Educ Stud Math (2008) 68:263–272
DOI 10.1007/s10649-007-9099-8

An international comparison using a diagnostic testing model: Turkish students' profile of mathematical skills on TIMSS-R

Enis Dogan · Kikumi Tatsuoka

Published online: 29 November 2007

Abstract This study illustrates how a diagnostic testing model can be used to make detailed comparisons between student populations participating in international assessments. The performance of Turkish students on the TIMSS-R mathematics test was reanalyzed with a diagnostic testing model called the Rule Space Model. First, mathematical and cognitive skills ('attributes') measured by the test were determined. One hundred sixty-two items were coded in terms of their attribute involvement, creating an incidence matrix—the Q-matrix. Using the Q-matrix and the student response data, each student's attribute mastery profile was determined. Mean attribute mastery levels of Turkish students were computed and compared to those of their American peers. It was shown that Turkish students were weak in algebra and probability/statistics. They also demonstrated poor profiles in skills such as applying rules in algebra, approximation/estimation, solving open-ended problems, recognizing patterns and relationships, and quantitative reading.

Keywords Diagnostic testing · International assessments · Mathematical skills

1 Introduction

TIMSS and TIMSS-R are among the most comprehensive international studies that investigate science and mathematics achievement at the middle school level. Turkey participated in TIMSS-R (1999) and ranked 31st among 38 countries in mathematics achievement according to mean total score. Given this picture, researchers and educators have much to learn about the weaknesses of Turkish students. Reports that rely on total scores and rankings based on total scores do not provide enough about these weaknesses, however. One attempt to provide more information was the TIMSS 1999 International Mathematics

E. Dogan (✉)
American Institutes for Research, 2000 K Street NW, Suite 300, Washington, DC 20006, USA
e-mail: edogan@air.org

K. Tatsuoka
Teachers College, Columbia University, 525 West 120th Street, New York, NY 10027, USA

Report by Mullis et al. (2000). Acknowledging that "it is important that educators, curriculum developers, and policy makers understand what students know and can do in mathematics and what areas, concepts, and topics need more focus and effort" (pg. 57), Mullis et al. reported the performance of the top 10%, top quarter, top half, and lower quarter of students in terms of specific knowledge and skills using scale anchoring. The top 10% was composed of students who could organize information, make generalizations, and explain solution strategies in non-routine problem solving situations. Students in the top quarter were able to apply their understanding and knowledge in a wide variety of relatively complex situations. The median benchmark (top half) included students who could apply basic mathematical knowledge in straightforward situations. Finally, students in the lower quartile were those who can only accomplish basic computations with whole numbers. Sixty two percent of all Turkish students were in this final category, while 25% were in the median benchmark, 6% were in the top quarter and only 1% were in the top 10% (MEB-EARGED 2003). Mullis's report sheds *some* light on the achievement profile of the Turkish students. The current study aims to illustrate how a diagnostic testing model (i.e. the Rule Space Model) can be used to draw a more detailed picture of this profile and produce more meaningful international comparisons.

This paper first provides some background information of and a brief introduction to the basic concepts of the Rule Space Model (RSM). Then, a list of knowledge and sub-skill components ('attributes') that explain achievement on the test is introduced. The paper discusses the validity of the list and continues with a comparison of Turkish and American students' mastery levels for specific attributes. Finally, the distribution of specific knowledge states (attribute mastery patterns) and common learning paths derived from these knowledge states for both countries are discussed.

1.1 Introducing the rule space model

The RSM is a probabilistic model for cognitive diagnosis that analyzes whether a given student possesses a list of cognitive skills required to complete a task such as solving a problem. The College Board has used the model to generate scoring reports since October 2001 for the PSAT/NMSQT (Milewski and Baron 2002). The RSM was also applied to several other standardized tests such as the SAT (Tatsuoka 1993; Guerrero 2001), TOEFL (Scott 1998), TOEIC (Buck and Tatsuoka 1998), NAEP Science Assessment (Yepes-Baraya et al. 1998), GRE-Q (Tatsuoka and Gallagher 1998; Tatsuoka and Boodoo 2000), TIMSS and TIMSS-R (Tatsuoka et al. 2004; Xin et al. 2004).

The RSM analysis is conducted in several steps. The first step involves identifying the 'attributes' that the test measures. An attribute is a description of the procedures, skills, processes, strategies, and knowledge a student must possess to solve a test item. Once the attributes are specified, an incidence matrix, Q, is constructed. This matrix displays the relationship between the specified attributes and the test items. The rows represent the items, the columns correspond to the attributes, and the entries in each column indicate which attributes are involved in solving the items. Using this Q-matrix, an examinee's pattern of attribute masteries (i.e. her 'Knowledge State') can be inferred given her item response pattern with several steps. First, Boolean algebra is used to generate all possible combinations of attribute mastery patterns and the corresponding binary item response patterns from the Q-matrix. For example, consider the Q-matrix in Table 1. It displays the relationship between three hypothetical test items and two attributes. According to Table 1, Item 1 requires Attributes 1 and 2, while Item 3 includes Attribute 2 only.

Table 1 An example Q-matrix

	Attribute 1	Attribute 2
Item 1	1	1
Item 2	1	0
Item 3	0	1

The RSM assumes that in order for an examinee to successfully answer an item, she has to master all of the attributes that it involves. Possible attribute patterns and the corresponding item response patterns, derived from Table 1, are as follows:

Attribute mastery pattern		Associated item response pattern		
A1	A2	Item 1	Item 2	Item 3
0	0	0	0	0
1	0	0	1	0
0	1	0	0	1
1	1	1	1	1

The item response patterns associated with possible attribute mastery patterns are called 'ideal item response patterns'. After creating the list of ideal item response patterns, the RSM determines which of these patterns are close to a given examinee's observed item response pattern using the criterion of Mahalanobis distance. An acceptable close match between the list and the student's item response pattern determines the examinee's knowledge state (KS) classification. The probability that a given examinee belongs to a 'probable' knowledge state is computed as a function of the square of the Mahalanobis distance (D^2) between the observed and the ideal item response patterns. Using these probabilities as weights, an examinee's attribute mastery probabilities are then computed.

In summary, the following steps describe the process for Rule Space analyses:

1. Identify the attributes underlying the test and construct the Q-matrix.
2. Generate the ideal item response patterns according to the Q-matrix.
3. Compare the examinees' observed item response patterns to the ideal item response patterns and classify examinees into knowledge states.
4. Infer attribute mastery probabilities according to the classification results.

2 Method

2.1 Data

This study uses data from two samples of eight graders who participated in the 1999 TIMSS-R study, which consisted of 2,900 Turkish and 4,411 American students.

2.2 Analysis

First, a set of attributes that explain performance in TIMSS-R was developed (Table 2) using written student protocols, expert solutions and interviews with high school teachers.

These attributes were grouped into three categories: Content (five attributes), Process (nine attributes) and Skill/Item type (nine attributes). With this list, 162 mathematics items were coded in terms of attribute involvement, which results in the Q-matrix (Corter and

Table 2 Knowledge, skill, and process attributes derived to explain performance on mathematics items from the TIMSS-R (1999) for population 2 (grade 8)

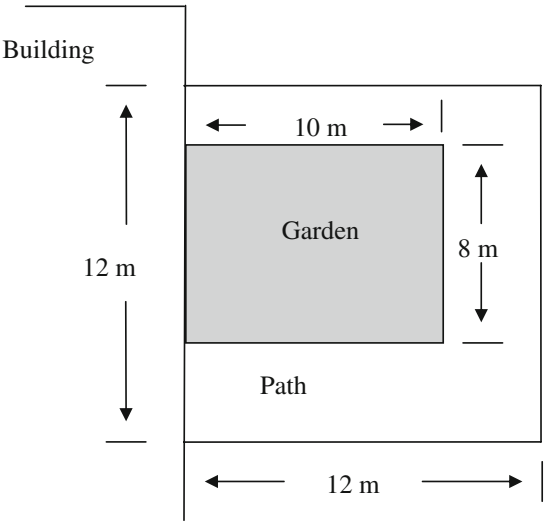
Attributes	
Content attributes	
C1	Basic concepts, properties and operations in whole numbers and integers
C2	Basic concepts, properties and operations in fractions and decimals
C3	Basic concepts, properties and operations in elementary algebra
C4	Basic concepts and properties of two-dimensional Geometry
C5	Data, probability, and basic statistics
Process attributes	
P1	Translate/formulate equations and expressions to solve a problem
P2	Computational applications of knowledge in arithmetic and geometry
P3	Judgmental applications of knowledge in arithmetic and geometry
P4	Applying rules in algebra
P5	Logical reasoning—includes case reasoning, deductive thinking skills, understanding necessary and sufficient conditions
P6	Problem Search; analytic thinking, problem restructuring and inductive thinking
P7	Generating, visualizing and reading figures and graphs
P9	Management of data and procedures in multistep problems
P10	Quantitative reading
Skill (item type) attributes	
S2	Apply number properties and relationships; number sense/number line
S3	Using figures, tables, charts and graphs
S4	Approximation/estimation
S5	Evaluate/verify/check options (for multiple choice items only)
S6	Recognize and extend patterns and relationships
S7	Using proportional reasoning
S8	Solving novel or unfamiliar problems
S10	Solving open-ended items
S11	Understanding language in word problems

Tatsuoka 2002). Figure 1 illustrates a released item from TIMSS-R and how it was coded using the attribute list in Table 2.

In order to see if the attribute content of an item explains its difficulty level, a linear regression analysis was conducted. Item difficulty was the dependent variable and the columns of the Q-matrix (attribute involvement for each item) were the independent variables in this analysis. Next, the Q-matrix and dichotomously scored student response data were used as input for the main RSM analysis. A special software called BUGSHELL (Tatsuoka et al. 1992) was used to perform the RSM analysis.

The RSM analysis yielded attribute mastery probability estimates and KS classification for each individual student. In order to further investigate the validity of the attribute list and the Q-matrix, a linear regression analysis was conducted to see if the attribute mastery probabilities can predict students' total scores (overall performance). Next, the mean mastery level for each attribute was computed for the Turkish sample and compared to those of the American students as determined in Tatsuoka et al. (2004). Finally, five composite attributes, and frequent knowledge states and learning paths derived from these composite attributes are introduced. Relative proportions of students from each country in each KS were computed. The hierarchical structure of the knowledge states revealed common learning paths.

Item J10: A rectangular garden that is next to a building has a path around the other three sides as shown.



What is the area of the path?
A. 144 m^2 B. 64 m^2 C. 44 m^2 D. 16 m^2

Attribute coding:

1. C4: Requires knowledge of rectangles
2. P2: Calculating the area
3. P6: Seeing that the area of path is the difference between areas of two rectangles (restructuring)
4. P9: Solution requires multiple steps
5. S3: Using the figure provided

Fig. 1 A sample TIMSS-R mathematics item with attribute coding

3 Findings

3.1 Q-matrix and its validity

One hundred sixty two items were coded as previously described. In order to assess the validity of the Q-matrix, a linear regression analysis was conducted to see if the columns of the Q-matrix can explain item difficulty. An adjusted R^2 value of 0.869 was obtained, indicating that nearly 87% of the variance in item difficulty levels was due to attribute involvement. RSM analysis was conducted using the Q-matrix and student response patterns on 162 items from four test booklets. Booklets one, three, five and seven of TIMSS-R were selected for these analyses because the other booklets showed an uneven distribution of attributes, i.e., few or no items measuring certain attributes.

A cutoff point of 4.5 on the squared Mahalanobis distance metric was used to define an acceptable close match between the observed and the ideal item response patterns for each knowledge state (see Tatsuoka et al. 2004, for the justification behind this cutoff point). Using this rule, a 99.5% classification rate was obtained for the Turkish sample. In other words, almost all observed item response patterns were classified into one of the logically derived knowledge states from the Q-matrix. The high classification rate also provided evidence for the validity of the attribute list and the Q-matrix.

3.2 Mean attribute mastery profiles of Turkish students

The RSM output provided KS classification and attribute mastery probabilities for each individual student. Mean attribute mastery probability levels for each attribute are displayed in Table 3.

The most difficult attributes for Turkish students were content attributes C3 (elementary algebra) and C5 (probability/statistics); process attribute P4 (applying algebraic rules), and skill/item type attributes S6 (patterns and relationships) and S10 (solving open-ended problems). Tatsuoka et al. (2004) determined the attribute mastery profiles of US students on the same set of attributes. In order to compare the performance of the Turkish and the US samples on these attributes, the standardized mean differences on attribute mastery probabilities were computed (under the z difference column in Table 3). US students outperformed their Turkish peers on 17 of all 23 attributes. The largest differences were

Table 3 Comparison of mean attribute mastery levels between the Turkish and the American samples

Attribute	Turkey		US		*Z* diff[a]
	Mean	SD	Mean	SD	
C1	0.86	0.23	0.93	0.18	−13.84
C2	0.78	0.33	0.86	0.32	−10.26
C3	0.59	0.30	0.76	0.25	−25.28
C4	0.70	0.31	0.68	0.31	2.70
C5	0.60	0.31	0.73	0.27	−18.45
P1	0.92	0.17	0.96	0.13	−10.77
P2	0.83	0.24	0.92	0.16	−17.77
P3	0.87	0.20	0.85	0.17	4.43
P4	0.47	0.34	0.59	0.29	−15.63
P5	0.74	0.25	0.65	0.31	13.67
P6	0.69	0.26	0.83	0.22	−23.91
P7	0.62	0.25	0.77	0.23	−25.90
P9	0.68	0.28	0.68	0.30	0.00
P10	0.58	0.28	0.87	0.20	−48.26
S2	0.61	0.30	0.78	0.24	−25.60
S3	0.80	0.26	0.95	0.14	−28.47
S4	0.62	0.27	0.88	0.21	−43.86
S5	0.83	0.28	0.97	0.11	−25.66
S6	0.34	0.39	0.54	0.36	−22.11
S7	0.95	0.15	0.91	0.20	9.75
S8	0.76	0.27	0.84	0.24	−12.94
S10	0.43	0.38	0.61	0.38	−19.81
S11	0.92	0.18	0.88	0.24	8.13

[a] Standardized difference in means

observed on attributes P10 (quantitative reading) and S4 (approximation/estimation), both in favor of US students.

3.3 Frequent knowledge states and learning paths derived from them

In order to explore frequent knowledge states and learning paths derived from them, five composite attributes were created[1]:

1. Algebra (ALG): C3, P4 and S6.
2. Geometry (GEO): C4, S3, P3 and P7.
3. Numbers (NUMB): C1, C2, S2 and S4.
4. Word problems (WORD): S11, P1, P2 and S7.
5. Advanced problem solving and thinking skills (ADV): P5, P6, P9, P10 and S10.

Students were reclassified according to their mastery of these five composite attributes. A student was assumed to have mastered a composite attribute if she has mastered all single attributes involved in it. In this way, a student's pattern of mastery across these five composite attributes determined her KS. Because the number of composite attributes is five, there were 32 (2^5) possible knowledge states. When Turkish and American data were combined, it was observed that nearly 85% of all students were classified into one of eight of these 32 possible knowledge states. Table 4 displays these knowledge states and the corresponding classification rates across the two samples.

As can be inferred from Table 4, these eight knowledge states are hierarchically related. For instance KS 4 is under KS 6, since attributes mastered under KS 4 are also mastered under KS 6 along with an additional attribute. Figure 2 displays the relationships among all eight knowledge states. There are three learning paths that can be derived from Fig. 2. A 'learning path' describes how a student population progresses from total nonmastery to perfect mastery. Learning paths are derived from the hierarchical relationship among the knowledge states. The three learning paths mentioned above are as follows:

Path 1 KS1 → KS2 → KS4 → KS6 → KS7 → KS8
Path 2 KS1 → KS3 → KS4 → KS6 → KS7 → KS8
Path 3 KS1 → KS3 → KS5 → KS6 → KS7 → KS8

The percentage of student on each path can be computed by adding the classification rate for each KS on that path, which was done for US and Turkish samples separately. The results indicate that Turkish students were mostly on Path 3 while US students were mostly on Path 2 (Table 5).

Most of both Turkish and American students tend to learn skills for solving word problems (Path 2 and Path 3) prior to learning any other skill. In contrast, Turkish students tend to learn geometry-related skills first (Path 3) while their American peers tend to learn number skills initially (Path 2). Note that C4 (basic concepts in geometry) was one of the few attributes where the Turkish students outperformed the US sample (Table 3). This occurrence might be explained because geometry-related skills appear earlier on the Turkish students' learning path. For both samples, students master advanced problem solving skills occurred prior to mastering algebra skills. This result implies that in order for the students

[1] BUGSHELL classified each student according to their mastery on each attribute. However, composite attributes were created to reduce the number of knowledge states and hence produce more manageable and interpretable results.

Table 4 Distribution of most frequent knowledge states

KS	Mastered composite attributes	Classification rate, Turkey (%)	Classification rate, US (%)	Classification rate, combined (%)
1	None	25.6	27.2	26.4
2	NUMB	0.5	2.1	1.4
3	WORD	46.5	16.0	30.3
4	NUMB and WORD	2.3	12.0	7.5
5	GEO and WORD	11.9	4.8	8.1
6	WORD and GEO and NUMB	1.6	7.0	4.5
7	All except ALG	0.3	7.8	4.3
8	All	0.2	4.1	2.3

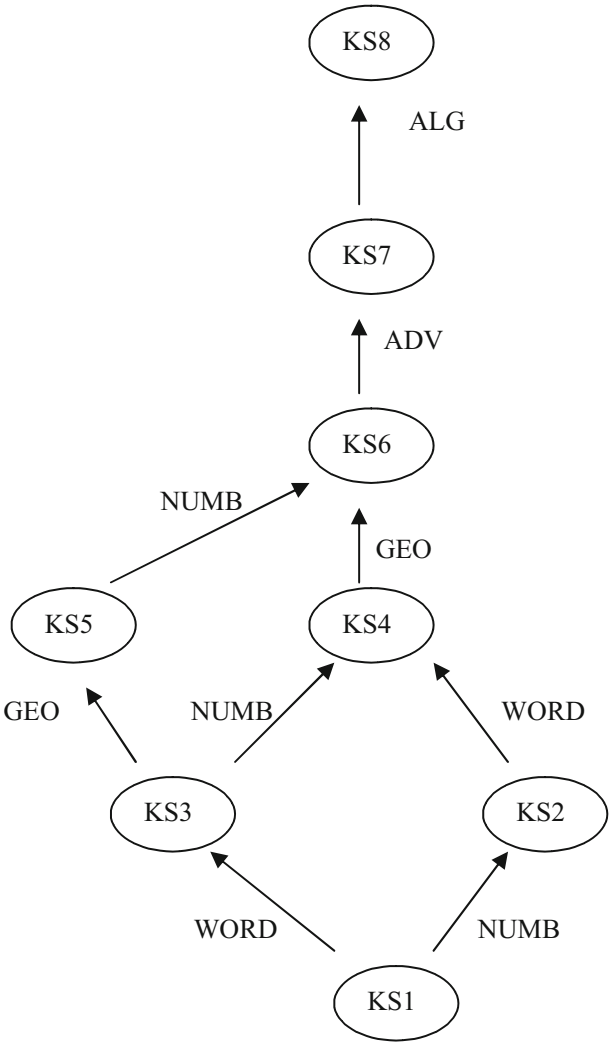


Fig. 2 Hierarchical structure among the knowledge states in Table 4

Table 5 Distribution of learning paths

Path	% of students (Turkey)	% of students (US)	% of students (overall)
1	4.9	33	18.6
2	50.9	46.9	48.6
3	60.5	39.7	49.1

to master algebra-related skills, they need to have mastered several thinking skills first (as reflected in ADV).

Looking at the hierarchical relationship among the attributes, it can be argued that mastery of thinking skills comes before algebra-related skills. Hence, teaching these skills via algebra might not be appropriate for this age group. Common learning paths indicate that learning about numbers and geometry occurs before the mastery of advanced problem solving and thinking skills. Teaching geometry- and number-related skills, consequently, may facilitate the development of these advanced skills, which in turn lead to success in algebra.

4 Discussion

As Van der Linden (1998) argued, international assessments are bound to represent multidimensionality rather than unidimensional knowledge. As a result, national populations should be expected to differ in terms of their performance on different dimensions. "In fact, international assessments are designed just to detect such differences" (Van der Linden 1998, p.574). This study illustrated how a diagnostic testing model can be used to uncover such differences among national populations. By using such a model, it is possible to unfold the multidimensional structure of international assessments such as TIMSS-R and, thereby, reveal the specific strengths and weaknesses of participating countries on existing dimensions.

In this study, eighth grade Turkish students' mathematics performance on TIMSS-R was reanalyzed with the above premise in mind, using the RSM. Students' mastery profiles on a set of 23 attributes were explored. Results indicated that when compared to their American peers, Turkish students were especially weak in mastering attributes P10 (quantitative reading), S4 (approximation/estimation), S6 (patterns and relationships) and S10 (solving open-ended problems). In other words, Turkish students comparatively did not perform well when asked to deal with uncertainty (S4), derive rules and generalize from cases (S6), construct answers as opposed to selecting an answer from given alternatives (S10), and read and understand suggestions that require logical thinking (P10).

Common learning paths of Turkish and American students were also explored by means of investigating frequent knowledge states and their hierarchical structure. Five composite attributes were introduced and students were classified according to their mastery of these composite attributes. The most important finding here was that both American and Turkish students tend to master advanced problem solving and thinking skills after mastering solving routine word problems, numbers and geometry and prior to mastering algebra-related skills. This result implies that teaching these advanced skills in the context of numbers and geometry might be developmentally more appropriate compared to teaching them via algebra for this student population.

As a result, this study illustrated how a diagnostic testing model can be used to achieve a detailed comparative description of performance of student populations. It is hoped that the educators from different countries can benefit from such detailed comparisons and learn from each others' strengths and weaknesses.

Acknowledgments This study has been supported by the National Science Foundation (REC NO. 0126064).

References

Buck, G., & Tatsuoka, 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.

Corter, J. E., & Tatsuoka, K. (2002). *Cognitive and measurement foundations of diagnostic assessment in mathematics* (Technical Report). New York, NY: College Board.

Guerrero, A. (2001). Cognitively diagnostic perspectives on English and Spanish versions of a test of mathematics aptitude. *Dissertation Abstracts International, 62*(08), 543B (UMI no. 3005725).

MEB-EARGED (2003). Üçüncü Uluslararası Fen ve Matematik Çalışması (TIMSS 1999) Ulusal Rapor [Third international mathematics and science study (TIMSS 1999) national report]. Ankara: Turkey.

Milewski, G. B., & Baron, P. A. (2002). Extending DIF methods to inform aggregate reports on cognitive skills. In *1998 National Council on Measurement in Education annual meeting*, New Orleans, LA.

Mullis, V. S., Martin, M. O., Gonzales, E. J., Gregory, K. D., Garden, R. A., O'Conner, K. M., et al. (2000). *TIMSS 1999 international mathematics report: Findings from IEA's repeat of the third international mathematics and science study at the eighth grade*. Chestnut Hill, MA: Boston College.

Scott, H. S. (1998). Cognitive diagnostic perspectives of a second language reading test. *Dissertation Abstracts International, 59*(11), 6113B (UMI no. 9912372).

Tatsuoka, K. K. (1993). *Proficiency scaling based on conditional probability functions for attributes.* (Technical rep. no. RR-93-50-ONR). Princeton, NJ: Educational Testing Service.

Tatsuoka, K. K., & Gallagher, A. (1998). *Variables that are involved in the underlying cognitive processes and knowledge of GRE quantitative* (Technical report). Princeton, NJ: Educational Testing Service.

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

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

Tatsuoka, K. K., Corter, J., & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 countries. *American Educational Research Journal, 11*(4), 901–926.

Van der Linden, W. J. (1998). A discussion of some methodological issues in international assessments. *International Journal of Educational Research, 29*, 569–577.

Xin, T., Xu, Z., & Tatsuoka, K. K. (2004). Linkage between teacher quality, student achievement, and cognitive skills: A rule-space model. *Studies in Educational Evaluation, 30*(3), 205–223.

Yepes-Baraya, M., Tatsuoka, K., Allen, N. L., O'Sullivan, C., Liang, J., & Hui, X. (1998, April). Application of rule space methodology to the 1996 NAEP science assessment: Grade 4 preliminary results. In *1998 National Council on Measurement in Education annual meeting*, San Diego, CA.