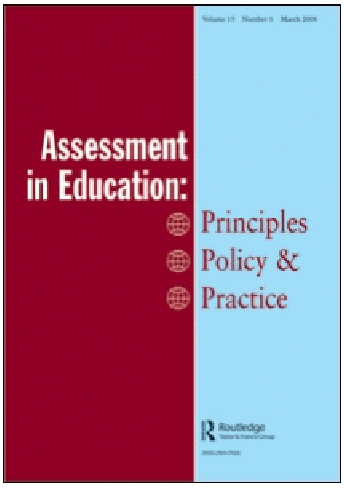


Assessment in Education: Principles, Policy & Practice
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713404048

Large-scale diagnostic assessment: comparison of eighth graders' mathematics performance in the United States, Singapore and Israel

Menucha Birenbaum [a]; Curtis Tatsuoka [b]; Tao Xin [c]
[a] Tel Aviv University, Israel [b] George Washington University, USA [c] Beijing Normal University, China

Online Publication Date: 01 July 2005

Assessment in Education
Vol. 12, No. 2, July 2005, pp. 167–181

Routledge
Taylor & Francis Group

Large-scale diagnostic assessment: comparison of eighth graders' mathematics performance in the United States, Singapore and Israel

Menucha Birenbaum[a]*, Curtis Tatsuoka[b] and Tao Xin[c]
[a]*Tel Aviv University, Israel*; [b]*George Washington University, USA*; [c]*Beijing Normal University, China*

A diagnostic model for large-scale assessment was applied to TIMSS data to compare mathematics performance of eighth graders from three countries—the US, Singapore, and Israel. Compared were attribute mastery probabilities for content, skills and cognitive processes underlying students' performance on the 1999 TIMSS-R mathematics test. Also compared were the proportions of students from the three samples in each of eight hierarchically ordered clusters of knowledge states. The results indicated significant differences in favour of the Singaporean sample on most attributes underlying the test. The results were discussed in light of the cultural context of education in the respective countries.

In view of the importance of imparting mathematical knowledge and thinking skills for establishing a nation's competitive advantage in an increasingly global economy, education systems in many countries are concerned about their students' performance in international tests and are seeking means for improving mathematics instruction (Wagemaker, 2002). Information regarding the attained curriculum 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).

The current study compared mathematics performance among three countries—the US, Singapore and Israel—which differ in culture and education system. Singapore was selected because its students excel in all international tests of

*Corresponding author. School of Education, Tel Aviv University, Ramat Aviv 69978, Israel. Email: biren@post.tau.ac.il

ISSN 0969-594X (print)/ISSN 1465-329X (online)/05/020167–15

DOI: 10.1080/09695940500143852

mathematics achievement, it was therefore interesting to compare the achievement of this small multi-ethnic country to that of the US—an economic superpower—and to Israel—a Middle Eastern country that shares some features with the other two countries. Like Singapore, Israel is a small country, surrounded by large and less developed countries and has a centralized education system. Like the US it maintains a western life style and has a 'spiral' curriculum in mathematics.

Reporting diagnostic feedback in large-scale assessments in general, and in international assessments in particular, is not a common practice but a much desired one (Atkin & Black, 1997). It can aid in interpreting test scores and at the same time guide curricular planning and instruction so that the diagnosed difficulties can be addressed promptly. Currently, 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 in the Third Mathematics and Science Study (TIMSS) allows for diagnostic feedback at four benchmarks, set at the 90th, 75th, 50th, and 25th percentiles of the international score distribution. To generate this feedback, the 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 such a group answered successfully and more than 50% of the lower benchmark group failed to answer correctly. The mastery profile for that benchmark was specified in terms of skills judged by experts to be necessary for successfully solving those particular items (Mullis *et al.*, 2001; Kelly, 2002). However, there are some inherent shortcomings to this method that preclude an accurate diagnosis on the individual level. As stated by Kelly (2002), the benchmark descriptions must be interpreted under the assumption that performance on the TIMSS scale is cumulative (i.e., students reaching a particular benchmark are assumed to have acquired the knowledge and skills described on the lower benchmark). Yet, this is not always the case, as is also implied in the other assumption that performance is continuous. Accordingly, it is recognized that students at the upper or lower ends of a given quartile may indeed know or understand some of the concepts that characterize a higher benchmark, or may not know or understand some concepts that characterize performance at a lower benchmark, respectively. A diagnostic approach that overcomes these shortcomings is the 'rule space' methodology developed by Taksuoka (1983, in press). Following is a brief account of this methodology.

Rule space methodology (RMS) is used to classify examinees' 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 a description of a procedure, skill, or content knowledge that a student must possess in order to complete the target task successfully. Binary attribute patterns that express mastery and non-mastery of attributes are termed *knowledge states*. Attributes and knowledge states are unobservable variables that RSM transforms into observable *attribute mastery probabilities*.

RSM belongs to the branch in statistics that deals with pattern recognition and classification problems, which has two stages—the design stage and the classification stage. At the design stage, an object is characterized by its feature variables and expressed by a pattern of feature variables. At the classification stage, the pattern is classified into one of predetermined classification groups.

However, attributes are usually impossible to measure because they are latent. RSM has extended this approach to deal with latent feature variables. This is done by introducing an item-by-attribute incidence matrix, referred to as *Q matrix* in RSM (Tatsuoka, 1990). Every column in Q matrix 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 specify the hypothesized underlying constructs measured by the test. The only assumption RSM uses at the design stage is that the right answer for a given item can be obtained if and only if all attributes involved in that item are used correctly. Then all possible combinations of attribute patterns from a given Q-matrix are mathematically generated by applying Boolean Algebra, and at the same time, knowledge states are also expressed by their corresponding item score patterns termed *ideal item score patterns* for differentiating them from students' observable item response patterns. Attribute patterns are not observable but corresponding ideal item score patterns are observable, which form predetermined classification groups in RSM. The classification space formulated by RSM thus contains a set of the possible knowledge states generated from a given Q matrix (Tatsuoka, 1991). A student's item response pattern now can be classified into one of the predetermined groups by applying Bayes' decision rules that provides us with the student's most plausible ideal item score pattern with the membership probability.

To recapitulate, a unique characteristic of RSM is the correspondence between attribute patterns and ideal item score patterns. This tie enables us making inference regarding an examinee's performance on latent attributes from his/her performance on observable item responses. Hence, RSM transforms a dataset of students by item scores into a dataset of students by attribute mastery probabilities, thus providing a methodology for large-scale diagnostic assessment.

RSM has been applied in various areas such as subtraction of fractions (Tatsuoka & Tatsuoka, 1992), signed numbers operations (Tatsuoka, 1990), algebra (Birenbaum *et al.*, 1993), the quantitative parts of the Scholastic Aptitude Test (SAT-M; Tatsuoka *et al.*, 1993), and the Graduate Record Examination (GRE; Tatsuoka & Boodoo, 2000), as well as in listening comprehension (Buck & Tatsuoka, 1998). Although RSM 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 RSM for analysing international data sets seems to hold great potential for significantly contributing to the teaching and learning of mathematics. The current study applied this methodology to examine between-countries differences in 8th grade mathematics knowledge.

Method

Participants

The study comprised three samples of eighth graders who participated in the 1999 TIMSS-R study, consisting of 4411 students from the US, 2490 from Singapore, and 2092 from Israel.

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 analysing TIMSS-R-1999 mathematics items for eighth graders (Tatsuoka *et al.*, 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: judgemental 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 *et al.*, 1996) was used to estimate the item response theory (IRT) a and b parameters for the items and the BUGLIB program (Tatsuoka *et al.*, 1992) was used for the rule space (RS) analysis.

Following the computations of students' attribute mastery probabilities, the mean probabilities for the three countries—the US, Singapore, 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.

Results

Test score distributions

Before presenting the results of the RS analyses, statistics regarding the test score distribution in the three counties are summarized. Singapore has the highest median (83.33) and the smallest dispersion of scores (Q3–Q1 = 23.08). The US 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 ($F_{2,\ 8990} = 1493.58$; $p < .0001$). The respective means (and standard deviations) for the US, Singapore, and Israel are 53.78 (22.00); 78.51 (17.91); and 48.34 (21.80). All means are significantly different ($p < .0001$) from each other according to Scheffé's procedure.

Attribute mastery probabilities

The results of one-way ANOVAs for the effect of country on the attribute mastery probabilities are presented in Table 1. 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. Singapore had the highest probabilities on 20 of the 23 attributes, on two attributes—Evaluate/verify/check options (S5), and Quantitative reading (P10)—Singapore and the US had the same mean probabilities. On one attribute—Approximation and estimation (S4)—the US had the highest mean probability. The US had the lowest mean probabilities on three attributes: Geometry (C4), Judgemental applications (P3), and Proportional reasoning (S7). On three attributes, the US and Israel had similar mean probabilities—Logical thinking (P5), Pattern recognition (S6), and Wordy problems (S11). On the remaining 17 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 Singaporean sample has mastered 22 attributes and has failed to master one attribute—Pattern recognition (S6); the average American and Israeli student mastered each 17 attributes and has failed to reach mastery on the six remaining attributes—five of which are common to the two samples: Pattern recognition (S6); Rule application in Algebra (P4); Logical reasoning (P5), Data and procedure management (P9), and Open-ended questions (S10). The sixth non-mastered attribute in the case of the US is Geometry (C4), and in the case of Israel, Data and statistics (C5).

Latent knowledge states

Clusters of hierarchically related latent knowledge states were derived from a cluster analysis[1] (using the K-means clustering method[2]) of students' attribute probability patterns in the combined sample. An eight-cluster solution is presented in Table 2, and a map of the transitional relations among the clusters is presented in Figure 1. A transition from one cluster of latent knowledge states to another is said to be possible

Table 1. One-way ANOVA's F tests, means and standard deviations of attributes and total score by country and Scheffé's test for multiple comparisons

Attribute	F values[a]	Israel		Singapore		USA		Scheffé [b]
		M	SD	M	SD	M	SD	
C1: Whole numbers and integers	19.11	.91	.19	.94	.15	.93	.18	2 > 1***; 2 > 3** ; 3 > 1***
C2: Fractions and decimals	216.50	.84	.32	.99	.09	.86	.32	2 > 1***; 2 > 3*** ; 3 > 1*
C3: Elementary algebra	194.60	.71	.27	.85	.24	.76	.25	2 > 1***; 2 > 3*** ; 3 > 1***
C4: Geometry	752.26	.78	.25	.93	.18	.68	.30	2 > 1***; 2 > 3*** ; 3 > 1***
C5: Data and statistics	68.06	.67	.27	.76	.21	.73	.26	2 > 1***; 2 > 3*** ; 3 > 1***
S2: Number sense	94.27	.74	.28	.83	.17	.78	.24	2 > 1***; 2 > 3*** ; 3 > 1***
S3: Figures, tables & graphs	122.31	.93	.17	.99	.08	.95	.14	2 > 1***; 2 > 3*** ; 3 > 1***
S4: Approximation & estimation	228.57	.76	.25	.85	.20	.88	.21	2 > 1***; 2 < 3*** ; 3 > 1***
S5: Evaluate/ verify options	23.47	.95	.15	.97	.10	.97	.11	2 > 1***; 3 > 1***
S6: Pattern recognition	322.77	.46	.38	.69	.28	.46	.33	2 > 1***; 2 > 3***
S7: Proportional reasoning	195.59	.94	.15	.99	.06	.91	.20	2 > 1***; 2 > 3*** ; 3 < 1***
S8: Unfamiliar problems	66.09	.85	.23	.92	.17	.87	.22	2 > 1***; 2 > 3*** ; 3 > 1***
S10: Open-ended items	701.49	.54	.40	.89	.24	.60	.38	2 > 1***; 2 > 3*** ; 3 > 1***
S11: Wordy problems	246.98	.87	.23	.99	.08	.88	.24	2 > 1***; 2 > 3***
P1: Translate	120.00	.94	.14	.99	.06	.96	.13	2 > 1***; 2 > 3*** ; 3 > 1***
P2: Computation application	343.19	.88	.21	.99	.04	.92	.16	2 > 1***; 2 > 3*** ; 3 > 1***
P3: Judgmental application	314.73	.90	.16	.94	.11	.85	.17	2 > 1***; 2 > 3*** ; 3 < 1***
P4: Rule application in Algebra	663.32	.53	.31	.81	.23	.59	.29	2 > 1***; 2 > 3*** ; 3 > 1***
P5: Logical reasoning	563.18	.66	.25	.87	.23	.65	.30	2 > 1***; 2 > 3***
P6: Problem search	325.46	.80	.25	.94	.12	.83	.22	2 > 1***; 2 > 3*** ; 3 > 1***
P7: Visualize Figures & Graphs	637.05	.72	.25	.93	.16	.77	.23	2 > 1***; 2 > 3*** ; 3 > 1***
P9: Data management	925.56	.64	.30	.94	.17	.68	.30	2 > 1***; 2 > 3*** ; 3 > 1***
P10: Quantitative reading	97.41	.80	.24	.87	.18	.87	.20	2 > 1** ; 3 > 1***
Total Test Score	1493.58	.48	.22	.79	.18	.54	.22	2 > 1***; 2 > 3*** ; 3 > 1***

*p<.05; **p < .01; ***p < .001

[a] For all F values p < .001

[b] 1 = Israel (n = 2092)

2 = Singapore (n=2490)

3 = USA (n = 4411)

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 centre 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 Table 2 along with the percentages of students from the US, Singapore, and Israel.

As can be seen in Table 2 and Figure 1, the cluster that comprised the lowest number of mastered attributes is cluster 8. The two attributes included in this cluster are Figures, tables, and charts (S3) and Evaluate/verify options (S5). The average score on the test (in term of percentage correct answers) for the 391 students in this cluster is 19.67. The percentages of students from Singapore, the US and Israel in this cluster are 0.60, 4.72, and 8.03, respectively. Cluster 5 is the highest in the hierarchy and comprises all 23 attributes. The average score on the test for the 3616 students in this cluster is 83.12. The percentages of students from Singapore, the US and Israel in this cluster are 74.86, 28.72, and 23.18, respectively. Presented in Figure 1 are three possible paths to progress from the lowest cluster to the highest one: (a) a path through clusters 3, 7, 4; (b) a path thought clusters 2 and 1; and (c) a path through cluster 6. As can be seen in Table 2 the percentages of students from the US and Israel in the clusters that define these paths are similar (a total of 34% and 35%, respectively in clusters 3, 7, and 4; 25% in clusters 2 and 1; and 7% and 10%, respectively cluster 6).

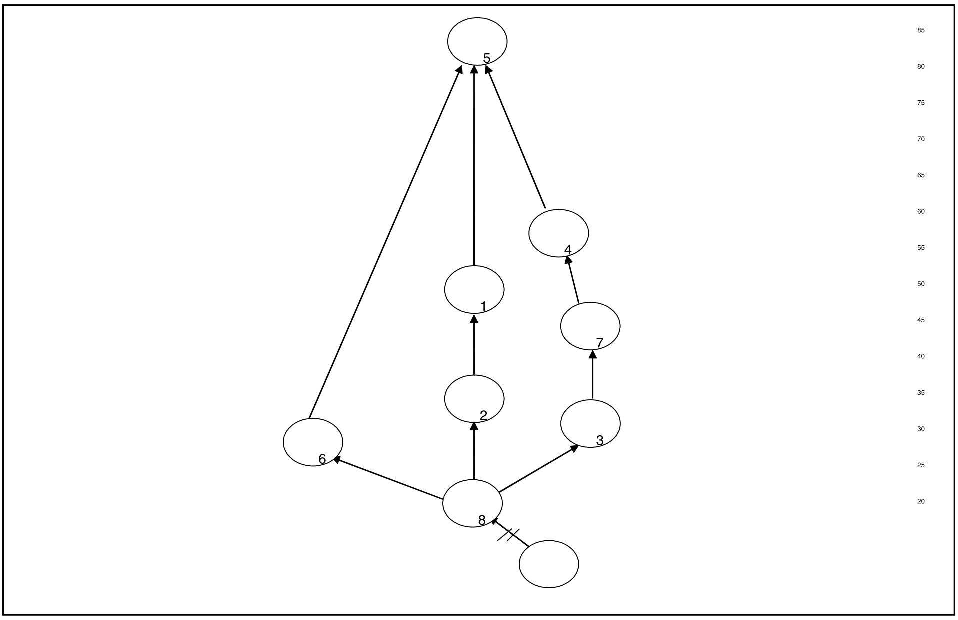
Discussion

The results of the current study indicate the superiority of Singapore's eighth graders over their US and Israeli counterparts in mathematics performance. This is evident not only in the total test score but also in the underlying dimensions of the TIMSS-R test, which capture content, process, and skill/item-type attributes, as well as in clusters of knowledge states that represent hierarchically ordered attribute mastery patterns. A close examination of the attribute profile of each country revealed similar patterns for the US 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 and procedure management (P9). It seems that the average student in the US and Israel is taught relatively well in content knowledge (except for Geometry [C4] in the US and Data and statistics [C5] in Israel) and special mathematics skills but not mathematical thinking skills. Singapore's students, on the other hand, are well taught in these skills (though less so with respect to Pattern recognition [S6]), as well as in content and special skills.

Two questions follow from the results of the current study: (a) what factors explain the excellent mathematics performance of the Singaporean sample? and (b) how can the results be used by educators and policy-makers in the US and Israel to promote the mathematics performance of their students? Since the investigation of

Table 2. Cluster analysis results (N = 8993)

Attribute	Cluster 5	4	1	7	2	3	6	8
C1: Whole numbers and integers	.98	.93	.94	.99	.66	1.00	.95	.57
C2: Fractions and decimals	1.00	.99	.98	.97	.97	.06	.23	.55
C3: Elementary algebra	.91	.74	.86	.49	.47	.68	.70	.47
C4: Geometry	.94	.86	.72	.33	.39	.56	.63	.63
C5: Data and statistics	.82	.70	.82	.42	.60	.37	.83	.51
S2: Number sense	.85	.76	.81	.79	.94	.71	.40	.51
S3: Figures, tables & graphs	1.00	.99	.98	.96	.82	.78	.82	.83
S4: Approximation & estimation	.88	.78	.93	.88	.96	.97	.60	.49
S5: Evaluate/ verify options	.99	.99	.98	.97	.88	1.00	.86	.79
S6: Pattern recognition	.81	.60	.11	.33	.02	.02	.16	.17
S7: Proportional reasoning	.99	.97	.95	.95	.89	.68	.92	.60
S8: Unfamiliar problems	.96	.87	.98	.60	.95	.59	.73	.55
S10: Open-ended items	.98	.83	.28	.37	.31	.07	.22	.39
S11: Wordy problems	1.00	.96	.97	.90	.68	.59	.63	.61
P1: Translate	1.00	.98	.98	.98	.90	.91	.92	.62
P2: Computation application	1.00	.91	.95	.98	.76	.99	.86	.61
P3: Judgmental application	.93	.92	.87	.85	.98	.74	.67	.69
P4: Rule application in Algebra	.84	.45	.80	.45	.56	.37	.24	.27
P5: Logical reasoning	.93	.62	.62	.23	.68	.45	.55	.60
P6: Problem search	.96	.80	.83	.85	.88	.94	.54	.45
P7: Visualize Figures & Graphs	.96	.80	.85	.60	.54	.51	.56	.50
P9: Data management	.96	.77	.51	.56	.34	.55	.67	.44
P10: Quantitative reading	.92	.84	.91	.89	.58	.87	.76	.58
Number of Students	3616	1748	1200	590	585	340	523	391
Percentages: SGP	74.86	12.57	5.54	3.21	2.21	0.56	0.44	0.60
Percentages: USA	28.72	21.33	17.50	9.18	7.84	4.90	6.94	4.72
Percentages: ISR	23.18	23.61	16.25	5.02	8.80	5.26	9.85	8.03
Mean of total score	83.12	55.77	49.81	44.76	31.69	30.16	25.08	19.67



Cluster	Mean	Mastered Attributes																						
5	83.12	C1	C2	C3	C4	C5	S2	S3	S4	S5	S6	S7	S8	S10	S11	P1	P2	P3	P4	P5	P6	P7	P9	P10
4	55.77	C1	C2		C4		S2	S3	S4	S5		S7	S8	S10	S11	P1	P2	P3			P6	P7	P9	P10
1	49.81	C1	C2	C3		C5	S2	S3	S4	S5		S7	S8		S11	P1	P2	P3	P4		P6	P7		P10
7	44.76	C1	C2				S2	S3	S4	S5		S7			S11	P1	P2	P3			P6			P10
2	31.69		C2				S2	S3	S4	S5		S7	S8			P1	P2	P3			P6			
3	30.16	C1						S3	S4	S5						P1	P2				P6			P10
6	25.08	C1				C5		S3		S5		S7				P1	P2							P10
8	19.67							S3		S5														

Figure 1. A map of transitional relations among clusters of latent knowledge states (N = 8993)

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 Singapore including cultural expectations, intended and implemented curricula, teacher status, professional development, supplementary education as well as students' and parents' attitudes towards the study of mathematics.

There is a strong relationship between economy and education in Singapore (Bracey, 1997; Gopinathan, 1999; Kaur, 2001). This small multi-ethnic country has become one of the 20 richest countries in the world by being intensely concerned with education and putting much emphasis on examinations, hence producing an internationally competent workforce (Bracey, 1997; Cheah, 1998, Gopinathan, 1999; Nevo & Weimann, 2001). The examination culture, as it is occasionally referred to (Cheah, 1998), reflects a nationwide obsession with excelling in exams. Cheng (quoted in Cheah, 1998) contends that examinations are 'the soul of the ethos about education in East Asian societies' (p. 192), emphasizing that examinations are not merely selection mechanisms but they also afford 'training opportunities for competition, adaptation, endurance, perseverance, and so forth'. (p. 192).

Meritocracy is the guiding principle in this credential society and therefore parents value education highly and encourage their children to invest much effort in preparing for examinations. Those who can afford it also invest in private tutoring for extra coaching, starting from kindergarten (Cheah, 1998; Gopinathan, 1999). In the Singaporean school system high-stake tests are given frequently and tracking based on their results starts as early as fourth grade, offering relatively few second chances (Gopinathan, 1999). According to the TIMSS-R background questionnaire, time spent on mathematics in class is higher in Singapore than in the other two countries as there are fewer discipline problems and interruptions during class (Beaton *et al.*, 1999). Singaporean students also reported spending more out-of-school time studying mathematics or doing mathematics homework (Beaton *et al.*, 1999; Leung, 2002). Moreover, they reported spending less time playing sports than did their western counterparts (Beaton *et al.*, 1999; Leung, 2002). No wonder that Singaporean students have been characterized as highly competitive and stressed (Nevo & Weimann, 2001).

As to curriculum, unlike the school mathematics curricula in the US and Israel that are 'spiral' (i.e., the same topics are repeatedly introduced in various grade levels yet with different depth), Singapore's curriculum is non-repetitive and focused (Bracey, 1997; Cogan & Schmidt, 1999). Schmidt and his colleagues have described the intended eighth grade mathematics curriculum in Singapore as coherent and challenging (Schmidt *et al.*, 2001). The US curriculum on the other hand has been described by Cogan and Schmidt (1999) as 'a mile wide and an inch deep' indicating that it is fragmented, unfocused, repetitive, and unchallenging. These authors also claim that instruction in the US is primarily driven by textbooks rather than standards. The mathematics curriculum in Israel, like that of the US, is less rigorous, less focused and less coherent than that of Singapore (Zuzovsky, 2001).

As to the implemented curriculum, Singaporean teachers invest much effort in coaching for the frequent exams and promote test-taking strategies. It is claimed that by doing so they encourage their students to practise rote memorization (Nevo & Weimann, 2001; Ramakrishnan, 2000). Although Singaporean students have excelled in international assessments, the Singaporean authorities do not seem to be satisfied with the graduates their education system produces, whom they describe as lacking in creativity and enterprise, exhibiting dependency and conformism (Kaur, 2001). Realizing the importance of workers' creativity, autonomy and flexibility for a successful economy in the twenty-first century, the Singaporean government has recently started to reconstruct the education system under the slogan 'Thinking schools—learning nation' in order to promote the desired competencies (Atkin & Black, 1997; Gopinathan, 1999; Kuar, 2001).

Another relevant distinction between Singapore and the other two countries concerns the teaching profession. Unlike in the US and Israel, teachers' salaries and status in Singapore are relatively high (Barro & Lee, 1986; Barro & Suter, 1988). Compared to teachers in the US and Israel, Singaporean teachers also have a better mathematics background (Ramakrishnan, 2000). Moreover, as part of their job, Singapore's teachers are regularly engaged in extensive professional development Gopinathan, 1999).

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? Mapping areas of strengths and weaknesses for each country on comparable underlying dimensions, as shown in the current study, can be highly informative for policy-makers and educators. Singapore's mathematics profile, as it emerged in the present study, can be used by policy-makers and educators in the US and Israel as an indication of what is educationally possible. Yet, this does not imply a recommendation to adopt Singapore's math textbooks, as several countries including the US and Israel have done (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 be thoughtfully analysed 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 Singapore's attribute mastery profile, which emerged in the current study, stimulate educators in the US 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 Singaporean students.

Mapping clusters of hierarchically related latent knowledge states based on each country's data can provide useful diagnostic information for evaluating national or district-wide curricula. Such maps present various progression paths towards mastery, which may reflect different emphases in curricula or different teaching methods (Birenbaum *et al.*, 1993). The diagnostic information extracted from individual patterns of attribute mastery probabilities can further serve classroom teachers in their effort to tailor remedial instruction to the specific weaknesses of each student.

Finally, two inferences from a psychometric stance: firstly, the current study demonstrated the utility of RS methodology for large-scale diagnostic assessment targeted at between-countries 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 patterns 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 define a relevant set of attributes first and then write items that tap that set of attributes.

Acknowledgement

This study has been supported in part by the National Science Foundation. REC. NO. 0126064. Opinions, findings, and conclusions or recommendations are those of the authors and do not necessarily reflect those of the National Science Foundation.

Authors' note

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 *et al.*, 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 Israeli desired population covered only 74% of the international desired population (Zuzovsky, 2001).

Notes

1. Cluster analysis is a statistical tool for solving classification problems. The object is to sort cases into groups or clusters so that the degree of similarity between cases in the same group is high and between members of different group is low.
2. This method of clustering produces k different clusters of greatest possible distinction.
3. 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.)

Notes on contributors

Menucha Birenbaum is Professor of Education at Tel Aviv University, Israel. She received her Ph.D. in educational psychology from the University of Illinois at Urbana-Champaign. Her current research focuses on consequential validity of assessment, and on analysis of learning interactions.

Curtis Tatsuoka received his Ph.D. in statistics from Cornell University. He is currently an Assistant Professor of Statistics at George Washington University, USA.

Tao Xin is a Professor in the Institute of Developmental Psychology, Beijing Normal University, China. His research interests focus on educational testing, and how schooling effects students' development.

References

Atkin, M. & Black, P. (1997) Policy perils of international comparisons: the TIMSS case, *Phi Delta Kappan,* 79(1), 22–28.

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.* Survey report (Washington DC, National Center for Education Statistics).

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, MA, Boston College, Center for the Study of Testing, Evaluation and Policy).

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.

Bracey, G. W. (1997) More on TIMSS, *Phi Delta Kappan,* 78(8), 656–657.

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. (1999) Middle school math reform, *Middle Matters,* 8(1), 2–3.

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

Gonzalez, E. J. & Miles, J. A. (Eds) (2001) *TIMSS 1999 user guide for the international database* (International Association for the Evaluation of Educational Achievement, Chestnut Hill, MA, Boston College, The International Study Center). Available online at: http://www.timss.org (accessed February 2004).

Gopinathan, S. (1999) Preparing for the next rung: economic restructuring and educational reform in Singapore, *Journal of Education and Work,* 12(3), 295–308.

Kaur, B. (2001) Singapore's school mathematics curriculum for the 21st century, paper presented at the *Qualifications and Curriculum Authority Conference on Reasoning Explanation and Proof in School Mathematics and Their Place in the Intended Curriculum*, Cambridge, 4–6 October.

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.

Leung, F. K. S. (2002) Behind the high achievement of East Asian students, *Educational Research and Evaluation,* 8(1), 87–108.

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: TIMS—eighth grade. Achievement for US States and districts in an international context* (Chestnut Hill, MA, International Study Center, Boston College).

Nevo, B. & Weimann, G. (2001) *The Singaporean puzzle* (Jerusalem, Tzivonim). (In Hebrew).

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.

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).

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. (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* (Hillsdale, NJ, Erlbaum), 543–588.

Tatsuoka, K. K. (1991) *Boolean algebra applied to determination of universal set of knowledge states.* Research report ONR–1 (Princeton, NJ, Educational Testing Service).

Tatsuoka, K. K. (in press) *Statistical pattern recognition and classification of latent knowledge states: cognitively diagnostic assessment* (Mahwah, NJ, Erlbaum).
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. & 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* (Mahwah, NJ, Erlbaum), 821–857.
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*, Chicago, April 22.
Tatsuoka, K. K., Corter, J., Guerrero, A. & Grossman, J. (in press) *Manual of attribute-coding for general mathematics in TIMSS studies* (New York, Columbia University, Teachers College).
Tatsuoka, K. K. & Tatsuoka, M. M. (1992) *A psychometrically sound cognitive diagnostic model: effect of remediation as empirical validity.* Research report (Princeton, NJ, Educational Testing Service).
Wagemaker, H. (2002) TIMSS in context: assessment, monitoring, and moving targets, in: D. F. Robitaille & A. E. Beaton (Eds) *Secondary analysis of TIMSS data* (Dordrecht, Kluwer), 3–10.
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 *TIMSS1999* (Tel Aviv, Ramot) (in Hebrew).

Appendix: List of Content, Process and Skill/Item-Type Attributes[3] 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.