DOCUMENT RESUME

ED 372 090 TM 021 752

AUTHOR	Wang, Yu-Chung Lawrence; Hocevar, Dennis
TITLE	Effects of Mathematics Test Content Specificity on Essential Dimensionality in U.S. and Japan Data.
PUB DATE	Apr 94
NOTE	22p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE	Reports - Research/Technical (143) -- Speeche_/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Ability; Comparative Analysis; Elementary Education; Elementary School Students; Estimation (Mathematics); Foreign Countries; *International Studies; Mathematics Achievement; *Mathematics Tests; Scores; *Scoring; Statistical Studies; *Test Content; Test Items; Test Results
IDENTIFIERS	Calibration; DIMTEST (Computer Program); Japan; Second International Mathematics Study; *Specificity; *Unidimensionality (Tests); United States

ABSTRACT

The major goal of this study is to apply the essential unidimensionality statistic of W. Stout and the corresponding computer program (DIMTEST) to a hierarchical level mathematics achievement data set and to determine the extent to which the undimensional assumption can be accurately applied to mathematics achievement data. The study also ascertains if the unidimensionality assumption is more tenable when applied to specific subsets of items than to broader categories of items. A comparison of the essential unidimensionality structure across cultures is also performed. Results indicate that in the Japanese and U.S. data form the Second International Mathematics Study (SIMS), there are several subscales in SIMS mathematics tests, and that individual scores should be calibrated on each subscale rather than on a total score in the SIMS test. Essential unidimensionality estimates for the four tests were not the same in the two countries, calling into question the equivalence of dimensionality of the four tests. Either items on the test are more unidimensional in Japan, or the ability spaces among Japanese students are more homogeneous than for U.S. students. Eleven tables are included. (Contains 10 references.) (Author/SLD)

Effects of Mathematics Test Content Specificity on Essential Dimensionality in U.S. and Japan Data

Yu-Chung Lawrence Wang and Dennis Hocevar

University of Southern California

Paper presented at annual meeting of the American Educational Research Association, April 1994, New Orleans, LA.

The authors wish to thank Ian Westbury for providing access to the SIMS data. This research was partially supported by the Educational Policy Fellowship during the first year of the study and by Chiang Ching-Kuo Research Scholarship in the second year of the study.

Effects of Mathematics Test Content Specificity on Essential Dimensionaiity in U.S. and Japan Data

Abstract

Stout (1987, 1990) has provided a weaker *essential dimensionality* assumption and argued that the IRT model fits when Lord's assumption of unidimensionality is replaced by the assumption of essential dimensionality. The major goal of this study is to apply Stout's essential dimensionality statistic and the corresponding computer program (i.e., DIMTEST) to a hierarchical level mathematics achievement data set, and, based on the result, to determine the extent to which the unidimensional assumption can be accurately applied to mathematics achievement data. The study also ascertains if the unidimensionality assumption is more tenable when applied to specific subsets of items (e.g., arithmetic, algebra, geometry and measurement) rather than broader categories of items (e.g., eighth-grade general mathematic achievement). A comparison of the essential unidimensionality structures across cultures (i.e., Japan vs. U.S.) also is performed.

Results indicate that the assessment of essential dimensionality in the Second International Mathematics Study (SIMS) Japan and U.S. data implies that there are several subscales in SIMS mathematic tests, and that individual scores should be calibrated on each of the mathematics subscales rather than on a total score in the SIMS Test. The essential dimensionality estimates for the four tests in the U.S. and Japan study were not the same. This result questions the equivalence of the dimensionality for the four SIMS tests which share 40 common items and 35 rndomly assigned unique items. According to the results of every possible comparison of the essential dimensionality between the U.S. and Japan, tests in the Japan study tend to be more essentially unidimensional than their U.S. counterparts. This result implies either the items on the test are more unidimensional in Japan than in the U.S., or that the ability spaces among Japanese students are more homogeneous than the U.S. students. Many restrictions in using DIMTEST on real data were encountered and discussed at the end of the study.

Item response theory has been widely used in bias or DIF study across cultures due to the unique invariance property of item parameters. The invariance property of IRT holds only when its two major assumptions, unidimensionality and local dependence, hold. The unidimensionality assumption (Lord, 1980) assumes that every individual taking the test uses the same single cognitive skill to respond to the whole set of items.

Lord's unidimensionality assumption has been criticized as unrealistic and lacking an appropriate statistical test (Traub, 1983). Humphery (1982) warned that a dimensionally narrowed test would weaken the validity of the test. Stout (1987, 1990) has provided a weaker *essential dimensionality* assumption and argued that the IRT model fits when Lord's assumption of unidimensionality is replaced by the assumption of essential dimensionality. The essential dimensionality assumption assumes that multidimensional item characteristics and examinee ability are suitable to unidimensional IRT as long as there is a dominant trait. Stout also provided a statistical test which has been refined by Nandakumar to assess whether or not essential dimensionality holds for a set of items. One should refer to Stout (1987, 1990) and Nandakumar (1993) and Nandakumar and Stout (1993) for a detailed definition of essential dimensionality. Though Stout and his colleagues have done many Monte-Carlo studies on the essential dimensionality measures using simulated data, few investigators have used a real test. This study should fill this gap by using four different SIMS mathematics achievement tests.

The major goal of this study is to apply Stout's essential dimensionality statistic and corresponding computer program (i.e., DIMTEST) to a hierarchical level mathematics achievement data set, and, based on the result, to determine the extent to which the unidimensional assumption can be accurately applied to mathematics achievement data. The study also ascertains if the unidimensionality assumption is more tenable when applied to specific subsets of items (e.g., arithmetic, algebra, geometry and measurement) rather than broader categories of items (e.g., eighth-grade general mathematic achievement). A comparison of the essential unidimensionality structures across cultures (i.e., Japan vs. U.S.) also is performed.

The results of this study have important implications to the area of mathematics achievement testing because contemporary test developers routinely use IRT methods to develop and refine tests. This study concentrates on the unidimensionality assumption since studies have found that violating the unidimensionality assumption produces a substantial lack of item parameter invariance (Ackerman, 1991; Oshima & Miller, 1990).

Methodology

Data.

The data for this study were taken from parts of the Second International Mathematics Study (SIMS) sponsored by the International Association for the Evaluation of Educational Achievement (1985). During 1980 and 1982, the Second IEA International Mathematics Stuᵈʸ researchers collected data on mathematics curricula, teaching practices, and achievement from samples of students, teachers, and schools in 20 countries. SIMS was conducted at two levels: (1) Population A in which students were (typically) in the national grade in which the modal age was 13; and (2) Population B where students were taking the most advanced pre-university mathematics course(s) offered in their school systems. Only the population A data set from the United States and Japan were used in this study.

Subjects.

Population A is defined as all eighth graders of mainstream public and non-public schools. Mentally, physically, emotionally, or learning disabled students who were placed in special education classes were excluded. Stratification variables which were used in the SIMS study were: School type (i.e., public vs. private), regional standard metropolitan statistical area (SMSA), location (i.e., east-central vs. south-west) and metropolitan status (i.e., city, suburb, other or district outside SMSA).

Instruments.

There were two major SIMS study designs: longitudinal and cross-sectional. Both U.S. and Japan were in the longitudinal study design, but in the longitudinal design, Japan used the cross-sectional instrument. For the U.S. instrument, an eighth-grade mathematics achievement test that consisted of a total of 180 items which were selected from the international bank of 196 items divided into a 40-item core subtest and four 35-item "rotated forms" was used. For the Japanese instrument, the first 40 items of the international bank of 196 items were assigned to the core test, the next 34 items were assigned to form A, and so on. It is important to mention here that the test construction strategies for the cross sectional and longitudinal studies were significantly discrepant which introduces a certain degree of nonequivalence to begin with. Table 1 shows the difference between the cross-sectional and longitudinal designs. The 40 core items (different for the U.S. and Japan) were administered to all examinees, and rotated forms were randomly assigned to students (approximately one- fourth of the students taking each form). In other words, each student was administered the core and one rotated form for a total of 74 or 75 of the 196 items in the pool. Tables 2.1 to 2.4. shows the number of items in each content area for both the U.S. and Japan. Subcontent areas with an asterisk (*) were selected for the essential unidimensionality examination reported herein.

Table 1 Test Construction Strategies for Both the SIMS Cross-sectional and Longitudinal Designs.

Cross Sectional Design (Japan)					Longitudinal Design (U.S.)				
Core	A	B	C	D	Core	A	B	C	D
40	34	34	34	34	40	35	35	35	35
The first 40 items of the international bank of 196 items were assigned to the core form. the next 34 items were assigned to form A, and so on. The total number of items in this study design was 176.					Items in each test form (i.e., core form and form A etc.) were selected from an international bank of 196 items and the total number of items in this study design was 180.				

Table 2.1. Content Table of Arithmetic Items for Japan and the U.S.

	Japan				U.S.			
Arith. Subarea	FORM A	FORM B	FORM C	FORM D	FORM A	FORM B	FORM C	FORM D
001	4	4	4	4	3	2	2	3
002*	4	4	3	3	6	6	6	6
003*	4	5	5	5	6	7	6	5
004*	4	3	3	4	11	10	10	11
005	0	1	1	1	0	1	1	1
006	2	1	2	1	1	1	1	0
008	1	2	1	2	1	1	1	0
009	0	0	0	1	0	0	0	1
# of total	19	20	19	22	28	28	27	27

Table 2.2 Content Table of Algebra Items for Japan and the U.S.

	Japan				U.S.			
Algba. subarea	FORM A	FORM B	FORM C	FORM D	FORM A	FORM B	FORM C	FORM D
101*	4	3	4	2	2	3	2	4
102	2	2	1	2	1	0	0	0
103	0	1	1	0	1	1	0	0
104*	3	3	4	4	3	3	4	2
105	1	1	0	1	1	0	1	0
106*	3	4	4	3	5	5	5	6
107	2	1	2	2	1	2	2	1
110	2	1	2	2	0	0	0	1
# of total	17	16	18	16	14	14	14	14

Table 2.3 Content Table for Geometry Items for Japan and the U.S.

	Japan				U.S.			
Geomet. subarea	FORM A	FORM B	FORM C	FORM D	FORM A	FORM B	FORM C	FORM D
201*	3	3	3	3	3	2	2	1
202*	3	2	4	2	3	4	3	4
203	2	3	3	2	1	0	1	1
204	2	2	2	2	1	2	2	2
205	0	1	0	0	1	0	0	0
206	1	2	1	1	0	2	1	0
207*	4	3	4	2	3	2	3	3
208	1	1	1	1	1	2	1	1
209	2	2	1	2	1	1	1	1
212	1	1	1	2	2	1	2	1
215	1	1	0	2	1	1	0	2
# of total	20	21	20	19	17	17	16	16

Table 2.4 Content Table of Measurement Items for Japan and the U.S.

	Japan				U.S.			
Measure. subarea	FORM A	FORM B	FORM C	FORM D	FORM A	FORM B	FORM C	FORM D
401*	2	3	2	2	3	2	3	3
402*	2	2	2	3	3	4	3	5
403	1	0	1	0	2	2	1	1
404*	5	5	4	5	4	4	6	4
# of total	10	10	9	10	12	12	13	13

This achievement test covered five major content areas: arithmetic, algebra, geometry, statistics and measurement. Only statistics items were excluded from the present study due to a small pool of items. There were several subcontent categories under each major content area. For example, there were eight subcontent areas within the arithmetic area: Natural Numbers (001), Common Fractions (002), Decimal Fractions (003), Ratios, Proportions, and Percent (004), Number Theory (005), Power and Exponents (006), Square Roots (008), and Dimensional Analysis (009). As shown in Tables 2.1 to 2.4, three subcontent areas, which contained a sufficient number of items, were selected from each of the four major content areas. They were Common Fractions (002), Decimal Fractions (003) and Ratio, Proportion and Percent (004) in arithmetic; Integers (101), Formulas and Algebraic Expressions (104), and Equations and Inequations (106) in algebra; Classification of Plane Figures (201), Properties of Plane Figures (202), and Coordinates (207) in geometry; and Standard Units (401), Estimation (402), and Determination of measures (404) in measurement. Table 3 displays one sample item for each area` in this study. Omits and "not reaches" were treated as wrong answers. Readers who are interested in the complete content of all SIMS items should refc.' to "Technical Report I" of SIMS (Chang & Ruzicka, 1985).

Table 3. Sample Items with the International Item Code in Twelve SIMS Areas.

Math. Area	Sample Item
Common Fraction. (002)	YS003. 2/5+3/8 is equal to (a) 5/13 (b) 5/40 (c) 6/40 (d) 16/15 (e) 31/40
Decimal Fraction. (003)	YS005. 0.40 x 6.38 is equa' to (a) 0.2552 (b) 2.452 (c) 2.552 (d) 24.52 (e) 25.52
Ratio, Proportion. (004)	YS008. In a school of 800 pupils, 300 are boys. The ratio of the number of boys to the number of girls is (a) 3:8 (b)5:8 (c) 3:11 (d) 5:3 (e) 3:5.
Integer (101)	YS012. (-2) x (-3) is equal to (a) -6 (b) -5 (c) -1 (d) 5 (e) 6
Formulas (104)	YS015. Simplify: 5x + 3y + 2x - 4y (a) 7x + 7y (b) 8x - 2y (c) 6xy (d) 7x - y (e) 7x + y
Equation Inequation. (106)	YS017. If P = LW and if P = 12 and L = 3, then W is equal to (a) 3/4 (b) 3 (c) 4 (d) 12 (e) 36
Classify Panle Figure. (201)	YS021. A quadrilateral MUST be a parallelogram if it has (a) one pair of adjacent sides equal (b) one pair of parallel sides (c) a diagonal as axis of symmetry (d) two adjanct angles equal (e) two pairs of parallel sides.
Properties of Panel Figures (202)	YS023*. The length of the circumference of the circle with center O is 24, and the length of arc RS is 4. What is the measures in degrees of the central angle ROS ? (a) 24 (b) 30 (c) 45 (d) 60 (e) 90
Coordinates (207)	YS028*. What are the coordinates of point P? (a) (-3, 4) (b) (-4, -3) (c) (3, 4) (d) (4, -3) (e) (-4, 3)
Standard Unit (401)	YS036. Which of the following is the most likely to be nearest to the weight of a normal man ? (a) 8.5 kg (b) 85 kg (c) 185 kg (d) 850 kg (e) 1850 kg
Estimation (402)	YS038*. On the above scale the reading indicated by the arrow is between (a) 51 and 52 (b) 57 and 58 (c) 60 and 62 (d) 62 and 64 (e) 64 and 66
Determination of measures (404)	YS037*. The total area of the two triangles is (a) 6X8 cm^2 (b) 6X8 / 2 cm^2 (c) 10X6 / 2 cm^2 (d) 16X12 / 2 cm^2 (e) 20X12 / 2 cm^2

Note. * denotes the accompanying figure for an item was omitted in this table.

Four test forms of eighth grade mathematic tests (form A, B, C and D) were investigated in both the U.S. and Japan study and were labled as test 1 to test 4 for U.S. tests A, B, C and D and 5 to 8 for Japan tests A, B, C and D, respectively (see Table 4). As mentioned earlier, every SIMS mathematics achievement test covers five mathematic areas: arithmetic, algebra, geometry, measurement, and statistics. But statistic subtests were excluded due to their small number of items and low reliability. Every subtest was relabeled for this study and these lables are displayed in Table 5. The first number of the subtest denotes the resource of the subtest, and the extension number 1, 2, 3, and 4 denotes arithmetics, algebra, geometry, and measurement.

The reliability coefficients of all tests for both countries were computed and displayed in Table 6. The standardized coefficient α ranged from .89 to .65 in the U.S. study and from .83 to .64 in the Japan study. Measurement tests were the least reliable among the four SIMS areas. Three measurement content tests out of a total of twelve tests had reliabilities lower than 0.70 in the U.S. study and one measurement test out of twelve tests had a reliability lower than 0.70 in Japan.

Table 4 Labels of 8 SIMS Tests.

Test	Items	N	Description
1	75	1652	U.S. test.A
2	75	1610	U.S. test B
3	75	1668	U.S. test C
4	75	1619	U.S. test D
5	74	1986	Japan test A
6	74	1982	Japan test B
7	74	1965	Japan test C
8	74	1851	Japan test D

Table 5. Labels of Subtests in SIMS Study.

Test	Items	N	Description
U.S. tests			
1.1	28	1652	Subtest of test 1 (Arithmetic)
1.2	14	1652	Subtest of test 1 (Algebra)
1.3	17	1652	Subtest of test 1 (Geometry)
1.4	12	1652	Subtest of test 1 (Measurement)
2.1	28	1610	Subtest of test 2 (Arithmetic)
2.2	14	1610	Subtest of test 2 (Algebra)
2.3	17	1610	Subtest of test 2 (Geometry)
2.4	12	1610	Subtest of test 2 (Measurement)
3.1	27	1668	Subtest of test 3 (Arithmetic)
3.2	14	1668	Subtest of test 3 (Algebra)
3.3	16	1668	Subtest of test 3 (Geometry)
3.4	13	1668	Subtest of test 3 (Measurement)
4.1	27	1619	Subtest of test 4 (Arithmetic)
4.2	14	1619	Subtest of test 4 (Algebra)
4.3	16	1619	Subtest of test 4 (Geometry)
4.4	13	1619	Subtest of test 4 (Measurement)
Japan tests			
5.1	19	1986	Subtest of test 5 (Arithmetic)
5.2	17	1986	Subtest of test 5 (Algebra)
5.3	20	1986	Subtest of test 5 (Geometry)
5.4	10	1986	Subtest of test 5 (Measurement)
6.1	20	1982	Subtest of test 6 (Arithmetic)
6.2	16	1982	Subtest of test 6 (Algebra)
6.3	21	1982	Subtest of test 6 (Geometry)
6.4	10	1982	Subtest of test 6 (Measurement)
7.1	19	1965	Subtest of test 7 (Arithmetic)
7.2	18	1965	Subtest of test 7 (Algebra)
7.3	20	1965	Subtest of test 7 (Geometry)
7.4	9	1965	Subtest of test 7 (Measurement)
8.1	22	1851	Subtest of test 8 (Arithmetic)
8.2	16	1851	Subtest of test 8 (Algebra)
8.3	19	1851	Subtest of test 8 (Geometry)
8.4	10	1851	Subtest of test 8 (Measurement)

Table 6. The Reliability Coefficients of the Four Major Mathematical Scales in Both the U.S. and Japan Data in SIMS

	Test Form	FORM A		FORM B		FORM C		FORM D	
	Test Content	KR-20	ST.α	KR-20	ST α	KR-20	ST α	KR-20	ST.α
	Arithmetic	.88	.88	.88	.88	.89	.89	.86	.86
U. S.	Algebra	.80	.80	.78	.78	.77	.78	.73	.73
	Geometry	.72	.72	.75	.74	.74	.74	.75	.75
	Measure	.68	.67	.68	.66	.66	.65	.72	.71
	Arithmetic	.79	.79	.77	.76	.78	.78	.78	.78
Japan	Algebra	.82	.82	.77	.78	.81	.82	.83	.83
	Geometry	.81	.81	.75	.74	.76	.75	.74	.74
	Measure	.74	.74	.72	.74	.64	.65	.71	.72

Assessing Essential Dimensionality.

The essential test of unidimensionality (Stout, 1987, 1990), which is available in a computer program DIMTEST (Stout, Douglas, Junker, & Roussos, 1992), was applied in the present study. A brief summary reference of the steps of Stout's procedure for assessing unidimensionality is described below:

1. The resource test items are divided into three subtests, that is, assessment subtest 1 and 2 and a partitioning subtest. The items in the AT1 subtest (i.e., the assessment subtest 1) are selected to be as unidimensional as possible and to be dimensionality distinct from the remaining items. Selecting AT1 can be done by expert opinion or by using factor analysis to choose the items with the highest same-sign loadings on the second extracted factor. Items in assessment subtest 2 or AT2 are selected by the DIMTEST computer program from the rest of the test so that their difficulty level is similar to the AT1 items. (The function of AT2 is to correct for pre-asymptotic statistical bias in Stout's statistic T.) Items in the PT subtest (i.e., partitioning subtest which are the remaining items after selecting AT1 and AT2 items) are used for the purpose of grouping examinees based on their PT score.

2. Examinees are assigned to k different subgroups according to their PT score. Examinees who answer all PT items correctly or incorrectly are excluded. A PT subgroup with too few examinees (less than 5 in this study) is deleted.

3. The variance estimate $\delta^{\wedge}{}_{k}{}^{2}$ for each PT subgroup on AT1 is computed.

4. The unidimensional variance estimate $\delta^{\wedge}{}_{ud,k}{}^{2}$ for each PT subgroup on AT1 is computed. See Nandakumar (1993) for computational formulas.

5. The two variance estimates are used to obtain the T_L statistic:

$$T_L = \frac{1}{(k)^{1/2}} \sum_{k=1}^{k} \left(\frac{\hat{\delta}_k^2 - \hat{\delta}_{ud,k}^2}{S_k} \right) \qquad [1]$$

where S_k is the standard error of estimate.

6. Compute a similar statistic T_B on AT2.

7. The essential dimensionality statistics are performed with a null hypothesis that the degree of essential unidimensionality of the whole test is equal to 1 (i.e., the test is undimensional) in contrast to the alternative hypothesis which assumes the degree of the essential dimensionality is greater than 1 (i.e., the test is not unidimensional). Stout's unidimensionality test statistic T is given by

$$T = \frac{(T_L - T_B)}{\sqrt{2}} \qquad [2]$$

The null hypothesis, which assumes unidimensionality (or H_0: $d_E=1$), is rejected when T is statistically significant greater than an upper percentile of the standard normal distribution.

In this study, the essential dimensionality of eight SIMS test forms (test 1 to test 8) were assessed four times by treating all items within the same major content area (i.e., arithmetic, algebra, geometry and measurement) as AT1 items. AT2 and PT items were selected from the remaining three major content areas. In other words, the degree of the essential dimensionality for a single test was estimated four times using four different sets of AT1 items. For example, to test the essential dimensionality of test 1 using 28 arithmetic items (or test 1.1) as AT1 items, the remaining algebra, geometry, statistic and measurement items were treated as AT2 and PT items (a total of 47 items). Similarly, the essential dimensionality of test 1 can be assessed using 14 algebra items (or test 1.2) as AT1 items and the AT2 and PT items were selected from the remaining items.

Along the same lines, in this study, the essential dimensionality of four mathematic contents (i.e., arithmetic, algebra, geometry, and measurement) were assessed using three possible homogenous AT1 items in a particular content. For example, in this study, to assess the essential dimensionality of arithmetic in test 1 (or test 1.1), three possible groups of AT1 items (i.e., 3 common fraction, 6 decimal fraction and 6 ratio proportion items) were used separately as AT1 item pool and the remaining arithmetic items in test 1 were treated as AT2 and PT items. The effects of the size of AT1 items also was investigated in the study.

DIMTEST can be run by a user friendly interactive subprogram called irtgo. Users should type *irtgo* under the directory containing DIMTEST and specify all the parameters correctly. Two essential dimensionality statistics are printed on the screen and saved in the last section of the output file. Readers should refer to the manual of DIMTEST for detailed information on running DIMTEST.

Results

The top section of Table 7 shows the summary results for Stout's essential dimensionality statistics provided by the computer program DIMTEST for four SIMS test forms across the U.S. and Japan datasets. It was found that the degree of the essential dimensionality of four SIMS test forms (1, 2, 3, and 4) vary when different AT1 items were used. For instance, in the U.S. data, test 1 was identified as essentially unidimensional when arithmetic and measurement items were used as AT1 items but it was identified as essentially multidimensional when algebra and geometry items were used as AT1. This association between the degree of the essential dimensionality statistics and the characteristics of AT1 items was consistently found in three other tests (2, 3, and 4). For example, test 2 was flagged as essentially unidimensional when arithmetic or geometry items were used as AT1, but flagged as multidimensional when algebra and measurement items were the AT1 items. This table also shows that none of the SIMS test was identified as essentially unidimensional across four different AT1 situations. In conclusion, SIMS tests 1, 2, 3, and 4 held the essential dimensionality assumption 2, 2, 2, 3 times, each out of a total of four trials, respectively in the U.S. data.

A similar pattern of essential dimensionality for the Japan data was found and is displayed in the bottom section of Table 7. Specifically, an association between the degree of the essential dimensionality and the choice of AT1 items was found, and again, none of the four SIMS tests in the Japan study were consistently detected as unidimensional across four AT1 cases. However, the SIMS Japan tests presented a slightly higher degree of essential unidimensionality than the U.S. tests. That is, for all possible pairs of comparisons, the degree of the essential dimensionality for the Japan tests tended to be higher than in the U.S. tests. For instance, for three tests (5, 6, and 8) using arithmetic AT1 items, Japanese data showed better essential dimensionality results than the U.S. data.

It is also noteworthy that the arithmetic items in test 3 in the U. S. data, and the measurement items in test 5 and 6 in the Japan data were found to be too easy and not appropriate as AT1 items. Also, it was found that SIMS tests were more likely to be identified as multidimensional when the AT1 items were algebra items in both samples. This result may imply that the dimensionality structure of the algebra items is significantly different from the other three mathematic areas. Finally, all eight SIMS tests were identified as multidimensional, at least once, in this analysis.

The four SIMS U.S. general tests which have 40 common items and 35 randomly assigned unique items did not have the same essential dimensionality estimates. For instance, test 1 was flagged as moderately unidimensional with arithmetic AT1 items while test 2 and test 4 were identified as boardline unidimensional. Similar inconsistencies were found in the Japan data.

Table 8 provides a summary of Stout's essential dimensionality statistics for the four different arithmetic tests using three arithmetic subcontents as AT1 in both countries. The top section of this table shows first, the U.S. natural number items in tests 1.1, 2.1 and 3.1 are not appropriate as AT1 items and second, arithmetic tests are flagged as essentially unidimensional in eight out of nine trails. This result indicates that the degree of essential dimensionality improved significantly when the arithmetic subcontent area was considered as a test rather than the whole 75-item mathematics achievement test.

The bottom section of Table 8 shows Stout's essential dimensionality statistics for the four arithmetic tests in the Japan study. Four subcontent areas in arithmetic were used as AT1. Similar to the U.S. data, the degree of the essential unidimensionality for a content specific test is better than a general test. For instance, only two of fourteen arithmetic tests (tests 5.1 and 7.1), violated the essential unidimensionality assumption in the Japan study.

Another finding from this analysis is that the essential dimensionality structure of the tests are not the same for the U.S. and Japan. This result may imply that the cognitive ability of the two groups is different, which results in the discrepancy in the interaction between respondents and SIMS items across the two distinctive cultures. However, since the two studies used only partially common items, the discrepancy of the essential dimensionality structure is confounded. Particularly because of the initial analyses (Table 7), the authors suspect the stability of Stout's essential dimensionality statistics. In other words, the discrepancies of the degree of the essential dimensionality across different test forms and cultures may reflect the fact that Stout's two essential dimensionality statistics are not reliable. Further study is needed to investigate the reliability of the essential dimensionality statistics and the validity of replacing Lord's unidimensionality assumption with Stout's essential dimensionality assumption.

The ratio proportion items in the U.S. were not used here due to their inappropriately large item numbers. Stout (1992) suggested the best range for the size of the AT1 items is greater than one fourth but less than one third of total items for the best essential dimensionality estimates. There were 11, 10, 10, 11 ratio items in forms A, B, C, D respectively out of a total of 28, 28, 27, 27 arithmetic items which violated Stout's rule. As a result, the use of ratio proportion items for the essential dimensionality tests were skipped in the U.S. data.

Similar results were found in algebra, geometry and measurement using three different AT1 assignments within each content area. Results are presented in Table 9, Table 10 and Table 11, respectively. In Table 9, only test 2.2 using integer AT1 items for the U.S. data and test 5.2 and 8.2 using equation AT1 items for the Japan data indicated a lack of essential dimensionality. This result indicates that the degree of the essential unidimensionality for the algebra content improves significantly in comparison to the four achievement tests taken as a whole (i.e., Table 7). The essential dimensionality statistics for many geometry and measurement tests were not calculated due to an inappropriate number of AT1 items. However, only one geometry test (4.3) in the U.S. and one in the Japan data (6.3) were found to be essentially multidimensional (notably both used "coordinates" as AT1 items). The essential dimensionality statistics for many geometry and measurement conte..ts were not calculated due to an inappropriate number of AT1 items which is one restriction in applying DIMTEST to real-life data. Nevertheless, the trend is for better support for essential dimensionality when contents rather than the whole tests are analyzed.

Table 7. Essential Dimensionaiity Statistics for 8 SIMS General Tests.

	Target Test #AT1/#Total	AT1	Stout's T	P-value	Refined T	P-value
	1 (28/75)		.73	.23	.60	.27
	2 (28/75)	Arithmetic	1.36	.09	1.35	.08
	3 (27/75)		AT 1 items failed the difficulty test			
	4 (27/75)		1.50	.07	1.48	.07
	1 (14/75)		3.75	.00**	4.22	.00**
	2 (14/75)	Algebra	4.54	.00**	4.96	.00**
	3 (14/75)		4.58	.00**	5.09	.00**
U.S.	4 (14/75)		1.01	.16	1.16	.12
	1 (17/75)		2.15	.02*	2.32	.01*
	2 (17/75)	Geometry	1.58	.06	1.60	.06
	3 (16/75)		1.50	.07	1.53	.06
	4 (16/75)		3.00	.00**	3.41	.00**
	1 (12/75)		-.42	.66	-.50	.69
	2 (12/75)	Measure.	1.92	.03*	2.19	.01*
	3 (13/75)		.35	.36	.46	.32
	4 (13/75)		.18	.43	.18	.43
	5 (19/74)		.20	.42	.17	.43
	6 (20/74)	Arithmetic	-.59	.72	-.67	.75
	7 (19/74)		2.28	.01*	2.58	.00**
	8 (21/74)		.81	.20	.99	.16
	5 (17/74)		1.19	.12	1.11	.13
	6 (16/74)	Algebra	.78	.22	.73	.23
	7 (18/74)		1.54	.06	1.70	.04*
	8 (16/74)		2.36	.01**	2.55	.01**
Japan						
	5 (20/74)		1.90	.03*	2.06	.02*
	6 (21/74)	Geometry	.13	.45	.08	.47
	7 (20/74)		.79	.22	.90	.18
	8 (19/74)		1.28	.12	1.24	.11
	5 (10/74)		AT 1 items failed difficulty test.			
	6 (10/74)	Measure.	AT 1 items failed difficulty test.			
	7 (09/74)		1.03	.15	1.19	.12
	8 (10/74)		1.90	.03*	2.28	.01*

Table 8. Essential Dimensionality Statistics for 8 Arithmetics Tests

	Target Test #AT1/#total		AT1	Stout's T	P-value	Refined T	P-value
	1.1	(3/28)		AT 1 items fail the difficulty test.			
	2.1	(2/28)	Natural	Too few AT1 items for DIMTEST.			
	3.1	(2/27)	Numbers	Too few AT1 items for DIMTEST.			
	4.1	(3/27)		-.26	.60	-.35	.63
	1.1	(6/28)		.53	.30	.60	.27
U. S.	2.1	(6/28)	Common	.06	.48	.06	.48
	3.1	(6/27)	Fractions	1.36	.09	1.64	.05*
	4.1	(6/27)		-.11	.55	-.19	.58
	1.1	(6/28)		.78	.22	.83	.20
	2.1	(7/28)	Decimal	.44	.33	.56	.29
	3.1	(6/27)	Fractions	-.19	.58	-.25	.60
	4.1	(5/27)		1.34	.09	1.63	.05
	5.1	(4/19)		AT 1 items failed the difficulty test			
	6.1	(4/20)	Natural	-.29	.61	-.31	.62
	7.1	(4/19)	Numbers	-1.17	.88	-1.31	.90
	8.1	(4/21)		-1.83	.97	-2.26	.99
	5.1	(4/19)		.50	.31	.52	.30
	6.1	(4/20)	Common	1.14	.13	1.36	.09
	7.1	(3/19)	Fractions	AT 1 items failed the difficulty test			
	8.1	(3/21)		1.19	.12	1.54	.06
Japan							
	5.1	(4/19)		-.68	.75	-.78	.78
	6.1	(5/20)	Decimal	-.66	.75	-.83	.80
	7.1	(5/19)	Fractions	.11	.46	.11	.46
	8.1	(5/21)		1.41	.08	1.59	.06
	5.1	(4/19)		1.87	.03*	2.09	.02*
	6.1	(3/20)	Ratio	.22	.41	.28	.39
	7.1	(3/19)	Proportions	1.54	.06	1.94	.03*
	8.1	(4/21)		.94	.17	.11	.14

Table 9. Essential Dimensionality Statistics for 8 Algebra Tests

	Target Test #AT1/#Total	AT1	Stout's T	P-value	Refined T	P-value
	1.2 (3/14)		.76	.22	.96	.17
	2.2 (3/14)	Integers	1.82	.03*	2.18	.01*
	3.2 (3/14)		.30	.38	.35	.36
	4.2 (4/14)		.96	.17	1.00	.16
	1.2 (3/14)		.29	.39	.37	.36
U.S.	2.2 (3/14)	Formulas	-.01	.50	-.07	.53
	3.2 (4/14)		.16	.44	.14	.44
	4.2 (3/14)		-.49	.69	-.69	.75
	1.2 (5/14)		-.83	.80	-.76	.78
	2.2 (5/14)	Equations	-.16	.56	-.11	.54
	3.2 (5/14)		.38	.35	.44	.33
	4.2 (6/14)		-.78	.78	-.65	.74
	5.2 (4/17)		-.21	.58	-.38	.65
	6.2 (3/16)	Integer	-1.90	.97	-2.36	.99
	7.2 (4/18)		-1.60	.95	-1.93	.97
	8.2 (2/16)		Too few AT1 items for DIMTEST.			
	5.2 (3/17)		-2.42	.99	-2.92	.99
Japan	6.2 (3/16)	Formulas	.25	.40	.31	.38
	7.2 (4/18)		.55	.29	.72	.23
	8.2 (4/16)		.58	.28	.63	.26
	5.2 (3/17)		2.81	.00**	3.47	.00**
	6.2 (4/16)	Equations	-.83	.80	-.91	.82
	7.2 (4/18)		.07	.47	.11	.46
	8.2 (3/16)		1.89	.03*	2.30	.01*

Table 10. Essential Dimensionality Statistics for 8 Geometry Tests

	Target Test #AT1/#Total	AT1	Stout's T	P-value	Refined T	P-value
U.S.	1.3 (3/17)	Classificat. of Plane Figure	1.06	.14	1.09	.14
	2.3 (2/17)		Too few AT1 items for DIMTEST.			
	3.3 (2/16)		Too few AT1 items for DIMTEST.			
	4.3 (1/16)		Too few AT1 items for DIMTEST.			
	1.3 (3/17)	Properties of Plane Figure	-.85	.80	-1.17	.88
	2.3 (4/17)		-.80	.79	-.93	.82
	3.3 (3/16)		.38	.35	.53	.30
	4.3 (4/16)		.37	.35	.45	.32
	1.3 (3/17)	Coordinates	1.29	.10	1.60	.05
	2.3 (2/17)		Too few AT1 items for DIMTEST.			
	3.3 (3/16)		.13	.45	.23	.41
	4.3 (3/16)		2.16	.02*	2.50	.01**
Japan	5.3 (3/20)	Classificat. of Plane Figure	-1.25	.89	-1.60	.95
	6.3 (3/21)		.16	.44	.26	.40
	7.3 (3/20)		-.46	.68	-.60	.73
	8.3 (3/19)		-.68	.75	-.89	.81
	5.3 (3/20)	Properties of Plane Figure	.55	.29	.64	.26
	6.3 (2/21)		Too few AT1 items for DIMTEST.			
	7.3 (4/20)		-.04	.52	-.01	.50
	8.3 (2/19)		Too few AT1 items for DIMTEST.			
	5.3 (4/20)	Coordinates	.92	.18	1.12	.13
	6.3 (3/21)		3.05	.00**	3.87	.00**
	7.3 (4/20)		-.63	.74	-.74	.77
	8.3 (2/19)		Too few AT1 items for DIMTEST.			

Table 11. Essential Dimensionality Statistics for 8 Measurement Tests

	Target Test #AT1/#Total	AT1	Stout's T	P-value	Refined T	P-value
	1.4 (3/12)		2.39	.01**	2.97	.00**
	2.4 (2/12)	Standard	Too few AT1 items for DIMTEST.			
	3.4 (3/13)	Units	1.09	.14	1.22	.11
	4.4 (3/13)		2.01	.02*	2.49	.01**
	1.4 (3/12)		-.49	.69	-.54	.70
U.S.	2.4 (4/12)	Estimations	-.59	.72	-.75	.77
	3.4 (3/13)		AT1 items failed the difficulty test.			
	4.4 (5/13)		Too many AT 1 items for DIMTEST			
	1.4 (4/12)		-2.32	.99	-2.52	.99
	2.4 (4/12)	Determinat	-1.80	.96	-1.99	.98
	3.4 (6/13)	of Measure	Too many AT 1 items for DIMTEST			
	4.4 (4/13)		-1.21	.89	-1.28	.90
	5.4 (2/10)		Too few AT1 items for DIMTEST.			
	6.4 (3/10)	Standard	-1.94	.97	-2.42	.99
	7.4 (2/09)	Units	Too few AT1 items for DIMTEST.			
	8.4 (2/10)		Too few AT1 items for DIMTEST.			
	5.4 (2/10)		Too few AT1 items for DIMTEST.			
Japan	6.4 (2/10)	Estimations	Too few AT1 items for DIMTEST.			
	7.4 (2/09)		Too few AT1 items for DIMTEST.			
	8.4 (3/10)		.40	.34	.37	.35
	5.4 (5/10)		Too many AT 1 items for DIMTEST			
	6.4 (5/10)	Determinat	Too many AT 1 items for DIMTEST			
	7.4 (4/09)	of Measure	Too many AT 1 items for DIMTEST			
	8.4 (5/10)		Too many AT 1 items for DIMTEST			

Conclusions and Discussion

The assessment of essential dimensionality in the SIMS Japan and U.S. data implies that there are several subscales in SIMS mathematic tests, and that individual scores should be calibrated on each of the mathematical subscales rather than on a total score in the Second International Mathematical Achievement Test. In other words, scores reported separately based on subscales (such as arithmetic, algebra, geometry, and measurement) are more appropriate than a single general scale (such as, general eighth grade mathematic achievement). Furthermore, when unidimensional IRT is applied to calibrate items, items within the same content area should be calibrated on the same scale. More importantly in reference to the use of IRT, it is readily apparent that the routine use of unidimensional IRT methods to calibrate mathematics achievement items deserves further scrutiny. On the type of tests to which IRT methods are sometimes applied (e.g. a general eighth grade mathematics achievement test), a substantial lack of the unidimensionality assumption was uncovered in the present analysis.

Secondly, in the analysis of the whole test, it has been found that the degree of the essential dimensionality depended on the choice of AT1 items and on the particular form selected. Also, effects of the size ot the AT1 item pool as well as the sample size might be considered.

The essential dimensionality estimates for the four general tests in each study (U.S. and Japan) were not the same. This result questions the equivalence of the dimensionality for the four U.S. tests and the four Japan tests. In other words, comparing scores across four U.S. tests and four Japan tests may be inappropriate when the dimensionality of the tests varies significantly.

According to the results of every possible comparison of the essential dimensionality between the U.S. and Japan, tests in the Japan study tend to be more essentially unidimensional than their U.S. counterparts. This result implies either the items on the test are more unidimensional in Japan than in the U.S., or that the ability spaces among Japanese students are more homogeneous than the U.S. students. However, one limitation to these conclusions is that the U.S. and Japan tests were not identical to begin with.

Many restrictions in using DIMTEST on real data were encountered. The first restriction is related to the unclear definition of AT1 items. According to the analyses in this study, the essential dimensionality estimates are highly associated with the selection of AT1. Stout has suggested the DIMTEST users select AT1 items to be as dimensionally homogeneous as possible and to be as dimensionally distinct from other items as possible. Hence, the degree of essential dimensionality for a specific group of items may vary when the dimensionality of the AT1 items changes. For example, in the present study, the degree of the essential dimensionality for a particular test was found to be different when the AT1 items were changed. Table 7 shows test 1 in the U.S. as essentially unidimensional when arithmetic items were treated as the AT1 items. When the algebra or geometry items were AT1, the same test was flagged as multidimensional. This discrepancy may have resulted from the variation of the degree of the essential dimensionality across the four SIMS mathematical contents.

The second restriction in using DIMTEST on a real test is related to the requirement on the size and difficulty of AT1 items. Stout suggested the most appropriate AT1 size is one fourth of the total items which is fairly hard to satisfy in a real achievement test with less than three subcategories or with a subcategory having very few items. An equal distribution of items across three subcategories of a test may induce the problem of having too many AT1 items if items within the same subcategory are selected as AT1 items. DIMTEST cannot be performed in the situation above.

The last defect of DIMTEST results are caused by its sample dependent characteristic. The essential dimensionality statistics themselves, as discussed earlier, measure the interaction between a group of subjects and items using original item responses for the analysis. The degree of essential dimensionality therefore may change when the degree of the homogeneity of the respondents' cognitive ability space changes. Hence, to validly generalize the result of the essential dimensionality for a test across samples, the homogeneity of the cognitive space across groups of examinees needs to be confirmed.

References

Ackerman, T. A. (1991). The use of unidimensional item parameter estimstes of multidimensional items in adaptive testing. Applied Psychological Measurement, 15, 13-24.

Chang, L. C., & Ruzicka, J. (1985). Second International Mathematics Study: United States Technical report l. University of Oklahoma, National Research Coordinator.

Humphrey, L. G. (1982). Systematic heterogeneity of items in tests of meaningful and important psychological attributes: A rejection of unidimensionality. Unpublished manuscript, University of Illinois.

Nandakumar, R (1993). Assessing essential unidimensionality of real data. Applied Psychological Measurement, 17, 29-38.

Nandakumar, R. & Stout, W. F. (1993). Refinment of Stout's procedure for assessing latent trait dimensionality. Journal of Educational Statistics, 18, 41-68.

Oshima, T. C., & Miller, M. D. (1990). Multidimensionality and IRT-based item invariance indexes: The effect of between-group variation in trait correlation. Journal of Educational Measurement, 27, 273-283.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. Psychometrika, 52, 589-617.

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. Psychometrika, 55, 293-325.

Stout, W., Douglas, J., Junker, B., Roussos, L. (1992). DIMTEST [Computer Program]. Champaign, IL: Department of Statistics, University of Illinois-Urbana-Champaign.

Traub, R. E. (1983) A prior considerations in choosing a item response model. In R. K. Hambleton,. (Ed.). Applications of item response theory (pp. 57-70). Vancouver, BC: Educational Research Institute of British Columbia.