DIAGNOSTIC ASSESSMENT IN TIMSS-R: COMPARISON OF EIGHTH GRADERS' MATHEMATICS KNOWLEDGE STATES IN THE UNITED STATES, JAPAN, AND ISRAEL

Menucha Birenbaum, Tel Aviv University, Israel
Curtis Tatsuoka, George Washington University, Israel
Tomoko Yamada, Columbia University, Israel

Abstract

The study combines between-countries and within-country comparisons of performance on the 1999 TIMSS-R mathematics test using a probabilistic model for cognitive analysis known as "rule space" (RS) (Tatsuoka, 1983, 1995, in press). Samples of 8th graders from three countries – the United States, Japan, and Israel – were compared with respect to mean probabilities of mastering attributes of content, skills, and cognitive processes underlying their performance on the test. Also compared were the proportions of students from the three samples in each of eight hierarchically ordered clusters of knowledge states. The same analyses were employed for an intra-country comparison on the Israeli sample between the two culturally diverse populations of Jews and Arabs, both of whom study according to the same intended curriculum but in separate schools. The results of the between-countries comparisons indicated the superiority of the Japanese sample in mathematics knowledge and thinking skills. The comparisons of Jewish and Arab students in Israel indicated significant differences in favor of the Jewish group on all attributes. The results are discussed in light of prevalent instruction – learning – assessment cultures in the respective countries and sectors. Also discussed is the utility of RS methodology for large-scale diagnostic assessment targeted at between-countries and within-country comparisons.

BACKROUND

The importance of international tests has been widely acknowledged. 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 (Wagemaker, 2002). 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.

The current study combines between-countries and within-country comparisons. First, a comparison of mathematics achievement is made among three countries – the United States, 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 – who study mathematics according to the same intended curriculum issued by the Israeli Ministry of Education but in separate schools.

Diagnosing individual performance in large-scale assessment is not a common practice but a much desired one. Rule-space (RS) is a probabilistic model for cognitive diagnosis (Tatsuoka, 1983, 1995, in press) that can be used to diagnose individual performance in large-scale assessments. It employs pattern analysis to classify students' item responses according to their profile of strengths and weaknesses on the underlying constructs measured by a test that are termed *attributes.* An attribute is thus a description of a procedure, skill, or content knowledge that a student must possess in order to successfully complete the target task. Performing a 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:* Student's 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 would be 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), ξ (zeta), and generalized ξs (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 ξs have been introduced in order to have orthogonal coordinates in a multidimensional RS (Tatsuoka, 1997). While ξ measures the unusualness of *n* item-score patterns, generalized ξs measure the unusualness of item-score patterns in subsets of *n* items. 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 architecture (Katz, Martinez, Sheehan & Tatsuoka, 1998), and 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).

The current study applied RS methodology to examine the between-countries and within-country differences in attribute mastery probabilities and in clusters of knowledge states.

METHOD

Participants: 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).

Instruments: Only four booklets (1, 3, 5, and 7) of the 1999 TIMMS-R mathematics test were used in this 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's* include attributes such as: applying number properties and relationships (number sense); approximation/estimation; recognizing patterns and sequences; solving open-ended items.

The test items for each test booklet were coded according to the set of 23 attributes. 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.

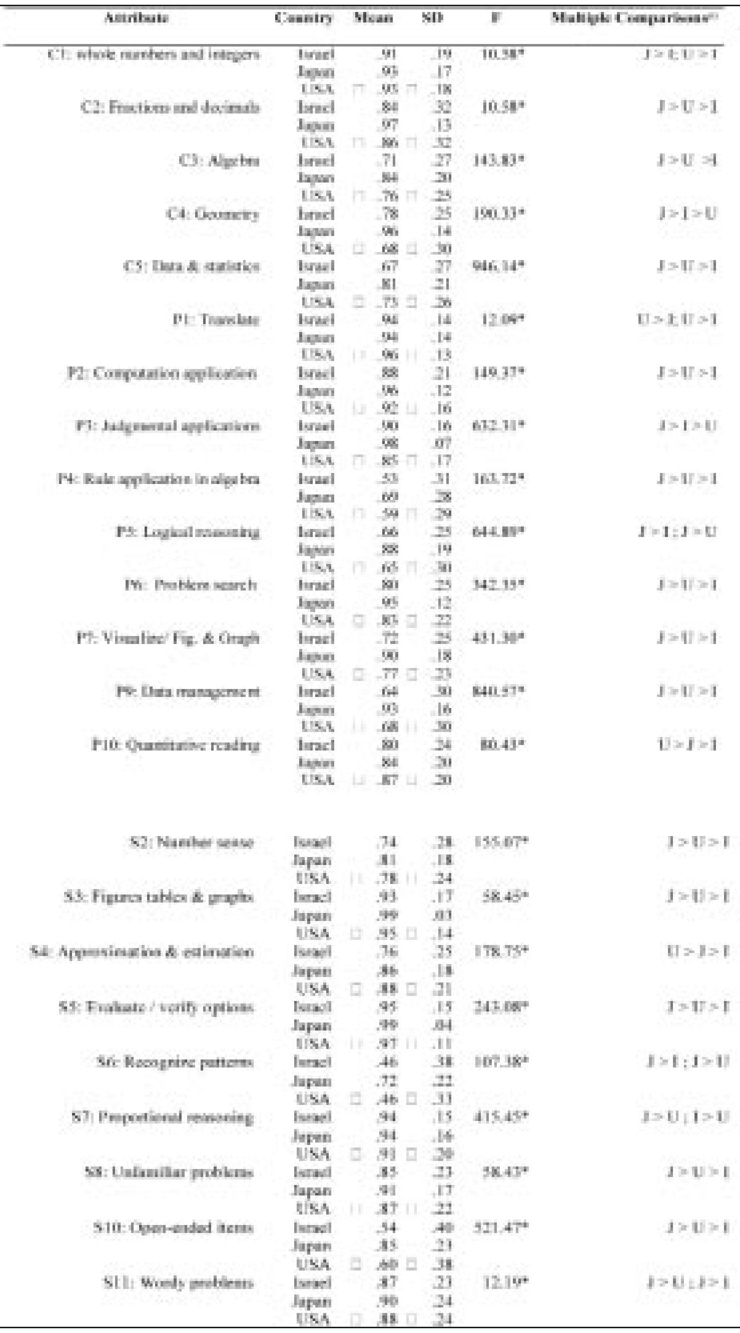
RESULTS

A. Between-countries comparisons

The results indicate the superiority of Japanese 8th graders in mathematics knowledge over their U.S. and Israeli counterparts. This is evident in the attribute mastery probabilities, presented in Table 1. As can be seen in the table, Japan had the highest probabilities on 20 of the 23 attributes and the U.S. had the highest mean on the other three attributes —Translation (P1), Quantitative reading (P10), and Approximation and estimation (S4). Setting the cut-off point for mastery probability at 0.70 reveals 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; Open-ended item type [S10], which involves divergent thinking; and data management [P9].

The superiority of the Japanese sample is also evident in terms of latent knowledge states. Clusters of hierarchically related knowledge states were derived from a cluster analysis of students' attribute probability patterns in the combined sample.

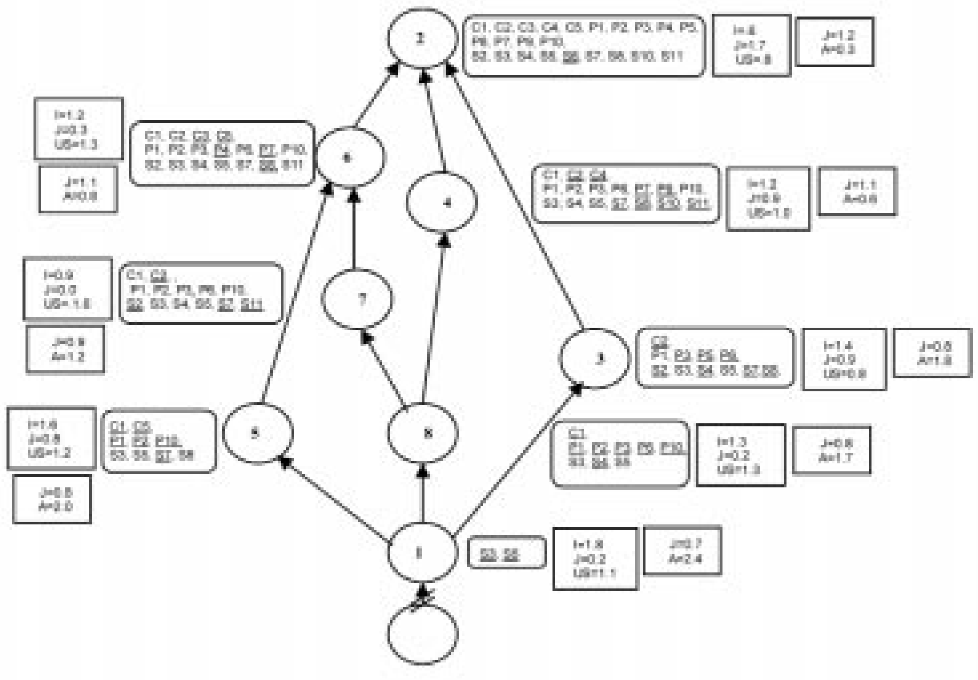
Table 1: One-way ANOVA results for the Effect of Country on the Attribute Mastery Probabilities



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 > I; 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		
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)

Figure 1: A Map of Transitional Relations among Clusters of Latent Knowledge States (N=8874)


2
C1, C2, C3, C4, C5, P1, P2, P3, P4, P5, P6, P7, P9, P10, S2, S3, S4, S5, S6, S7, S8, S10, S11
I=.6
J=1.7
US=.8
J=1.2
A=0.3
I=1.2
J=0.3
US=1.3
J=1.1
A=0.8
C1, C2, C3, C5, P1, P2, P3, P4, P6, P7, P10, S2, S3, S4, S5, S7, S8, S11
6
4
C1, C2, C4, P1, P2, P3, P6, P7, P8, P10, S3, S4, S5, S7, S8, S10, S11
I=1.2
J=0.9
US=1.0
J=1.1
A=0.8
I=0.9
J=0.0
US=1.6
C1, C2, P1, P2, P3, P6, P10, S2, S3, S4, S5, S7, S11
7
J=0.9
A=1.2
3
C2, S6, P1, P3, P5, P6, S2, S3, S4, S5, S7, S8
I=1.4
J=0.9
US=0.9
J=0.8
A=1.8
I=1.6
J=0.8
US=1.2
C1, C5, P1, P2, P10, S3, S5, S7, S8
5
8
C1, P1, P2, P3, P5, P10, S3, S6, S8
I=1.3
J=0.2
US=1.3
J=0.8
A=1.7
J=0.8
A=2.0
1
S3, S8
I=1.8
J=0.2
US=1.1
J=0.7
A=2.4

Notes: 1. Listed are attributes with loadings > 0.75 on a cluster.
2. Ratios of count to expected count are presented in squares (I=Israel; J=Japan; J=Jews; A=Arabs).

A map of the transitional relations among the clusters is presented in Figure 1. A transition from one cluster of 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 1. The numbers of students included in each cluster are presented in the boxes 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 the figure, the cluster that comprised the lowest number of mastered attributes is cluster 1. 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. The cluster that comprised all 23 attributes is cluster 2. 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.

B. Within-country comparisons

Table 2 presents the results of mean differences in attribute mastery probabilities between Jews and Arabs in Israel. As can be seen in the table all 23 attributes yielded significant differences in favor of the Jewish population. From a mastery standpoint, the content, skills, and process attributes that the average Arab student, compared to his/her Jewish counterpart, seems to be failing to master are: Use of basic concepts and operations in fractions and decimals [C2], in Algebra [C3], and in Geometry [C4]; Use of prior knowledge of number properties and relationships (number sense) [S2]; Use of approximation and estimation [S4]; and Visualize figures and graphs [P7]. Similarly, the comparisons between the two populations with respect to clusters of knowledge states, as presented in Figure 1, indicate a severe under-representation of Arab students in the highest cluster in the hierarchy that included mastery of all attributes. (The ratios for Arabs and Jews on cluster 2 are 0.3, and 1.2, respectively.)

DISCUSSION

The results of the between-countries comparisons indicated the superiority of Japanese 8th graders in mathematics knowledge over their U.S. and Israeli counterparts. This is evident in attribute mastery probabilities, as well as in clusters of knowledge states.

The following characteristics of the Japanese educational context can, at least partially, account for the superiority of their 8th graders in mathematics: The Japanese society places a high value on mathematics achievement (Schümer, 1999); in Japan, like in other east Asian societies, testing is considered the ethos of education (Cheah, 1998); the Japanese mathematics curriculum has been described as coherent and challenging (Schmidt et al., 2001); classroom practice in Japan's middle school focuses on developing mathematical thinking rather than mathematical skills (Hiebert et al., 2003; Sawada, 1999; Schümer, 1999; Stigler, et al., 1999); Japanese students receive supplementary instruction both at the "juku" school and at home. Teachers' salaries in Japan are relatively high (Barro & Lee, 1985; Barro & Suter, 1988), and Japan's teachers are regularly engaged in extensive professional development through "study lessons" (Frenandez & Chokshi, 2002; Stigler & Hiebert, 1999).

Yet, the high achievement level of the Japanese students does not imply a recommendation to adopt Japan's mathematics teaching materials and textbooks as was the case with those of Singapore, which were adopted by countries including the U.S. and Israel (Ramakrishnan, 2000). Emphasizing the importance of the cultural context in which instructional materials are developed and operated, Cogan and Schmidt (2002) warn against "the folly of adopting in a wholesale fashion the curricular patterns observed in an alien culture" (p. 38). They do advocate learning from other countries, but argue that their instructional materials and methods must

Table 2: Means and SD for Jewish (n = 1684) and Arab (n = 408) 8th Graders on 23 Attributes, t-Values and Effect Size (d) Values

Attribute	*Group*	*Means*	*SD*	*t-Value*	*d-valu*
C1: whole numbers and integers	Jews	.93	.17	7.67*	.54
	Arabs	.83	.24		
C2: Fractions and decimals	Jews	.88	.28	10.53*	.072
	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 application	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: Wordy problems	Jews	.90	.21	9.64*	.66
	Arabs	.75	.30		

*p < .0001

be thoughtfully analyzed and creatively translated into each country's unique cultural context for education. As they state, "failing to recognize the cultural nature of schooling and measures of it precludes useful insights and conclusions being developed for improving educational policies and practice" (p. 38). Along this line it is recommended that Japan’s attribute mastery profile, which emerged in the current study, 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 Japan's students.

Comparison between the attainments of Jewish and Arab students in Israel is of high national importance, given the centralized education system in this country. Previous research had pointed out the substantial discrepancy in mathematics achievement between the two subpopulations (Aviram, Cfir & Ben-Simon, 1998) but the nature of this difference in terms of cognitive processes has not been investigated before. The results of the current study indicate significant differences in favor of the Jewish group on all attributes. Differences in prevalent classroom practices between these two culturally diverse subpopulations can, at least partially, account for our results. 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 activities, mostly challenging ones, 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.

Finally, two inferences from a psychometric stance: First, the current study demonstrated the utility of RS methodology for large-scale diagnostic assessment targeted at between-countries and within-country comparisons. This methodology seems to overcome the deficiency inherent in benchmark descriptions according to which performance must be interpreted as cumulative and continuous (Kelly, 2002). As was apparent in the map of knowledge-state clusters, various patters of strength and weakness are likely to exist at similar test score levels. Secondly, 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.

References

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.

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.

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.

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.

Cheah, Y. M. (1998). The examination culture and its impact on literacy innovations: The case of Singapore. *Language and Education, 12*(3), 192-209.

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

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

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

Katz, I. R., Martinez, M. E., Sheehan, K. M., & Tatsuoka, K. K. (1998). Expanding the rule space methodology to a semantically-rich domain: Diagnostic assessment in architecture. *Journal of Educational and Behavioral Statistics, 24*(3), 254-278.

Kelly, D. (2002). The TIMSS 1995 International benchmarks of mathematics and science achievement: Profiles of world class performance at fourth and eighth grades. *Educational Research and Evaluation, 8*(1), 41-54.

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

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.

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

Stigler, J. W, Gonznales, 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 Books.

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

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

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. (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. Nichol, S. F. Chipmanm, & 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., & 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., & Guerrero, A. (2003). *Manual of attribute-coding for general mathematics in TIMSS studies.* New York: Columbia University, Teachers College.

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., & 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.

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

NOTE

This study was supported in part by the National Science Foundation under Grant REC-0126064. Opinions, findings, and conclusions or recommendations are those of the authors and do not necessarily reflect those of the National Science Foundation.