Rolf V. Olsen is a research fellow at the Department of Teacher Education and School Development at the University of Oslo. He is in the process of completing his Phd thesis. In his work he is particularly concerned about how the data in large scale international comparative assessments, such as PISA and TIMSS, may be used in secondary analysis to address research questions in science education. He has a background in physics and prior to his Phd he wrote a Masters thesis about students understanding of quantum physics. Parallel to this he worked as physics teacher in upper secondary school. Beyond his work at the university he is the leader of the Norwegian association for physics teachers.

ROLF V. OLSEN
Department of Teacher Education and School Development, University of Oslo
rolfvo@ils.uio.no

An exploration of cluster structure in scientific literacy in PISA: Evidence for a Nordic dimension?

Abstract
The cognitive items covering the domain of scientific literacy in the Programme for International Student Assessment (PISA) are explored through an analysis of the item residuals (item-by-country interactions) with the aim of looking for a distinct Nordic pattern. The findings of a cluster analysis indicate that the profile across the Nordic countries is not very distinct. However, stable profiles are established for a number of other groups of countries, and the Nordic countries are shown to be members of a larger group of countries which is labelled North-West European countries. Furthermore, item characteristics are used to find possible explanations for the profiles.

Introduction

In this article patterns across cognitive items in scientific literacy (sometimes referred to as 'science' throughout the article) from the OECD study Programme for International Student Assessment implemented in 2003 (PISA 2003) are explored. From prior research on similar data it is reasonable to expect that countries with geographical, linguistic, political or economical similarities cluster together. Of specific interest in this paper are the Nordic countries that in prior studies have been shown to have profiles across cognitive items that are relatively similar to each other. Indications for such a Nordic cluster have been established in analysis of reading items from PISA 2000 (Lie & Roe, 2003), analyses of mathematics items from the Third International Mathematics and Science Study (TIMSS 1995) (Grønmo et al., 2004b; Lie et al., 1997; Zabulionis, 2001) as well as in analyses of science items from TIMSS 1995 (Angell et al., in press; Grønmo et al., 2004b; Lie et al., 1997) and science items in PISA 2000 (Kjærnsli & Lie, 2004). A Nordic profile is particularly present in the analysis of items from TIMSS 1995, while in the analysis of PISA 2000 items the indications are weaker. The latter may be due to the fact that science and mathematics were minor domains in PISA 2000, and as a consequence the number of items was quite low. It is also worth commenting here that Finland did not participate in TIMSS 1995 while all the Nordic countries participate in PISA. In the analysis of the PISA 2000 data referred to above, it was especially Finland that did not cluster together with the other Nordic countries, followed by Denmark that also had a profile that to some degree was drawn away from the Nordic cluster.

In the above mentioned analyses of data from PISA 2003 and TIMSS 1995 other clusters of countries were even more strongly present. In the analyses of the science data in TIMSS 1995 (Angell et al., in press; Grønmo et al., 2004b; Lie et al., 1997) the English speaking countries had the most distinct profile. Furthermore, in this analysis the German speaking countries, East-European countries and East-Asian countries clustered together. In addition some strong pairs of countries (France & Belgium French and the Netherlands & Belgium Flemish) were present. Lastly, in TIMSS 1995 there was a cluster of underdeveloped countries (Columbia, Philippines and South Africa). In the cluster analyses of science items in the PISA 2000 data (Kjærnsli & Lie, 2004) the English group of countries was again a dominant cluster in the solution, and also a German speaking cluster (including Denmark) and a cluster consisting of the countries Portugal, Brazil and Mexico were quite distinct. In addition there were indications for an East European cluster.

Although the above mentioned studies applied a method similar to the one used in this article, none of them used the items themselves in order to give a more detailed description of the profiles. This article will therefore seek to reconfirm the cluster structure found in these studies, including a more thorough evaluation of the stability of the solution. Furthermore, broad descriptors of the items are used to establish the main characteristics for the clusters. Specifically this exploration is aimed at studying to what degree there is evidence for a Nordic cluster in the PISA data.

The article sets out to answer several interrelated questions:

1) What groups or clusters of countries are indicated by the cognitive items in the science domain of PISA 2003?
2) To what degree does the cognitive data in the science domain of PISA 2003 suggest that there is a common Nordic profile?
3) To what degree can some very broad item descriptors be used to describe unique aspects of the profiles across the cognitive items for the established clusters of countries?

Given that scientific literacy was a minor domain in PISA 2003, this article can not reach any solid conclusions regarding these questions. However, the analysis will point forward to what is feasible when data have been collected in 2006, this time with science as the major domain of the PISA assessment.

The data analysed are so called item-by-country interaction. These data are measures of how much the achievement for a country on an item deviates from what could be expected given the overall achievement of the country and the overall difficulty of the item (more will be said about this later). In projects like TIMSS and PISA efforts are made to minimize such interactions (Adams & Wu, 2002). In the discussion of the results presented this aspect will be brought up again and some recommendations and consequences for international large scale assessment of student achievement will be discussed.

Scientific literacy in PISA

PISA is a study organised through the Organisation for Economic Co-operation and Development (OECD). The study is repeated every three years, and different domains are emphasised in the test material each time. The study has so far been implemented twice, in 2000 and in 2003. Reading was the main emphasis in 2000 and mathematics in 2003. Scientific literacy was a minor domain in both studies. In 2006, science will be the major domain.

In 2003 some 270 000 students from 41 countries participated. The main results from the study have been reported in the international report (OECD-PISA, 2004) and in national reports (eg. Björnsson et al., 2004; Kjærnsli et al., 2004; Kupari et al., 2004; Mejding, 2004; Skolverket, 2004).

The framework for the study gives comprehensive descriptions of the domains, including scientific literacy (OECD-PISA, 2003, p. 133):

> *Scientific literacy is the capacity to use scientific knowledge, to identify questions and to draw evidence-based conclusions in order to understand and help make decisions about the natural world and the changes made to it through human activity.*

This definition is further developed and operationalised through the three main dimensions that the items should cover:

A. The *content* dimension that identifies several areas within science seen as particularly relevant given the overall definition
B. The *competency* dimension that identifies three scientific competencies:
   I. Describing, explaining and predicting scientific phenomena
   II. Understanding scientific investigation
   III. Interpreting scientific evidence and conclusions
   The first of these competencies involves *understanding scientific concepts*, while the second and third can be relabelled as *understanding scientific processes*. The item share across these three competencies is 50 % in competency I and 50 % in competencies II and III.
C. The *situation* dimension identifies three contexts or major areas of application; 'Life and Health', 'The Earth and the Environment', and 'Science in Technology'.

All items are categorised within these three dimensions. When describing the unity and diversity of clusters of countries, B and C will be used to characterise the items. The reason for not using the content dimension is twofold; this dimension has not been equally important in the item development, and it has too many categories to be useful in the analysis.

Method

A full account of the method used to analyse the data is given in Olsen (In progress). A short and compact overview is given in the following.

The residual matrix

The data input for the analyses presented in this article is a matrix with the percentages of students receiving score (p-value) on each science item in the PISA 2003 cognitive test for each of the participating countries. The number of items is 34 and the number of countries included in the analysis is 41. The p-values across items for high performing countries will in general be relatively high as compared to those for low performing countries. Similarly; the p-values for hard items will in general be low across countries as compared to easier items. These overall patterns can be regarded as not very interesting when we seek to find country specific patterns across items. The p-value matrix is therefore transformed to cancel out these general effects by subtracting the item specific part and the country specific part from each item respectively. These *residuals* represent the achievement for a country on a specific item, beyond what can be expected from the item and country averages alone.

Cluster analysis

Cluster analysis is a generic term for methods aiming to cluster individual cases (or variables) into larger groups which at the same time are a) similar to cases within the group and/or b) dissimilar to cases outside the group. These properties of a cluster will in the following be referred to as *internal cohesion* and *external isolation* respectively.

The results presented in this article are based on an *agglomerative hierarchical cluster analysis* using *average linkage* to cluster cases (countries), and with correlations between the cases as the *measure of proximity*. 'Agglomerative hierarchical' refers to that the overall aim of this type of cluster analysis is to group the cases (or the variables) in successive steps. In other words starting with *n* cases, the cases are merged in *n - 1* steps. As a result the data originally organised into *n* clusters (countries) ends up in one overall cluster containing all the individual countries. 'Average linkage' refers to that during the sequence when already formed groups of cases are clustered with new cases or merged with another group, this decision is based on average values of the 'proximities' or similarities for the groups. The specific method used is based on pragmatic reasons. In general there is no clear advice in the literature on how to choose an appropriate measure of proximity or clustering method. Everitt et al. (2001, p. 62) have reported that the *average linkage method* is relatively robust and that it takes account of cluster structure. While proximities in the final solution are measured by correlations and the clustering method used is average linkage, other proximity measures and clustering procedures have been used to study the *stability* of the final solution presented.

The result of a cluster analysis is commonly presented by a *dendrogram* (as seen in Figure 1). Dendrograms represent the hierarchical structure in the data. They illustrate when and how, in the stepwise procedure from *n* single cases to one single metacluster, cases merge to form the clusters. On the left all cases are separated and then proceeds with lines showing the clustering. The points where lines meet, that is where cases or groups of cases are merged, are referred to as *nodes*. The important decision to be made is when to stop, while reading the diagram from the left to the right. If there is an interesting clustering pattern in the data this will obviously lie somewhere in between the two extremes. In an SPSS application, which is used in the analyses presented below, the cases are sorted from top to bottom so that the countries belonging to the same cluster are placed near each other, and also they are placed so that the lines in the diagram do not cross, and thus the readability of the diagram is enhanced. Also, the dendrograms are shown with a standardized metric for the proximity measures in a range from 0 to 25. In this metric the ratios of the original proximity measures are preserved, whether they were Euclidian distances or a measure of similarity such as correlations (Norusis, 1988). Thus, this metric facilitates comparisons between solutions using different proximity measures.

Stability and validity

There is no clear advice in the literature for how to document the validity of a cluster analysis. A point of departure for establishing a validation procedure is that no valid interpretations can be done if the solution is not robust or stable. In the literature *stability* is not uniquely defined and this term refers variously to a property of the data themselves, the properties of a technique or to a property of the proposed solution. The last meaning of the word applies here. Gifi (1990) has given a comprehensive overview of different types of stability considerations in multivariate analysis, and especially relevant here is the type of stability labeled as *statistical stability under selection of technique*:

> *If we apply a number of techniques that roughly tries to answer the same question to the same data, then the result should give us roughly the same information. As the use of 'roughly' indicates, this form of stability is somewhat complicated to study. However, if nine out of ten techniques point to the same important characteristic of a data set, then the tenth technique is disqualified if it does not show this characteristic (p. 38).*

A solution should not simply be an artifact of the method used. Besides, it should not be very sensitive to specific data points, or in other words, the solution should show *stability under data selection* (Gifi, 1990). A robust solution is therefore obtained if applying different methods give similar solutions or if removing parts of the dataset do not alter the solution substantially. Accordingly, the data were analysed using other clustering methods and other measures of proximities, and the analyses were replicated omitting selected parts of the data to see whether the solution was affected by specific data points.

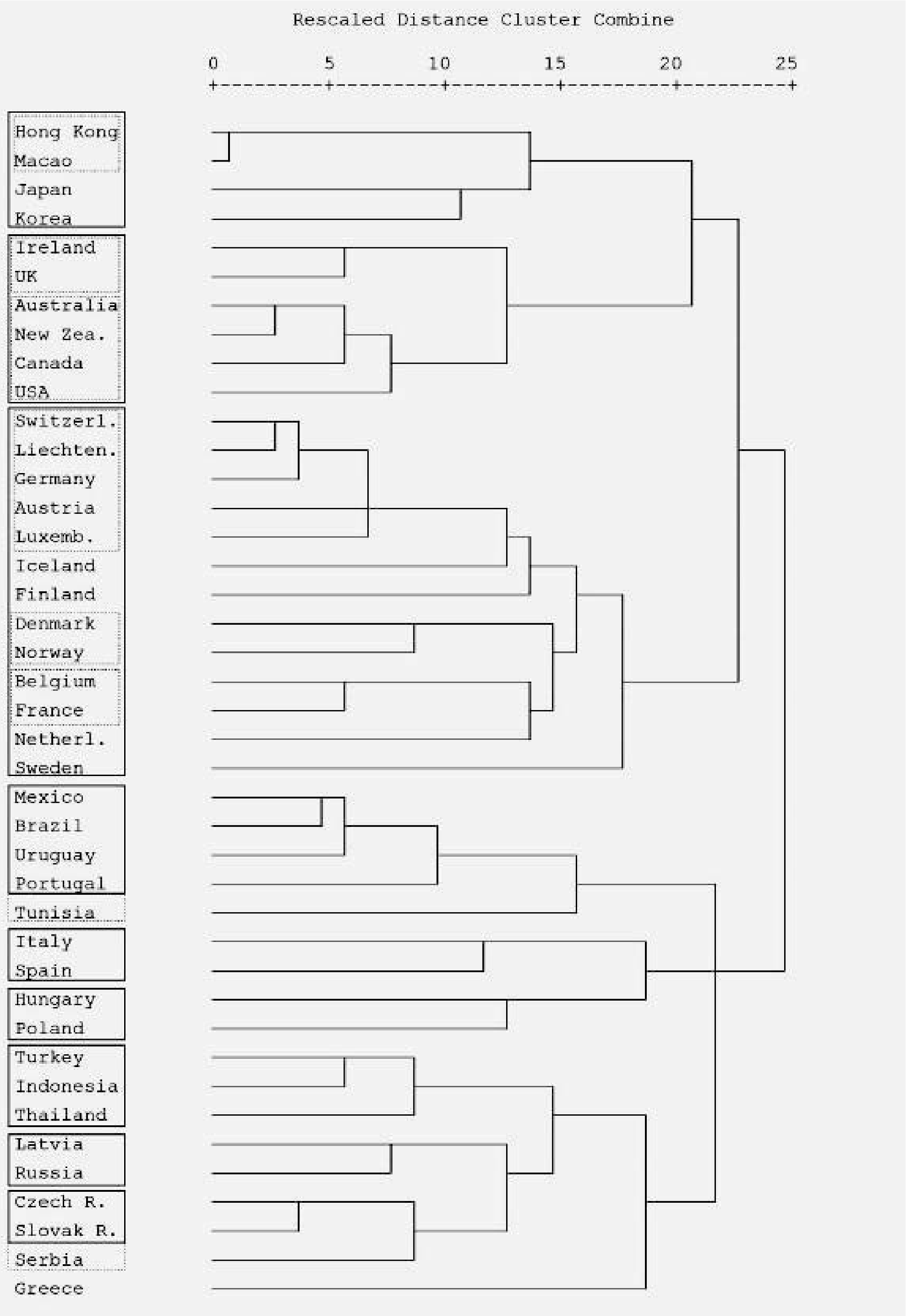

Figure 1: The dendrogram for country clustering. The groups with high degree of external isolation are framed. Subgroups or possible extensions are shown with dotted lines

Results

Identifying the main clusters

Figure 1 shows the dendrogram representing the solution of the cluster analysis. In this figure some groupings of countries have been marked by frames. These groups are externally isolated from the rest of the countries. This is seen by the relatively large distance from the node where they are merged to the next node up in the hierarchy. These groups are therefore the initial candidates for being perceived as clusters. However, some of these groups are very small (e.g. Italy and Spain). Some other possible groups in the hierarchy are difficult to label, which indicates that they are difficult to interpret as meaningful units (e.g. Turkey, Indonesia and Thailand). Four groups of countries remain as distinct and meaningful clusters and these can be labelled as:

1) 'East Asian countries' (short label 'EastAsia')
2) 'English speaking countries' (short label 'English')
3) 'North-West European countries' (short label 'NorthEur')
4) 'South American countries + Portugal' (short label 'SouthAm')

The cluster of North-West European countries

Figure 1 gives little support for the hypothesis of a distinct Nordic cluster. Instead the Nordic countries are merged into the largest groups of countries. This larger cluster is to some degree a linguistic cluster. In all these countries, except for France, Finland and parts of Belgium, Luxembourg and Switzerland, the majority language is a Germanic language. It is also a geographical cluster (countries in the North-West corner of Europe). In such a large cluster it can not be expected that all pairs of countries are similar. Sweden is for instance included in the group at a very late stage. However, all the countries share similarities with the average profile for the countries within the group. The average correlation coefficient between the countries within the cluster is 0.32 and the coefficient alpha is relatively high (see Table 1), both taken as indicators of moderate to high internal coherence. In addition the cluster is externally well isolated from the rest of the countries. It could therefore be accepted as a cluster despite the fact that there are some negative correlations in this group.

Within this group there is one distinct subgroup that could also be perceived as a cluster by itself, the 'German speaking countries' (short label 'German') with a high degree of internal cohesion. Many of the countries' residuals are strongly correlated with one or more countries in this subgroup of German speaking countries, or in other words, the subgroup 'German' is not totally isolated from the other countries in cluster 3. In the larger cluster it seems as though this subgroup is a "centre of gravitation" attracting the other countries.

As is evident, the criteria for what counts as a cluster is not definite. Although the Nordic countries did not stand out as a cluster in this analysis, it is still considered worthwhile looking into the internal clustering mechanisms between the Nordic countries (short label 'Nordic'), based on the average correlations and alpha coefficients in Table 1 and with the references given to prior studies documenting a Nordic unity across cognitive items. Several Nordic countries have moderately positively correlated residuals. Denmark stands out as the country with the most prominent overall Nordic profile with correlations with the other Nordic countries in the range 0.2-0.5. Sweden is at the other extreme with weak overall correlations with the Nordic neighbours. The latter is quite surprising given the fact that Sweden has been very much centrally placed in the Nordic cluster in similar analysis of other datasets. This is for instance not consistent with the findings in similar analysis of the science items in the PISA 2000 data (Kjærnsli & Lie, 2004).

Throughout the article these two clusters will be included. The German speaking group of countries is natural to include given that this cluster satisfies all criteria given above for what constitutes a cluster. The reason for also including the Nordic countries as a cluster is mainly that the unity or diversity among these countries is the object of study for this article. Dividing the large North-West European cluster into two smaller groups of countries is also based on that in some instances it is warranted to have about equal group sizes when comparing properties of different clusters (for instance in evaluating and comparing averages or the coefficient alphas).

Table 1: The coefficient alphas and the average correlations within the clusters.

Cluster	Coefficient alpha	Average correlation within cluster
EastAsia	0.77	0.45
English	0.85	0.50
NorthEur	0.86	0.32
German	0.89	0.65
Nordic	0.59	0.24
SouthAm	0.83	0.57
EastEur	0.81	0.45

The other clusters

The three other clusters will not be discussed in detail. They are all well established and fulfil the requirements stated above for what defines a cluster. For the cluster of South-East Asian countries it should be noted that Hong Kong and Macao are very close to each other. The correlations between the residuals for these countries is almost 0.9! This is the strongest relationship between any pairs of participating school systems in PISA, and this is most probably related to the fact that they are school systems within two regions of the same country, China. However, all the correlations between the countries in this group are positive and fairly high. The English speaking countries are also split into two subgroups but all countries (except USA and Ireland) have residuals that are moderately or highly correlated with each other. The fourth group is a bit more problematic to label. The countries in this group all have Latin languages, but at least two other countries with similar languages (Italy and Spain) do not belong to the cluster. The label South-America + Portugal is therefore a better suggestion, indicating also that it might be more meaningful to reduce this cluster to only the three Latin-American countries.

In the cluster analyses of data from TIMSS 1995 (Angell et al., in press; Grønmo et al., 2004b; Zabulionis, 2001) a distinct Eastern-European cluster of countries was present. The dendrogram (and also the total correlation matrix not given here) suggests that Hungary and Poland differ most markedly from their partners in what Zabulionis (2001) labels the 'post-communist' group of countries. Still, the group of five countries (Czech Republic, Slovak Republic, Russia, Latvia, and Serbia and Montenegro) in the lower end of the dendrogram is much more coherent than for instance the Nordic group (see average correlations and alphas given in Table 1). This group will therefore be included and treated as a cluster (short label 'EastEur'). Further arguments for including this cluster are given when discussing the stability of the analysis.

Stability and validity

To verify the stability of the solution the data have been analysed combining other proximity measures (Block distance or Manhattan distance and ordinary and squared Euclidian distance) and other clustering methods (single linkage and Wards method) in order to reveal if the clusters mapped in Figure 1 could possibly be artefacts of the specific method used. Also, the analysis has been done excluding some countries one at a time. Furthermore, in order to study possible floor- and ceiling effects, a matrix of logistically transformed residuals has been analysed (see Olsen, In progress for details). In short it can be concluded from these analyses that the solution presented is robust, and it is unlikely that the clusters reported are artefacts of the specific method chosen. In addition, varying the measures and the methods has identified the possibility of an East European cluster and some weak indications of a Nordic profile. As a last element of validation, this cluster structure has been compared to other studies using a similar method (Angell et al., in press; Grønmo et al., 2004b; Kjærnsli & Lie, 2004; Lie & Roe, 2003; Zabulionis, 2001). In short, the correspondence between these analyses is quite remarkable. Many of the same clusters reappear from study to study, a phenomenon that will be brought up when discussing possible implications. In result, there are six approximately equally sized clusters that will be referred to throughout the article.

Explaining the clusters by looking at the items

It is not evident what is required for, and what counts as, an explanation for these findings. While for instance Grønmo et al. (2004b) and Zabulionis (2001) in their discussion gave a focus on wider cultural factors, or in other words, explanations at a deep and fundamental level of understanding beyond the items themselves, this article is delimited to explanations based directly on the analyses of the items that actually produced the cluster structure. This is not to dismiss the fact that the labels chosen for the groups refers to factors beyond the items themselves, such as geographical or linguistic commonalities between the countries in a group. Rather, it is to emphasise that the clusters are based on *profiles across highly specific items*. These clusters represent groups of countries with similar item-by-country interactions. An equivalent statement is that countries within a group perform better or worse than what could be expected on many of the same items. In order to gain further insight into these patterns it is worthwhile looking more closely at the patterns across items in order to identify differential weaknesses and strengths related directly to the items. In the analysis below the Nordic perspective will be given priority.

In order to identify the items with explanatory power, the single item data can be explored one item at a time. We could for instance proceed by identifying items favoured by specific clusters or items separating clusters effectively. This has been attempted with the science items in PISA 2003, but this approach was eventually abandoned since the number of items characterising the Nordic countries were too few to make meaningful suggestions for generalisations. This line of analysis will therefore be postponed until data from the 2006 study is available. The number of science items will be about three times as high in 2006 and the potential for such analyses will be much better.

Instead of using single items by themselves the relationship between the item residuals and descriptors for some more overall characteristics shared by many items has been analysed. With only 34 items available for analysis it is important to use descriptors of a general character not splitting the items into more than two groups, such as competency, context, format, relationship to text and item difficulty. In this analysis several ways of classifying the items will be used:

1. *Competencies*: Item analyses from PISA clearly demonstrates that countries perform differently on items testing mainly factual knowledge or understanding of concepts (competency I) and items testing the understanding of and skills in some fundamental scientific processes (competencies II and III) (Kjærnsli et al., 2004; Lie et al., 2001). The variable 'Comp" in Table 2 is coded 1 for items in Competency I and 2 for items in competency II or III.
2. *Context*: Also, countries have different emphases in science curricula (Cogan et al., 2001; Martin et al., 2004) which means that items from different *areas of application* might work differently in countries. The PISA framework operates with three *situations* or *contexts* describing the areas of application. These contexts have after an initial screening of the data been recoded into two distinct *areas of application*; 'Life and Health' (coded 1) and 'Physical World' (coded 2). The latter is a combination of the two original situations labelled 'Earth and Environment' and 'Science in Technology'.

3. *Format*: Previous research is ambiguous regarding the differential effects of item format across countries. This will be explored through the dichotomy given by constructed response items (coded 1) vs. selected response items (coded 2).
4. *'Textdist' (text distance)*: It is evident from the science items in PISA that they are very much related to textual stimulus, and the items have therefore been dichotomously classified according to their *closeness to the text*. To some degree items differs in the way that they are dependent on the textual material. Some items could more or less be answered by skilful reading (coded 2), while others require to a much larger degree that external information is brought into the solution (coded 1).
5. *p-value*: In addition the difficulty of the item, indicated by the percentage of credited responses (p-value) will be used. Analysis of the Norwegian data revealed for instance that students performed relatively better on easier items in mathematics (Kjærnsli et al., 2004). This item descriptor differs from the other descriptors (1-4 above) in that it is represented by a continuous variable.

Table 2: Correlation between some item characteristics and the average item residuals in clusters of countries. Statistically significant correlations are boldfaced.

	Item characteristics				
	Comp	Context	Format	Textdist	Pvalue
EastAsia	-0,27	0,18	**-0,42**	-0,15	-0,21
English	**0,44**	-0,23	-0,11	0,16	-0,05
NorthEur	0,19	-0,14	-0,04	**0,36**	0,08
German	-0,06	-0,15	-0,08	0,19	0,03
Nordic	0,23	-0,04	0,07	**0,45**	0,08
SouthAm	0,12	-0,13	0,22	0,14	0,13
EastEur	-0,34	0,29	0,31	**-0,34**	-0,01

What stands out in Table 2 from a Nordic perspective is that the textual aspect of the PISA items is very important. This means that Nordic countries perform relatively better on items where careful analysis of the text in the stimulus material is vital to reach a solution than on items were the solution is more independent of the textual material itself. Furthermore, a relatively strong relationship is found with the 'Competency' variable, which means that the Nordic countries perform relatively better on items testing understanding and mastery of scientific processes (competency II and III). Response format is not very important, but to the extent that Nordic students perform relatively better on a format the positive sign for this correlation tell us that the Nordic countries on average have small positive residuals for selected response items. In other words; there does not seem to be a bias disfavouring the Nordic countries on tests that include multiple choice items, despite the fact that this format is not very common in the Nordic countries. The same has been documented for the mathematics items in PISA 2003 (Kjærnsli et al., 2004). There is also a similar weak tendency for the Nordic countries to perform better on easier items. Finally, the contexts do not seem to play any significant role in explaining the Nordic profile.

A compact characterisation of the other clusters is that the English students are favoured by items testing their understanding and mastery of scientific process skills, and they perform relatively better on items set in a context related to life and health; the East Asian students perform relatively well on difficult items testing conceptual understanding where they have to formulate answers themselves, preferably in contexts relating to the physical world; East European countries are favoured by multiple choice items testing conceptual understanding related to the physical world, and where interpretation of the text is not crucial in order to reach a qualified solution; while the German and the South American profile is relatively even across these item characteristics. Many of these characteristics are consistent with what was found in TIMSS 1995 (Lie et al., 1997) and TIMSS 2003 (Grønmo et al., 2004a).

Conclusions and implications

Is there a Nordic profile of science achievement?

The results presented are not conclusive regarding the Nordic aspect of the research questions. It is evident from inspecting the item residuals themselves that there are many items with similarities between the Nordic countries. For many items the residuals in the Nordic countries are very close to each other. In an attempt to "explain" the Nordic profiles, it was found that it was particularly related to an index of how closely the items were linked to the textual material in the stimulus. Nordic students did particularly well on items were the correct response was highly dependent on reading and interpreting the textual material. The other item characteristics were not strongly linked to the Nordic profile of residuals. Furthermore, the average correlation between the Nordic countries' residuals was moderate. Sweden was particularly weakly linked to the other Nordic countries.

On the other hand, it is not evident that the profiles across items for the Nordic countries have more in common than they have with other North-West European countries (Figure 1). The countries within this larger cluster are similar in many respects; they have predominantly Germanic languages, they are geographical neighbours, they are wealthy countries belonging to the same cultural sphere. It is interesting to note that, within this cluster, the countries with predominantly German speaking students have moderate to high correlations with all the other countries in the cluster. In other words, the German subcluster of countries functions as the "glue" holding this larger cluster together. This effect cannot be fully understood as a linguistic effect since several of the countries within this larger group have non-Germanic languages. It is tempting to suggest that this empirical finding might somehow be an effect of the historical ties between these countries, where especially Germany has been a dominant country, not only politically and economically, but also within educational theory. The extent to which this has had an effect on educational policy and curriculum is not easy to specify. It is therefore not easy to relate such wider cultural factors to the concrete cluster analysis presented in this article.

Consistency across studies

In all studies reported so far using a version of the same method to explore clustering across cognitive items the English, the East Asian and the German clusters are always more or less clearly present (Angell et al., in press; Grønmo et al., 2004b; Kjærnsli & Lie, 2004; Lie & Roe, 2003; Zabulionis, 2001), independent of the domain tested and the specific nature of the study, and these clusters are stable over time and invariant of the specific method used in the analysis. Furthermore, the larger metacluster of North-West European countries has been present in all the studies analysing science items. In addition an East European cluster has been clearly present, especially in studies of the TIMSS 1995 data which included a large number of countries from East Europe. In addition, a Nordic cluster has been more clearly present in the analyses of TIMSS 1995 items. All in all, the consistency across the reported analyses gives further reassurance to the conclusion that the clusters of countries presented above are indeed countries or school systems with common characteristics.

It is reasonable to suggest that a further investigation of this phenomenon is warranted. Central to such an investigation would be theoretical contributions with reviews of possible antecedents for such clusters. Furthermore, one should find ways to include items from the questionnaire describing the school systems as explanatory variables for the profiles. Also, a more distinct science educational perspective should be possible to develop when more items are included. This would make it possible to use more refined item characteristics, and it would be possible to identify relatively large pools of items characterising each cluster. Moreover, it has been argued elsewhere that TIMSS and PISA are studies which can be seen as complementary to each other (Olsen, 2004), and done with care the pool of items characterising the clusters of countries could be supplemented by similar analysis of TIMSS 2003 data.

Residuals and fair tests

Tests such as those in PISA are developed to measure a well defined cognitive trait. In order to do this with some level of precision it is necessary to have many items in a test. When developing the PISA test considerable efforts are made to produce items with minimal item-by-country interactions. Items with large interactions are consciously tossed out after the field trial. A test with no such interactions would in many ways be considered the perfect test since such interaction could threaten the aim of the test; to compare countries by measuring the same trait in all countries. If there are large item-by-country interactions, this could imply that some of the items measure different concepts in different countries, and as such the item is not very effective since the item then would provide little information to the overall measure. Seen from a didactical or subject centred perspective these procedures mean that highly interesting information is consciously *not* collected. Fortunately, as is evident in the results presented, there is so much variation in countries' profiles across items that distinct clusters of countries still could be detected.

Wolfe (1999) studied profiles of residuals across content categories in mathematics in the Second International Mathematics Survey (SIMS). He concluded that when the profiles of achievement are too discrepant, the overall comparison is either "fundamentally unfair or essentially random" (Wolfe, 1999, p. 225). He continued that as a consequence of this regional designs are required to enhance the validity of international studies so that countries more similar to each other are compared. His conclusion is not totally relevant for PISA. This study does not, as SIMS and the sequels TIMSS 1995 and 2003, intend to be a 'fair test'. PISA intends to measure cognitive traits that the international community of policy makers and researchers to some extent agree on are central for being "prepared for life". His argument related to the error component is on the other hand just as important for PISA as in any other international comparative assessment. However, the residuals that Wolfe reports from SIMS are on average higher than the residuals in PISA. It is likely that this is due to the increased focus on quality found in later international assessment studies (Porter & Gamoran, 2002), including a thorough screening of the item-by-country interactions in the field trials (Adams & Wu, 2002). Still, if the residuals had been computed once more, but this time in a matrix consisting only of countries with similar profiles, they would have been reduced. As such the information that each item provides is higher for a scale produced across countries with comparable profiles.

This is an argument for giving priority to Nordic comparisons, given that the profiles are comparable across the Nordic countries. This Nordic perspective has been central in the Norwegian reports from TIMSS 1995 (Lie et al., 1997) and PISA (Kjærnsli et al., 2004; Lie et al., 2001) and in a supplementary Nordic PISA report (Lie et al., 2003).

From a psychometrical perspective the residuals used for analysis in this article are regarded as "errors" or random fluctuations around the true score. Since these residuals are systematically linked to characteristics of the items other than the trait being measured, they are clearly not random fluctuations. If PISA is perceived to be a 'fair test', different weighting of items with special item characteristics could be regarded as a bias. In general, the distribution of items across different characteristics is always to some extent arbitrary. This implies that when interpreting the results of an international test, particularly when discussing the results as seen from a specific national context, the operationalisation of the trait being tested must be evaluated according to a national frame of reference. If for instance a science test is loaded with items in mechanics one has to evaluate whether this is a representative test for a country, given the national priorities in the curriculum.

Scientific literacy and reading

The main characteristic of the Nordic profile is that students in our region tend to do well on items involving careful reading, or as an alternative and more pessimistic interpretation, the Nordic students achieves poorly on items not directly dependent on reading of the text. In the analysis done here these two possible interpretations cannot be distinguished. Moreover, this textual characteristic of the items was in general the item characteristic that most successfully could account for differences in the clusters achievement profiles.

Norris and Phillips (2003) have described the relationship between scientific literacy in a *fundamental sense* as being able to read/write science texts and a *derived sense* as being knowledgeable and competent in science. The effect of the textual descriptor implies that the link between aspects of scientific literacy in its fundamental sense is indeed a vital part of scientific literacy as operationalised in PISA. Fang (2005) has from a systemic functional linguistic perspective (e.g. Halliday & Martin, 1993) and by providing examples of material from textbooks in school science demonstrated that these two types of scientific literacy are not only interrelated, but also inseparable.

In the framework for PISA (OECD-PISA, 1999, 2003) this and related linguistic perspectives are not explicitly linked to the overall trait of scientific literacy. In these documents it is clearly stated that scientific literacy as measured by PISA should be set in contexts with some degree of authenticity. This has introduced what is a 'fingerprint' for many PISA items; they are organised in groups of items relating to the same stimulus material (examples are provided in OECD-PISA, 2002 and many more are found in Norwegian at www.pisa.no). For many of these units the stimulus material is an extended piece of text, and the texts have the same characteristics as those analysed by Fang (2005). Many of these texts have a high informational density, processes and phenomena observed in nature or laboratory are abstracted by use of nouns (nominalisation), and they include specialised technical language.

The amount of research and theoretical discussions related to this linguistic characteristic of talking, writing and reading scientific texts is significant (e.g. Bisanz & Bisanz, 2004; Fang, 2005; Lemke, 1990; Norris & Phillips, 2003; Roth & Lawless, 2002; Wallace et al., 2004; Wellington & Osborne, 2001) and in science education research this is now a mature subject of study. The rough indicator for how the text is related to the solution could effectively account for differences in profiles across countries, and given the available theoretical discussions on how learning science in many respects is learning to talk, write and read science, and that being scientific literate in many ways is to know and understand the language of science, this aspect deserves closer attention in the future frameworks of PISA. Furthermore, this link between this emerging field of science education and PISA's operational definition of scientific literacy deserves closer inspection and discussion. One way to proceed would be to analyse some of the stimulus material more closely, for instance using the framework of systemic functional linguistics. The arguments for treating the connection between literacy in a wider sense and scientific literacy in more detail is further strengthened by the fact that PISA also includes reading literacy as well as mathematical literacy as test domains. Applying a common linguistic approach on items across these domains could give valuable insights into how these domains relate.

References

Adams, R., & Wu, M. (Eds.). (2002). *PISA 2000 technical report*. Paris: OECD Publications.

Angell, C., Kjærnsli, M., & Lie, S. (in press). Curricular and cultural effects in patterns of students' responses to TIMSS science items. In S. J. Howie & T. Plomp (Eds.), *Contexts of learning mathematics and science: Lessons learned from TIMSS*. Lisse: Swets & Zeitlinger Publishers.

Bisanz, G. L., & Bisanz, J. (2004). *Research on everyday reading in science: Emerging evidence and curricular reform.* Paper presented at the National Association for Research in Science Teaching (NARST), 1.-3. April Vancouver.

Björnsson, J. K., Halldórsson, A. M., & Ólafsson, R. F. (2004). *Stærðfræði við lok grunnskóla. Stutt samantekt helstu niðurstaðna úr PISA 2003 rannsókninni*. Reykjavik: Námsmatsstofnun.

Cogan, L. S., Hsingchi, A. W., & Schmidt, W. H. (2001). Culturally specific patterns in the conceptualization of the school science curriculum: Insights from TIMSS. *Studies in Science Education, 36*, 105-134.

Everitt, B. S., Landau, S., & Leese, M. (2001). *Cluster analysis* (4 ed.). London: Arnold.

Fang, Z. (2005). Scientific literacy: A systemic functional linguistic perspective. *Science Education, 89*(2), 335-347.

Gifi, A. (1990). *Nonlinear multivariate data analysis*. New York: John Wiley & Sons.

Grønmo, L. S., Bergem, O. K., Kjærnsli, M., Lie, S., & Turmo, A. (2004a). *Hva i all verden har skjedd i realfagene? Norske elevers prestasjoner i matematikk og naturfag i TIMSS 2003*. Oslo: Institutt for lærerutdanning og skoleutvikling, Universitetet i Oslo.

Grønmo, L. S., Kjærnsli, M., & Lie, S. (2004b). Looking for cultural and geographical factors in patterns of response to TIMSS items. In C. Papanastasiou (Ed.), *Proceedings of the IRC-2004 TIMSS* (Vol. 1, pp. 99-112). Lefkosia: Cyprus University Press.

Halliday, M. A. K., & Martin, J. R. (1993). *Writing science: Literacy and discursive power*. Pittsburgh: University of Pittsburg Press.

Kjærnsli, M., & Lie, S. (2004). PISA and scientific literacy: Similarities and differences between the Nordic countries. *Scandinavian Journal of Educational Research, 48*(3), 271-286.

Kjærnsli, M., Lie, S., Olsen, R. V., Roe, A., & Turmo, A. (2004). *Rett spor eller ville veier? Norske elevers prestasjoner i matematikk, naturfag og lesing i PISA 2003*. Oslo: Universitetsforlaget.

Kupari, P., Välijärvi, J., Linnakylä, P., Reinikainen, P., Brunell, V., Leino, K., Sulkunen, S., Törneroos, J., Malin, A. & Puhakka, E. (2004). *Nuoret osaajat: PISA 2003 - tutkimuksen ensituloksia*. Jyväskylän yliopisto: Koulutuksen tutkimuslaitos.

Lemke, J. (1990). *Talking science: Language, learning and values*. Norwood: Ablex.

Lie, S., Kjærnsli, M., & Brekke, G. (1997). *Hva i all verden skjer i realfagene? Internasjonalt lys på trettenåringers kunnskaper, holdninger og undervisning i norsk skole*: Institutt for lærerutdanning og skoleutvikling, UiO.

Lie, S., Kjærnsli, M., Roe, A., & Turmo, A. (2001). *Godt rustet for framtida? Norske 15-åringers kompetanse i lesing og realfag i et internasjonalt perspektiv.* Oslo: Institutt for lærerutdanning og skoleutvikling. Universitetet i Oslo.

Lie, S., Linnakylä, P., & Roe, A. (Eds.) (2003). *Northern lights on PISA: Unity and diversity in the Nordic countries in PISA 2000*: Department of Teacher Education and School Development, University of Oslo.

Lie, S., & Roe, A. (2003). Unity and diversity of reading literacy profiles. In S. Lie, P. Linnakylä & A. Roe (Eds.), *Northern lights on PISA* (pp. 147-157) Oslo: Department of Teacher Education and School Development, University of Oslo.

Martin, M. O., Mullis, I. V. S., Gonzales, 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*. Boston: TIMSS & PIRLS International Study Center, Lynchs School of Education, Boston College.

Mejding, J. (Ed.) (2004). *PISA 2003 - danske unge i international sammenligning*. København: Danmarks Pædagogiske Universitet.

Norris, S. P., & Phillips, L. M. (2003). How literacy in its fundamental sense is central to scientific literacy. *Science Education, 87*(2), 224-240.

Norusis, M. J. (1988). *SPSS/PC+ advanced statistics v2.0*. Chicago: SPSS Inc.

OECD-PISA (1999). *Measuring student knowledge and skills*. Paris: OECD Publications.

OECD-PISA (2002). *Sample tasks from the PISA 2000 assessment: Reading, mathematical and scientific literacy*. Paris: OECD Publications.

OECD-PISA (2003). *The PISA 2003 assessment framework: Mathematics, reading, science and problem solving knowledge and skills*. Paris: OECD Publications.

OECD-PISA (2004). *Learning for tomorrow's world. First results from PISA 2003*. Paris: OECD Publications.

Olsen, R. V. (2004). *The OECD PISA assessment of scientific literacy: How can it contribute to science education research?* Paper presented at the National Association for Research in Science Teaching (NARST), 1.-3. April, Vancouver.

Olsen, R. V. (In progress). *Phd-thesis. To be submitted 2005.* Oslo: University of Oslo.

Porter, A. C., & Gamoran, A. (Eds.) (2002). *Methodological advances in cross-national surveys of educational achievement*. Washington, DC: National Academy Press.

Roth, W.-M., & Lawless, D. (2002). Science, culture and the emergence of language. *Science Education, 86*(3), 368-385.

Skolverket. (2004). *PISA 2003 - svenska femtonåringars kunskaper och attityder i ett internationellt perspektiv. Rapport 254*. Stockholm: Skolverket.

Wallace, C. S., Yore, L. D., & Prain, V. (2004). *The fundamental sense of science literacy: Implications for non-English speaking and culturally diverse people.* Paper presented at the National Association for Research in Science Teaching (NARST), 1.-3. April, Vancouver, Canada.

Wellington, J., & Osborne, J. (Eds.) (2001). *Language and literacy in science education*. Philadelphia: Open University Press.

Wolfe, R. G. (1999). Measurement obstacles to international comparisons and the need for regional design and analysis in mathematics surveys. In G. Kaiser, E. Luna & I. Huntley (Eds.), *International comparisons in mathematics education*. London: Falmer Press.

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