International Journal of Psychological Studies; Vol. 4, No. 3; 2012
ISSN 1918-7211 E-ISSN 1918-722X
Published by Canadian Center of Science and Education

Challenges of the Fennema-Sherman Test in the International Comparisons

Jari Metsämuuronen[1]

[1]Faculty of Behavioral Sciences, Helsinki University, Finland

Correspondence: Jari Metsämuuronen, Faculty of Behavioral Sciences, Helsinki University, Siltavuorenpenger 5 A, P.O.B. 9, FIN-00014, Finland. E-mail: Jari.Metsamuuronen@methelp.com

Received: May 3, 2012 Accepted: May 21, 2012 Online Published: July 16, 2012

doi:10.5539/ijps.v4n3p1 URL: http://dx.doi.org/10.5539/ijps.v4n3p1

Abstract

The shortened version of Fennema-Sherman test is used to measure attitudes toward mathematics in several international testing settings like TIMSS and PISA. On the basis of Classical Item Analysis and Confirmatory Factor Analysis in the different achievement levels and sets of countries, it is suggested that there are two items on the Fennama-Sherman test which should be discarded due to cultural and achievement considerations. Items "Mathematics is more difficult for me than for many of my classmates" and "Mathematics is not one of my strengths" are too complicated for test takers who belong to the lowest quartile of achievement. These items also seem to carry culturally sensitive elements especially in East Asian countries where fairly good students answer illogically due to such negative wordings. Alternative possibilities for test items are recommended.

Keywords: student attitudes, item analysis, test reliability, confirmatory factor analysis, mathematics achievement, international testing, international assessment, international comparison

1. Introduction

The correlation between mathematics achievement and attitudes toward mathematics is widely studied (e.g., House & Telese, 2008; Shen& Tam, 2008; Kadijevich, 2006; 2008). Some researchers have noticed remarkable differences in correlation between countries (e.g., House &Telese, 2008; Kadijevich, 2006; 2008; Wilkins, 2004; Shen, 2002; Papanastasious, 2000; 2002; Stevenson, 1998): in some countries, the correlation between attitudes and achievement may can be near zero, like in Macedonia (Kadijevich, 2008), Philippines (Wilkins, 2004), Indonesia (Shen, 2002) or in Moldova (Shen, 2002) whereas in some other countries, the correlation can be as high as 0.60 (e.g., in Korea, Shen, 2002).

A shortened version of Fennema–Sherman Mathematics Attitude Scales (Fennema & Sherman, 1976) are used in several international comparisons, like in Trends in International Mathematics and Science Study 2007 (TIMSS, Mullis, Martin, & Foy, 2008) and its predecessors 1995, 1999, and 2003 as well as in Programme for International Student Assessment (PISA). Original scales include nine dimensions but in these international comparisons only three dimensions with four items in each (see Table 1) and two negative items in each of the first two dimensions are used. The names of the factors can be "Liking Math", "Self-concept in Math", and "Experiencing utility in Math" (compare naming in, e.g., Kadijevich, 2006; 2008). This kind of "Expected factor structure" can be found in all Western countries including European countries (except in Bulgaria and Romania), Australia, Canada, Israel, the United States and Russia. However, after performing exploratory factor analysis (EFA) with Principal Axis Factoring, 3 factors, and Promax rotation with Kaiser Normalization separately in all countries, it is notable that in several countries, this structure cannot be found. Instead, an unexpected factor structure ("Totally unstructured", see Table 2) characterized by one factor of pure negative items can be found in almost all countries in the Middle East and several countries in East Asia. A "Moderately unstructured" factor structure characterized by fragmentation of the factor "Experiencing Utility in Math" (see Table 3) is found in Bahrain, Bulgaria, Georgia, Iran, Japan, Korea, Romania, Singapore, and Turkey. Note that three of the countries are highly performing East Asian countries.

Table 1. Factor structure in North America (combined N = 22278) in TIMSS 2007 (Expected factor structure)

Pattern Matrix[a]			
	Factor		
	1	2	3
MATH IS MORE DIFFICULT FOR ME…	-.804[b]		
MAT IS NOT ONE OF MY STRENGTHS	-.778		
USUALLY DO WELL IN MATHS	.768		
I LEARN THINGS QUICKLY IN MATH	.725		
I HATE MATH		-.878	
I ENJOY LEARNING MATHEMATICS		.865	
MATH IS BORING		-.740	
WOULD LIKE TO TAKE MORE MATH		.615	
NEED MAT TO GET INTO THE <UNI>…			.706
NEED MAT TO GET THE JOB I WANT			.698
NEED MAT TO LEARN OTHER SUBJ			.555
WILL HELP IN MY DAILY LIFE			.552
Extraction Method: Principal Axis Factoring.			
Rotation converged in 5 iterations.			
Loadings > .30 are seen			

Table 2. Factor structure in Syria in TIMSS 2007 (Totally unstructured)

Pattern Matrix[a]			
	Factor		
	1	2	3
I ENJOY LEARNING MATH	.846[b]		
I HATE MATHEMATICS	-.783		
WOULD LIKE TO TAKE MORE...	.588		
I LEARN THINGS QUICKLY IN…	.442		
USUALLY DO WELL IN MATHS			
NEED MAT TO GET INTO THE…		.659	
NEED MAT TO GET THE JOB I…		.609	
NEED MAT TO LEARN OTHER…		.494	
WILL HELP IN MY DAILY LIFE		.396	
MAT IS NOT ONE OF MY…			.696
MATH IS MORE DIFFICULT …			.662
MATH IS BORING	-.312		.364
Extraction Method: Principal Axis Factoring.			
a. Rotation converged in 5 iterations.			
b. Loadings over .30 are seen			

Table 3. Factor structure in Korea in TIMSS 2007 (Moderately unstructured)

Pattern Matrix[a]			
	Factor		
	1	2	3
USUALLY DO WELL IN MATHS	.883[b]		
MATH IS MORE DIFFICULT FOR…	-.832		
MAT IS NOT ONE OF MY…	-.774		
I LEARN THINGS QUICKLY IN…	.676		
I ENJOY LEARNING MATH		.663	
WOULD LIKE TO TAKE MORE…		.636	
I HATE MATHEMATICS	-.359	-.634	
MATH IS BORING		-.619	
WILL HELP IN MY DAILY LIFE		.589	
NEED MAT TO LEARN OTHER…		.520	.335
NEED MAT TO GET THE JOB I...			.789
NEED MAT TO GET INTO THE...			.764
Extraction Method: Principal Axis Factoring.			
a. Rotation converged in 7 iterations.			
b. Loadings over .30 are seen			

Instead of cultural matters, this article finds the answer to the fragmentation in the factor structure from the characteristics of the Fennema-Sherman test itself. Two items are shown to be too difficult for the students at the lowest achievement level. Additionally, the test is culturally biased. This is shown by answering four research questions:

1) How the item discrimination differ between different achievement levels and different cultures?

2) What kind of connection there is between the test reliability and achievement level of the respondents?

3) How the expected factor structure of Fennema-Sherman test fits for different cultural settings?

4) How well does the structure fit for the different achievement levels and cultural settings of the test takers?

Finally, some exemplar items from the relevant test battery used in Finland are suggested in order to replace two poorly behaving items from the original Fennema-Sherman test.

2. Data and Methods

2.1 Data

All the countries in TIMSS 2007 (N = 57) were combined into a dataset consisting of a total of 248,160 eight grader students. For analysis, the dataset was divided into 20 percentiles (N ≈ 12, 000 in each) on the basis of the first plausible value of Mathematics achievement (Table 4). In some analyses, the quartiles are also used; obviously the lowest quartile includes percentiles 1–5 and highest quartile includes percentiles 16–20. Two points of the data and percentiles are worth noting: 1) the range in 1st and 20th percentile is much wider than with other groups because of representing the tail populations and 2) none on the percentile shows normal distribution; in percentiles 2–19 the population is merely uniform than normally distributed. There may thus be some estimation error in the parameters of CFA and EFA. However, because of the robust procedures with large sample sizes (Principal axis factoring with EFA and Maximum likelihood estimation in CFA), the results can be taken stable.

Table 4. Descriptive statistics for 20 percentiles

Percentile of Achievement[1]	N	Mean1	Minimum[1]	Maximum[1]	Range	Std. Deviation	Skewness	Kurtosis
20	12241	674.4	632.1	898.4	266.3	37.2	1.376	2.126
19	12358	611.4	594.3	632.1	37.8	10.8	0.193	-1.147
18	12372	580.9	568.6	594.3	25.7	7.4	0.092	-1.19
17	12347	558.4	548.8	568.6	19.8	5.7	0.061	-1.197
16	12356	540.0	531.4	548.8	17.4	5.0	0.017	-1.189
15	12349	523.6	515.9	531.5	15.6	4.5	0.039	-1.232
14	12352	508.5	501.3	515.9	14.6	4.2	0.016	-1.19
13	12336	494.1	487.1	501.3	14.2	4.1	0.018	-1.178
12	12334	480.2	473.2	487.1	13.9	4.0	-0.002	-1.209
11	12315	466.1	459.0	473.2	14.2	4.1	0.013	-1.198
10	12329	451.9	444.6	459.0	14.4	4.1	-0.021	-1.195
9	12327	437.1	429.5	444.6	15.2	4.4	-0.019	-1.201
8	12306	421.9	414.2	429.5	15.3	4.4	-0.02	-1.202
7	12319	406.2	398.1	414.2	16.1	4.7	-0.011	-1.219
6	12287	389.4	380.5	398.1	17.6	5.1	-0.036	-1.201
5	12272	370.8	360.8	380.4	19.6	5.6	-0.05	-1.187
4	12269	350.0	338.4	360.8	22.5	6.5	-0.06	-1.198
3	12251	324.9	310.2	338.4	28.2	8.1	-0.078	-1.192
2	12198	291.5	269.1	310.2	41.1	11.8	-0.194	-1.154
1	12190	222.5	5.0	269.1	264.1	40.9	-1.378	2.143

1. 1st plausible value (PV) of mathematics in TIMSS 2007

2.2 Fennema-Sherman Test in TIMSS 2007

Usually in the internationally setting the shortened Fennema-Sherman Mathematics attitude scale is divided into two sets of questions (in TIMSS 2007 questions 8 and 9) with the same question: *"How much do you agree with these statements about learning mathematics?"* The statements in Question 8 are as follows:

a. I usually do well in mathematics,

b. I would like to take more mathematics in school,

c. * Mathematics is more difficult for me than for many of my classmates,

d. I enjoy learning mathematics,

e. * Mathematics is not one of my strengths,

f. I learn things quickly in mathematics,

g. * Mathematics is boring,

h. *I hate mathematics.

The items with asterisk (*) are opposite to the scale, and thus they are reversed before scoring. The last question in the set "I hate mathematics" was originally negative but the item was reversed ("I like mathematics") before releasing the data. For the analysis it was reversed back to reach the original structure of the test: two negative and two positive items for both Dimension 1 and 2. The statements in Question 9 are as follows:

a. I think learning mathematics will help me in my daily life,

b. I need mathematics to learn other school subjects,

c. I need to do well in mathematics to get into the <university> of my choice,

d. I need to do well in mathematics to get the job I want.

Later shortened versions of the items are used to shorten the narrative, and texts on tables and figures. The abridged versions are obviously recognized. As seen on Table 1, three dimensions are constructed as follows:

1) Liking MATH: Question 8, items b, d, g*, and h*,

2) Self-concept in MATH: Question 8, items a, c*, e*, and f, and

3) Experiencing utility in MATH: Question 9, items a to d.

Alpha reliabilities for the scales are respectively 0.72, 0.70, and 0.74 in the whole dataset. In what follows, it is seen that reliabilities for the scale "Self-Concept in MATH" are very low in the lowest achievement groups.

2.3 Statistical Methods

Four research questions were set concerning 1) the item discrimination at different achievement levels and cultural settings; 2) the test reliability at different achievement levels of the students; 3) the expected factor structure of the Fennema-Sherman test in different cultural settings, and 4) the fit of the test structure at the different achievement levels and cultural settings. Correspondingly, the Fennema-Sherman test in TIMSS 2007 dataset is analyzed four ways in different ability groups specified as 20 percentiles of mathematics achievement: 1) with Classical Item Analysis (CIA), 2) with reliability estimates, 3) with traditional Exploratory Factor Analysis (EFA) and factor loadings, and 4) with Confirmatory Factor Analysis (CFA).

To answer the first research question, the test items are evaluated on the basis of CIA, particularly with the index to item discrimination. A procedure suggested first by Henrysson (1963) and discussed by Cureton (1966) is to calculate the item-rest correlation ρ_{gXC} (C for "Corrected") rather than the item-total correlation ρ_{gX} (that is, the traditional Pearson product moment correlation between an item and the score). Classically, the lower boundary for item-total correlation is given as $\rho_{gX} = 0.20$ because when the item is either extremely easy or demanding, the mathematical procedure cannot produce much higher values even though the item itself would be perfectly discriminating. However, this low value can be accepted only for extreme items (in achievement testing for extremely demanding or extremely easy items); when an item has somewhat average difficulty level, much higher values (near 0.40–0.60) should be expected for an item to show a discriminative power. In practice, such items with $\rho_{gXC} < 0.20$ are usually discarded because of lack of accuracy.

The second research question is tackled by using the Classic Alpha reliability (Kuder & Richardson, 1937; Gulliksen, 1950; Cronbach, 1951). When the values for item discrimination for single items are high, on the basis of Lord and Novick (1968, formula 15.3.8, where the item-total correlation is in-built in the formula of alpha reliability), the reliability of the score is high. Though it is known that Alpha reliability always underestimates the real reliability (Gulliksen, 1950; Lord & Novick, 1968; Vehkalahti, 2000), it is, in practice, the most used indicator for the general reliability (Hogan, Benjamin, & Brezinski, 2000). The challenge with the classical reliability is that when knowing the overall reliability of the score, it is not possible to know how good the test is in the *extremes* of the scale. This is an important matter because we tend to be interested in the lower performers (as well as the high performers) with negative attitudes. If an attitude test is not reliable in the low-performer's group, then the attitude results for these low performers do not mean anything. When the reliability of the score is lower than $\alpha = 0.60$, it is traditionally taken as too low to indicate sufficient test reliability. Two types of reliabilities are used: first, the reliability for the total sum of the attitude scale (combining all the 12 items) and second, the reliability for single dimensions. For the total sum, Alpha model gives certainly too low an estimation because it cannot utilize the information of the structure in the test; Alpha model always assumes unidimensionality (see critical discussion about the assumptions, e.g., in Tarkkonen, 1987; Vehkalahti, 2000).

To answer the third research question, the traditional Exploratory Factor Analysis is used and factor loadings are extracted. Because it is theoretically expected to find a factor solution of three factors (see Table 1), Principal Axis Factoring (PAF) with the 3-factor solution is selected at each achievement level; the analysis is done in SPSS environment. During the process, it was easily seen that the factors correlate moderately with each other. Thus an oblique rotation (Promax) is used.

The fourth research question is tackled by using Confirmatory Factor Analysis (CFA) in AMOS environment (Arbuckle 2007). The modern CFA is based on Karl K. Jöreskog's early works in the late 1960s (Jöreskog, 1967; 1969; 1970) and later his (Jöreskog, 1973), Keesling's (1972) and Wiley's (1973) works (see Bollen, 1989a, 6). CFA and SEM analysis are known by its own notation (see e.g., Jöreskog et al., 2003; Bollen, 1989a; Bentler, 1995; Byrne, 2001; Ullman, 2001) and several – even confusing number of – indices for model fit. The only statistical test, however, for testing a null hypothesis concerning the model fit is the Chi Square test comparing the observed and expected covariance structures. One simple test is to divide the Chi Square coefficient by its degrees of freedom (CMIN in AMOS). In practice, though, the Chi Square test is sensitive for large sample size; when sample size is larger than 300–400, the Chi Square test rejects the hypotheses of the model fit too easily. In this massive dataset ($N > 240000$), Chi Square test definitely rejects the hypotheses of a good fit. Thus in the case, the incremental fit indices are more preferable to use. In AMOS outputs, the following incremental fit

indices for the model fit are in use: Normed Fit Index (NFI, Bentler & Bonett, 1980), Relative Fit Index (RFI, Bollen, 1986), Incremental Fit Index (IFI, Bollen, 1989b), Tucker-Lewis Index (TLI, known also as Non-Normed Fit Index, NNFI, Bentler & Bonnet, 1980), Comparative Fit Index (CFI, Bentler, 1988). AMOS also uses the Root mean square error of approximation (RMSEA, Steiger & Lind, 1980; Browne & Cudeck, 1993) and several indicators based on information criteria. Except RMSEA, the rule of thumb in evaluating the goodness-of-fit between the theoretical and the observed covariance structure is that the values for the incremental fit indices should be over .90 (Bentler & Bonnet, 1980); values below .90 indicate that the dataset does not match the model and thus the model should be developed. The rule of thumb for RMSEA is that the values below .05 show good fit, values .08 and less show moderate discrepancy for the model and data, and values over .10 show the models which should be rejected (Browne & Cudeck, 1993). The upper boundary of RMSEA = .05 is tested in AMOS (in what follows, PCLOSE). Item-wise accuracy or item reliability is measured by Squared Multiple Correlation coefficients (SMC); the higher value the better. In the analysis, Maximum likelihood estimation method is used.

3. Results

3.1 Item Discrimination of the Negative Items of Fennema-Sherman Test

On the basis of the TIMSS data it is obvious that there are some challenges in the negatively expressed items of the Fennema-Sherman test especially in the East Asian countries. Incrementally growing values for the item discrimination of the negative items in different achievement levels (see Figures 1 and 2) reveal several things. First, the short and strict negative statement *"I hate Mathematics"* behaves nicely at all the achievement levels. However, all the other negative statements seem to embed challenges in the lowest achievement level. Second, because of the low item discrimination (less than .20) in the lowest quartile (Figure 1, here East Asian dataset as an example), two negative items *"Mathematics is not one of my strengths"* and *"Mathematics is more difficult for me than for many of my classmates"* are not recommendable to be used in a test that is intended for the low performing students. These items would most probably be discarded if the test was originally constructed with international scope. The latter item seems to discriminate consistently in the North American dataset and moderately well in European and Middle Eastern datasets (Fig. 2). Thirdly, though the item *"Mathematics is boring"* has also quite a low value for item discrimination in the lowest percentile, otherwise it behaves quite well, albeit not as well as the item *"I hate mathematics"*. Fourth, the different levels of item discrimination seen in Figure 1 are in the same vertical order also in North American, European and Middle Eastern datasets: the highest item-rest correlations are attained with item *"I hate Mathematics"*, the second highest with *"Mathematics is boring"*, the third highest with *"Mathematics is not one of my strengths"* and the lowest with item *"Mathematics is more difficult for me than for many of my classmates"*. This seemingly gradual increase of complexity is discussed in the final section.

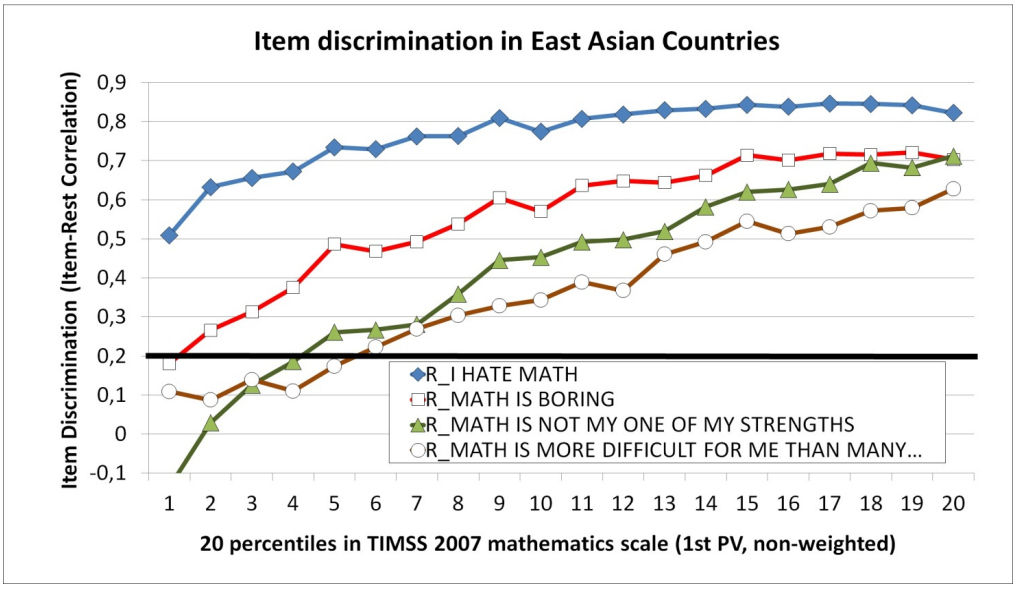

Figure 1. Item discrimination for negative items in the East Asian sub-dataset

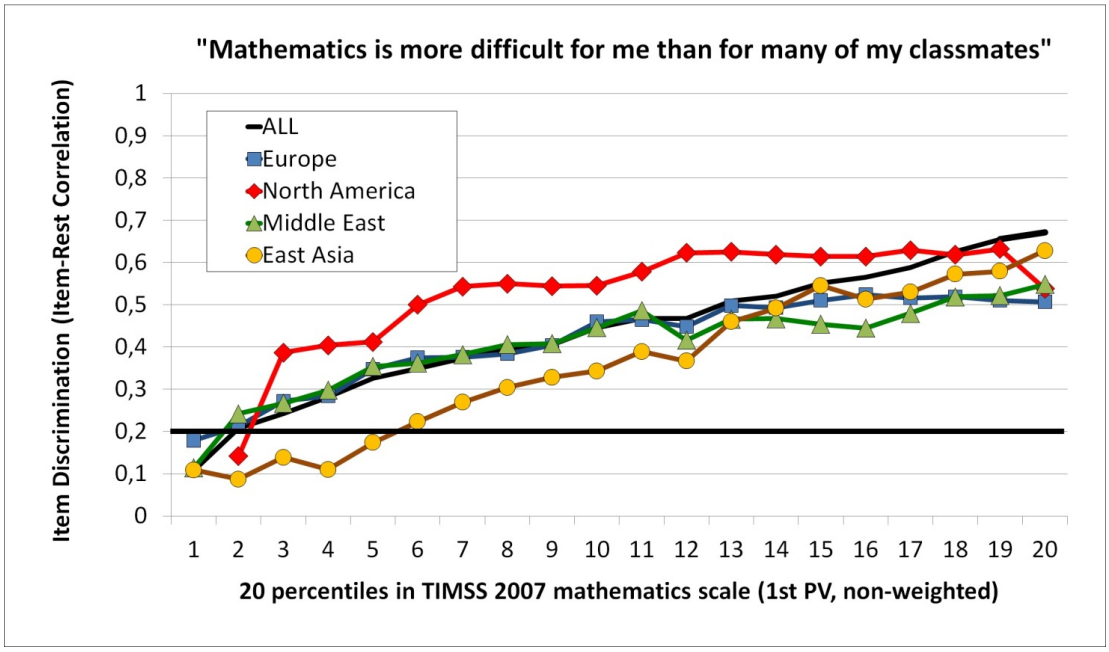

Figure 2. Item discrimination for item "Mathematics is more difficult for me than for many of my classmates" in different sub-populations in different achievement groups

3.2 Reliability of the Attitude Test in Different Achievement Levels

Two out of three dimensions – "Liking MATH" and "Experiencing Utility on MATH" – do not appear to contain any serious problems in test accuracy in the group of the lowest level students. Contrarily, it is evident that the reliability of the scale of "Self-Concept in MATH" is very low when it comes to students in the lowest quartile (Figure 3): reliabilities are remarkably below .60 whereas with the other dimensions reliabilities stay steadily over .70 even in the lowest ability groups.

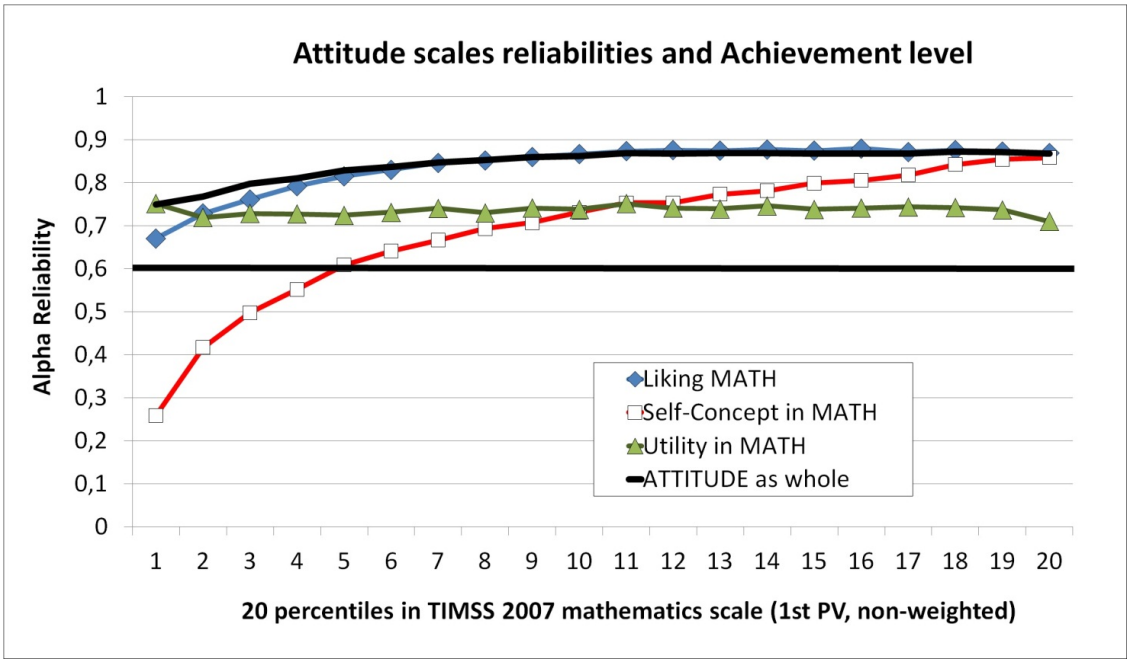

Figure 3. Reliabilities of Fennema-Sherman test scores in different achievement levels

From the international viewpoint, the score for "Self-Concept in MATH" appears to be very problematic: the score is less reliable in East Asian countries, Middle Eastern countries and in Europe than in North America (Figure 4). Especially in the East Asian countries, even the mediocre students (in percentiles 6–8) are not tested accurately: reliabilities in the lowest quartile are lower than 0.50 and in percentiles 6–8 reliabilities are 0.60 or less. It seems that the score of "Self-Concept in MATH" is the most coherent in North America where the test has been developed (and pretested).

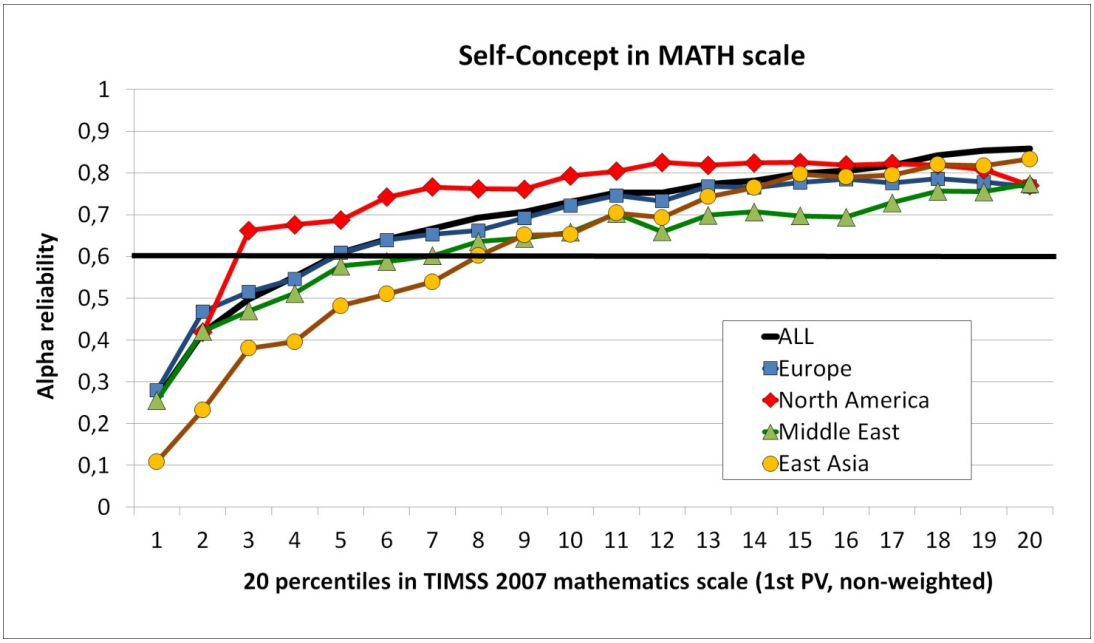

Figure 4. Reliabilities for score of "Self-Concept in MATH" in different achievement levels in different sets of countries (N > 240000)

3.3 Factor Structure of the Fennema-Sherman Test in Different Achievement Levels

A standard EFA with three factor solution reveals one reason why the factor structure deconstructs as seen earlier in Table 3. The factor structure is strictly dependent on the achievement level of the test takers and there seems to be logical phases of a kind of abstract thinking (Table 5).

Table 5. Factor loadings of selected items in the different achievement level groups

Achievement level in Math[1)]	Loading[2)] of "HATE"[3)]	Loading[2)] of "BORING"[4)]	Loading[2)] of "DIFFICULT"[5)]	Loading[2)] of "WEAK"[6)]	Loading[2)] of "I LEARN"[7)]	Loading[2)] of "I DO WELL"[8)]
20	-0.844	-0.794	-0.782	-0.818	0.719	0.813
19	-0.848	-0.782	-0.781	-0.784	0.738	0.811
18	-0.851	-0.786	-0.770	-0.752	0.747	0.788
17	-0.858	-0.784	-0.752	-0.713	0.716	0.754
16	-0.856	-0.775	-0.742	-0.696	0.69	0.735
15	-0.854	-0.755	-0.748	-0.677	0.694	0.705
14	-0.876	-0.741	-0.733	-0.646	0.668	0.669
13	-0.856	-0.73	-0.749	-0.603	0.664	0.654
12	-0.888	-0.695	-0.717	-0.618	0.628	0.579
11	-0.864	-0.683	-0.735	-0.616	0.606	0.540
10	-0.875	-0.672	-0.699	-0.593	0.592	0.531
9	-0.897	-0.649	0.699	0.586	-0.492	-0.429
8	-0.878	-0.580	0.688	0.574	-0.449	-0.366

7	-0.901	-0.554	0.658	0.553	-0.329	a
6	-0.917	-0.502	0.639	0.554	a	a
5	-0.849	-0.440	0.609	0.512	a	a
4	-0.842	-0.363	0.560	0.536	a	a
3	-0.841	0.349	0.547	0.575	a	a
2	-0.810	0.403	0.437	0.637	a	a
1	-0.751	0.491	0.426	0.536	a	a

1) 20 percentiles on the basis of unweighted 1st Plausible value (PV) in TIMSS 2007 dataset

2) The extraction method: Principal Axis Factoring; the rotation method: Promax with Kaiser Normalization; 3-factor solution

3) – 8) Variables: "I hate math" as original (negative), "Math is boring", "Math is more difficult for me than for many of my classmates", "Math is not one of my strengths", "I learn things quickly in Math" in expected factor, and "I usually do well in Maths" in expected factor

a) loading < .30 in the expected factor

At the lowest level of achievement, the factor structure is characterized by *one clear factor of negative items*. Technically speaking, at this level all the negative items—except "I hate mathematics"—correlate with each other more than with the expected positive counterparts. The loadings of the negative items are highly positive. At this level of achievement, the absolute value of the bilateral correlation coefficient of variable "*Math is not one of my strengths*" and its positive counterpart "*I usually do well in mathematics*" is r = 0.12 or less, and for variables *"Math is boring"* and *"I enjoy learning mathematics"* it is r = 0.30 or less. At this level, the reliabilities for the score of "Self-Concept in MATH" are less than $\alpha = 0.50$ as also seen in Fig. 1. One plausible reason for low correlation is that the *general reading comprehension of low-ability level students may be inadequate for them to understand the statements.* This may also be connected with their poor mathematics skills: they were not able to read the stems of the test items. Another explanation may be that at this low level of achievement, the students' *level of abstract thinking may be low:* many of the lowest ability level test takers seem to comprehend the negative wordings inadequately. One may hypothesize that these students have enough general ability to understand the positive sentence and react adequately, but not enough abstract level thinking to understand the relevance of the negative wording and to judge whether they have a positive or a negative opinion of this negative sentence. Thus, this lowest level of abstract thinking is called *Concrete level.*

The second-lowest achievement level is, technically speaking, characterized by the fact that the variable *"Math is boring"* has a growing loading in the correct factor—and thus, the intended factor structure has started to develop. Reliabilities for the score of "Self-Concept in MATH" range $\alpha = 0.55–0.64$, which shows very low reliability for the test. However, the other negative variables are still correlating positively with each other without corresponding positive variables in the same factor. This level of abstract thinking is present in the percentiles 4–6. This second-lowest level of abstract thinking is called *Developing level.*

The third level is called *Formed level* because all the factors are formed but they are still immature in comparison with the expected structure. Technically speaking, the factors "Liking MATH" and "Experiencing Utility in MATH" are formed as they are intended. However, the third factor with negative items, "Self-Concept in MATH", is characterized by negative loadings for the positive items and positive loadings for the negative items; hence, the negative items are dominating the factor loadings. This level of abstract thinking is present in the percentiles 7–9. Reliabilities for the score of "Self-concept in MATH" in these percentiles range $\alpha = 0.67–0.71$.

The *Matured level* of abstract thinking is characterized by the expected factor structure: the higher the achievement levels, the higher loadings and correlations between the corresponding variables in the expected factor. This level includes percentiles 10–20 and it requires around 445 points or more in TIMSS mathematic scale to achieve this matured level of abstract thinking. Reliabilities for the score of "Self-Concept in MATH" range $\alpha = 0.73–0.86$.

These four levels of abstract thinking—Concrete, Developing, Formed, and Matured—are readily observed from a large number of international students. The categories seem to have a slight connection to Jean Piaget's theory of growing abstract thinking (Piaget, 1970) with concrete and formal operations. Though it would be tempting to generalize the results to individual growth, the design gives no possibilities to draw conclusions that the individual maturation of the abstract thinking would follow these steps. Nevertheless, these different levels of

abstract thinking have a strict connection to deconstruct of the expected factor structure and it may also explain why the Fennema-Sherman test does not work in certain countries. The deconstruction of the factor structure (seen at Tables 1−3) is not limited to differences in academic achievement but seem also be due to cultural differences. There may be an inseparable connection of cultural-based and achievement-based factors behind fragmentation in factor structure.

Because the factor structure is strictly dependent on the achievement level of the test takers, the version of Fennema-Sherman test used in TIMSS-and PISA settings is evidently biased so that the attitudes of the lowest achievement groups cannot be measured in a reliable way.

3.4 Model Fit in Different Sets of Countries and in Different Achievement Levels

Generally speaking, the Confirmatory Factor Analysis suggests that the model of three factors with four items on each, explains quite well the covariance in the data. Figure 5 demonstrates the simplest factor model created on the basis of the expected factor model in North American data. The loadings and SMCs are seen in the graph. Note that in this simple model, the error terms are assumed to be non-correlated. The factor loadings are quite high (Ranging .60 — .87) and the item reliabilities are sufficient or high (SMCs range .36—.83). Reliability for the total sum is α = .845. Normed Fit Index (NIF = 0.96) tells that the model is 4% away from the most perfect model, that is, the model is acceptable, as indicated also by the other incremental fit indices (IFI = .96 and CFI = .96).

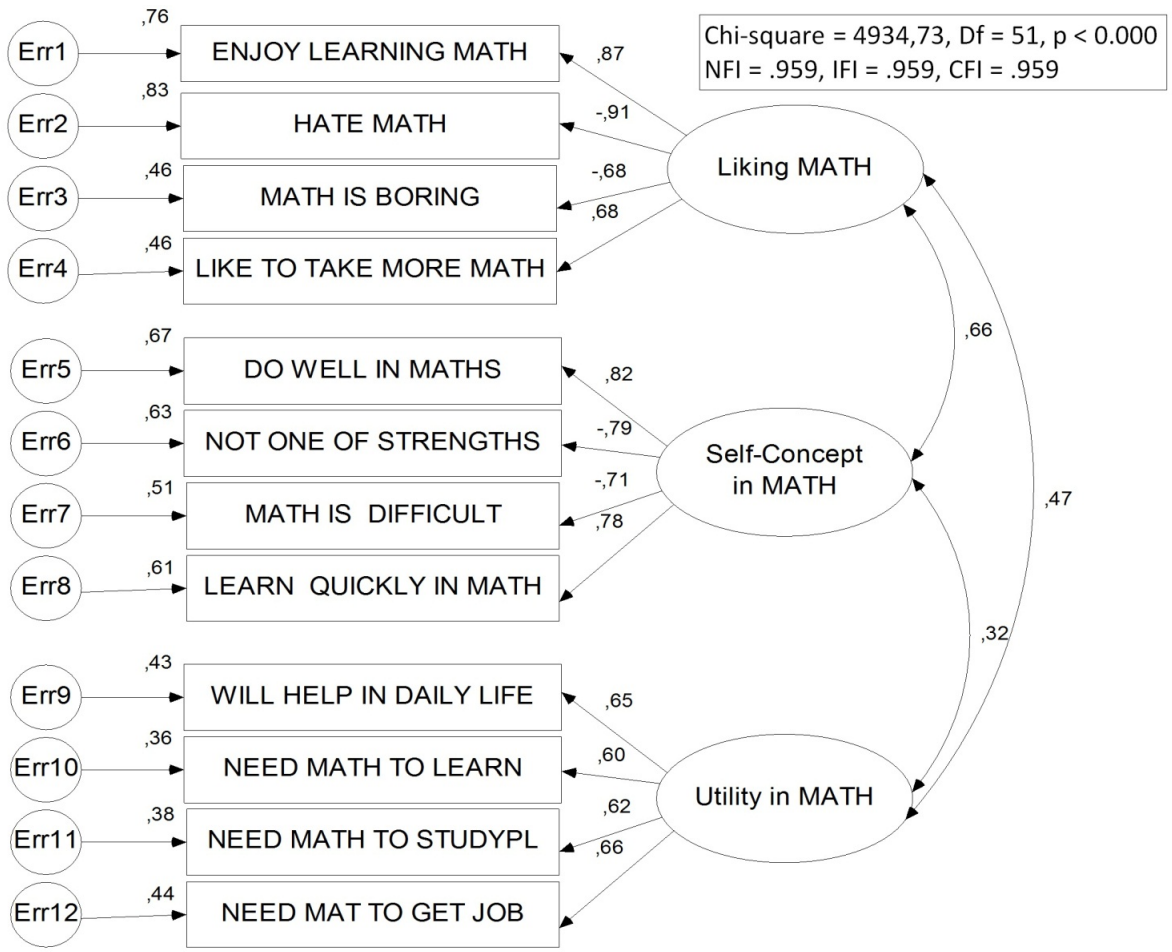

Figure 5. Measurement model of mathematical attitudes of eight-graders in North America (N = 27 711) in the TIMSS 2007 Dataset – "Expected model"

From the international viewpoint, this general model can be accepted very well in Europe and to some extent in East Asia though in the East Asian set of data some indicators (RFI = .87, TLI = .87 and RMSEA > .08) hint that there are notable discrepancy between the data and the model (Table 6). The model does not fit the Middle

Eastern data (all values of goodness of fit indices fall below .90 and RMSEA > .08). This alone indicates that the Fennema-Sherman test should be revised for international comparisons. On the basis of Section 3.1, the problematic items are already known; in Section 4, some recommendations are given to on the basis of experiences from the Finnish version of the test.

Table 6. Baseline comparisons for "North American", "European", "Middle Eastern" and "East Asian" models (All students)

	Goodness of Fit Index						
	NFI	RFI	IFI	TLI	CFI	RMSEA	PCLOSE
North America	.959	.937	.959	.937	.959	.065	.000
Europe	.953	.927	.953	.928	.953	.064	.000
Middle East	.889	.831	.890	.831	.890	.082	.000
East Asia	.915	.870	.915	.871	.915	.095	.000

There is no vast discrepancy between the model and the data when it comes to the highest performing eight graders (the highest quartile, percentiles 16–20, Table 7) though some indicators for Middle Eastern population are just below the thumb rules (NFI = .89 and TLI = .89). On the contrary, the discrepancy of the model and data is drastic when focusing on the lowest quartiles in different sets of countries (Table 8): in Europe and East Asia, the model can be said to fit the data to some extent – otherwise practically all the indicators show the need for remodeling the construct.

Table 7. Baseline comparisons of high-achieving students in North American, European, Middle Eastern and East Asian populations (highest quartile)

	Goodness of Fit Index						
	NFI	RFI	IFI	TLI	CFI	RMSEA	PCLOSE
North America	.959	.937	.960	.939	.960	.062	.000
Europe	.961	.940	.962	.941	.962	.059	.000
Middle East	.926	.886	.927	.889	.927	.073	.000
East Asia	.940	.908	.940	.909	.940	.085	.000

Table 8. Baseline comparisons of low-achieving students in North American, European, Middle Eastern and East Asian populations (lowest quartile)

	Goodness of Fit Index						
	NFI	RFI	IFI	TLI	CFI	RMSEA	PCLOSE
North America	.887	.827	.902	.848	.901	.084	.000
Europe	.922	.881	.926	.886	.925	.065	.000
Middle East	.887	.828	.888	.829	.888	.077	.000
East Asia	.922	.881	.925	.885	.925	.076	.000

When focusing on the most suspicious part of the construct, "Self-Concept in MATH" (see Fig. 4 above), it can be noted that there is no problem when it comes to measuring students' attitude in the highest performing quartile: the model fits almost perfectly with the data (NFI = .996, Figure 6, Table 9). There is some unexplained variability in Middle Eastern and East Asian models (RMSEA > .15). However, on average, the fit is good.

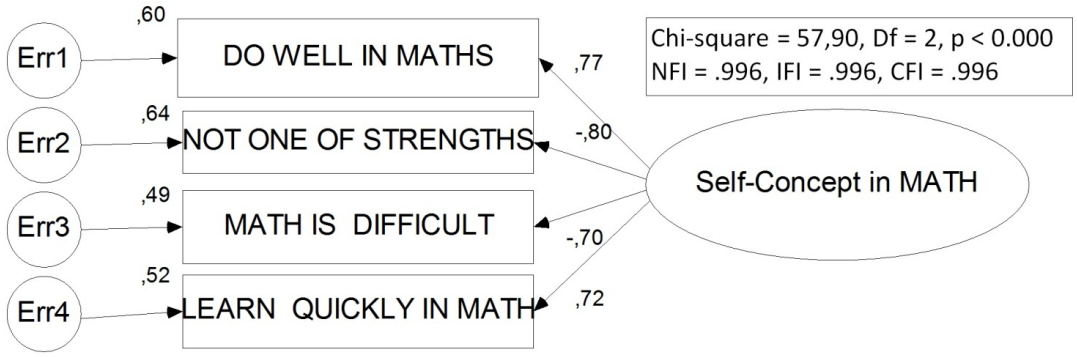

Figure 6. Measurement model of "Self-Concept in Math" of HIGHEST quartile eight-graders in North American sub-population (N = 9700) in the TIMSS 2007 Dataset

Table 9. Baseline comparisons for "Self-Concept in Math" of HIGHEST quartile (highest-performing) students in different sets of countries

	Goodness of Fit Index						
	NFI	RFI	IFI	TLI	CFI	RMSEA	PCLOSE
North America	.996	.980	.996	.981	.996	.052	.282
Europe	.992	.958	.992	.959	.992	.071	.000
Middle East	.965	.826	.965	.827	.965	.118	.000
East Asia	.983	.913	.983	.913	.983	.115	.000

The lowest end of the achievement scale shows an opposite fact. The model does not fit at all in the datasets of the lowest quartile students (Figure 7 and Table 10). The apparent reason, based on Figure 7, is that two negative items, *"Mathematics is not one of my strengths"* and *"Mathematics is more difficult for me than for many of my classmates"*, show very low SMCs (in European sub-population 0.04 and 0.06) indicating very low item reliability. In the Middle Eastern dataset, some indices for fit are negative. Also, in the East Asian dataset, there appears to be negative variance for variable Err4, indicating extremely poor model structure.

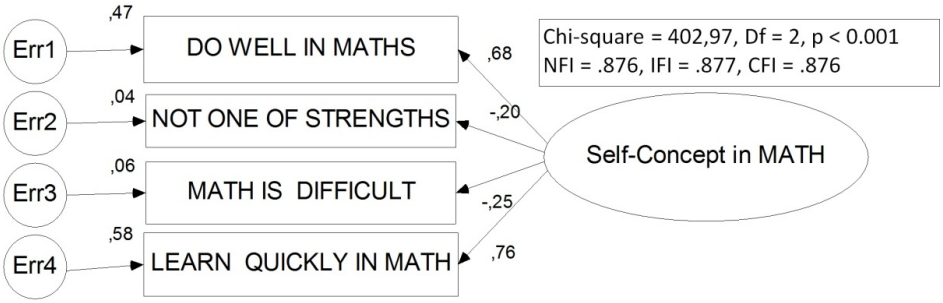

Figure 7. Measurement model of "Self-Concept in Math" of LOWEST quartile eight-graders in European sub-population (N = 4 620) in the TIMSS 2007 Dataset

Table 10. Baseline comparisons for "Self-Concept in MATH" of LOWEST quartile (lowest-performing) students in different sets of countries

	Goodness of Fit Indices						
	NFI	RFI	IFI	TLI	CFI	RMSEA	PCLOSE
North America	.893	.466	.897	.475	.895	.175	.000
Europe	.876	.380	.877	.382	.876	.156	.000
Middle East	.659	-.706	.659	-.707	.659	.189	.000
East Asia	.823	.114	.824	.115	.823	.184	.000

The bottom-line is that on the basis of Classical item analysis (Section 3.1), Reliability estimates (Section 3.2), EFA (Section 3.3), and CFA (Section 3.4), the structure of Fennema-Sherman test is challenged. Especially arguable is its use in international comparisons and especially when using the test to measure the attitudes of the lowest performing students.

4. Alternative Items on the Basis of the Finnish Version

The shortened Fennema-Sherman test has been used in all four TIMSS and PISA rounds and thus it has its own value in giving the trend information on changes in the attitudes toward mathematics. The factor structure seems attractive and justifiable: without any doubt, the dimensions of "Liking MATH", "Self-Concept in MATH", and "Experiencing Utility in MATH" are important elements of the construct. However, on the basis of Sections 3.1 – 3.4, there seems to be some challenges in the structure which evidently have to do with the abstract thinking of the students as well as with cultural issues. First issue is the use of *too complicated* negative items. Second, there seems to be *too many* negative items for international testing purposes. Third, it may be possible that concentrating on a *too general concept* of "mathematics" rather than, e.g., more concrete "mathematics lessons" or "mathematics as a school subject" may be too abstract to many of the low performing students. Alternative test items to consider are given to on the basis of Finnish experiences. Though the Finnish version of the test is described in detail in this Section, it is not necessarily advisable to change the whole test construction in TIMSS- and PISA settings.

In the Finnish national achievement testing, student attitudes are an essential part – as in TIMSS and PISA. A modified Fennema-Sherman test with the same dimensions as in the international settings has been used in numerous assessment questionnaires in several subjects (e.g., in Mathematics, Mother tongue, Science, Languages, Arts, and Physical education tests) in different grades (grades 4, 6, 7, and 9). The original Fennema-Sherman test has been amended with the following principles: 1) to include less negative items (just one for each dimension), 2) to include simpler wordings and 3) to focus – not in "mathematics" but – more concrete "mathematics lessons" and "mathematic as a school subject". The first point is based on the observation that some of the items in the original Fennema-Sherman test are ethically and morally questionable. For example, the item *"I hate mathematics"* was changed to a positive form of *"I like mathematics lessons"* and *"I like to study Mathematics"*, because it is not intended to take a stand for the fact that the students could "hate" some school subjects. The second point is based on the need to use the test also with younger children below eight- or nine grades. For example, instead of wording *"Mathematics is more difficult for me than for many of my classmates"* much shorter and perhaps more straightforward alternatives are in use: *"Many things in Mathematics are difficult"* and *"Mathematics is an easy subject"*. The third point is intended to help students to think of situations in the classroom more concretely. For example, instead of *"Mathematics is boring"*, more concrete alternatives such as *"Mathematics is a boring subject"* and *"Mathematics is one of my favorite subjects"* are in use. Though the dimensions are the same, the item-wise changes are so radical that the Finnish test is no more Fennema-Sherman test but rather "loosely based on Fennema-Sherman test" as described by Metsämuuronen (2009, 20). While the TIMSS- and PISA versions use four point Likert scale without value 0 (scale is actually −2, −1, +1, +2 though the numbers 1 to 4 are in use), in the Finnish test, the 5-point Likert scale is in use (−2, −1, 0, +1, +2).

The characteristics of the Finnish test are not discussed in depth in this article. Instead some ideas are laid out as to what kind of changes could be done to raise the quality of the Fennema-Sherman test for international testing settings. The indicators for the construct validity (measurement model and the related indicators for the model fit) are briefly shown, and the values for the item discrimination of negative items are compared in different achievement levels on the basis of fractions of European students' achievement in TIMSS 2007. The items and

the basic factor structure in the Finnish attitude test are found in Table 11. Alpha reliabilities for the dimensions seem to vary somewhat between different samples; in Mattila (2005) they were 0.90 (Liking Math as a school subject), 0.87 (Self-Concept in Math), 0.79 (Experiencing Utility in Math), and 0.915 for the total score.

Table 11. Items and factor loadings in the Finnish attitude test based on the Fennema-Sherman attitudes test (Mattila 2005, Metsämuuronen 2009, N = 4511)

Pattern Matrix[a]			
	Factor		
	Liking Math as a school subject	Self-Concept in Math	Experiencing Utility in Math
5. I like Mathematics lessons	.898[b]		
14. I like to study Mathematics	.732		
4*[c] Mathematics is a boring subject	.720		
8. Usually we have interesting tasks in Mathematics lessons	.701		
6. Mathematics is one of my favorite subjects	.696		
10. I think I'm good in Mathematics		.883	
12. I can manage even the difficult tasks in Mathematics		.762	
1. Mathematics is an easy subject		.748	
3.*[c] It is impossible for me to get good results in Mathematics		.700	
11.*[c] Many things in Mathematics are difficult		.573	
13. I believe I need Mathematical knowledge and skills in my work-life			.848
7. Mathematical knowledge and skills are important in everyday-life situations			.721
15. I think that Mathematical skills are important			.696
2. I will need Mathematics in my studies to come			.622
9.*[c] In the future, I will not need the matters I have learned in Mathematics			.514

Extraction Method: Principal Axis Factoring.

Rotation Method: Promax with Kaiser Normalization.

a) Rotation converged in 6 iterations.
b) Loadings > .30 shown
c) Items with *-sign are negative

From the technical viewpoint, the test construction of the Finnish version is as good as the TIMSS 2007 version (Figures 8 and 9); indicators for the model fit show a good fit with the model and data. Especially noteworthy is the fit of the data with the model in the lowest quartile. Although Finland did not participate in the TIMSS 2007, it can be assumed that the Finnish student's achievement in mathematics does not differ radically from the average European population, considering that the PISA results in mathematical literacy have been quite high. Where the fractions of the European students in TIMSS are known, these fractions are used to divide the Finnish population into 20 percentiles in the national test. The item discriminations of the Finnish test are then estimated for in these fractions. Table 12 shows the fractions, reliabilities, and item discrimination for selected items which could be used as alternative substitutes for low discriminating, negative items in the TIMSS 2007 set of questions. Note that the reliability of the construct "Self-Concept in MATH" is clearly more constant (0.68 − 0.82) over the ability levels compared with the construct used in TIMSS 2007 setting (Alpha reliability ranges 0.26 − 0.86). When comparing the construct and indicators of the ultimately lowest performing Finnish students,

(the lowest quartile on the basis of European fractions in TIMSS 2007, see Table 12) the construct of "Self-Concept in MATH" fits quite much higher in the Finnish version than in TIMMS 2007 test for the European students (Fig. 7).

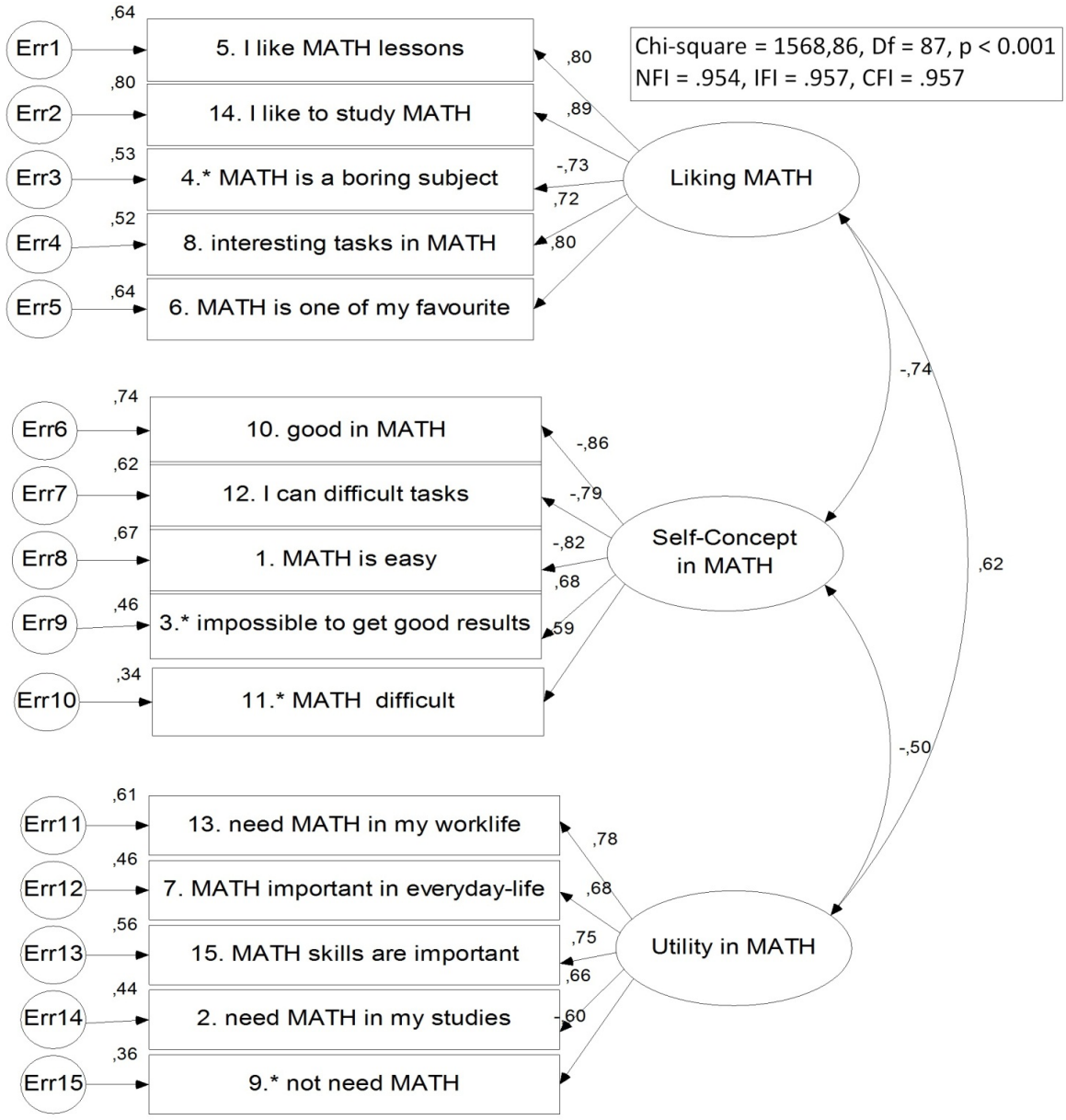

Figure 8. Measurement model of the Finnish version of Fennema-Sherman test in the Finnish population (N = 4511)

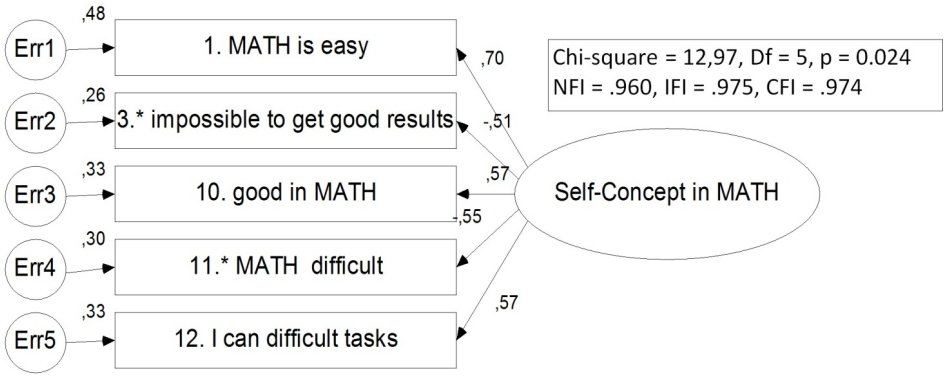

Figure 9. Measurement model for the Finnish version of "Self-Concept in MATH" in the LOWEST quartile of the Finnish population (N = 379)

Table 12. Item discrimination for low-discriminating items of Fennema-Sherman test and their alternatives in the Finnish version

20 percentiles of International European fractions	Fractions for Europe as a whole (%)	Finnish student population[1] N = 4 511	Reliability for "Self-Concept in Math"	European benchmark R_NOT STRENGTHS[2]	12. I can manage even with difficult…[3]	3.*[2] It is impossible to get good results… [4]	European benchmark R_MORE DIFFICULT[2]	1. Math is easy subject[5]	11.*[2]Mathis difficult subject [6]
1	0.5	23	0.73	-0.01	0.40	0.34	0.15	0.66	0.53
2	1.1	49	0.75	0.20	0.48	0.58	0.20	0.62	0.42
3	1.7	77	0.68	0.26	0.68	0.32	0.28	0.60	0.37
4	2.2	99	0.75	0.21	0.49	0.45	0.28	0.53	0.26
5	2.9	131	0.69	0.31	0.22	0.29	0.34	0.41	0.34
6	3.5	159	0.76	0.35	0.53	0.32	0.39	0.57	0.35
7	4.2	190	0.75	0.41	0.58	0.37	0.38	0.64	0.15
8	4.7	212	0.79	0.38	0.58	0.44	0.38	0.54	0.33
9	5.6	254	0.80	0.45	0.55	0.52	0.41	0.63	0.40
10	5.9	267	0.76	0.48	0.52	0.39	0.48	0.62	0.32
11	6.5	293	0.80	0.50	0.57	0.41	0.46	0.59	0.37
12	6.9	313	0.82	0.52	0.57	0.52	0.45	0.65	0.34
13	7.2	325	0.80	0.56	0.56	0.39	0.51	0.59	0.38
14	7.6	344	0.81	0.59	0.55	0.47	0.49	0.62	0.38
15	7.4	335	0.77	0.60	0.54	0.38	0.51	0.65	0.41
16	7.7	348	0.80	0.61	0.57	0.40	0.53	0.62	0.44
17	7.5	339	0.77	0.61	0.48	0.40	0.52	0.61	0.41
18	6.9	311	0.77	0.63	0.51	0.36	0.53	0.52	0.41
19	6.2	279	0.75	0.63	0.58	0.36	0.51	0.59	0.33
20	3.6	163	0.71	0.62	0.57	0.36	0.51	0.42	0.35

1) 9th graders 2004 (Mattila 2005)

2) Reversed items, Benchmarking items: "Mathematics is not one of my strengths", "Mathematics is more difficult for me than for many of my classmates".

3) – 6) 12. I can manage even with the difficult tasks in Mathematics, 3* It is impossible for me to get good results in Mathematics, 1. Mathematics is an easy subject, and 11.* Many things in Mathematics are difficult.

Figures 10 and 11 illustrate three features from Table 12. First, especially in the lowest quartile, both the positive and negative version of the alternative items discriminate the students much better (item-rest correlation > .40) than the original items (.00 −< .30). Second, the item-rest correlations of the alternative items are much more stable in each achievement group compared with the original items. Third, the positive alternative is consistently higher discriminative than the negative alternative. Obviously, there is no evidence that the alternative items would operate in different cultures. Thus, there will be a need for a pre-test process in different cultural settings if the proposed changes will be considered.

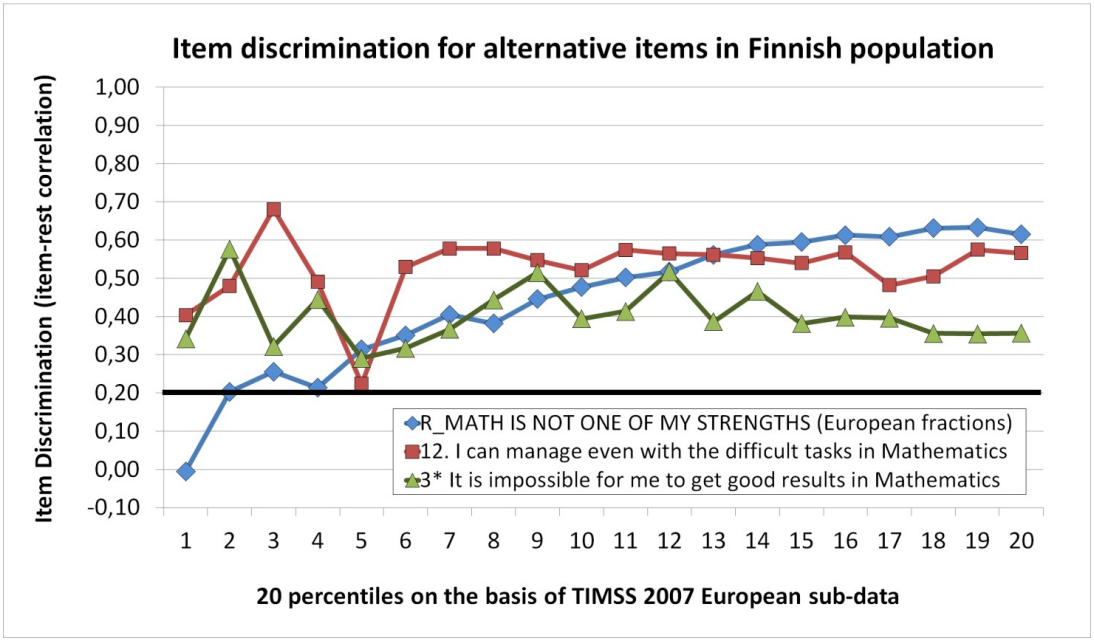

Figure 10. Item discrimination for item "Math is not one of my strengths" and its possible alternatives in the Finnish version

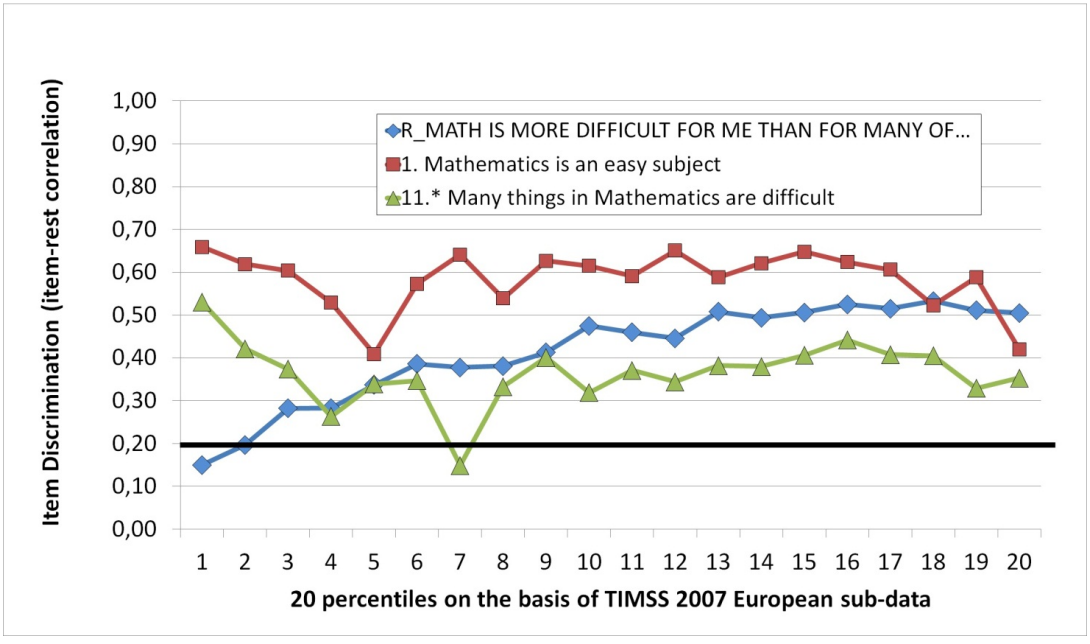

Figure 11. Item discrimination for item "Math is more different for me than for many of my classmates" and its possible alternatives in the Finnish version

The Finnish data suggests two alternatives: instead of the item *"Math is not one of my strengths"*, either a positive one *"I can manage even the difficult tasks in Mathematics"* or a negative one *"It is impossible for me to get good results in Mathematics"* may be used. Both operationalize *"not my strengths"* and both are consistently better in the lower achievers' groups. Out of two alternatives, the positive one is recommendable from the accuracy perspective. Instead of the item *"Mathematics is more difficult for me than for many of my classmates"*, the data suggests either a positive one *"Mathematics is an easy subject"* or a negative one *"Many things in Mathematics are difficult"*, which are simpler and would generate the same kind of information.

5. Discussion and Suggestions

If there were no differences in the test characteristics between different cultures and achievement levels, the test can be seen culturally and achievement-wise unbiased. However, in several countries more than 25% of the students will get an inconsistent and biased test result because the test does not fit the intellectual level or culture of the students. On the basis of the results, it is evident that the shortened version of Fennema-Sherman Mathematic Attitude test as it is in use in TIMSS- and PISA datasets was found to be achievement-levelly and culturally biased. The low item-rest correlations (gXC< .20) and reliabilities (< .60) in score of "Self-Concept in MATH" as well as the low values of the model fit indices (e.g., NFI < .90) in CFA and fragmented factor structures in EFA suggest strongly to modify the Fennema-Sherman test when using it in the international testing settings. Notably, two complex negative items, "Mathematics is not one of my strengths" and "Mathematics is more difficult for me than for many of my classmates", should be changed in order to maintain the good standard of testing. The items seem to be too complicated for the students in the lowest achievement groups.

Though the results are based on large data and the analysis has been quite detailed, they carry three weaknesses. First, the data includes oversampling in some countries. For example, compared with the sample sizes in Taiwan (N = 4, 046) and Hong Kong (N = 3, 470) in the original set of data there are actually three samples from Canada (British Columbia, Ontario, and Quebec, N = 11, 660) and three samples from USA (USA, Massachusetts, and Minnesota, N = 11, 051) which produces strict over-sampling in dataset for some countries. This evidently has an effect to the results. Second, as discussed in Section 2.1 it is worth noting that when dividing the original dataset into 20 percentiles, none on the percentile shows Normal distribution; in percentiles 2–19 the population is merely uniform than normally distributed and the range in 1st and 20th percentile is much wider than with other groups because of representing the tail populations. There may thus be some estimation error in the parameters of CFA and EFA. However, because of the robust procedures with large sample sizes, the results can be taken stable. Third limitation in generalizing the results of the Finnish version of the test is that the Finnish version uses five-point Likert scale and the TIMSS- and PISA versions use four point scale. This causes that the measurement error in the Finnish version is lower than in the original version. This is not a drastic challenge because in any case there is need to pretest the new items before using them.

The results raise some ideas for further development. First, on the basis of the systematic order of the item discrimination values analyzed in Section 3.1 (see Fig. 1), it seems that there is a kind of a hierarchy of complexity in negative items which could be utilized when constructing attitude tests in the international settings. A simplistic analysis of the differences between the items suggests a classification of negative items into five categories on the basis of their complexity:

1) The simplest type of negative attitude item is a *short and straight negative statement with extreme wording,* such as "I hate mathematics". It is notable that this item correlates quite well with the expected factor structure even in the lowest ability groups (see Table 5 in Section 3.3). These types of items are pure opinions with simple wordings and thus easy to judge even by the reasonably low level of abstract thinking.

2) Somewhat more complex statements are such where there is a *short and straight negative wording with a more abstract non-extreme expression,* such as "Mathematics is boring". Though the stem is short, the item includes an ambiguous word of "boring": What is boring? How does one discriminate and evaluate the level of "boring"? Mathematics in school can be boring for those who are extremely *good*; on the other hand, mathematics can be boring for those who are extremely *poor* in understanding what is taught. Teaching can be boring; exercises can be boring – what are we seeking with an ambiguous term? Compared with the simplest stem, there is a need for more intellectual processing to judge whether the opinion is the same or the opposite.

3) The third level complexity comes when the statement is *short but includes ambiguous wordings and opposing trigger "not"* like in "Mathematics is not one of my strengths" which really is a complex statement. How to measure strength? What actually is meant by "is NOT strength"? Is it weakness? Or is it in the middle range of my abilities? To make the judgment, there is a need for several decisions whereas the same kind of information can be gathered from much simpler sentence constructions, such as "I am weak in mathematics" or "I am (not)

good in mathematics". Other options are discussed in Section 4.

4) The fourth level complexity is found in the most complex and intellectually demanding wording in the TIMSS 2007 attitude test; *a long sentence with contradicting double expression and ambiguous wording* coining a positive expression (such as "more") with a negative expression (such as "difficult") like in "Mathematics is more difficult for me than for many of my classmates". Complexity comes first, from comparative trigger "more than" which needs ability to high level comparison, second, from several ambiguous words, such as, "difficult", "many", "more", and their combinations which need several cognitive processes, and third, the long sentence which requires good working memory. The item would certainly be much more unambiguous to respond though giving the same information with just a straightforward wording "Mathematics is difficult/easy to me". Some other alternatives are discussed in Section 4.

5) The fifth level complexity is not in Fennema–Sherman scale as suggested, for example, in Mehrens and Lehmann (1991, 108, 201―202 and originally Edwards, 1957). This level complexity could be found in a *double negative statement with ambiguous wording* such as "Mathematics is not one of my weaknesses"—which obviously would have been at rick or quip, rather than a real statement.

This hierarchy is, obviously, just a civilized guess of psychological processes in human brains. It would be interesting to try to verify the logic by an experimental design – or set of designs; the tools of cognitive psychology could be used in the process, for example.

Another area to develop is more urgent: to amend the two ill-behaving negative items in the Fennema-Sherman test used in the international settings. In Section 4, relevant empirical evidence is given, on the basis of comparable modified Fennema-Sherman test, of items which could be used when replacing the challenging items. Above, some heuristic suggestions are also given. The alternatives are collected here.

In the international testing settings, simpler items seem to operate better than complex ones. The following principle might be worth consideration as a basic rule of thumb: in the international testing settings, items from complexity Categories 1 or 2 should be used instead of items from Categories 3 and 4 (see description of categories above). When following the rule, the item "Mathematics is not one of my strengths" could be replaced by

1) "I'm weak in mathematics" or

2) "I'm (not) good in mathematics" discussed above, or by

3) "I can manage even with the difficult tasks in Mathematics", or

4) "It is impossible for me to get good results in Mathematics" suggested in Section 4 on the basis of tested Finnish items.

The item *"Mathematics is more difficult for me than for many of my classmates"* could be replaced by either

1) "Mathematics is difficult for me" or

2) "Mathematics is (not) easy for me" as discussed above, or by

3) "Mathematics is an easy subject" or

4) "Many things in Mathematics are difficult" as suggested in Section 4 on the basis of tested Finnish items.

All eight alternatives belong to complexity categories 1 or 2. It is understood that any serious changes require that the suggested items be pre-tested with the international data. It is especially recommended that *the pre-testing begins in East Asian- or Middle Eastern countries* because the discrepancy between the models is largest in those areas. Another suggestion is that at least some of *the negative items should be changed to positive ones.* However, from the contemporary (North American-European?) psychometrical viewpoint, the negative items are important to ensure the consistency of the respondent. Though the positive alternatives to substituting the poor behaving items seem to be more discriminative (see Fig. 10 and 11), it is therefore recommended to reduce the number of negative items as low as possible—in practice to one negative item per dimension.

Third possible area of further development is to perform deeper study of cultural issues behind the results – here only the technical aspect of the test was addressed. Especially, the East Asian well-performing countries (Korea, Japan, Taiwan, Hong Kong, and Singapore) would be interested focus of further analysis: what kind of connection there are between the cultures and mental processes of the students in these countries where the learning results do not explain the possible unexpected fragmentation on the factor structure.

The results raise two serious issues. First, the results challenge the whole idea of using one common test across

the globe to test mental structures. It especially challenges the idea of using an excessive number of negative items in the international attitude tests. No problem appears when testing and comparing students from the same types of cultures—like in students in the United States and in Europe. However, caution is wisdom when it comes to the comparison and inferring something over the datasets of Asia and Middle East or of Europe and Africa. Second, the connection of achievement level and factor deconstruction raises question of testing and reporting the results of the lowest achievement level students. If the respondent does not understand the abstract meaning in the attitude statement, or there are some (unknown) cultural elements connected to the pattern of response to the attitude items, the score does not mean anything when reliability remains low. When the reliability of the score is lower than $\alpha = 0.60$ (as it is in the lowest achievement groups), one could use any Internet test to achieve the same, if not better, consistency than by using well-tested and well-documented Fennema–Sherman test. In any case, closer analysis is needed in interpreting the correlations between attitude scales and achievement scales in international comparisons; a more in-depth analysis of lower group connections should be carried out.

References

Arbuckle, J. L. (2007). *AmosTM 16 User's Guide.* Amos Development Corporation. USA.

Bentler, P. M. (1988). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238–246. http://dx.doi.org/10.1037/0033-2909.107.2.238

Bentler, P. M. (1995). *EQS: Structural Equations Program Manual.* Multivariate Software Inc, Encino, CA.

Bentler, P. M., & Bonnet, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*(3), 588–606. http://dx.doi.org/10.1037/0033-2909.88.3.588

Bollen, K. A. (1986). Sample size and Bentler and Bonett's non-normed fit index. *Psychometrika, 51*(3), 375–377. http://dx.doi.org/10.1007/BF02294061

Bollen, K. A. (1989a). *Structural Equations with Latent Variables.* John Wiley & Sons, New York.

Bollen, K. A. (1989b). A new incremental fit index for general structural equation models. *Sociological Methods and Research, 17*(3), 303–316. http://dx.doi.org/10.1177/0049124189017003004

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.) *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage Publications.

Byrne, B. M. (2001). *Structural Equation Modeling with AMOS. Basic Concepts, Applications, and Programming.* Lawrence Erlbaum Associates, Publishers. London.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334. http://dx.doi.org/10.1007/BF02310555

Cureton, E. F. (1966). Corrected item-test correlations. *Psychometrika 31*(1), 93–96. http://dx.doi.org/10.1007/BF02289461

Edwards, A. L. (1957). *The social desirability variable in personality assessment and research.* New York: Hols, Rinehart & Winston.

Fennama, E., & Sherman, J. A. (1976). Fennema-Sherman mathematics attitude scales: Instruments designed to measure attitudes toward the learning of mathematics by females and males. *Journal for Research in Mathematics Education, 7*(5), 324-326.

Gulliksen, H. (1987/1950). *Theory of Mental Tests.* Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Henrysson, S. (1963). Correction of item-total correlations in item analysis. *Psychometrika, 28*(2), 211–218. http://dx.doi.org/10.1007/BF02289618

Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement, 60*(4), 523–531. http://dx.doi.org/10.1177/00131640021970691

House, J. D., & Telese, J. A. (2008). Relationships between Student and Instructional Factors and Algebra Achievement of Students in the United States and Japan: An Analysis of TIMSS 2003 Data. *Educational Research and Evaluation, 14*(1), 101—112. http://dx.doi.org/10.1080/13803610801896679

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika, 32*(4), 443–482. http://dx.doi.org/10.1007/BF02289658

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34*(2), 183–202. http://dx.doi.org/10.1007/BF02289343

Jöreskog, K. G. (1970). A general method for analysis of covariance structures. *Biometrika, 57*(2), 239–251. http://dx.doi.org/10.1093/biomet/57.2.239

Jöreskog, K. G. (1973). A general method for estimating a linear structural equating system. In A. S. Goldberger & O. D. Duncan (Eds.), *Structural Equation Models in the Social Sciences* (pp. 85–112). New York: Academic Press.

Jöreskog, K. G., Sörbom, D., Du Toit, S., & Du Toit, M. (2003). *LISREL 8: New Statistical Features* (3rd Ed.). Lincolnwood, IL: Scientific Software International.

Kadijevich, Dj. (2006). Developing trustworthy TIMSS background measures: A case study on mathematics attitude. *The Teaching of Mathematics, 9*(2), 41-51. Retrieved from http://elib.mi.sanu.ac.yu/journals/tm/17/tm924.pdf

Kadijevich, Dj. (2008). TIMSS 2003: Relating Dimensions of Mathematics Attitude to Mathematics Achievement. *ZbornikInstitutazapedagoskaistrazivanja, 40*(2), 327-346. http://dx.doi.org/10.2298/ZIPI0802327K

Keesling, J. W. (1972). Maximum Likelihood Approaches to Causal Analysis. Ph.D dissertation. Department of Education. University of Chicago.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*(3), 151–160. http://dx.doi.org/10.1007/BF02288391

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores.* Reading, Mass.: Addison – Wesley Publishing Company.

Mehrens, W. A., & Lehman, I. J. (1991). *Measurement and evaluation in education and psychology* (4th ed.). Fort Worth: Harcourt Brace College Publishers.

Mattila, L. (2005). National Achievement Results in Mathematics in Compulsory Education in 9th grade 2004. Oppimistulostenarviointi 2/2005.Opetushallitus. Helsinki: Yliopistopaino. [In Finnish.]

Metsämuuronen, J. (2009). Methods Assisting Assessment; Methodological solutions for the National Assessments and Follow-Ups in the Finnish National Board of Education. Oppimistulosten arviointi 1/2009. Opetushallitus. Helsinki: Yliopistopaino. [In Finnish.]

Mullis, I. V. S., Martin, M. O., Foy, P., Olson, J. F., Preuschoff, C., Erberber, E., Arora, A., & Galia, J. (2008). TIMSS 2007 International Mathematics Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grade. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from http://pirls.bc.edu/timss2007/mathreport.html

Papanastasiou, C. (2000). Effects of Attitudes and Beliefs on Mathematics Achievement. *Studies in Educational Evaluation, 26*(1), 27-42.

Papanastasiou, E. (2002). Factors That Differentiate Mathematics Students in Cyprus, Hong Kong, and the USA. *Educational Research and Evaluation: An International Journal on Theory and Practice, 8*(1), 129—146. http://dx.doi.org/10.1076/edre.8.1.129.6919

Piaget, J. (1970). *The principles of genetic epistemology.* London: Routledge & Kegan Paul.

Shen, C. (2002). Revisiting the relationship between students' achievement and their selfperceptions: A cross-national analysis based on the TIMSS 1999 data. A*ssessment in Education: Principles, Policy & Practice, 9*(2), 161—184. http://dx.doi.org/10.1080/0969594022000001913

Shen, C., & Tam, H. P. (2008). The Paradoxical Relationship between Student Achievement and Self-Perception: A Cross-National Analysis Based on Three Waves of TIMSS Data. *Educational Research and Evaluation, 14*(1), 87-100. http://dx.doi.org/10.1080/13803610801896653

Steiger, J. H., & Lind, J. C. (1980). Statistically-based tests for the number of common factors. Paper presented at the Annual Spring Meeting of the Psychometric Society, Iowa City.

Stevenson, H. W. (1998). A Study of Three Cultures: Germany, Japan and the United States - An Overview of the TIMSS Case Study Project. *Phi Delta Kappan, 79*(7), 524-529.

Tarkkonen, L. (1987). On Reliability of Composite Scales. An Essay on the measurement and the properties of the coefficients of reliability-an unified approach. Tilastotieteellisiätutkimuksia 7. Helsinki: Finnish

Statistical Society.

Ullman, J. B. (2001). Structural Equation Modeling. In B. G. Tabachnick & L. S. Fidell, *Using Multivariate Statistics* (4th Ed.) (pp. 653–771). Boston: Allyn and Bacon.

Vehkalahti, K. (2000). Reliability of Measurement Scales. Statistical Research Reports 17. Finnish Statistical Society. Retrieved from http://ethesis.helsinki.fi/julkaisut/val/tilas/vk/vehkalahti/

Wiley, D. E. (1973). The identification problem for structural equation models with unmeasured variables. In A. S. Goldberger & O. D. Duncan (Eds.), *Structural Equation Models in the Social Sciences* (pp. 85–112). New York: Academic Press.

Wilkins, J. L. M. (2004). Mathematics and science self-concept: An international investigation. *The Journal of Experimental Education, 72*(4), 331—346. http://dx.doi.org/10.3200/JEXE.72.4.331-346