Profiles of scientific competence in TIMSS 2003: Similarities and differences between countries

Marit Kjærnsli and Svein Lie
Department of Teacher Education and School Development,
University of Oslo, Norway

Keywords: science, cultural differences, item analysis, cluster analysis, cognitive domains

Abstract

The aim of the present contribution is to investigate similarities and differences of strengths in science competences between countries, based on TIMSS 2003 data. Analyses have been based on systematic investigation of patterns of p-values (percentage correct) for individual science items. Hierarchical cluster analysis has been applied to establish meaningful groups of countries. The resulting pattern of how countries cluster together into groups of increasing size, based on similarities of strengths and weaknesses are presented and discussed. As a measure of similarity between countries we have applied the Pearson correlation coefficient between the p-value residuals (i.e. each country's set of p-values, corrected for the country's average of all items and the international item difficulty).

For each of the groups of countries, average p-value residuals have been calculated to investigate characteristic features. These features are described in terms of separate measures of relative strengths according to item format, subject domain, and cognitive domain. Finally, data on relative emphases in the intended curriculum (curriculum documents) and in the implemented curriculum (percentage of topics taught) is shown to explain to a considerable degree the patterns of achievements within the different content domains.

Introduction

In a number of earlier papers, similarities and differences between countries concerning cognitive strengths and weaknesses have been analysed based on achievement data from TIMSS 1995 (Grønmo, Kjærnsli & Lie, 2004; Angell, Kjærnsli & Lie, 2006). Similar analyses have been carried out based on data from the OECD PISA study (Lie & Roe, 2003; Kjærnsli & Lie, 2004; Olsen, 2005). These analyses have been based on systematic investigations of patterns of p-values (percentage correct) for individual achievement items. Following a method proposed by Zabulionis (2001), hierarchical cluster analysis has been applied in these studies as a tool to establish meaningful groups of countries based on similar areas of relative strengths and weaknesses.

The aim of the present contribution is to further investigate these patterns of cognitive strengths in science based on TIMSS 2003 data (Martin, Mullis, Gonzalez & Chrostowski, 2004). In 2003 more countries participated than in earlier studies, which allows us to say more about how countries seem to group together. Further, we will investigate the characteristic features for each of these country groups, and the findings will be discussed in light of cultural traits and traditions in science education. Finally we will present more evidence to help us understand which factors that lie behind the mechanism of this clustering of countries. In particular, we will try to understand how curricular factors are influencing this pattern.

In large-scale international studies like TIMSS, uni-dimensional models are applied for scaling student competencies. Consequently, pronounced different item functioning (DIF) across countries is often regarded as a source of measurement error, thus used as criteria for item exclusions. On the other hand, it has been repeatedly shown by test-curriculum analyses (for TIMSS 2003, see Martin et al.,

2004, Appendix C) that the exact selection of items for the test have only minor influence on the countries' relative standing. The position in the present study is that the differential item functioning brings some very interesting information on strengths and weaknesses of individual countries. We will investigate this effect in a systematic but simple way based on the p-value residuals mentioned above. In order to prevent too many cases (countries), we have used country groups as our unit of analysis. In an earlier analysis based on data from TIMSS 1995 (Angell et al, 2006), we applied a similar strategy to construct country groups, but further analyses used one country from each group as the unit of analysis. Since more detailed data were available on curricular factors in 1995 (Schmidt, Raizen, Britton, Bianchi, & Wolfe, 1997) we could then go into more details for these countries. For our present investigation we will apply data from questionnaires to science teachers and national research coordinators.

Clusters of countries

The basis for our analysis a complete matrix of p-values by country, covering 190 items (or rather: score points) and 50 countries (including a few regions within a country). For each cell in this matrix we have calculated the p-value residual, i.e. how much better or worse (in percentage correct) the particular country achieved on the particular item compared to what is expected from the over-all achievement of the country (for all items) and the over-all difficulty of the item (for all countries). By applying hierarchical cluster analysis to the p-value residual matrix, we obtain a pattern of relations between countries. This pattern can be displayed in a so-called "dendrogram" that from left to right illustrates how countries cluster together into groups of increasing size, based on similarities of strength and weaknesses. By moving from left to right, that is from high positive to negative correlations, countries are clustered into even larger groups until they all are united. As a measure of similarity between two countries or between already established groups of countries we have applied the (Pearson) correlation coefficient between the p-value residuals. There are alternative criteria (Olsen, 2005), but the results are similar, even if details depend on the exact method being applied. Thus, the picture presented in the following represents a reasonably stable solution.

The dendrogram in figure 1 immediately draws our attention to the remarkable pattern of meaningful groups. We are therefore able to define certain groups of countries that can be identified and labelled according to either location (political or regional unit) or cultural trait (e.g. language). For our further analysis we have concentrated on the following rather distinct country groups with at least three countries:

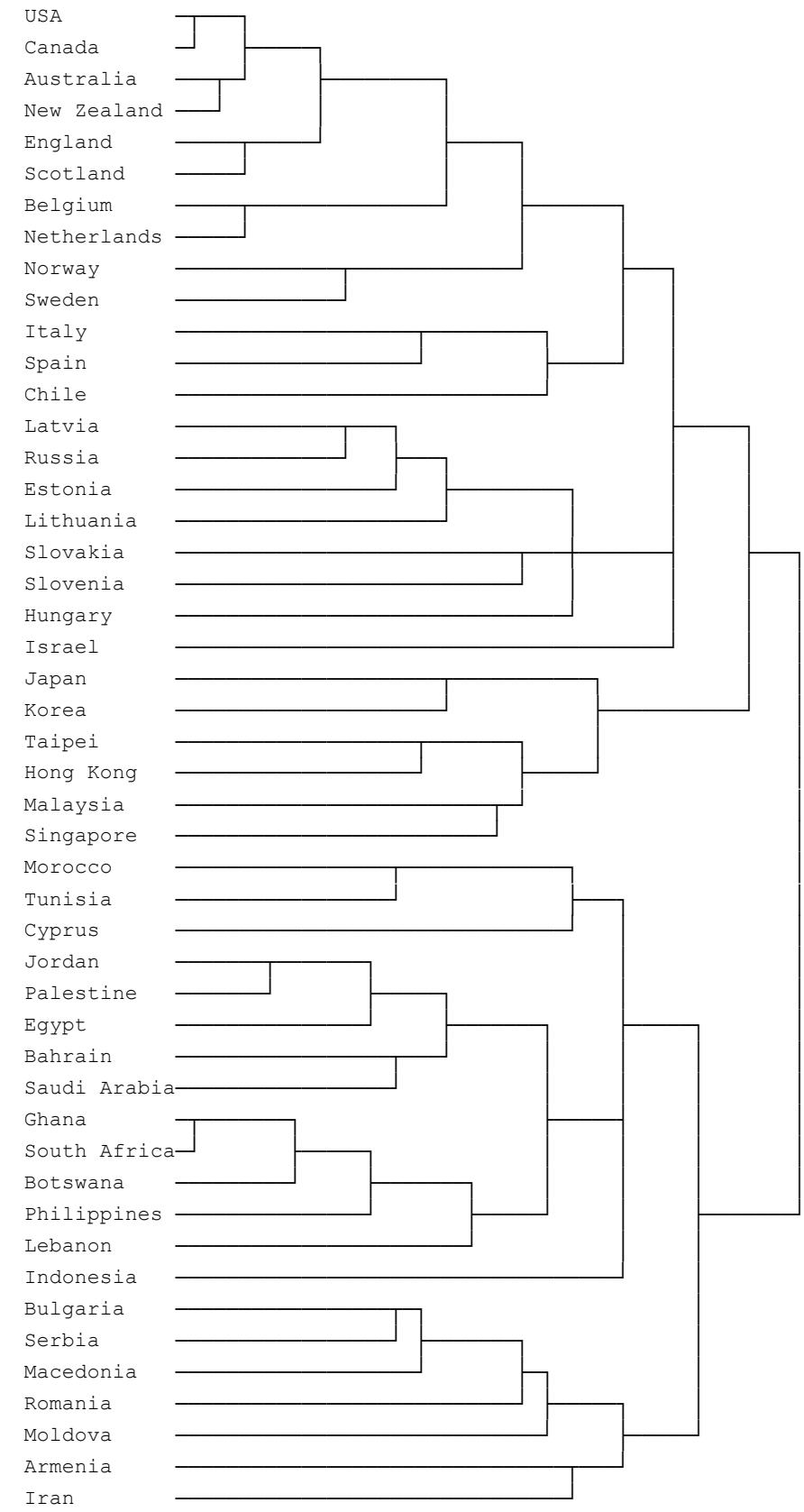
- English-speaking: Australia, Canada (Ontario and Quebec), England, New Zealand, Scotland, USA
- East-Central Europe: Estonia, Hungary, Latvia, Lithuania, Russia, Slovakia and Slovenia
- East-Asia: Chinese Taipei, Hong Kong SAR, Japan, Korea, Malaysia and Singapore
- South-East Europe: Bulgaria, Macedonia, Moldova, Rumania and Serbia
- Arabic: Bahrain, Egypt, Jordan, Palestine and Saudi-Arabia
- Southern Africa: Botswana, Ghana and South Africa
- Roman: Italy, Spain (Basque province) and Chile

In addition, we want to include two pairs of countries of particular interest to us:

- Nordic: Norway, Sweden
- Dutch: The Netherlands and Flemish Belgium

The labels used above should not be taken too literally, but rather represent labels of reference. These nine groups of countries will be in our focus for the rest of this paper. By calculating average p-value residuals for each group we can analyse how these values relate to characteristic features for items. In this way we are able to investigate some main characteristics of cognitive strengths and weaknesses for each country group.

Figure 1: Dendrogram for clustering of countries according to similarities between countries in patterns across science items



Firstly, we will compare the pattern described above with findings by the same method from other data sets. The main message is that the general patterns established from earlier analyses of science achievement data from TIMSS 1995 (Angell et al., 2006), PISA 2000 (Kjærnsli & Lie, 2004, Grønmo et al., 2004) and PISA 2003 (Olsen, 2005) are confirmed here. Even if details are different, mainly due to the fact that countries' participation in the various studies varies, the pronounced linkages within each of the English-speaking countries, the East-(Central) European countries, the East Asian countries, the Nordic countries and the Dutch "countries" are confirmed. In addition, three other distinct groups appear, the Arabic countries, the Southern African countries and the South East European (or Balkan) countries. The group of "Roman" countries, linked together by our data are, however, not so easily labelled, and the label applied should not be taken literally.

Characteristic features for country groups

Now we will turn our attention to the essential features that are characteristic of each of the groups discussed above: What do countries in each group have in common? Our approach consists of classifying all science items according to some selected criteria, and investigating how these classifications relate to the patterns of p-value residuals for each country groups (Olsen, 2005).

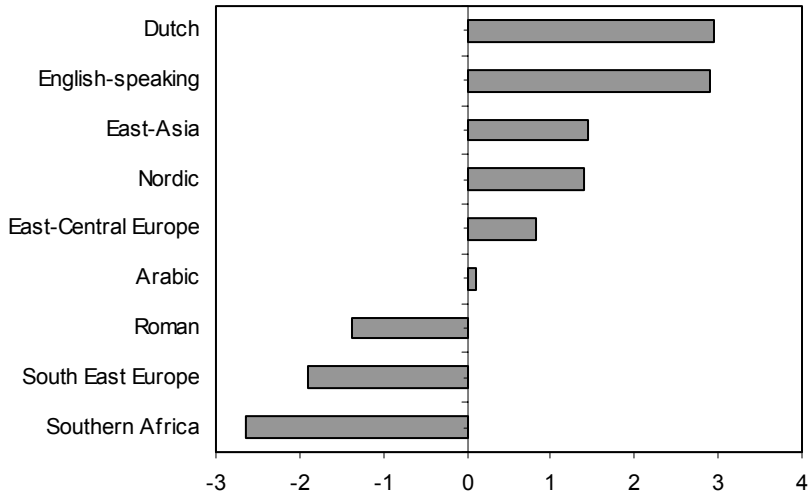
The following item criteria will be applied:

- Item format: constructed response vs multiple choice
- Science content: Life science, Chemistry, Physics, Earth science, or Environmental science
- Cognitive domain: Factual knowledge, Conceptual understanding, or Reasoning and analysis

Item Format

The TIMSS achievement test consists of both multiple choice items and constructed response items. The distribution between these two formats was about 60 percent of multiple choice and 40 percent constructed response items. In the following we will look closer to how the country groups performed within these two item formats. Figure 1 compares the relative strengths within the two formats.

Figure 2: Constructed response items versus multiple choice items for each country group. Positive values in favor of constructed response items.

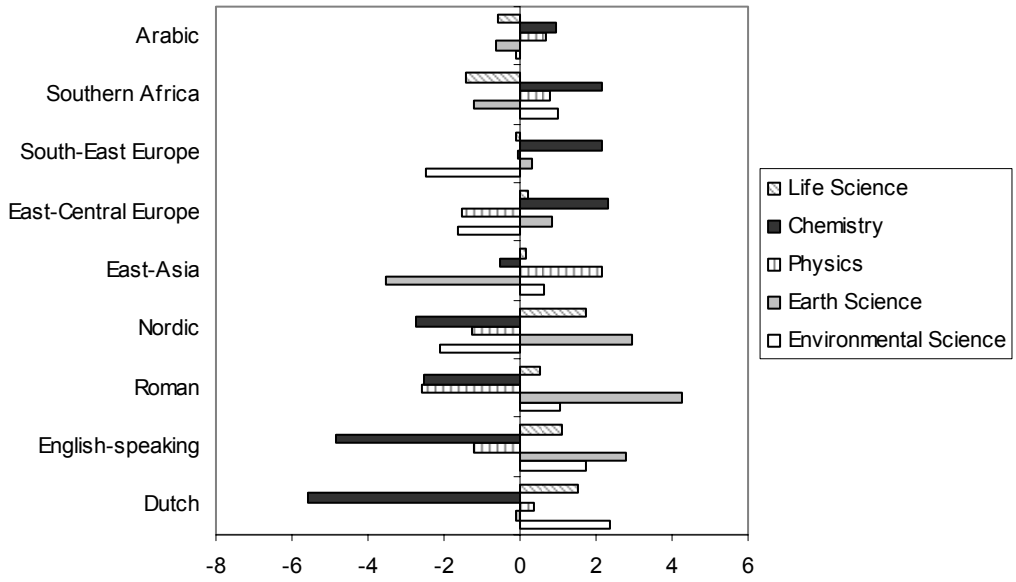


From figure 2 it can be seen that in particular the Dutch and the English-speaking groups of countries perform relatively better on constructed response items. On the other hand, the groups from Southern Africa and South East Europe perform particularly better on multiple choice items.

Science content domains

TIMSS framework defines five content domains in population 2*, Life Science, Chemistry, Physics, Earth Science and Environmental Science*. Each content domain has several main topic areas that are presented as a list of specific assessment objectives. For a more detailed description, see TIMSS Assessment Framework (Mullis, Martin, Smith, Garden, Gregory, Gonzales, Chrostowski & O'Connor, 2001).

Figure 3: Achievement in science content domains for each country group, sorted by increasing spread among the domains



In figure 3 the relative strengths in each content domain are displayed. The country groups have been sorted by increasing spread among the domains, so that the country groups with the most distinguished profiles appear towards the bottom. Some remarkable characteristics stand out in this figure. The most pronounced is the fact that the variation between groups is much larger in chemistry and partly in earth science than in life science and physics. The Dutch and English-speaking countries perform relatively much worse in chemistry than in the other content domains. Concerning earth science, a pronounced strength is seen in the Nordic, Roman and English-speaking groups, and a weakness appears in the East-Asian group.

One extreme case may illustrate the situation for chemistry in the Dutch and English-speaking countries. The following item (S022202) represents the main topic "Particular structure of matter" within the cognitive domain of factual knowledge.

What is formed when a neutral atom gains an electron?

A. A mixture
B. An ion
C. A molecule
D. A metal

The p-value residuals for this item for the Dutch and the English-speaking groups are as low as -30 and -25, respectively. This means as many percentage points as 30 and 25 respectively lower than what is expected based on the over-all abilities for these country groups and the over-all difficulty of this item. The item requires just recognizing the correct term, and this is a signal that such information is not regarded as an important part of the chemistry curriculum in these countries.

Cognitive domains

All science items in TIMSS 2003 have been classified into one of three cognitive domains according to how the students are expected to act or type of cognitive activity required, to reach a correct response. These three domains are assessed across the science content domains: *Factual Knowledge, Conceptual understanding* and *Reasoning and understanding* (Mullis et al., 2001).

Within the category *Factual knowledge* the students need to demonstrate knowledge of relevant science facts, information, tools and procedures. Thus, the concept Factual knowledge involves more than just memorization and recall of isolated bits of information.

Conceptual understanding requires students to extract and use scientific concepts and principles to find solutions and develop explanations, support of statements of facts or concepts, demonstrate relationships, equations and formulas in context. The problems in this cognitive domain are designed to involve more straightforward applications of concepts and require less analysis than items that are categorized in the domain Reasoning and analysis.

Reasoning and analysis covers challenges like to solve problems, develop explanations, draw conclusions, make decisions and extend knowledge to new situations. The students are for example expected to evaluate and make decisions based on their conceptual understanding. Some items require students to bring knowledge and understanding from different areas and apply it to new situations.

Figure 3: Achievement in the cognitive domains for each country group, sorted by increasing spread among the domains

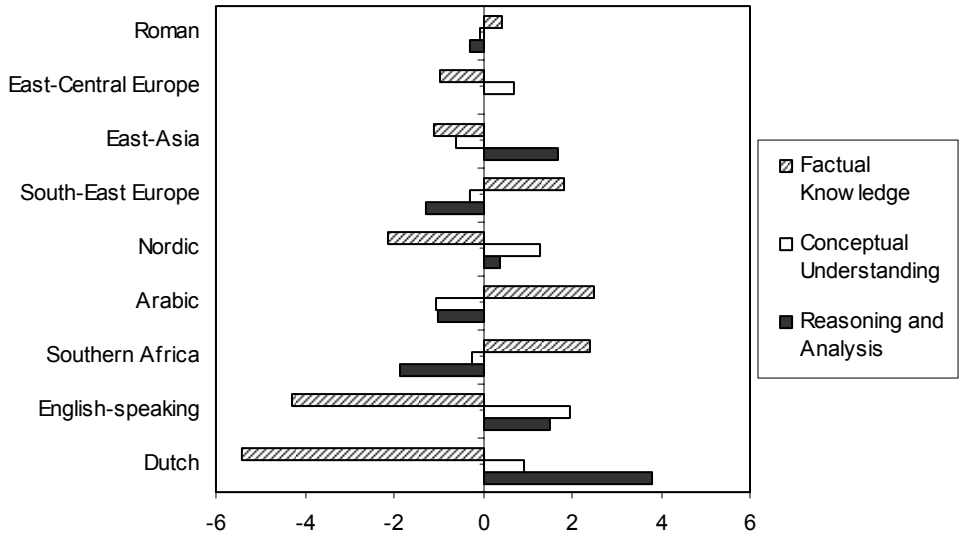


Figure 3 displays the relative strengths and weaknesses concerning cognitive domains among the country groups. Like in figure 2 the most distinguished profiles appear near the bottom of the figure. And again, the Dutch and English-speaking groups stand out and with similar profiles: a particular relative weakness in *Factual knowledge*. It further appears that the *Factual knowledge* domain has the most variation between the groups, whereas the *Conceptual understanding* domain varies much less.

The role of the intended and the implemented curriculum

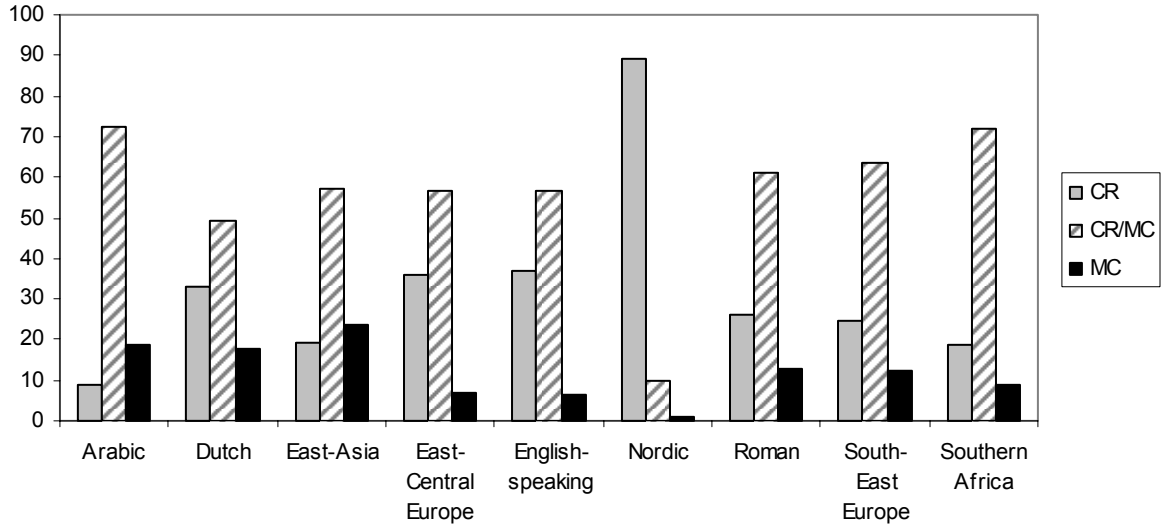
After having discussed some characteristics of each country group, we will for the remaining part of this paper focus on to what extent other data from TIMSS can explain these characteristics. In particular we will investigate the role of what in TIMSS are called the *intended* and the *implemented* curriculum, respectively. By intended curriculum is here meant the curricular documents and how these give prescriptions for distribution of emphasis across subject and cognitive domains. The

implemented curriculum refers to what is actually taught and the emphasis given to the different aspects.

Item format

The science teacher questionnaire included a question on the relative frequency of different item formats in the assessment of students in science. These data (given in Exhibit 7.13 in Martin et al., 2004) is graphically displayed for our country groups in figure 4.

Figure 4: The relative distribution of item formats used in science assessments (CR = mostly Constructed response, MC = mostly Multiple response, CR/MC = about half of each)



From figure 4 we can see that the dominant response is an even distribution of the two item formats. However, the Nordic group (i.e. Norway and Sweden) stands out with a very different profile with multiple choice items playing essentially no role in assessment practice. This is interesting information in itself, but it does not provide any explanation to the pattern shown in figure 2. Neither does the other features in the above figure offer much explanation to figure 2. Thus we conclude that the item formats applied do not seem to play any strong role in shaping the results in the TIMSS science test.

Science content domains

The intended curriculum

The national research coordinators in TIMSS 2003 responded to a questionnaire on the national context for mathematics and science in their respective countries. They were to respond to specific questions on curricular coverage for a series of science (and mathematics) topics as they were described in the framework (Mullis et al., 2001). For each of these 44 science topics, there is data on whether or not it is expected to be taught up to and including the actual grade (eighth grade for most countries). We have taken these data as our measure of the intended curriculum, and applied them in the form of the *percentage of the given topics covered*, such as these are given for each subject domain in Martin et al. (2004, exhibit 5.7).

Table 1 shows the intended curriculum by the above method averaged within each country group. It should be stated that the Environment domain contains only three topics, thus the data in table 1 are less reliable for this domain than the others. The far right column gives the (Pearson) correlation of these five numbers with the corresponding p-value residuals displayed in figure 2. These correlations are all positive, of medium size for most groups. The average across all groups is 0.36. Thus we notice a clear and positive relationship between (relative) achievement and curricular emphasis in the data.

Table 1: Percent of science topics in the TIMSS framework covered by the national intended curriculum (to be taught up to and including grade 8) and correlation with achievement

Groups of countries	Life science	Chemistry	Physics	Earth science	Environ-mental science	Correlation with p-value residuals
Arabic	93	83	96	75	100	0.13
Dutch	81	41	36	25	83	0.54
East-Asia	69	68	79	55	56	0.71
East-Central Europe	79	92	91	92	90	0.03
English-speaking	76	67	75	78	61	0.27
Nordic	92	67	71	78	100	0.16
Roman	82	73	63	67	78	0.10
South-East Europe	82	93	96	87	80	0.65
Southern Africa	56	66	73	64	72	0.64

The implemented curriculum

Next we will investigate the parallel relationship between achievement and the *implemented* curriculum. In the science teacher questionnaire the teachers were asked about which out of the list of 44 framework topics that actually would be taught up to and including the present school year. Exhibit 5.8 in Martin et al. (2004) gives this information in the form of the percentage of students that has been taught the topics within each of the content areas. Table 2 gives this information for each of the country groups, in addition to the Pearson correlation with p-value residuals. A few countries did not provide comparable data for environmental science, so there are three empty cells in the table. The correlations are in these cases calculated for the four other domains only.

Table 2: Average percent of students taught the TIMSS science topics and correlation with achievement

Groups of countries	Life science	Chemistry	Physics	Earth science	Environ-mental science	Correlation with p-value residuals
Arabic	74	76	83	61	54	0.55
Dutch	72	33	39	42		0.68
East-Asia	59	71	71	34	38	0.59
East-Central Europe	73	80	61	88		0.80
English-speaking	65	65	64	66	57	-0.28
Nordic	54	54	48	68	33	0.70
Roman	83	72	67	76	69	0.46
South-East Europe	87	93	92	91		0.55
Southern Africa	51	44	42	28	45	0.23

Not unexpectedly we notice that the correlations are generally somewhat higher in table 2 than in table 1. The average is 0.48. Our data on emphasis in the classrooms explain more of the relative strengths and weaknesses than does the intended curriculum.

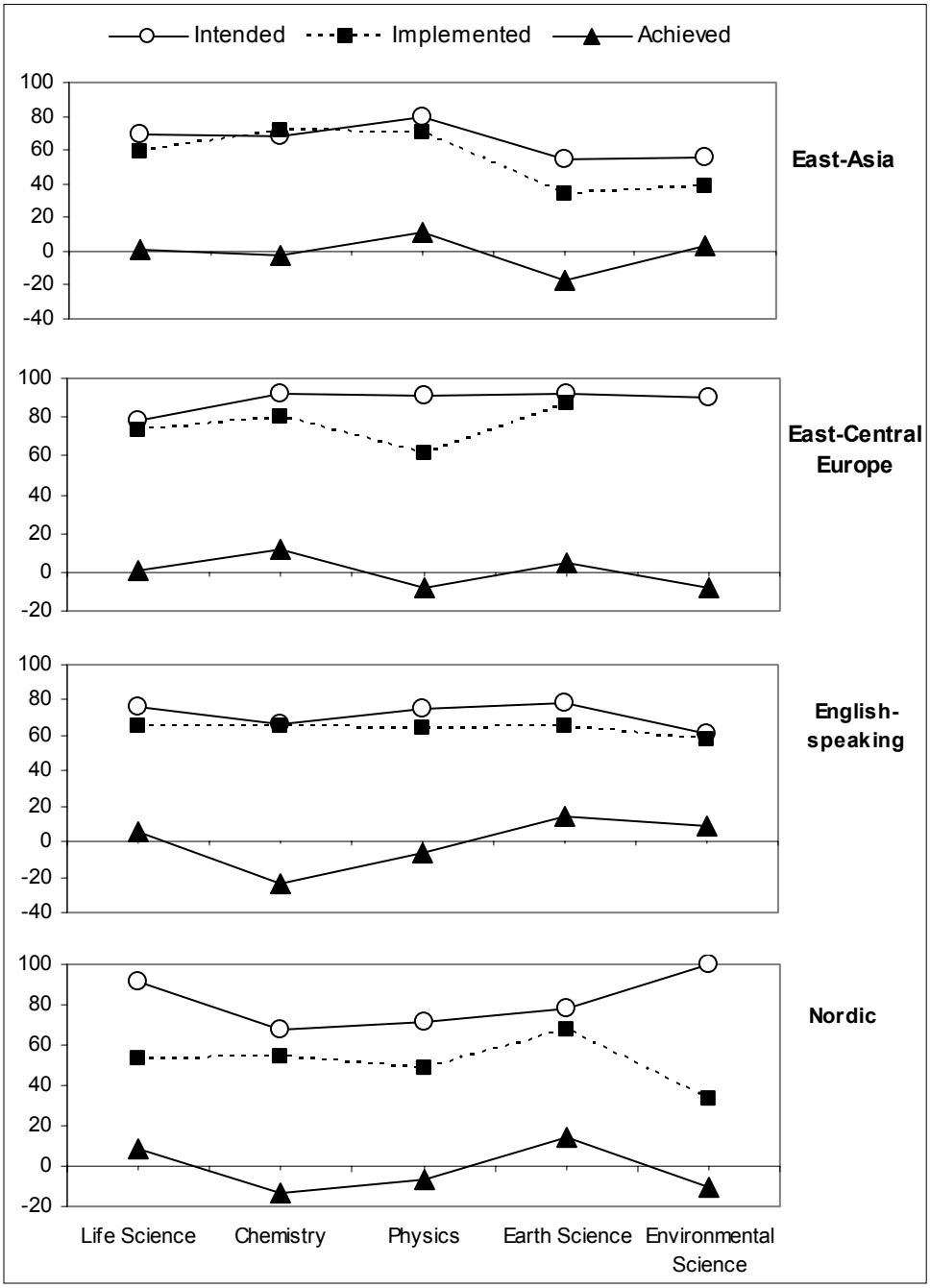
A comment should be given on the negative correlation for the English-speaking group in table 2. Whereas the low coverage for chemistry in table 1 is reflected in figure 3, this is not paralleled for table 2. There the English-speaking "profile" does not show any particular drop for chemistry. It appears that even if the chemistry topics are reasonably well covered by teaching, the characteristics of how the chemistry topics are addressed may to some degree be at odds with what is required by the TIMSS chemistry items. The item discussed above (S022202) appears to be an extreme example.

Intended, implemented and achieved curriculum

In figure 5 we have tried to illustrate all three levels, *Intended, Implemented and Achieved curriculum* for some of the country groups. Here we have applied the TIMSS notation of *Achieved curriculum* for the assessment results, in the form of p-value residuals. In order to simplify comparison between the

shapes of the three curves, the achieved curriculum data have somewhat arbitrarily been multiplied by 5. Although the numbers are not directly comparable, it is nevertheless interesting to compare the shape of the three curves for some selected groups of countries. We have here selected the three clusters with most countries in addition to the Nordic group which is of special interest to us.

Figure 5: Comparison between intended, implemented and achieved curriculum for four groups of countries (see text for explanation)



It clearly appears that there are similarities between the three curves for each of the displayed country groups. The situation for the other groups of countries is similar. However, there are some interesting differences, especially when looking at the results for the domain Environmental Science for the Nordic group. Here one can see a clear difference between the intended curriculum on one hand and the implemented and achieved curriculum on the other hand. The gap between the intended and implemented can not be explained by our data, but it is interesting to follow this question up. It could

well be that it is easier to give this area emphasis in the intended curriculum than to follow up in the classroom.

Cognitive domains

Finally we will study how much emphasis each group of countries put on the three cognitive domains in their intended science curricula. The data is taken from Exhibit 5.6 in Martin et al., (2004). In table 3 the first two categories are identical with what is reported by Martin et al. For the third category we have with some doubt collapsed the three "Writing explanations about what was observed and why it happened", "Formulating hypotheses or predictions to be tested" and "Designing and planning experiments or investigations" into one category, called "Reasoning and Analysis". This merging of categories is not obvious and contributes to the somewhat less credibility of the data in table 3 than in the other tables. Also the simple scale applied to measure "emphasis" (A lot of -, Some -, Very little –, and No emphasis) contributes to the lower data quality. However, in table 3 we nevertheless do calculate the correlations with the achievement data for the pattern of the three different cognitive domains.

Table 3: Emphasis in intended curriculum for each group of country

Groups of countries	Factual knowledge	Conceptual understandings	Reasoning and analysis
Arabic	3.80	4.00	3.27
Dutch	3.50	3.00	2.67
East-Asia	4.00	4.00	3.28
East-Central Europe	3.57	3.57	2.67
English-speaking	3.71	4.00	3.67
Nordic	3.50	3.50	3.33
Roman	4.00	4.00	3.00
South-East Europe	3.60	3.40	2.47
Southern Africa	3.67	3.67	2.67

As we might have suspected, these data cannot to any extent explain the profiles in figure 3. The average correlation coefficient with p-value residuals is as low as 0.07.

In an analysis of country differences regarding cognitive profiles in mathematics in TIMSS, Klieme & Baumert (2001) classified each item according to a set of cognitive demands. For each item the dependence on each of these categories were coded by a group of coders. By this multi-dimensional approach, data on differential item functioning were obtained and compared to what was expected from national analyses of various sources. Their analysis was applied to a few countries with meaningful results. It seems, on the other hand, that in our analysis with items classifies according to three mutually exclusive categories, we cannot obtain a meaningful relationship between curriculum and achievement information.

Conclusion

The aim of this article has been threefold. Firstly, we have identified some country groups of similar profiles of relative strengths from item to item. Secondly, we have described some characteristic features for each of these country groups. And thirdly, we have looked into some other TIMSS data, i.e. on emphases in the intended curriculum as well as implemented by teachers in the classrooms, and found that these to some extent provide explanations for the above patterns of features for country groups. However, our search for relating the cognitive profiles to curriculum factors did not lead to any further understanding.

In order to go deeper into the country (or group) profiles, one would need stronger tools to handle the differential functioning aspect. Firstly, instead of (residuals of) p-values, IRT measures of item difficulties would be more appropriate. And further, in the future it may be possible to scale the items

by applying within-item multidimensionality to model the inner complexity of individual items. Possibly one may then be able to build the cognitive profiles directly into the scaling procedure.

References:

Angell, C., Kjærnsli, M., Lie, S. (2006). Curricular Effects in Patterns of Student Responses to TIMSS Science Items. In Howie, S.J. & Plomp, T. (eds.): *Contexts of Learning Mathematics and Science*: Routledge. p. 277-290.

Grønmo, L. S., Kjærnsli, M., Lie, S. (2004). Looking for cultural and geographical factors in patterns of responses to TIMSS items. I: *Proceedings of the IRC-2004 TIMSS*: Cyprus University Press, Kailas Printers and Lithographers Ltd. p. 99-112.

Kjærnsli, M., Lie, S. (2004). PISA and Scientific Literacy: similarities and differences between the Nordic countries. *Scandinavian Journal of Educational Research* 2004 (Vol. 48, no. 3). p 271-286.

Klieme, E., Baumert, J. (2001). Identifying national cultures of mathematics education: Analysis of cognitive demands and differential item functioning in TIMSS. *European Journal of Psychology of education* Vol. XVI, no 3, p. 385-402

Lie, S., Roe, A.. Exploring unity and diversity of Nordic reading literacy profiles. I: *Northern lights on PISA. Unity and diversity in the Nordic countries in PISA 2000.* 2003. p. 147-157.

Martin, M.O., Mullis, I.V.S., Gonzalez, E.J., & Chrostowski, S.J. (2004). *TIMSS 2003 International Science Report: Findings From IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades.* International Study Center, Lynch School of Education, Boston College.

Mullis, I.V.S., Martin, M.O., Smith, T.A., Garden, R.A., Gregory, K.D., Gonzales, E.J., Chrostowski, S.J. & O'Connor, K.M. (2001): *TIMSS Assessment Frameworks and Specifications 2003.* International Study Center, Lynch School of Education, Boston College.

Olsen, R. V. (2005). An exploration of cluster structure in scientific literacy in PISA: Evidence for a Nordic dimension?. *NorDiNa* 2005;1 (1) p. 81-94

Schmidt, W. H., Raizen, S.A., Britton, E., Bianchi, L.J., and Wolfe, R. G. (1997): *Many visions, many aims. A Cross-National Investigation of Curricular Intensions in School science.* Kluwer Academic Publishers.

Zabulionis, A. (2001). Similarity of mathematics and science achievement of various nations. *Educational Policy Analysis Archives 9* (33)