.87	.75
P4: Rule application in algebra	.26	.80	.49	.35	.26	.78	.47	.39
P5: Logical reasoning	.57	.89	.80	.67	.56	.63	.26	.47
P6: Problem search	.42	.97	.87	.78	.55	.84	.84	.95
P7: Visualize/ Fig. & Graph	.49	.95	.64	.76	.57	.85	.55	.54
P9: Data management	.44	.94	.47	.78	.68	.52	.43	.54
P10: Quantitative reading	.58	.90	.66	.82	.76	.91	.80	.89
S2: Number sense	.51	.84	.91	.74	.39	.81	.81	.72
S3: Figures tables & graphs	.81	1.00	.92	.99	.84	.98	.88	.81
S4: Approximation & estimation	.45	.89	.96	.75	.63	.93	.88	.98
S5: Evaluate / verify options	.75	1.00	.96	.99	.88	.98	.93	1.00
S6: Recognize patterns	.17	.79	.23	.61	.15	.14	.35	.05
S7: Proportional reasoning	.58	.98	.80	.97	.92	.95	.95	.68
S8: Unfamiliar problems	.49	.93	.96	.91	.76	.97	.70	.60
S10: Open-ended items	.38	.95	.29	.79	.22	.28	.40	.10
S11: Wordy problems	.61	.98	.51	.95	.62	.97	.90	.56
Number of students	369	3679	583	1563	498	1136	701	345
Ratio of count to expected cont								
Israel	1.78	.63	1.43	1.17	1.58	1.22	.91	1.29
Japan	.21	1.74	.94	.87	.08	.31	.02	.23
USA	1.05	.78	.84	.99	1.22	1.27	1.57	1.28
Within Israel:								
Jewish students	.66	1.18	.81	1.11	.75	1.06	.94	.83
Arab students	2.42	.26	1.77	.56	2.01	.75	1.23	1.71

A transition from one cluster of latent knowledge states to another is said to be possible whenever the set of mastered attributes associated with the lower cluster is a proper subset of the higher connected cluster. Attributes yielding a coefficient of 0.75 or larger were considered meaningful for defining a cluster center of latent knowledge states in terms of mastery. Those are the attributes that appear in Figure 2. The numbers of students included in each cluster are presented in Table 3 along with a ratio that indicates the proportion of students from the U.S., Japan, and Israel in each cluster, computed as the ratio of count to expected count based on the marginal distributions.

As can be seen in Table 3 and Figure 2, the cluster that comprised the lowest number of mastered attributes is Cluster 1. The two attributes included in this cluster are Figures, tables, and charts (S3) and Evaluate, verify, and check options (S5). The average score on the test (in term of percentage correct answers) for students in this cluster is 18.53. Of the 369 students grouped in this cluster, the number of Japanese students is only about a fifth of their expected number, whereas the number of Israeli students is 1.78 times larger than expected, and that of the U.S. students is almost as expected. Next in hierarchy is Cluster 8, comprising nine attributes, including the ones of the lower connected cluster. The additional attributes in this cluster are Whole numbers and integers (C1), Translate (P1), Computation application (P2), Judgmental Application (P3), Problem search (P6),

Quantitative reading (P10), and Approximation and estimation (S4). The average score on the test for students in this cluster is 31.14, and the ratios of U.S., Japanese, and Israeli students are 1.28, 0.23 and 1.29, respectively. Cluster 5 is also connected to Cluster 1. It is defined by mastery of nine attributes, the additional ones being two content attributes – Whole numbers and integers (C1) and Data & statistics (C5); three process attributes – Translate (P1), Computation application (P2), and Quantitative reading (P10); and two skill/item-type attributes—Proportional Reasoning (S7), and Unfamiliar problems (S8). The average score on the test for students in this cluster is 25.82, and the ratios of U.S., Japanese, and Israeli students are 1.22, 0.08, and 1.58, respectively. Cluster 3 is also connected to Cluster 1 and is defined by mastery of 11 attributes, the additional ones being: a content attribute – Fractions and decimals (C2); four process attributes – Translate (P1), Judgmental applications (P3), Logical reasoning (P5), and Problem search (P6); and four skill attributes – Number sense (S2), Approximation and estimation (S4), Proportional reasoning (S7), and Unfamiliar problems (S8). The average score on the test for students in this cluster is 32.31, and the ratios of U.S., Japanese, and Israeli students are 0.84, 0.94, and 1.43, respectively. Next in hierarchy appear Clusters 7 and 4, each comprising the attributes of Cluster 8 and a few additional ones. Cluster 7 comprises 13 attributes, the additional ones being Fractions and decimals (C2), Number sense (S2), Proportional reasoning (S7), and Word problems (S11). The average score on the test for students in this cluster is 40.97, and the ratios of U.S., Japanese, and Israeli students are 1.57, 0.02, and 0.91, respectively. Cluster 4 comprises 17 attributes, the additional ones being Fractions and decimals (C2), Geometry (C4), Visualize figures and graphs (P7), Data management (P9), Proportional reasoning (S7), Open-ended items (S10), and Word problems (S11). The average score on the test for students in this cluster is 53.86, and the ratios of U.S, Japanese, and Israeli students are 0.99, 0.87, and 1.17, respectively. Cluster 6 is connected to Clusters 7 and 5, and comprises 18 attributes, the additional ones being Algebra (C3), Rule application in algebra (P4), and Visualize figures and graphs (P7). The average score on the test for students in this cluster is 50.15, and the ratios of U.S., Japanese, and Israeli students are 1.27, 0.31, and 1.22, respectively. Cluster 2 is the highest one. It is connected to Clusters 3, 4, and 6, and comprises all 23 attributes. The additional attribute in this cluster is Pattern recognition (S6). The average score on the test for the 3679 students in this cluster is 79.56. Students from the U.S. and Israel are underrepresented in this cluster, as they constitute only 0.78 and 0.63, respectively, of their expected count, whereas Japanese students are overrepresented by 1.74 of their expected count.

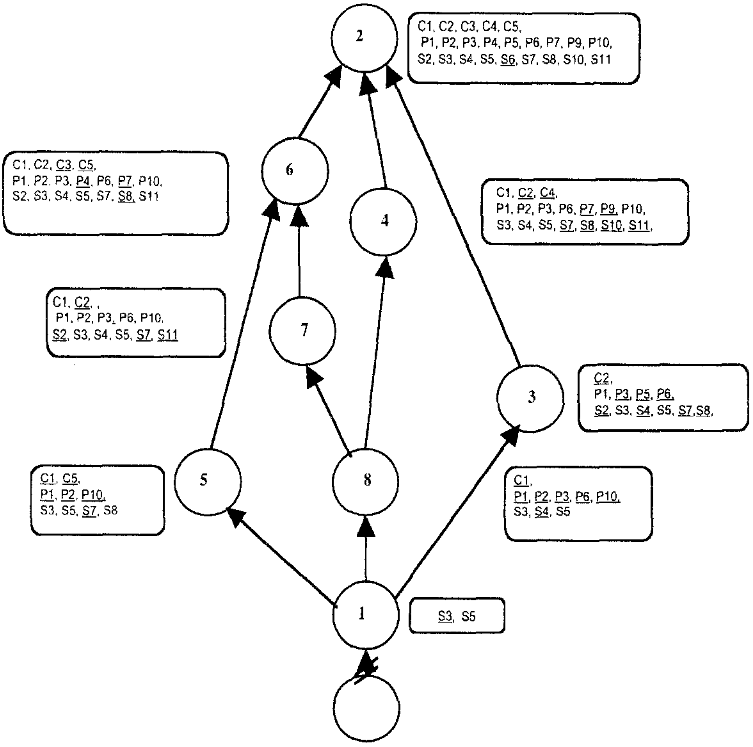

Figure 2: A Map of Transitional Relations among Clusters of Latent Knowledge States (N=8874)

Comparisons Between Jews and Arabs in Israel

A. Test Score Distribution

Figure 3 displays box plots for the Jewish and Arab groups. As can be seen in the figure, the median score for the Arab group (28.89) is considerably lower, and the score dispersion (Q3 - Q1 = 21.03) is considerably smaller than in the Jewish group (50.00 and 34.95, respectively). A t-test for independent samples indicated a significant difference in favor of the Jewish group (t_{766} = 20.05; p <.001) with an effect size of about one standard deviations (d = 0.96).

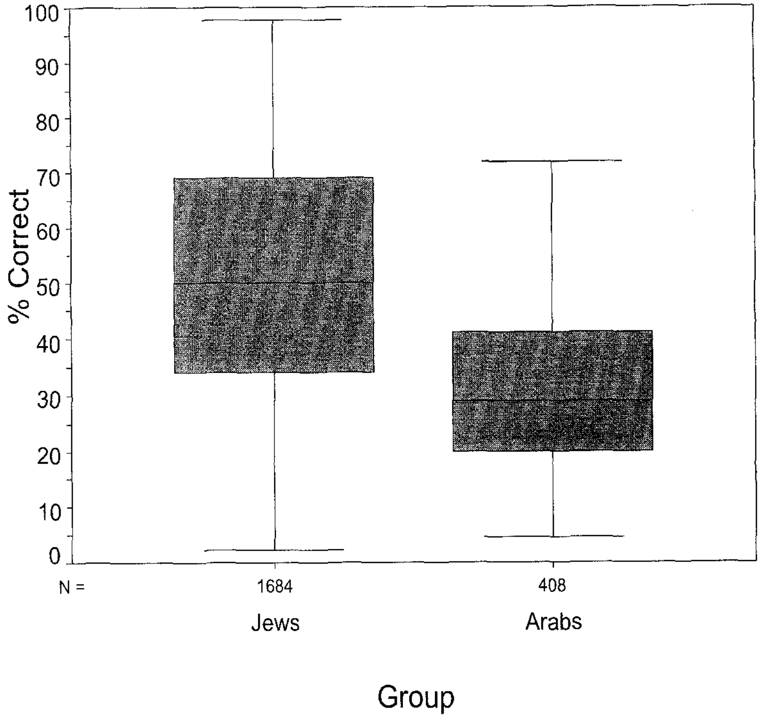

Figure 3: Box Plots for the Jewish and Arab Groups on the Total Score

B. Attribute Mastery Probabilities

Table 4 presents *t*-test results for the Jewish and Arab groups with respect to each attribute. As can be seen in the table, all mean differences are significant, with effect sizes ranging from 0.28 to 0.72 standard deviations with a mean *SD* of 0.53. A mastery cut-off point of 0.70 indicates that the average Arab student, as compared to his/her Jewish counterpart, has not mastered six attributes: Fractions and decimals (C2), Algebra (C3), Geometry (C4), Visualize figures and graphs (P7), Number sense (S2), and Approximation and estimation (S4). Yet, the average Jewish and Arab students both have failed to master six attributes: Data and statistics (C5), Rule application in algebra (P4), Logical reasoning (P5), Data management (P9), Pattern recognition (S6), and Open-ended items (S10).

Table 4: Means and *SD* for Jewish (n=1684) and Arab (n=408) 8th Graders on 23 Attributes, Effect Size (d) Values

Attribute	Group	Means	*SD*	*t*-Value	*d*-value
C1: Whole numbers and integers	Jews	.93	.17	7.67*	.54
	Arabs	.83	.24		
C2: Fractions and decimals	Jews	.88	.28	10.53*	.72
	Arabs	.66	.41		
C3: Algebra	Jews	.74	.26	9.10*	.60
	Arabs	.58	.29		
C4: Geometry	Jews	.81	.23	11.37*	.67
	Arabs	.65	.27		
C5: Data & statistics	Jews	.69	.26	5.82*	.34
	Arabs	.60	.29		
P1: Translate	Jews	.95	.13	7.83*	.50
	Arabs	.88	.18		
P2: Computation application	Jews	.90	.19	7.79*	.54
	Arabs	.79	.26		
P3: Judgmental applications	Jews	.92	.14	7.81*	.59
	Arabs	.83	.21		
P4: Rule application in algebra	Jews	.56	.31	10.06*	.55
	Arabs	.39	.30		
P5: Logical reasoning	Jews	.67	.25	4.97*	.28
	Arabs	.60	.24		
P6: Problem search	Jews	.82	.23	7.90*	.50
	Arabs	.70	.27		
P7: Visualize/ Fig. & Graph	Jews	.75	.24	11.19*	.66
	Arabs	.59	.25		
P9: Data management	Jews	.66	.30	7.93*	.44
	Arabs	.53	.29		
P10: Quantitative reading	Jews	.81	.23	7.23*	.42
	Arabs	.71	.26		
S2: Number sense	Jews	.76	.27	5.66*	.36
	Arabs	.66	.31		
S3: Figures tables & graphs	Jews	.94	.16	7.81*	.52
	Arabs	.85	.22		
S4: Approximation & estimation	Jews	.78	.23	5.86*	.37
	Arabs	.69	.29		
S5: Evaluate / verify options	Jews	.97	.13	7.74*	.61
	Arabs	.88	.22		
S6: Recognize patterns	Jews	.50	.38	10.77*	.64
	Arabs	.26	.34		
S7: Proportional reasoning	Jews	.95	.14	5.81*	.40
	Arabs	.89	.19		
S8: Unfamiliar problems	Jews	.88	.21	8.92*	.58
	Arabs	.75	.28		
S10: Open-ended items	Jews	.59	.40	12.14*	.64
	Arabs	.34	.35		
S11: Word problems	Jews	.90	.21	9.64*	.66
	Arabs	.75	.30		

C. *Latent Knowledge States*

The ratios for Jewish and Arab students in the eight clusters of knowledge states described above appear in Table 3. As can be seen in the table, in the lower clusters (1, 5, 8, and 3) Arab students are overrepresented (with ratios of actual count to expected count of 2.42, 2.01, 1.71, and 1.77, respectively), whereas in the higher clusters (4, 6, and 2) they are underrepresented (with ratios of actual count to expected count of 0.56, 0.75, and 0.26, respectively.)

Discussion

Between-Countries Perspective

The results of the current study indicate the superiority of Japanese 8th graders in mathematics knowledge over their American and Israeli counterparts. This is evident not only in the total test score but also in the underlying dimensions of the TIMSS test, which capture content, process, and skill/item-type attributes, as well as in hierarchically ordered clusters of knowledge states. A close examination of the attribute mastery profile of each country revealed similar patterns for the U.S. and Israel, with relative strength in most content and special skills but with considerable deficiency in mathematical thinking skills such as Logical thinking (P5); Pattern recognition (S6), which involves inductive thinking, and open-ended item type (S10), which involves divergent thinking.

Two questions follow from these results: a) What factors explain the excellent mathematics performance of the Japanese sample? and b) How can the results be used by educators and policy makers in the U.S. and Israel to promote the mathematics performance of their students? Since the investigation of any explanatory variable was beyond the scope of the current study we can only be speculative in our attempt to answer the first question. In this regard, we will briefly address various aspects of the educational and cultural context of Japan including intended curriculum, classroom instruction, teacher characteristics, teacher professional development, teacher status, and supplementary education as well as students' and parents' expectations and attitudes.

As to curriculum, the intended 8th grade mathematics curriculum in Japan was described as coherent and challenging (Schmidt et al., 2001), yet it includes fewer topics than those of the U.S. and Israel (12 compared to 44 and 21, respectively) as reported by Cogan and Schmidt (2002). This supports the conclusion, stated by the same authors elsewhere, that U.S. math is "a mile wide and an inch deep" (Cogan and Schmidt, 1999), indicating that it is fragmented, unfocused, repetitive, and unchallenging. These authors also claim that what primarily drive instruction in the U.S. are textbooks rather than standards. The mathematics curriculum in Israel, like that of the U.S., is spiral and less focused and coherent than that of Japan (Zuzovsky, 2001). Our finding that the average U.S. student and his/her Israeli counterpart failed to master content attributes included in their intended curriculum (Geometry in the U.S. and Data and statistics in Israel) also attests to the common wisdom of "less is more."

The way the curriculum is taught in Japan's middle schools is also quite different from that of the U.S. and Israel, as it focuses on developing mathematical thinking rather than mathematical skills (Sawada, 1999; Schümer, 1999). TIMSS video studies have shown that a typical script of Japanese lessons advances as follows: the teacher poses a complex thought-provoking question, the students struggle with the problem, several students present ideas or solutions to the class, the teacher leads a class discussion of the various solutions, then the teacher summarizes the conclusions and makes connections to mathematical concepts (Hiebert et al., 2003; Stigler, Gonzales, Kawanaka, Knoll & Serrano, 1999). Typical mathematics classes in the U.S. and Israel, on the other hand, focus on promoting skill acquisition and are characterized by the following scripts: the teacher explains a theorem and then uses a sample problem to show step-by step how to apply the formula in concrete situations; or the teacher presents a problem and demonstrates how to solve it followed by students' practice (Kawanaka & Stigler, 1999; Hiebert & Stigler, 2000; Schmidt, et al., 2001; Zuzovsky, 2001).

A comparison of teachers' use of questions in the 8th grade mathematics classroom in Japan and the U.S. indicated that Japanese teachers more frequently asked higher order questions than U.S. teachers and they did so in different situations and for achieving different goals (Kawanaka & Stigler, 1999). While Japanese teachers tended to ask those kind of questions when the class was sharing the solution methods that students generated while working at their desks, the U.S. teachers tended to ask such questions when they guided students to use principles and procedures. The latter also tended to use such questions for assessment purposes, judging students' responses as right or wrong, and rarely asked students to reflect on their peers' responses. Moreover, Kawanaka and Stigler (1999) observed two kinds of problem-solving activities in Japanese classrooms, which they term "divergent" and "convergent." The former refers to open-ended problem solving in which the students are asked to solve a non-routine problem on their own using any method they wish or just think about how to solve the problem without actually solving it. The latter refers to solving a given problem when the students know what solution method is required. This observation, especially with regard to divergent problem solving, lends support to the results of the current study which indicate the superiority of the Japanese sample in dealing with open-ended questions (S10).

However, it should be noted that there is more to mathematics education in Japan than what is offered in the regular classroom instruction. In addition to this type of instruction, the majority of Japanese students in post-elementary education receive supplementary instruction both at the *juku* school and at home through parental tutoring. Public school teachers in Japan count on external sources for skill acquisition and therefore do not assign homework. They expect students to voluntarily review the material taught in class and engage in drill and practice at home and at the *juku* (Schümer, 1999). Reviewing thus becomes the student's responsibility. Moreover, having to pass entrance exams at various school levels seems to increase students' motivation to do well at school, as does their parents deep concern about their school attainments, especially in mathematics which is highly valued nation-wide. As noted by Schümer (1999), mathematics education in Japan is a composite of three elements: upbringing, regular classroom experience, and supplementary schools.

Another relevant distinction between Japan and the other two countries concerns the teaching profession. Unlike in the U.S. and Israel, teachers' salaries and status in Japan are relatively high (Barro & Lee, 1986; Barro & Suter, 1998). Moreover, as part of their job, Japanese teachers are regularly engaged in extensive professional development through *lesson study* (Fernandez & Chokshi, 2002; Lewis & Tsuchida, 1998; Stigler & Hiebert, 1999), which takes the form of a community of practice within a school where teachers create a lesson for introducing a particular topic, implement it, and analyze the teaching–learning process and its consequences. They then suggest improvements, try them out, and repeat the process until they have reached their goal, and then move to the next one. Engaging in situated learning experiences of this kind not only improves instructional strategies, but is also likely to shape teachers' epistemological beliefs regarding knowledge and the way it is acquired (Hoffer & Pintrich, 1994).

Having speculated on possible cultural and contextual effects on the results of the current study we proceed now to our second question – How could the national profiles, obtained by means of the RS, be used to promote students' mathematics performance? Identifying areas of strength and weakness for each country on comparable underlying dimensions, as shown in the current study, can be highly informative for policy makers and educators. Japan's mathematics profile that emerged in the present study can be used by policy makers and educators in the U.S. and Israel as an indication of what is educationally possible. Yet, this does not imply a recommendation to adopt Japanese math textbooks, as was done with Singapore's math textbooks, which were adopted by other countries including the U.S. and Israel (Ramakrishnan, 2000). Rather, it is recommended that Japan's attribute mastery profile stimulate educators in the U.S. and Israel to reflect on their own intended mathematics curriculum and the way it is being taught in order to find out what needs to be changed and how to get to what is educationally possible, as exemplified by the Japanese students (Cogan & Schmidt, 2002; Wagemaker, 2002). In this regard, identifying and mapping clusters of hierarchically related latent knowledge states at the country level can be useful; such maps (as demonstrated in this article) can help tracing developmental paths towards mastery which are valuable for evaluating both the intended and the implemented curricula and can aid in designing remedial instruction.

Within-Country Perspective

Inter-country comparisons in the context of international assessment allow comparable results that are not susceptible to invariance of the intended curriculum, as is the case in between-countries comparisons. Comparison between the attainments of Jewish and Arab students in Israel is of high national importance, given the centralized education system in this country. Moreover, the comparison between the Jewish majority and the Arab minority in Israel, which represent two culturally diverse populations with almost no inter-group contact, is particularly interesting from a sociological point of view. Such a comparison is made possible in the Israeli context due to the fact that both groups study according to the same intended mathematics curriculum issued by the Israeli Ministry of Education but in separate schools.

Previous research had pointed out the substantial discrepancy in mathematics achievement between the Jewish and Arab subpopulations in Israel (Aviram, Cfir, & Ben-Simon, 1998; Bashi, Kahan, & Davis, 1981; Birenbaum & Nasser, 2002; Zuzovsky, 2001), but the nature of this difference in terms of cognitive processes has not been investigated before. The findings of the current study indicated that all 23 attributes that were used to define the underlying construct measured by the 1999 TIMSS-R mathematics test for 8th graders, and which successfully accounted for the test score variance of the Israeli sample, yielded significant differences in mean mastery probabilities in favor of the Jewish population. From a proficiency standpoint the average Arab student, compared to his/her Jewish counterpart, seems to be failing to master basic topics that are learned in earlier grades. Similarly, the comparisons between the two populations with respect to clusters of knowledge states indicated an alarming under-representation of Arab students in the highest cluster in the hierarchy, which included mastery of all attributes. These findings emphasize the deficient prior mathematical knowledge of the average Arab student, as compared to his/her Jewish counterpart. Because mastery of these attributes is fundamental to achievement in higher mathematics they should be targeted as "prime candidates" for remedial interventions.

Again, we can only speculate on the causes of the large achievement gap between the two populations as the examination of any explanatory variable was beyond the scope of the current study. In the context of previous research the gap can be attributed, at least partially, to discrepancy in educational resources available to Jewish and Arab municipalities in Israel and to differences in teachers' qualifications in the two sectors (Mazawi, 1997; Yogev & Ayalon, 1996). As argued by Al-Haj (1995) the effect of the shortage in educational resources is evident in the quality of the implemented curriculum; that is, the instruction supplied by Arab schools. A recent study compared the instruction, learning, and assessment culture in 8th grade mathematics classes in Arab and Jewish, low, medium-, and high-achieving schools (Birenbaum & Nasser, 2002). The most noticeable differences were with respect to the extent and amount of engaging students in mathematics-related activities. Jewish students in the higher achieving schools were engaged in more, mostly challenging activities than any of the Arab classes. They were encouraged to attempt to solve problems collaboratively before the teacher discussed the solution and to compare various solutions. In the Arab classes, students were kept more passive and were not encouraged to work collaboratively; rather, the teacher was engaged in writing problems and their solution on the backboard and explaining them. Following the teacher's presentation, students were given working sheets of the drill-and-practice type, consisting of problems that were adopted from the textbook. In the Jewish classes, especially the high-achieving ones, various sources were used to introduce a variety of tasks, and strategies of how to address the problem and how to evaluate the solution were taught. Compared to the Arab classes, more frequent assessments (quizzes and tests) were administered in the Jewish classes, and more detailed feedback was provided.

We speculate that such differences in instructional practice can account, at least partially, for the significant mean differences encountered in the current study between the Jewish and Arab samples on all the attribute mastery probabilities. Accordingly, to improve instruction in the Arab sector, we recommend to allocate resources for teacher professional development that will expose teachers to current views of math teaching (NCTM, 2000) and address teachers' epistemological beliefs about mathematical knowledge and the way it is acquired (Ernest, 1999). Cultivating communities of practice (Wenger, McDermott, & Snyder, 2002) that will consist of Jewish and Arab mathematics teachers could be a productive means to meet this end, and to further improve instruction in the Jewish mathematics classes whose students are still lacking in some higher-order thinking as was evident in our results. These communities of practice could use electronic conferencing (Bonk & King, 1998) to examine and discuss actual examples of teaching and assessing higher-order mathematical thinking. Additionally, *lesson study* – the Japanese way – should be encouraged at the level of the individual school. The information provided by the RS analyses could be used in these discussions to spur relevant remedial instruction tailored to students' knowledge states as well as to tailor further assessment to the targeted attributes.

Finally an inference from a psychometric stance: the current study demonstrated the utility of RS methodology for large-scale diagnostic assessment targeted at between-countries and within-country comparisons. However, due to the nature of this study – a secondary data analysis – the attributes were defined post-hoc rather than at the stage of test design, which resulted in uneven distribution of items across the various attributes. In order to increase the validity and reliability of future international comparisons, it is recommended to first define a relevant set of attributes and then write items that tap that set of attributes.

Notes

1. This article is based in part on work supported by the National Science Foundation under Grant No. REC-0126064. Opinions, findings, and conclusions or recommendations are those of the authors and do not necessarily reflect those of the National Science Foundation.
   An earlier version of this article was presented at The 1st IEA International Research Conferences (IRC-2004), May 11-13, Lefkosia, Cyprus.
2. Israel is marked in the 1999 TIMSS-R report as one of the countries that did not comply with the guidelines for sample implementation (Beaton, Mullis, Marton, Gonzalez, Kelly, & Smith, 1999). Excluded from the Israeli sample were special education students; students of the extreme orthodox independent school system; students in religious schools where science is not taught; students in regular classes who suffer from severe physical, mental, or emotional problems, and students who lack proficiency in the language in which the test was written. As a result, the desired Israeli population covered only 74% of the desired international population (Zuzovsky, 2001).

References

Al-Haj, M. (1995). *Education, empowerment and control: The case of the Arabs in Israel.* Albany: State University of New York Press.

Aviram, T., Cfir, R., & Ben-Simon, A. (1998). *The national feedback to the education system–mathematics for 8th grade.* Jerusalem: National Institute for Testing and Evaluation. (Hebrew)

Barro, S.M., & Lee, J.W. (1986). A comparison of teachers' salaries in Japan and the United States. Washington, DC: SMB Economic Research.

Barro, S.M., & Suter, L. (1988). *International comparisons of teacher's salaries: An exploratory study.* Washington, DC: Survey report. National Center for Education Statistics.

Bashi, Y., Kahan, S., & Davis, D. (1981). *Achievement of the Arab elementary school in Israel.* Jerusalem: The Hebrew University, School of Education. (Hebrew)

Batrice, Y. (2000). *The Palestinian women in Israel: Reality and challenges: An empirical study.* Ako: Dar Alaswar. (Arabic)

Beaton, A.E. Mullis, I.V.S., Martin, M.O., Gonzales, E.J., Kelly, D.L., & Smith, T.A. (1999). *Mathematics achievement in the middle school years: IEA's Third International Mathematics and Science Study - Repeated (TIMSS-R).* Boston College, Center for the Study of Testing, Evaluation and Policy.

Bechger, T.M., van-Schooten, E, De Glopper, C., Hox-Joop J.J. (1998). The validity of international surveys of reading literacy: The case of the IEA reading literacy study. *Studies in Educational Evaluation, 24* (2), 99-125.

Birenbaum, M., & Nasser, F. (2002). *Mathematics achievement in the Jewish and Arab sectors and their relationships to student and teacher characteristics and educational context.* Research report 99-02 (submitted to the Chief Scientist of the Israeli Ministry of Education.) Tel Aviv University, School of Education. (Hebrew)

Birenbaum, M., Kelly, A.E., & Tatsuoka, K. (1993). Diagnosing knowledge states in algebra using the rule-space model. *Journal for Research in Mathematics Education, 24* (5), 442-459.

Bonk, C.J., & King, K.S. (Eds.) (1998*). Electronic collaborators: Learner-centered technologies for literacy, apprenticeship, and discourse.* Mahwah, NJ: Erlbaum.

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.

Cogan, L.S., & Schmidt, W.H. (1999). Middle school math reform. *Middle Matters, 8* (1), 2-3.

Cogan, L.S., & Schmidt, W.H. (2002). "Culture shock" – Eight-grade mathematics from an international perspective. *Educational Research and Evaluation, 8* (1), 13-39.

Ernest, P. (1999). Forms of knowledge in mathematics and mathematics education: Philosophical and rhetorical perspectives. *Educational Studies in Mathematics, 38,* 67-83.

Fernandez, C., & Chokshi, S. (2002). A practical guide to translating lesson study for U.S. setting. *Phi Delta Kappan, 84* (2), 128-134.

Gonzalez, E.J., & Miles, J.A. (Eds.) (2001). *TIMSS 1999 user guide for the international database.* International Association for the Evaluation of Educational Achievement (IEA). Chestnut Hill, MA: Boston College, The International Study Center. (Available at http://www.timss.org)

Hiebert, J., & Stigler, J.W. (2000). A proposal for improving classroom teaching: Lessons from TIMSS video study. *The Elementary School Journal, 101* (1), 3-20.

Hiebert, J., et. al, (2003). *Teaching mathematics in seven countries: Results from the TIMSS 1999 video study.* Washington DC: National Center for Education Statistics.

Hofer, B., & Pintrich, P. (1997). The development of epistemological theories: Beliefs about knowledge and knowing and their relation to learning. *Review of Educational Research, 67*(1), 88-140.

Jaeger, R.M. (1994). Evaluating policy inferences drawn from international comparisons of students' achievement test performances. *Studies in Educational Evaluation, 20* (1), 23-39.

Jaeger, R.M., & Hattie, J.A. (1996). Artifact and artifice in education policy analysis: It's not all in the data. *School Administrator, 53* (5), 24-25, 28-29.

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* (4), 255-278.

Lewis, C., & Tsuchida, I. (1998). A lesson is like a swiftly flowing river: Research lessons and improvement of Japanese education. *American Educator, 4*, 14-17, 30-52.

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

Martin, M.O., Mullis, I.V.S., Gregory, K.D., Hoyle, C., & Shen, C. (2000). *Effective schools in science and mathematics: IEA's third international mathematics and science study.* Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Mazawi, A.E. (1997). Concentrated disadvantage and access to educational credentials in Arab and Jewish localities in Israel. *British Educational Research Journal, 25* (3), 355-370.

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 – eight grade. Achievement for U.S. States and districts in an international context.* Chestnut Hill, MA: International Study Center, Boston College.

NCTM (2000). *Principles and standards for school mathematics.* National Council of Teachers of Mathematics. Available at: http://standards.nctm.org/document/index.htm

Ramakrishnan, M. (2000). Should the United States emulate Singapore's education system to achieve Singapore's success in the TIMSS? *Mathematics Teaching in the Middle School, 5* (6), 345-348.

Sawada, D. (1999). Mathematics as problem solving: A Japanese way. *Teaching Children Mathematics*, (Sept.), 54-58.

Schmidt, W., McKnight, C.C., Houang, R.T., Wang, H.C., Wiley, D.E., Cogan, L.S. & Wolfe, R.G. (2001). *Why schools matter: A cross-national comparison of curriculum and learning.* Indianapolis IN: Jossey-Bass.

Schommer, M. (1994). An emerging conceptualization of epistemological beliefs and their role in learning. In R. Garner, & P. A. Alexander (Eds.), *Beliefs about text and interaction with text.* (pp. 25-40). Hillsdale, NJ: Erlbaum.

Schümer, G. (1999). Mathematics education in Japan. *Journal of Curriculum Studies, 31*(4). 399-427.

Stigler, J. W, Gonzales, P., Kawanaka, T., Knoll, S., & 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.* Washington, DC: National Center for Education Statistics. (http://nces.ed.gov/timss).

Stigler, J.W., & Hiebert, J. (1999). *The teaching gap: Best ideas from world's teachers for improving education in the classroom.* New York: Summit.

Tatsuoka, C.M., Varadi, F., & Tatsuoka, K.K. (1992). *BUGLIB.* Unpublished computer program, Trenton, NJ.

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. (1984). Caution indices based on item response theory. *Psychometrika, 49* (1), 95-110.

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 universal set of knowledge states.* Research Report ONR-1. Educational Testing Service, Princeton, NJ.

Tatsuoka, K.K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P.D. Nicols, S.F. Chipman & R.L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327-359). 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* (1), 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 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., & Linn, R.L. (1983). Indices for detecting unusual patterns: Links between two general approaches and potential applications. *Applied Psychological Measurement, 7* (1). 81-96.

Tatsuoka, K.K., & Tatsuoka, M.M. (1992). *A psychometrically sound cognitive diagnostic model: Effect of remediation as empirical validity.* Research Report. Educational testing Service, Princeton, NJ.

Tatsuoka, K.K., Birenbaum, M., Lewis, C., & Sheehan, K.K. (1993). *Proficiency scaling based on conditional probability functions for attributes.* (Research report 39-50). Princeton, NJ: Educational Testing Service.

Tatsuoka, K.K., Corter, J., Dean, M., & Grossman, J. (2003). *Exploring mathematical thinking skills in TIMSS.* Paper presented at the annual meeting of the National Council of Measurement in Education (NCME). (Chicago, IL, April 22).

Tatsuoka, K.K., Corter, J., & Guerrero, A., (2003). *Manual of attribute-coding for general mathematics in TIMSS studies.* New York: Columbia University, Teachers College.

Tatsuoka, K.K., & Tatsuoka, M.M. (1987). Bug distribution and pattern classification. *Psychometrika, 52* (2), 193-206.

Wagemaker, H. (2002). TIMSS in context: Assessment, monitoring, and moving targets. In D.F. Robitaille, & , A.E Beaton (Eds.). *Secondary analysis of TIMSS data* (pp. 3-10). Dordrecht, The Netherlands: Kluwer.

Wenger, E., McDermott, R., & Snyder, W.M. (2002). *Cultivating communities of practice.* Boston, MA: Harvard Business School Press

Yogev, A., & Ayalon, H. (1996). Between policy and research considerations: Constructing a welfare index for Arab schools in Israel. *Israel Social Science Research*, 115-142.

Zimowski, M.F., Muraki, E., Mislevy, R., & Bock, R.D. (1996). *BILOG-MG.* Chicago, Il: Scientific Software International.

Zuzovsky, R. (2001). *Learning outcomes and the educational context of mathematics and science teaching in Israel: Findings of the third international mathematics & science study TIMSS-1999.* Tel Aviv: Ramot. (Hebrew)

The Authors

MENUCHA BIRENBAUM received her Ph.D. in educational psychology from the University of Illinois at Urbana-Champaign. She is currently an associate professor at Tel Aviv University.

CURTIS TATSUOKA received his Ph.D. in statistics from Cornell University. He is currently an assistant professor of Statistics at George Washington University.

TOMOKO YAMADA is a Ph.D. candidate at Teachers College, Columbia University. She earned her BA degree in psychology from the University of California at Berkeley and her MA degree in educational psychology from New York University.

Correspondence: <biren@post.tau.ac.il>

Appendix

List of Content, Process and Skill/Item-Type Attributes* Used in the Current Study

Content attributes

C1: Use basic concepts and operations in whole numbers.
C2: Use basic concepts and operations in fractions and decimals.
C3: Use basic concepts and operations in elementary algebra.
C4: Use basic concepts and properties in geometry.
C5: Read data and use basic concepts in probability and statistics.

Process attributes

P1: Translate/formulate equations and expressions to solve a problem.
P2: Apply computational knowledge in arithmetic, algebra and geometry.
P3: Apply knowledge in arithmetic, algebra and geometry to identify true relationships, properties and/or to set new goals in solving a problem.
P4: Apply rules in solving equations.
P5: Use logical reasoning (case reasoning, deductive thinking, generalizations).
P6: Apply problem search, analytic thinking, problem restructuring and inductive thinking.
P7: Generate and visualize figures and graphs.
P9: Manage numerical information, procedures, goals, and conditions.
P10: Apply quantitative and logical reading.

Skill/item-type related attributes

S2: Use prior knowledge regarding number properties (number sense) and relationships.
S3: Comprehend various representations and use them interchangeably (e.g., written instructions, figures, tables, charts and graphs).
S4: Use approximation/estimation.
S5: Evaluate/verify/check options in a multiple-choice item.
S6: Recognize patterns of various representations (numeric, geometric, algebraic).
S7: Use proportional reasoning.
S8: Solve problems that appear unfamiliar.
S10: Work with open-ended items.
S11: Work with verbally loaded items.

Note: *Adapted from Tatsuoka et al., 2003. (Four attributes are missing from the original list [C6, P8, S1, and S9), they were dropped because of insufficient item involvement in each booklet.)