

ACHIEVEMENT TESTS FROM AN ITEM PERSPECTIVE

An exploration of single item data from the PISA and TIMSS studies, and how such data can inform us about students' knowledge and thinking in science

Rolf Vegar Olsen

Thesis submitted in partial fulfilment for the degree
Doctor Scientiarum

Department of Teacher Education and School Development
Faculty of Education
University of Oslo

Acknowledgements

The present work has been financed by the Norwegian PISA project, which in turn receives funds from the Ministry of Education and Research. Without this financial support it would not have been possible to conduct the work presented in this dissertation. I would therefore like to extend my sincere thanks to the Ministry for their support. I would also like to thank the Department of Teacher Education and School Development at the University of Oslo for support over the last years.

In the documents regulating the doctoral programmes at the University of Oslo it is mentioned in several places that a dissertation should present the *independent* research contribution of the candidate. In general, I do not sympathise with the view on research implicitly contained in such statements. It is therefore important to thank all of you who, in one way or the other, have participated in my *dependent* research contributions. It is not possible to give a full account of who all of you are, and it is not always easy to specify explicitly what the nature of your contribution has been. Nevertheless, I would like to identify and thank those of you who probably deserve it the most. Needless to say, of course, that all the errors, flaws, omissions and weak arguments that are still present, are entirely my own responsibility.

First of all I would like to acknowledge the role of my colleagues in the Norwegian PISA and TIMSS projects, who I have had the pleasure of working closely with over the last five to six years. I would particularly like to thank Svein Lie – not only because he is my supervisor – but also because he is such a fine role model and mentor in my professional life. I would also like to give special thanks to Are Turmo with whom I have co-authored a number of papers over these years, and who I have had the opportunity to share my thoughts and uncertainties with – besides sharing an office during all of this period. I would also like to give special thanks to Marit Kjærnsli who, as the project manager of PISA, has had a key role in my professional life and development. Hopefully, I can look forward to many more years of exciting and challenging work on these projects with all of you.

Special thanks also to Camilla Schreiner, Svein Sjøberg and Andreas Quale who read some of my material towards the end of the process, and who gave valuable criticism and specific contributions as to how this dissertation, or parts of it, could improve.

My thanks furthermore go to all other colleagues in the field of science education, with whom I have had numerous talks and discussion that I believe have had a major impact on my work and my ‘Bildungsreise’ into the science education community.

Finally, I have to thank my family – particularly my wife Karen, and my two kids Evert and Ingrid – for your patience and continued support, particularly in the short recurring periods when I did not believe that I could complete my work.

Oslo, June 2005



1 INTRODUCTION	3
1.1 BACKGROUND, RATIONALE AND AIMS.....	3
1.1.1 TIMSS, PISA and their test theory rationale.....	3
1.1.2 The item-specific information	4
1.1.3 The nominal information.....	6
1.1.4 Three levels of aggregation of the item data.....	7
1.1.5 The aim of this thesis.....	9
1.2 THE RESEARCH SEQUENCE	9
1.3 THE STRUCTURE OF THE THESIS.....	10
1.3.1 The chapters.....	10
1.3.2 The papers.....	11
1.4 REFERENCES	12
2 LINCAS AS A FRAME FOR RESEARCH.....	13
2.1 INTRODUCTION TO THE CHAPTER	13
2.2 THE GROWTH OF LINCAS: FROM FISS TO PISA.....	13
2.2.1 Purpose I: The research purpose.....	14
2.2.2 Purpose II: The effective policy purpose	16
2.2.3 From FISS to TIMSS.....	19
2.2.4 OECD and the need for new types of assessments.....	20
2.3 COMPARING PISA WITH TIMSS	22
2.3.1 PISA in a nutshell.....	23
2.3.2 TIMSS in a nutshell.....	23
2.3.3 Why compare the two?	24
2.3.4 Different concepts measured.....	25
2.3.5 Different international frame of reference.....	26
2.3.6 Different organisations behind the studies	28
2.3.7 Different sampling designs	30
2.3.8 Complementary or incompatible?.....	32
2.4 LINCAS, POLICY MAKERS AND RESEARCHERS IN SCIENCE EDUCATION	34
2.4.1 Three observations framing the discussion.....	34
2.4.2 A shift in how research is financed and organised.....	36
2.4.3 The role, function and position of the policy makers in PISA and TIMSS.....	37
2.4.4 LINCAS as a link between policy makers and researchers?	38
2.5 EXPLORING THE POSSIBILITIES FOR SECONDARY RESEARCH.....	40
2.5.1 Arguments for secondary analysis of LINCAS.....	42
2.5.2 Targeting research questions in science education	46
2.6 CLOSING REMARK	49
2.7 REFERENCES	49
3 METHODS AND METHODOLOGICAL REFLECTIONS	63
3.1 INTRODUCTION TO THE CHAPTER	63
3.2 A FUNDAMENTAL ISSUE REGARDING THE USE OF QUANTITATIVE METHODS: DATA OR MODEL?	64
3.2.1 Exploration and descriptive statistics versus confirmation and inferential statistics.....	64
3.2.2 Pure versus applied data analysis or statistics.....	67
3.3 MARKING RUBRICS AND CODES USED FOR THE ITEMS IN PISA AND TIMSS.....	70
3.3.1 The nature of making categories.....	70

3.3.2 The general nature of the codes and marking rubrics used for the items in PISA and TIMSS.....	73
3.3.3 The codes used in PISA and TIMSS – the double digits.....	79
3.3.4 The use of double digit codes in analyses: some examples.....	84
3.4 ANALYSIS OF NOMINAL VARIABLES.....	85
3.4.1 General introduction.....	86
3.4.2 Correspondence analysis.....	87
3.4.3 Homogeneity analysis.....	92
3.5 REFERENCES.....	97
4 SUMMARY AND DISCUSSION	102
4.1 SUMMARY OF CHAPTER 1.....	102
4.2 SUMMARY OF CHAPTER 2.....	103
4.3 SUMMARY OF CHAPTER 3.....	103
4.4 SUMMARY OF PAPER I.....	105
4.5 SUMMARY OF PAPER II	106
4.6 SUMMARY OF PAPER III	109
4.7 CONCLUDING SUMMARY.....	112
PAPER I	
PAPER II	
PAPER III	
APPENDIX 1	

1 Introduction

1.1 Background, rationale and aims

1.1.1 TIMSS, PISA and their test theory rationale

The work presented in this thesis is mainly related to the Programme for International Student Assessment (PISA), a large-scale international comparative achievement study in education initiated by the Organisation for Economical Co-operation and Development (OECD). The overall question asked by the OECD, that motivates PISA, is to what degree 15-year-olds are prepared to meet the challenges of the future. The study produces several measures supposed to function as indicators for this overall and rather open-ended question. In the broad spectrum of indicators developed in the study, we find measures of concepts that are cognitive, meta-cognitive or affective in nature. The major cognitive indicators are assessments of students' reading, mathematical and scientific literacy. Central to my thesis is the measure of scientific literacy, as defined in the framework and operationalised in the test material.

The Third International Mathematics and Science Study (TIMSS¹) is another large-scale international comparative achievement study in education that also includes a science component. My thesis and the rationale for it grew out of work related to analyses of single items in this study. Although TIMSS also includes measures of students' competency in science, it has a fundamentally different departure point than PISA. One of the aims of TIMSS was, and still is, to produce indicators of how successfully educational systems implement the intended science curriculum. A somewhat detailed comparison of PISA as compared to TIMSS will be returned to in chapter 2.

Assessments like these are developed within the framework of test theory. Test theory is a collection of methods (and a rationale for these) for developing and analysing measures of psychological entities, which are theoretical and literally hypothetical entities. Such entities are often referred to as traits or constructs. Scientific literacy, as measured in PISA, is one such trait, and the achieved science curriculum, as measured in TIMSS, is another. The tests developed are in principle norm-referenced tests. This means that the achievement scores reported from the studies are standardised: they are calculated by first finding the international average and standard deviation, and then each student is given a score expressed as a deviation from this mean². This

¹ Since the TIMSS 2003 study the abbreviation stands for Trends in International Mathematics and Science Study. Nevertheless, TIMSS 2003 should be seen as a continuation of the studies in 1995 and 1999, and the change of name should not be taken to indicate a completely new design. The main development reflected in this change of name is that the cyclic design has been given more central role in the design of the study.

² The alternative to norm-referencing is to develop scales that are computed relative to a more absolute criterion than the norm, and such tests are often referred to as criterion-referenced. To some extent the PISA assessment instruments are also criterion-referenced test; the tests have been developed from a very detailed framework describing the concepts that the tests should be developed to measure. This is to some degree a description of what characterises achievement at different levels. These descriptions can therefore be conceived of as criteria. Nevertheless, technically speaking the test scores refer to a norm.

enables us to interpret students' scores by performing comparisons between students. By such comparisons it is possible to produce statements like 'student A is more scientifically literate than student B', or 'nation X has more successfully achieved the science curriculum than country Y'. Comparisons like these are, of course, the essence of comparative research. To be able to make comparisons, we need measures of differences where there is a difference. In other words, to compare students' achievement we need tests that represent the same trait for all students in all countries, and this is not a straightforward task. In statistical terms, the overall differences are represented by the variance. In order to get reliable achievement measures we need tests that produce a relatively large variance: variance between countries, between schools and between students. The aim of many of the analyses that use data from PISA and TIMSS will be to explain or account for this variance. Typically, one can use student background variables as explanatory variables in correlation/regression-based analyses to account for students' achievements.

However, the measures produced and the theoretical underpinnings of such measures do not help us to interpret what students actually *think*, *know* or *can do* within the domain measured. Mislevy (1993) commented on this in response to what could be called the 'quest for a new foundation for test theory':

Educational measurement faces today a crisis that would appear to threaten its very foundations. The essential problem is that the view of human abilities implicit in standard test theory – item response theory as well as classical truescore theory – is incompatible with the view rapidly emerging from cognitive and educational psychology. Learners increase their competence not by simply accumulating new facts and skills, but by reconfiguring their knowledge structures, by automating procedures, and chunking information to reduce memory loads, and by developing strategies and models that tell them when and how facts and skills are relevant. The types of observations and the patterns in data that reflect the ways that students think, perform, and learn cannot be accommodated by traditional models and methods. (Mislevy, 1993. pp. 19-20)

Developing a new type of test theory is definitely beyond the ambitions for my thesis. However, the idea underlying the work in this thesis was that the data collected in studies like PISA and TIMSS have the potential to provide information about important characteristics of students' thinking and knowledge in science. However, this information is typically not analysed and presented in the primary publications from such studies.

1.1.2 The item-specific information

Although I have indicated that analyses of students' cognitive traits through test theory have some limitations, this is not meant to be a criticism of test theory as such. Moreover, it is not meant to indicate that the use of test theory is inadequate for the purposes of TIMSS and PISA. Test theory provides some powerful tools for measuring psychological traits such as students' ability in science. In TIMSS and PISA the primary agenda is to study such traits.

And in the reporting there is much focus on low and high achieving countries. However, there is also development in PISA to use the described scales more in the reporting. The described scales are verbal descriptions of students' achievement profiles at different levels of the scale.

However, in this thesis I will demonstrate that there is additional information to be analysed at the item level. In short my argument presented in the following is that the measures or scales developed to represent the student's achievement or ability only exploit a small fraction of the information given by the data collected. Even though all the items to some degree measure the same trait, and thus contribute to a reliable overall score, there will be a major portion of item-specific variance. Let us take a simple example. For a typical test simple isolated right-wrong items have a (so-called *point-biserial*) correlation with the overall test score in the order of 0.30–0.40. In classical psychometrical terms this means that only 9–16% of the variance for an item can be seen as 'true' variance related to the common trait being measured, whereas the major portion of the variance is 'error' variance. Furthermore, the correlation between items is very low, with values of 0.10–0.20, again a sign of very low (1–4%) common variance.

Figure 1.1 illustrates that items are included in a test to measure a trait. In the model in Figure 1.1 the trait is represented by a single latent factor. The underlying idea for such factorial models is that there is a psychological trait, e.g. ability in science that, to some degree, causes or accounts for students' item scores on each single item. The degree to which this trait accounts for the item score is represented by the numbers besides the arrows pointing from the latent factor to the item. The fact that the direction of the arrow is from the latent factor intends to communicate the direction of the hypothesised causal relationship (Loehlin, 1998).

Although this model lacks detail regarding the way the latent variables (the total scores) are actually calculated in TIMSS or PISA, it illustrates the principle that even if each single item contributes to the overall score as they are intended to do, most of the item variance is not accounted for by the latent factor.

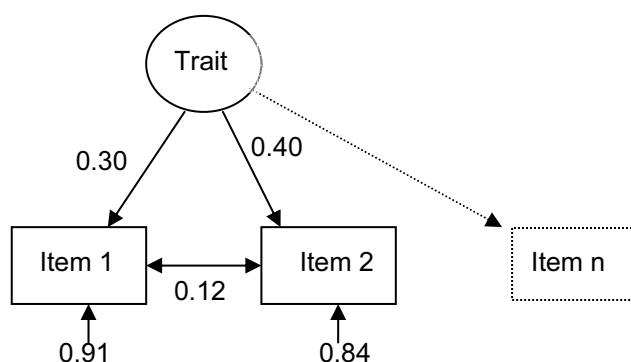


Figure 1.1: A hypothetical example of typical inter-item and item-trait correlations. In this example the trait is represented as a latent variable in a one-factor solution. The inter-item correlations, the factor loadings and the item residuals are presented for two of the n items. It is easy to calculate that the standardised item-specific variances (the residuals) are 0.91 (calculated as $1-0.3^2$) and 0.84 for these two items respectively.

Based on the above considerations alone, one can argue that from a cost-efficient perspective the major portion of information from a typical test is thrown away when only the overall score is analysed. From a science educator's perspective there will be potentially a lot of additional information within reach by engaging

in secondary analyses. The item-specific variance can be viewed as relevant and highly interesting information, while from a test theory perspective it is simply regarded as ‘error variance’. This item-specific variance may, for instance, be systematically linked to components such as the topic addressed in the item, the item format, and the context or situational factors specific to the item. If that is the case, it would in general be difficult to label this as ‘error variance’ or random noise.

1.1.3 The nominal information

So far the argument has concentrated on the ‘right-wrong’ dimension only. The argument is even stronger when one also takes into account the information about *what* the student responses were for each specific item, not only whether the responses were correct or not. The actual student responses, as originally coded and punched into the data file, contain information beyond the correctness information. In a multiple choice item for instance, the person who enters the code into the computer types a number representing the response selected by the student.

Question 4: **SEMMELWEIS’ DIARY**

Many diseases may be cured by using antibiotics. However, the success of some antibiotics against puerperal fever has diminished in recent years.

What is the reason for this?

- A Once produced, antibiotics gradually lose their activity.
- B Bacteria become resistant to antibiotics.
- C These antibiotics only help against puerperal fever, but not against other diseases.
- D The need for these antibiotics has been reduced because public health conditions have improved considerably in recent years

Figure 1.2: An example item from the PISA 2000 science assessment: ‘Semmelweis’ Diary’, question 4.

The example in Figure 1.2 is typical of the multiple choice items in PISA. Of four available responses, only one response alternative (in this case B) is regarded as a correct response³. In this situation students who select one of the wrong alternatives (A, C or D) have given different responses. This may be

³ Strictly speaking, it is not always the case that some answers are correct or wrong. It would be more precise to say that one of the responses in a multiple choice item is regarded as ‘appropriate’ or ‘better’ given the question, while the other choices are regarded as answers that are ‘not appropriate’ or ‘worse’ given the question. However, it is easier in the following to *speak* of these responses as wrong or incorrect.

considered as information that could be analysed and studied. For instance, it might be that the response labelled ‘A’ captures students with a very interesting view on the mechanisms of antibiotics, although it is not altogether evident what possible views could be behind such a response. It might for instance reflect some sort of a ‘half-time’ principle where a substance gradually changes properties through a physical or chemical process. However, the point here is only that the information punched into the data file is originally represented by a categorical variable, or in terms of measurement levels, it is a nominal variable.

This is evidently also the case for all other items in PISA. The items with more open-ended formats are also first coded, and of particular interest for this thesis is the fact that many of the open-ended science items are coded according to a marking rubric that for some items is quite detailed. I have included several examples of items and associated marking rubrics in Appendix 1, and the general principles for the marking is presented in some detail in chapter 3. However, it is enough for the introduction here to conclude that as the item data are punched into the computer, information exists that describes the type of response given by the student.

As the data are processed, however, the item is *scored*; one response in a multiple choice item receives 1 point, and all the rest receive 0 points. In other words, all students who selected one of the wrong answers receive the same score. The point which should be obvious by now is that by analysing the scored variables only, information is lost. This information could potentially be related to specific ways of representing knowledge, or specific solution strategies applied by the students.

1.1.4 Three levels of aggregation of the item data

My suggestion, proposed in this thesis, is that by analysing the cognitive item data at different stages in the processing of the data, and at different levels of aggregation, it is possible to explore information about the ways in which students typically represent scientific knowledge and make use of this when confronted by an item. Figure 1.3 illustrates how the relationship between three levels of aggregation of the items can be perceived. Since my argument also relates to the categorical information that is available before the item data are scored, from now on I will not use the term variance, but instead will continue to use the more imprecise term ‘information’.

The test score is obtained by a lengthy process starting with the coding of single items, continued by the scoring of the items. The item scores are then aggregated for each student into a total student score. This process is discussed in more detail in chapter 3. The point here is that the information represented by the overall score is not equal to the total information available from the single items. This is illustrated by the size of the areas in Figure 1.3, and this is exactly the point that is also argued in Figure 1.1. Moreover, as will be returned to in chapter 3, the information at the test score level does not have the same properties as the information available at a lower level of aggregation. For instance, depending on the test, this information is usually very reliable.

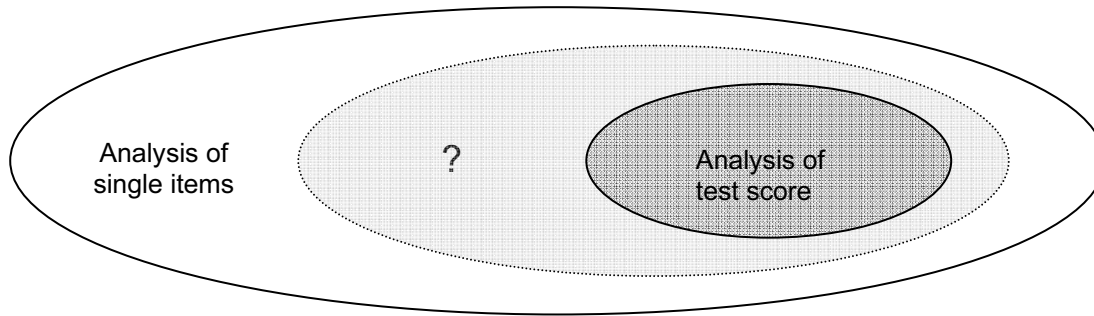


Figure 1.3: The information available for analysis at different levels of aggregation of item data.

On the other hand, at the lowest level of aggregation, items may be analysed one by one. This is information that by nature is specific to the item, and, as is clear from Figure 1.1, this information is only partly affected by the overall trait represented by the total test score. Many factors affect students' responses to a particular item: the format of the item, the wording of the stimulus material, specific knowledge related to the item, where the item is placed in the booklet, etc.

The third level of aggregation, labelled by a question mark in Figure 1.3, illustrates that the information in the items can be aggregated and analysed at an intermediate level between the two extremes. In general the intermediate level refers to the collective information represented by more than one item, and less than all items. Some possible intermediate levels are mentioned briefly below.

Both the TIMSS (Mullis *et al.*, 2001; Robitaille *et al.*, 1993) and the PISA framework (OECD-PISA, 1999, 2003c) define subdomains within each of the overall domains tested. Items have been categorised according to these subdomains. It is therefore possible to establish subscales representing more specific traits by aggregating item scores for the items categorised within the same subdomain. This would therefore aggregate item level data to an intermediate level. Results for, and analyses of, these subscales are presented in the international and national reports from studies like TIMSS and PISA. These measures provide more detailed information about students' strengths and weaknesses in different countries than the overall scores. One typically finds that countries that perform equally well on the test have different profiles across these subdomains. In principle, there are many ways to categorise the items, or in other words, information may be aggregated to an intermediate level in a number of ways: the items may be categorised by format, by content, by the cognitive processes involved, etc. Subdomains established by categorising items in this way can be regarded as theoretically established item clusters at an intermediate level.

The clustering of items could also be data driven. For instance, clusters could be formed by item difficulties. This approach has been used to analyse both TIMSS and PISA data, mainly as a step to establishing so-called described scales or proficiency levels, that is, verbal descriptions of what it means to have a score within some predefined intervals or at certain points along the scale (Kelly, 1999; Turner, 2002). In general, a great number of data-driven methods for grouping items (or respondents) into clusters exists: factor analysis, cluster analysis, discriminant analysis and correspondence analysis, to name only a few.

1.1.5 The aim of this thesis

From the discussion above my general rationale and motivation should be quite transparent: there is a large amount of information in the data from TIMSS and PISA that is typically not analysed as part of the primary agenda of these studies. The motivation for my work was to refine and extend this rationale, and furthermore to describe and discuss the nature of this information: (a) the nature of the information derived from students responses to single items (paper I); (b) the nature of the information when using a group of items as the unit of analysis (paper II); and (c) the nature of the information in the country-specific profiles across the items (paper III).

Even if the unit of analysis varies, the overall concept to be explored is the information in the single items. In the papers I explore this information, both to have an empirical base for the general discussions about the nature of the information in the single items, but also as examples of how the information in the single items may be analysed. At the end of the day, the aim of each analysis has been to find evidence for students' knowledge and thinking in science beyond their overall ability or achievement. The underlying and more general purpose of this work is to illustrate that through secondary analysis of the data from TIMSS and PISA contributions can be made to research in science education, and furthermore contributions can be made to the continuing improvement in the quality of information about subject-specific issues available from these studies.

1.2 The research sequence

One of the main characteristics of my work is that I have not been able to follow what is often labelled as 'the scientific approach' (see for instance Ary *et al.*, 1996), where the initial activity is to identify a legitimate research question. The science education community has been criticised for not building new research on the knowledge already established (Millar *et al.*, 2000). That is, one should initiate research that will contribute to the cumulative effect of all research. In this sense, research should ideally be done in a sequence that starts with the requirement that the researchers are familiar with the literature in their field, and then continues in the following sequence: (a) present a problem or an issue arising from the literature; (b) develop a design that would allow for the analysis of the stated problem; (c) develop the instruments to be used; (d) collect the data or material as required by the design; and (e) analyse the material. In reality this is a simplified model, and interactions between the different stages would frequently occur. However, this sequence describes the overall succession with which a research project often progresses.

In my case I have chosen to follow a somewhat different sequence. I started by arguing that the data already collected in PISA and TIMSS contain information that is worthwhile to explore. I have already presented part of this argument above, but the rationale will be developed further in section 1.3 and chapter 2. The initial activity in my research was therefore not to establish a very specific research question connected with the contemporary agenda of science education. Instead I have chosen the data at hand (stage 'd' above) as my starting

point. An alternative way to see this is that the data, in my case, replaced the role that the research literature in the field was given in the sequence above (stage ‘a’).

1.3 The structure of the thesis

This thesis consists of four chapters and three articles published in journals. In addition I have included an appendix with a collection of some of the items from PISA that have been made publicly available (Appendix 1). All the units included in this appendix are to some extent referred to in the thesis. This set of items functions as an illustration of the type of information that is available in the codes used to mark the items.

When cross-references are made to the chapters, they are labelled by their number. When referring to parts within a chapter, I will use the term ‘section’. For instance, this is section 1.3. The three articles are labelled as papers I, II, and III. Each of the chapters and papers has its own bibliography. In the following I will give very brief descriptions of the content of the chapters and papers in the thesis.

1.3.1 The chapters

Chapter 2 contains several elements: it presents the historical development of large-scale international comparative achievement studies in science education, it includes a comparison of TIMSS and PISA, it contains a discussion of the policy-relevance of TIMSS and PISA, it suggests why the data from these studies are valuable research resources in science education, and furthermore how the data may be analysed. Each of these main topics in the chapter may be regarded as isolated themes that target topics deserving closer attention. As such these parts may be considered as the beginning of several separate theoretical papers. However, the overall intention is that these parts, when taken together, constitute a line of reasoning for why researchers in science education could be motivated to take part in secondary research related to large-scale international comparative achievement studies. In many ways, therefore, this chapter is a vital contribution to the overall aim and purpose of this thesis, as presented in 1.1.5 above.

In chapter 3 some of the methodological issues relating to the empirical work in the three papers are presented. This includes a brief description of my general methodological position, which in short is captured by the two concepts *exploratory data analyses* and *pragmatism*. Furthermore, a central part of this chapter is to give a theoretical description of the character of the information in the single items in PISA, and the analytical potential of that information. This description is naturally affected by the findings of the analyses presented in the following papers, and as such, much of what I have found regarding the overall aim of the thesis is already presented in this chapter. In addition, a non-mathematical presentation of correspondence and homogeneity analysis, central to paper II, is given towards the end of the chapter. Chapter 3 is therefore a more fragmented chapter, as compared to chapter 2, and does not contain one overall line of argument.

Chapter 2 and 3 present the general background in more detail than journal article format allows. A central element in secondary data analysis (see more in section 2.5) is to become familiar with the data: the procedures used to collect the data; the intentions behind the instruments used to collect the data; and the historical, social and political context in which the data are embedded (Rew *et al.*, 2000). Chapters 2 and 3 may therefore also be considered as the documentation of my familiarity with the data that I have analysed in the three papers.

Chapter 4 is a summary and discussion of the main findings of both my theoretical and empirical work, and these summaries will particularly emphasise the overall aim of the thesis. However, the discussion will be kept fairly short since it will mainly be a repetition of statements in the chapters and the papers.

1.3.2 The papers

The three papers following these chapters are all examples of different types of analyses that have been done of the data in TIMSS and PISA. They are separate articles set in slightly different contexts and with different units of analysis. However, all the papers are introduced and framed within the general rationale presented above: to study the item-specific information, and how this may be used to gain insight into students' mental representations of scientific phenomena or concepts. Below is a list of the three papers:

- Paper I: Olsen, Turmo & Lie (2001): Learning about students' knowledge and thinking in science through large-scale quantitative studies, *European Journal of Psychology of Education*, 16(3), pp. 403-420.
- Paper II: Olsen (2004): The Search for Descriptions of Students' Thinking and Knowledge: exploring nominal cognitive variables by correspondence and homogeneity analysis, *Scandinavian Journal of Educational Research*, 48(3), pp. 325-341.
- Paper III: Olsen (2005): Item-by-country interactions in PISA 2003: Country-specific profiles of science achievement. *Not published*.

A condensed version with a more narrow focus has been published as Olsen (2005): An exploration of cluster structure in scientific literacy in PISA: Evidence for a Nordic dimension?, *NorDiNa*, 1(1), pp. 81-94.

Paper I explores, describes and generalises some important characteristics of the information in single items, using data from TIMSS 1995. However, the findings of this paper are not specific to TIMSS, and therefore this paper addresses the general aim of this thesis: to explore and describe the nature of information in the cognitive data from studies like PISA and TIMSS.

Paper II explores the methodologically challenging issue of how students' responses to a relatively small selection of science items in PISA 2000 may be used to develop profiles for the codes used to mark the students. Having

established a relationship between the different codes used for the items, the paper studies how this information can be used to compare the profiles for countries across these items.

Paper III explores how the information in the scored science items in PISA 2003 can be used to describe the typical strengths and weaknesses of countries beyond their overall level of achievement.

1.4 References

- Ary, D., Jacobs, L. C., & Razavieh, A. (1996). *Introduction to Research in Education* (5th ed.). Fort Worth: Harcourt Brace College Publishers.
- Kelly, D. L. (1999). *Interpreting the Third International Mathematics and Science Study (TIMSS) achievement scales using scale anchoring*. PhD-thesis, Boston College, Boston.
- Loehlin, J. C. (1998). *Latent variable models: factor, path and structural analysis* (3rd ed.). Mahway: Lawrence Erlbaum Associates, Publishers.
- Millar, R., Leach, J., & Osborne, J. (2000). Foreword to Section 1: Researching teaching and learning in science. In R. Millar, J. Leach & J. Osborne (Eds.), *Improving Science Education: the contribution of research* (pp. 7-10). Buckingham: Open University Press.
- Mislevy, R. J. (1993). Foundations of a New Test Theory. In N. Frederiksen, R. J. Mislevy & I. I. Bejar (Eds.), *Test Theory for a New Generation of Test* (pp. 19-40). Hillsdale: Lawrence Erlbaum Associates, Publishers.
- Mullis, I. V. S., Martin, M. O., Smith, T. A., Garden, R. A., Gregory, K. D., Gonzalez, E. J., Chrostowski, S. J., & O'Connor, K. M. (2001). *TIMSS Assessment Frameworks and Specifications 2003*. Boston: International Study Center, Lynch School of Education, Boston College.
- OECD-PISA. (1999). *Measuring Student Knowledge and Skills*. Paris: OECD Publications.
- OECD-PISA. (2003). *The PISA 2003 Assessment Framework: Mathematics, Reading, Science and Problem Solving Knowledge and Skills*. Paris: OECD Publications.
- Rew, L., Koniak-Griffin, D., Lewis, M. A., Miles, M., & Ann, O. S. (2000). Secondary Data Analysis: New Perspective for Adolescent Research. *Nursing Outlook*, 48, 223-229.
- Robitaille, D. F., Schmidt, W. H., Raizen, S., Mc Knight, C., Britton, E., & Nicol, C. (1993). *Curriculum Frameworks for Mathematics and Science*. Vancouver: Pacific Educational Press.
- Turner, R. (2002). Proficiency Scales Construction. In R. Adams & M. Wu (Eds.), *PISA 2000 Technical Report* (pp. 195-216). Paris: OECD Publications.

2 LINCAS as a frame for research

2.1 Introduction to the chapter

This chapter is a synthesis and further development of several papers I have been involved in writing over the last years (e.g. Olsen, 2004; Turmo & Olsen, 2003). The primary function of this chapter in the thesis is to present the broader background for the empirical papers in the thesis.

My thesis is framed within large-scale international comparative achievement studies in education (LINCAS⁴). I would like to acknowledge that the data I am analysing is a consequence of a collective effort by a great number of people throughout the world. This means that as a researcher using these data, I should be aware of and familiar with the historico-socio-political context in which the data were collected (Rew *et al.*, 2000). I therefore find it necessary to give a short introduction to and discussion of the history of LINCAS (section 2.2). This is done with a particular emphasis on science assessments as part of LINCAS. The aim of this first part is to provide a general background describing the origin of the TIMSS study, which in turn is used to describe why PISA entered the scene in the latter half of the 1990's. In particular this is useful in order to understand the wider rationale for the inclusion of scientific literacy as a component in PISA. My analyses presented in the three papers in this thesis are based on data from both these studies, but the majority of my analytical work has been related to PISA, and this study is therefore given priority in this chapter.

The chapter then turns to a more detailed presentation of PISA and TIMSS (section 2.3), mainly through a comparison of the two. Then follows a short section discussing the policy relevance of research in science education, and in light of this, a short discussion or characterisation of the relationship between research in science education and LINCAS is given (section 2.4).

This basis provides a meaningful background for the main purpose of this chapter: to discuss the potential for secondary analyses of the PISA/TIMSS data (section 2.5).

2.2 The growth of LINCAS: From FISS to PISA

This section is a general introduction to the genesis and growth of LINCAS. LINCAS as an abbreviation of large-scale international comparative achievement studies in education was first used by Bos (2002), and he defined studies under this label as

...studies in which both achievement of a certain age/grade group in one or more subjects is compared across education systems and effects of contextual factors at system, school, classroom and student level on achievement are studied. (p. 2)

This definition points to a number of characteristics for these studies. Primarily they measure *achievement* of some sort, and this has to be done by applying a method which allows for *international comparisons*. The definition also

⁴ This abbreviation is taken from Bos (2002), see below.

identifies that differences between countries can be studied as *effects of contextual factors*. Implicitly this part of the definition also implies that the aim of these studies is to be able to find measures which can be *generalised* to schools and educational systems. The definition does not explicitly identify the criterion for labelling a study as 'large-scale'. However, the fact that these studies should be able to say something about the effects of contextual factors at system level suggests that there should be more than just a few countries involved. Also, in order to get reliable measures which may be generalised to the system level, rigorous procedures for sampling a large number of schools/classes and students must be employed.

In the following presentation of the growth of LINCAS, I will suggest that although the aims of, the use of data from, the organisation of and the methodology applied by LINCAS have developed gradually, there have been two distinctly different and partly competing overall visions underlying the studies. LINCAS was first conceived as a specific design or method for conducting research into education with an across country comparative perspective. This initial idea behind the birth of LINCAS will be labelled *Purpose I – the research purpose*. Gradually, LINCAS have been adopted by policy makers. LINCAS are considered as instruments with which to monitor the outcomes of educational systems and the study of possible determinants of such outcomes which could be regulated by legislation or regulation of the educational system. In short, such regulations, either through laws or through other instruments (e.g. curriculum plans, economic incentives, overall plans for teacher education and certification, to mention just a few instruments that government can use as tools for systemic change) will be referred to in general as educational policy. This rationale for LINCAS will be labelled *Purpose II – the effective policy purpose*.

The labels Purpose I and II are only suggested as useful heuristic devices for understanding some of the ideological tensions that these studies have to live with⁵. However, in using this dichotomy I do not suggest that the research purpose and the effective policy purpose are incompatible. On the contrary, as will be clear in the following discussion, I will suggest that LINCAS may be considered as one of the arenas where researchers in education and educational policy makers can exchange ideas and develop mutual interest for and acceptance of each other's engagement in educational issues, both at national and international level.

2.2.1 Purpose I: The research purpose

The idea of comparative studies in education is not new (Howson, 1999; Postlethwaite, 1988), and some even claim that systematic studies of education from a comparative perspective have been around since the ancient Greeks (Kaiser, 1999; Shorrocks-Taylor, 2000). Comparative studies received a boost with the establishment of national educational systems during the 19th century (Bruhn, 1995; Howson, 1999) and with the establishment of firm national institutions with a responsibility for education at a national level. This was

⁵ This is to some degree inspired by the way Roberts (forthcoming) uses the terms Vision I and Vision II in his review of the concept of scientific literacy.

further strengthened in the latter half of the 20th century when educational reform was put high on the political agenda throughout the world (Husén & Tuijnman, 1994). During the 20th century other developments in society, which will be returned to shortly, greatly affected the volume and progress of comparative research in general and LINCAS in particular. Today the label ‘comparative studies in education’ refers to various types of research ranging from issues of the more philosophical and methodological aspects of comparing across cultures to very specific studies of narrowly defined aspects of education across countries, regions or classrooms. This label in general also covers studies with a great variety of designs and scales (see for instance Alexander *et al.*, 1999; Alexander *et al.*, 2000).

The idea of LINCAS as we know it today was materialised and defined as a research agenda with the establishment of IEA – the International Association for the Evaluation of Educational Achievement in 1961⁶ under the auspices of the UNESCO Institute for Education (Husén & Tuijnman, 1994; Keeves, 1992). The fundamental idea of the founders of IEA is very clearly expressed by one of them, Torsten Husén (1973):

We, the researchers who... decided to cooperate in developing internationally valid evaluation instruments, conceived of the world as one big educational laboratory where a great variety of practices in terms of school structure and curriculum were tried out. We simply wanted to take advantage of the international variability with regard both to the outcomes of the educational systems and the factors which caused differences in those outcomes. (p. 10)

The term laboratory in this quote is used only as a metaphor, since laboratory conditions with controlled experiments are not feasible in educational research, due to both practical and ethical considerations⁷. The alternative to the experiment would therefore be survey designs where the variables of interest could be studied under a great variation of different conditions. In this way “differences between education systems would provide the opportunity to examine the impact of different variables on educational outcome” (Bos, 2002, p. 5). “Thus the studies were envisaged as having a research perspective..., as well as policy implications” (Kellaghan & Greaney, 2001, p. 92). The assumption is, in other words, that educational organisation and practice affects educational opportunities and outcome, and this can be the subject of empirical research with the aim to:

...go beyond the purely descriptive identification of salient factors which account for cross-national differences and to explain how they operate. Thus the ambition has

⁶ From various sources it may be found that IEA was established in 1959. This was the year when the first study undertaken by and presented by the IEA started, but formally IEA was founded in 1961.

⁷ One classical application of experimental designs is to study the effect of a certain intervention in teaching. Such studies are still reported in educational research to some extent. However, today it is acknowledged that true experimental designs are not feasible, and instead such designs are most often labeled as quasi-experimental designs. Furthermore, when an effect is documented, it is rarely stated that teaching method A is better than teaching method B, rather that teaching method A is *more likely* to lead to the desired outcome than teaching method B. The aim of arriving at such conclusions is also acknowledged within the community of science education researchers, at least by some, as important research (Millar, 2003).

been the one prevalent in the social sciences in general, that is to say, to explain and predict, and to arrive at generalizations. (Husén, 1973, pp. 10-11)

The two quotes above from Husén should be seen as typical of the time and for the prevailing optimism regarding how the social sciences could contribute to the development of a better understanding of the causal relationship between different types of factors in society. Today social scientists are probably more reluctant to use phrases like “factors which caused differences” or “the ambition...in social sciences in general...is to predict, and to arrive at generalisations”.

2.2.2 Purpose II: The effective policy purpose

Policy makers are required to establish overall plans for the nation’s educational system; e.g. to

- establish systems for the training and certification of teachers;
- decide the amount of, and how to distribute, resources;
- specify the overall purpose of education as part of the wider social context and specific goals of achievement;
- decide how to organise the progression of schooling from childhood until adolescence and beyond.

To a large extent LINCAS and other internationally comparative data have been regarded by policy makers as providing information that is relevant in their continuous evaluation of such overall plans.

What was initially a formulation of a platform for comparative educational research coincided with a growing recognition among politicians, industrial leaders and others that education was one of the most central agents to realise long term political, societal or economic visions, such as to

- develop a society with a better distribution of resources across class, race, gender or any other social group;
- fulfil the need for a highly competent workforce in order to succeed in the international marketplace;
- enhance and further develop democracy by giving all citizens basic and further education so that they are enabled to fulfil their own life-agenda and become full-fledged participants in the democratic process.

These were just a few examples of the visions of the ideal society that to a large degree were, and still are, shared visions throughout large parts of the world. At the same time, during the post Second World War period, international organisations such as the UN⁸, the World Bank, the OECD and the European Union, were established and quickly grew in size and influence. These are organisations with different, and to some degree conflicting, agendas. But they all invest time and effort in the study of education in their member countries, and several of these organisations are linked to each other through joint projects.

⁸ United Nations

Examples of two joint OECD and UNESCO⁹ projects are the World Education Indicator programme (e.g. UNESCO & OECD, 2003) and the follow-up of PISA 2000 in 11 less developed countries in 2002 (OECD & UNESCO, 2003). UNESCO's role in the educational policy arena is mainly to give aid and support for educational development in the underdeveloped countries. This is particularly formulated in the Jomtien declaration (UNESCO, 1990), which eventually lead to the Education for All (EFA) initiative¹⁰. So far there have been no LINCAS directly related to EFA¹¹, but as part of their strategy they aim to implement international assessments by the year 2015 (UNESCO, 2002).

The role of the World Bank is also noteworthy as an example of how these organisations are intimately linked on the global educational arena. This institution does not participate in the development of the international studies, but provides funding for economically less developed countries to participate in, for instance, IEA studies such as TIMSS 1999 (Martin & Mullis, 2000), and parts of the EFA activities, for instance support for the establishment of national assessment systems (Kellaghan & Greaney, 2001).

In the post Second World War period the idea of central rational planning and policy making was very strong. At least in the Nordic social democratic countries this was the era of large-scale national initiatives to build the welfare state, using a top-down rational procedure¹² often referred to as social engineering. All in all, the main idea of LINCAS, to provide policy makers with solid empirical and quantitative descriptions and indicators of the educational systems, was perfectly aligned with the ideological climate of modernity in this period.

IEA became a provider of educational data and analyses not only to national policy makers, but also to several of the previously mentioned international organisations. In addition to UNESCO, which was involved in the establishment of the IEA, OECD (before PISA was established) used data from IEA studies in their publications *Education at a Glance* (eg. the use of TIMSS data in OECD, 1996, 1997, 1998). Since the first studies conducted in the early 1960's IEA has been in charge of a great number of LINCAS and other international comparative surveys in different subjects, and over the years the studies have grown to include a great number of countries throughout the world, and at the same time the methodological challenges that LINCAS are confronted with have been a driving force in the development of new designs and psychometrical procedures (Porter & Gamoran, 2002).

⁹ United Nations Educational, Scientific and Cultural Organization

¹⁰ For more information see <http://www.unesco.org/education/efa/>

¹¹ There has been an assessment of life skills, reading and numeracy in the Monitoring Learning Achievement (MLA) project implemented in several countries. However, so far the instruments have mainly been used in national contexts. Nevertheless, the data have proved to be useful in monitoring some of the aims of EFA (UNESCO, 2005).

¹² It is a paradox, however, given this ideological climate, that Norway joined the IEA quite late, unlike for instance Sweden which actually hosted the IEA for a period in the beginning. It is not easy to find any explanation for why Norway did not participate in the 1960's, but the most likely main reason why Norway did not participate in the 1970's is the fact that the 'battle against positivism' was very strong in Norway during this period, including a rejection of testing and measurement as such.

During the last few decades the growth of LINCAS has also probably been fuelled by the reform of public services that is often referred to as ‘new public management’. This is characterised by deregulation of the public sector and a drive towards a higher degree of privatisation of those parts of the public sectors that can be thought of as the infrastructure of society (e.g. the telecommunication sector, public transportation, postal services, health and education), although the ideals of privatisation are not equally present in all countries.

Deregulation implies a transfer of responsibility from the central government to the local authorities. Nevertheless, important decisions related to schools are to be made by policy makers at administrative levels above the local community level or local school level. Also, there is a need for information at the national level. A consequence in most countries where deregulation took place was therefore to reinforce the central government’s role by installing a national assessment system. This was a shift from the regulation of inputs (eg. specification of the use of the resources or number of students per class) to controlling the output (achievement, surveys of students and parents). In this way the service providers were made accountable both to the central government and to the users of the services. On the one hand the central government could control and direct the services by connecting measures of the output to incentives, or to intervene and manipulate the system to work as intended. On the other hand, the users could make use of the output measures in personal decisions regarding the public services.

In this context LINCAS provide many indicators considered as relevant, especially for the policy maker:

- A. They produce measures of some of the outputs, most importantly achievement measures.
- B. They produce indicators for systemic factors that may be directly linked to policy, such as average class size, availability of resources (e.g. computers) and allocation of time to different subjects. Moreover, LINCAS offers the possibility of relating such factors to achievement.
- C. They produce indicators of relationships between variables that policy seeks to change in a certain direction, e.g. the aim of schooling to provide an equal opportunity for all to learn, independently of background.
- D. Some LINCAS, for instance PISA and TIMSS, produce indicators for how A, B and C above changes over time by repeating the surveys with regular intervals.

Most importantly they provide a context for the interpretation of many of these indicators. In an assessment with no international component it would be very difficult to establish whether effects are small or large. Even though the international variation cannot be used in order to draw causal inferences, it provides a description of what is possible and a context in which national data can be compared. One example is related to the issue of equity. It is often expressed in policy documents that large systematic differences in achievement between pupils from different socio-economic levels indicate that school systems do not provide equal opportunities for all. Also, a large standard deviation in

achievement in the total population is often considered as an indicator of inequities. For both these examples the international context provides an opportunity for the policy makers to evaluate whether or not the differences between students or groups of students are large or small as compared to other systems perceived as relevant for comparison.

Furthermore, LINCAS have the advantage, as seen from the perspective of a person in charge of decision making, that all the indicators are produced through sampling plans and designs that make it possible to generalise to the system level. Such designs are most likely to be perceived by the policy makers as a reassurance that their decisions are the final step in a rational process. In this process informed decision could in principle be thought of as largely independent of the people in charge of making the decisions.

In summary then, the second purpose of effective policy development is in many respects compatible with the aims of the researchers who established IEA and conducted the first surveys (Purpose I). The difference is that within Purpose II LINCAS is not mainly considered as a basic research in education. This is not to say that LINCAS can no longer be used to study fundamental issues in educational research. However, such research issues are secondary to the primary purpose, which is to monitor the educational outcome of educational systems in order to inform policy makers. This issue of secondary research possibilities will be returned to in some more detail later.

2.2.3 From FISS to TIMSS

LINCAS primarily aim at studying school subjects (or competencies across school subjects) that are seen as instrumental to many of the societal or political aims briefly mentioned above. Science is a subject considered as important in that respect. In the first studies conducted it is also likely that the decision to include some school subjects, such as science and mathematics, and to exclude other subjects, was based on two other assumptions. Firstly, the subjects of science and mathematics were probably to some extent thought of as ideal contexts for evaluating students' cognitive competencies per se (e.g. general problem-solving skills), unlike for instance subjects related to the arts emphasising more aesthetic dimensions of students' competence. The outcomes in maths- and science-related achievement tests could therefore be taken as indicators of how successfully schools have fostered cognitive and metacognitive skills in general. Secondly, the subjects of science, mathematics, reading comprehension, English and French as foreign languages and parts of civic education are often considered as universal subjects which in principle could be identical in different countries, and indeed many countries do have these subjects in their secondary school curriculum (Husén, 1973). On the other hand a subject such as history would of course be very influenced by the national context. Therefore, science and mathematics were probably seen as promising candidates for across-country comparisons since it could be assumed that large parts of the curriculum across countries would be similar.

Science has therefore been one of the subjects that have been included from the very beginning, starting with the 1959–1961 pilot study conducted in 12

countries to investigate the feasibility of conducting large-scale assessments across countries. In this pilot study science items were included alongside items in other subjects (Foshay, 1962; Husén, 1967). This study eventually led to the establishment of the IEA, and the first study conducted by this association was the *Six Subject Study* with science as one of the subjects¹³ (Comber & Keeves, 1973). Then came SISS (*Second International Science Study*) (IEA, 1988; Keeves, 1992; Postlethwaite & Wiley, 1992; Rosier & Keeves, 1991) followed by TIMSS 1995 (Beaton *et al.*, 1996; Harmon *et al.*, 1997; Martin *et al.*, 1997; Mullis *et al.*, 1998) and the sequels TIMSS 1999¹⁴ (Martin *et al.*, 2000a) and TIMSS 2003 (Martin *et al.*, 2004b).

2.2.4 OECD and the need for new types of assessments

The OECD realised that they needed other types of measures than those typically obtained from the IEA studies. From an OECD perspective one problem with the IEA studies was primarily related to the simple fact that not all OECD countries participated in them (Kellaghan & Greaney, 2001; OECD-PISA, 1999). However, there is also reason to claim that the OECD evaluated the output measures from the IEA studies as not fitting very well to the OECD perspective on education. In a meeting by the educational ministries in 1996 it was decided that ‘lifelong learning for all’ should guide the organisation’s interest and engagement in issues related to education (OECD, 2001). In light of this decision, the available international achievement measures were evaluated as not suitable. This is evident in the introduction of the first PISA framework stating that

Underlying OECD/PISA is a dynamic model of lifelong learning in which new knowledge and skills necessary for successful adaptation to changing circumstances are continuously acquired over the life cycle. (OECD-PISA, 1999, p. 9)

And furthermore, after referring to other LINCAS performed over the years the framework continues:

The quality and scope of these surveys have greatly improved over the years but they provide only partial and sporadic information about student achievement in limited subject areas. (OECD-PISA, 1999, p. 10)

The use of the notion of ‘partial and sporadic information’ to characterise curriculum-based studies such as TIMSS is probably best understood in light of one of the concepts so regularly seen in policy publications related to education, and especially so in publications framed within ideas of lifelong learning or continued learning; the concept of human capital. This is defined by the OECD as

...the knowledge, skills, competencies and other attributes embodied in individuals that are relevant to personal, social and economic well-being. (Schleicher, 2000, p. 69)

¹³ The First International Science Study, or the abbreviation FISS is also used as a label for this study. However, this is a retrospective label of the science part of the study, and was not initially a label used before the advent of the follow-up studies.

¹⁴ Often referred to as TIMSS Repeat.

Human capital is seen as one of the most important resources in a post-industrial society often referred to as the ‘knowledge society’. Evaluating aspects of human capital for school leavers is therefore not only an evaluation of the educational system. Just as important is that these aspects are regarded as predictors of young people’s preparedness for life, or as predictors of the future resources in a country. Several of the indicators produced by the OECD are related to monitoring the human capital in countries.

Prior to PISA typical indicators or proxies of human capital were, in addition to the curriculum-based achievement measures, various measures of investments in education and research, or the absolute and relative numbers of students leaving with a qualification (degree) from different levels in the educational system. Specifically related to science is the numbers of students entering or leaving science-related studies. However, such indicators are relatively rude measures of the general level of human capital in the population, and in general, such indicators of human capital have not been successful predictors of economic growth or of other measures of social development (Steedman, 1999). Andreas Schleicher, the head of the OECD department responsible for producing such indicators, is very clear on this in stating that

Estimates of the stock of human capital or human skill base have tended, at best, to be derived using proxies such as level of education completed. When the interest in human capital is extended to include attributes that permit full social and democratic participation in adult life and that equip people to become ‘lifelong learners’, the inadequacy of these proxies becomes even clearer. (Schleicher, 2000, p. 69)

It may be argued that historically such indicators were probably very informative since the main objective of educational policy was to increase the volume of, and participation rate for, people in all levels of the educational system. In economically less developed countries the top priority is still to provide access to education for all (Reddy, 2005; UNESCO, 2005), but in general, the situation in the OECD countries and other economically affluent countries has been to focus less on these quantitative aspects, and more on the quality of the human capital. The measures of participation rates in education or public spending in education do not indicate, for instance, whether students are leaving school with cognitive and affective resources for continued learning.

By the inclusion of science as one of the central cognitive elements of human capital needed for lifelong learning, a message is given about the role of science in general education. The choice of measuring a science component as part of human capital is most often justified by reference to the science base of the knowledge society. Society is heavily influenced by technology, the assumption being that technological innovations are derivations of progress in science. Therefore, knowledge in, of, and about science and technology is needed. The argumentation may, however, be extended to also include a broader humanistic or ‘Bildung’ perspective on general education. In order for people to live up to the expectations of, and the possibilities offered by, a deregulated democratic society, i.e. to take responsible actions and to be informed users and decision makers, it is for instance necessary to be able to evaluate and make use

of empirical data, perhaps even to be able to collect such data in a proper manner¹⁵. Furthermore, an insight into scientific ways of relating to the world is also sometimes argued for as an important part of our cultural heritage, and as such, an understanding of scientific methods and the products and consequences of science is to understand the self and the world. Science can therefore be seen not only as an indicator of future human capital in science per se, but also as part of the general cognitive resources that everyone would need to make sense of the world and to participate in society.

The PISA framework and other OECD documents, i.e. those describing the project *Definition and Selection of Competencies* (DeSeCo) (see for instance Rychen & Salganik, 2003), provides further details into why the lifelong learning perspective introduces a need for indicators related to students' achievement other than those offered by for instance IEA studies such as TIMSS¹⁶. Some of the more specific aspects of a design deemed as suitable for the OECD will be returned to when presenting PISA through a comparison with TIMSS.

2.3 Comparing PISA with TIMSS

At the outset PISA and TIMSS are regarded as very similar types of study; they are both large-scale international comparative achievement studies in education and as such the previously cited definition of LINCAS by Bos (2002) summarises some of the important similarities between these two studies. They are large-scale surveys with a very similar methodological basis, e.g. they

- are sample-based studies of clearly defined populations;
- apply the same type of instruments (e.g. student questionnaire and cognitive booklets);
- process the data with similar psychometrical methods;
- are governed by a consensus-driven process from initial ideas to final instruments;
- enforce rigorous quality control, e.g. of translation or adaptation of the test material;
- have cyclic designs with a focus on measuring trends.

Furthermore, both studies include measurements of highly related constructs: e.g. mathematical and scientific competency, and student and school background characteristics and attitudes. However, in the following some of the important differences will be highlighted instead. Some of the differences have already been touched upon in the historical approach since this approach demonstrated

¹⁵ The ability to deal with empirical data is not an ability exclusively related to the school subject of science. It might be considered as equally relevant to the school subjects that in general may be referred to as civic education. However, I would claim that science as a school subject traditionally has had, and should continue to have, the main responsibility for fostering such competencies since the collection, analysis and evaluation of empirical data is essential in the natural sciences.

¹⁶ It should, however, also be noted that in the DeSeCo perspective the measures derived from PISA are also considered as too limited given the much broader description of competence developed in the project. In DeSeCo competence is conceived of through three key competencies; to be able to interact in heterogeneous groups, act autonomously, and use tools interactively (Rychen & Salganik, 2003).

how PISA is a product of prior LINCAS, that is, the IEA studies in general and TIMSS in particular.

2.3.1 PISA in a nutshell

A very broad description of the purpose of PISA is that it aims to regularly monitor student achievement in the participating countries by the end of compulsory education. Also, a growing number of countries outside the OECD participate in PISA. The project has a very wide scope, testing what has been phrased as *literacies* in different areas: reading literacy, mathematical literacy and scientific literacy. Together with this, a wide array of background information about the students and the schools is collected.

The data collection takes place every three years. The first testing cycle took place in 2000, with reading literacy as the core domain with 2/3 of the testing time (OECD-PISA, 2001, 2002a, 2003a, 2005; OECD & UNESCO, 2003; Willms, 2003). In 2003 mathematical literacy was the main domain (OECD-PISA, 2004a, 2004b), and in 2006 scientific literacy will be the major domain. The frameworks (OECD-PISA, 1999, 2002b, 2003b) and the international and national reports (e.g. Kjærnsli *et al.*, 2004; Mejding, 2004; Prenzel *et al.*, 2004) give much deeper presentations of the project. In this paper some of the most important characteristics of PISA will be given by comparing with TIMSS.

2.3.2 TIMSS in a nutshell

TIMSS is the successor of a series of surveys of science and mathematics achievement across the world. Like PISA, the cognitive measures involved in TIMSS, are contextualised by background variables. The achievement measures in TIMSS are related to the curriculum in the participating countries, and the format and substance in the items that students respond to are probably familiar to most students in many countries throughout the world.

TIMSS has been regularly administered every four years since 1995 (Beaton *et al.*, 1996; Harmon *et al.*, 1997; Martin *et al.*, 1997; Martin *et al.*, 1999; Martin *et al.*, 2000b; Mullis *et al.*, 1998; Mullis *et al.*, 2000a; Robitaille, 1997; W. H. Schmidt *et al.*, 2001; W. H. Schmidt *et al.*, 1997a; W. H. Schmidt *et al.*, 1997b; Stigler *et al.*, 1999). That is, the study was also implemented in 1999 (Martin *et al.*, 2000a; Mullis *et al.*, 2000b) and in 2003 (Martin *et al.*, 2004b; Mullis *et al.*, 2004). However, the extensive design/analysis plan of the 1995 study (Robitaille & Garden, 1996), including three populations and a complex set of instruments (several achievement tests, a performance test, several background questionnaires, a video study and analyses of curriculum documents and textbooks), was not repeated in the following studies. Only the achievement tests and background questionnaires in two of the populations were repeated. A new study is already planned in 2007.

The two domains covered in TIMSS are allocated equal testing time every year. The design and purpose of TIMSS is also described in detail in the frameworks (Mullis *et al.*, 2001; Robitaille & Garden, 1996; Robitaille *et al.*, 1993) and in the international and national reports (e.g. the Norwegian reports

Angell *et al.*, 1999; Brekke *et al.*, 1998; Kind *et al.*, 1999; Kjærnsli *et al.*, 1999a; Lie *et al.*, 1997)

2.3.3 Why compare the two?

There are several reasons for presenting the projects through comparisons. First of all, in the spirit of the comparative perspective, systematic comparison has the potential to describe features that would otherwise remain hidden. One prime example of this is that through this comparison it will be emphasised that the monitoring of educational systems through international comparison may be done from different perspectives, leading to different results and possibly different interpretations. Secondly, since this thesis is written in English and consequently with an international target group, presenting PISA by comparison with TIMSS is to take advantage of the fact that TIMSS is better known in some parts of the world. This is especially true for the USA where PISA in general is not widely known (Bybee, 2005). Thirdly, many countries participate or have participated in both studies. Of the 32 countries that took part in PISA 2000, 28 countries also participated in either TIMSS 1995 or TIMSS 1999, or both. This number is even higher for countries participating in PISA 2003.

Furthermore, some 20 countries, including Norway, participated in TIMSS 2003 and PISA 2003. Both these surveys were implemented in the spring of the year 2003, and the national and international reports from both studies in most countries were released in December 2004. This coincidence in time of two studies that, seen from the outside, are very similar may create confusion. This implies that there is a need for systematic comparisons of the two for those who make use of the results from both studies in these countries. To highlight the importance of this, OECD has commissioned a report on the comparison of TIMSS and PISA, intended to be published in the near future. The need to compare these two studies is also recognised in some recent and forthcoming publications addressing mainly the US context (Ferrini-Mundy & Schmidt, 2005; National Center for Education Statistics (NCES), 2005; Neidorf *et al.*, forthcoming-a; Neidorf *et al.*, forthcoming-b; Nohara, 2001)

Tables 2.1 and 2.2 below give a systematic presentation of the main similarities and differences between TIMSS 2003 and PISA 2003. Even if the comparisons for some minor issues are specific for the studies conducted in 2003, the main issues in the comparisons are valid for TIMSS and PISA in general. Furthermore, some relevant issues that vary across test years will also be referred to in the text discussing the condensed and simplified comparison given in these tables. The bullet points in the columns of Tables 2.1 and 2.2 are aligned so that statements referring to the same aspects can be read from left to right. I have no intention in this comparison to rate one as better overall than the other. Instead, my intention is to highlight how the differences may be understood by the different aims and purposes of the two studies. Even if this comparison aims at presenting both TIMSS and PISA, more weight is given to the latter.

2.3.4 Different concepts measured

First of all, what is most important for this thesis is the fact that the concept of scientific literacy in PISA is different from the concept of science in TIMSS in several important ways. This difference deserves an even closer inspection than what is offered in this section. In the following, the major differences presented in Table 2.1 are only briefly summarised.

	Similarities	Differences	
		TIMSS	PISA
Test domain	<ul style="list-style-type: none"> Both test knowledge and skills in mathematics and science. The competency perspective in PISA is quite similar to 'Scientific Inquiry' – a minor dimension in TIMSS. 	<ul style="list-style-type: none"> Framework aims at representing national curricula. Detailed content dimension. More focus on conceptual understanding. Descriptive rationale ('what school science is'). Exhaustive specification in framework. <p style="text-align: center;">⇓</p> <p>'Distance' between framework and items small.</p>	<ul style="list-style-type: none"> Includes also reading literacy (and in 2003 also general problem solving). Curriculum not explicitly treated in framework. Framework describes a few broadly defined competencies. Normative rationale ('what school science should be'). Inclusive specification in framework. <p style="text-align: center;">⇓</p> <p>'Distance' between framework and items large.</p>
Organisation and participation	<ul style="list-style-type: none"> Implemented through national ministries as assignments to research institutes. About 20 countries participated in both studies in 2003. Both projects have world class experts in psychometrics and educational sciences. 	<ul style="list-style-type: none"> Organised and initiated by IEA. Final decisions made by National Project Managers. <p style="text-align: center;">⇓</p> <p>Researchers in charge</p> <ul style="list-style-type: none"> About 50 participating countries (in 2003) representing a rich cultural diversity. 	<ul style="list-style-type: none"> Organised and initiated by OECD. Final decisions made by representatives from the ministries in the participating countries. <p style="text-align: center;">⇓</p> <p>Policy makers in charge</p> <ul style="list-style-type: none"> Increasing number and broader spectrum of countries (in 2000 about 30, in 2006 almost 60).

Table 2.1: Short and simplified description of similarities and differences between PISA and TIMSS regarding what they test and how they are organised.

Primarily what should be noted is that scientific literacy measured in PISA is not based on a curriculum analysis as the concept of science in TIMSS. This means that TIMSS intends to measure a country's achievement in what is commonly taught in school science. PISA has instead taken the challenge of defining what

scientific competencies adolescents need in their present and future life: as autonomous individuals dealing with challenges in their own lives, as citizens in a democratic society, and as professionals in a skilled work force. In other words, the lifelong learning perspective declared to be central by the OECD ministers has directed PISA towards measuring how successfully the students leaving lower secondary education can continue to learn throughout life. It is assumed that the school is a major source of influence for fostering the competencies needed, but also life outside school has a considerable impact. Fortunately, major parts of the testing material are still devoted to topics commonly covered in the science taught in the schools of many countries. This implies that the science courses in most of the participating countries to some degree aim at fostering scientific literacy as defined through the consensus process of PISA. However, this varies from country to country.

The difference is that TIMSS intends, as far as possible, to be a fair measurement between countries in the way that the measurement should reflect the intended curriculum equally fairly in all countries. TIMSS is therefore based on a descriptive rationale ('what school science is'), while PISA is based on a normative rationale ('what school science should be'). These differences are justified by the different aims and purposes of the studies, which will be returned to shortly.

The practical consequences of these different approaches is clearly seen in the framework specifications and in the actual items included in the tests. The TIMSS framework (Mullis *et al.*, 2001) is much finer grained than the PISA framework (OECD-PISA, 2003b). Some of the statements in the TIMSS framework are so precisely formulated that in the actual item writing one could almost rephrase the formulation into a question. PISA on the other hand puts a major focus on wider descriptions of broader categories, giving examples of concepts or scientific issues and specific items as illustrations of the principles in the framework. Furthermore, the main principle guiding item writing in PISA has been to find tasks and contexts that are seen as relevant to the broader issue of 'preparedness for life'. As a consequence items are framed by a relatively extended piece of stimulus material, followed by typically 2–4 items relating to this material. The cluster of items referring to the same material is referred to as 'units' in PISA. TIMSS on the other hand mostly consists of isolated items with no or very little stimulus material¹⁷.

2.3.5 Different international frame of reference

Another aspect that may be regarded by some (for instance politicians or journalists in the media) as a similarity between the two studies, is the fact that they are international measurements where each nation or educational system is ranked or otherwise compared to an international context. It is obvious, however, that the international context is quite different in the two studies. Even if many countries have participated in both studies, the PISA study (at least the PISA

¹⁷ The design of the items allocated to the minor dimension labelled 'Scientific Inquiry' in TIMSS 2003 are more similar to the PISA units. These items are given in small clusters relating to the same extended stimulus material.

2000 study) mainly included economically developed countries. TIMSS on the other hand, especially TIMSS 2003, to a much larger degree includes African countries, East and Central European countries and some economically less developed Asian and South American countries. However, the composition of countries is gradually changing in PISA as more and more countries are included (in 2006 the number of countries is nearly double the number in 2000)¹⁸. To consider the countries that are actually participating is important when comparing the two studies.

The standardised achievement scores for the students refer to the international sample in different ways. In TIMSS all nations or systems participating are included in developing the standardised student scores, and consequently the average scores for the countries are affected by the sample of countries included. The OECD has adopted another system in PISA where the standardised scores refer to the OECD countries only. This is especially relevant in light of the cyclic design aimed at measuring trends in the participating countries. The reporting of trends is made possible in both studies through the use of a set of link items that are common in consecutive studies¹⁹. Therefore, independent of the international context the achievement scores from two consecutive TIMSS or PISA studies can be used in the national context (e.g. comparing Norwegian students' achievement in TIMSS 1995 with that of Norwegian students in TIMSS 2003). However, the fact that in TIMSS the pool of countries varies considerably between the test years introduces some problems concerning how to interpret the national data against the international comparative background. In PISA, even if the number of countries participating increases, the core of OECD countries is always present.

The lists of results ranking countries, often referred to in a rhetorically condescending way as 'league tables' (Robinson, 1999) or as an international 'horse race' (Brown, 1998), are therefore very difficult to interpret, and especially so in TIMSS. Even if the two studies do have ways of linking and calibrating the scores that in principle ensure that the measures are independent on the countries actually participating, the likelihood of there being interpretational flaws, especially in statements trying to compare or summarise findings from both studies, should not be neglected. In particular this means that since the countries actually participating vary from test to test, presenting a country's achievement by its rank not only seriously reduces the information available, but is also misleading and largely meaningless, especially when the aim is to compare achievements for two different tests. This is equally true for both PISA and TIMSS. The main difference between the two studies is that presenting a country's achievement by its score relative to the reported international average is a more stable and meaningful parameter in PISA than in

¹⁸ The OECD-PISA web page claims that 43 countries participated in PISA 2000 (retrieved from www.pisa.oecd.org May 26th 2005). In fact only 32 countries participated in 2000 (OECD-PISA, 2001). However, 11 more countries took part in an additional study in 2002 (OECD & UNESCO, 2003), and these countries have been classified as 'countries participating in 2000'.

¹⁹ In order to link measurements something has to be in common for the measurements to be linked. In PISA and TIMSS some items are not published because they are used in the next study. These items can therefore be used to link the scores from one study to the next one.

TIMSS. That is, in PISA one can always have statements starting with for instance “relative to the OECD average, the Norwegian students...”. Similar statements are impossible to have in TIMSS since the international average refers to an international context that is not easily labelled, and since the composition of participating countries varies from one test year to the next.

2.3.6 Different organisations behind the studies

It is reasonable to expect that a study is influenced by the people and/or organisations that initiate, define, report, and pay for the study, and it is especially relevant to suggest that both the IEA and the OECD are parts of a larger political and/or ideological context. I have already discussed in section 2.2 how the OECD perspective on human capital and lifelong learning is probably one of the main reasons for the PISA orientation towards measuring wider literacy concepts. It is therefore very likely that the different agendas of the IEA and the OECD have directly led to the different operationalisations of, for instance, scientific competence in the two studies. This issue will be referred to again when discussing the policy relevance of LINCAS in section 2.4 below.

TIMSS and PISA differ in how the policy makers are involved in the projects. The simplified version of this is that PISA is more policy driven (Purpose II), while TIMSS is more research driven (Purpose I). However, this is a generalisation that rightfully may be criticised: the policy maker is also very much engaged in TIMSS, and the documents from TIMSS address policy-relevant issues. And vice versa; a lot of trained researchers are involved in PISA, and the reports that are issued from PISA may be regarded as typical research reports. Nevertheless, as will be returned to later, the policy makers are more actively involved in PISA than in TIMSS.

Another issue that may be considered as an ideological issue relating to the two studies is the composition of participating countries, and especially the issue of whether some countries dominate the projects – which countries participate, which countries are more involved in the development and where the finances come from.

Both PISA and TIMSS are based on a financial system where the participating countries pay a fee to cover the international expenses, in addition to carrying costs at the national level. In TIMSS a disproportionate amount of the international costs are covered by the USA. Some US institutions are major donors to the IEA²⁰. The World Bank is another important agent in the IEA studies. Its role is mainly to pay the costs for some of the less affluent countries that participate²¹. Although the World Bank is an international organisation, it is also to some extent dominated by the USA since this country is the largest shareholder in the bank. Furthermore, two of the institutions in the consortium in

²⁰ The minutes from the General Assembly 2004 (available from <http://www.iea.nl/iea/hq/>) identify the following institutions as major donors to the IEA: the Ford Foundation, National Science Foundation (NSF), the Department of Education's National Center for Education Statistics (NCES), and the United Nations Development Program (UNDP).

²¹ The World Bank covered the international costs for 21 of the countries participating in TIMSS 2003 (Gilmore, 2005).

charge of the international part of the study are located in the US.²² It is therefore reasonable to believe that the US not only covered much of the costs of TIMSS, but that this country also had a larger influence on TIMSS than any other participating country.

However, it is not easy to identify any particular effects of this dominance, which some researchers (e.g. Orpwood, 2000) have indicated to be an important underlying characteristic of TIMSS, possibly introducing bias in the measurements (Keitel & Kilpatrick, 1999; Linn, 2000). It has, for instance, been observed that none of the released items from TIMSS are related to sex or evolution (Sjøberg, 2005a, 2005b). However, it is evident from the framework that both of these topics are covered to some degree. Furthermore, there is a widespread belief that the introduction of multiple choice items, that are unfamiliar in many countries, may produce a country bias in favour of, for instance, the USA, where students are used to the format. However, inspecting the data, it is not possible to find any clear evidence for this claim. Lie *et al.* (1997) found that in TIMSS 1995 students in the USA were favoured by multiple choice items in science as compared to both the international average, and as compared to, for instance, Norwegian students. However, reviews of the mathematics items in the same study did not show this effect. The same has been demonstrated for mathematics items in PISA 2003 (Kjærnsli et al., 2004). Furthermore, I have analysed the science items in PISA 2003 through a cluster analysis in paper III of this thesis, and this also shows that format could not explain the relative strengths and weaknesses of the English-speaking countries.

Another ideological concern regarding the issue of who is involved in the studies relates specifically to PISA and the underlying values of the OECD. It has been claimed that the aim of the OECD is primarily to strengthen the economies of the member countries in the global competition of markets for products and services (Sjøberg, 2005b; Uljens in Mathiasson, 2005). Article 1 of the Convention (which is found as a cover sheet in most OECD publications) describes the purpose of the OECD with particular reference to economic growth. However, the message is more nuanced and has more facets than what is implied when only referring to the issue of global competition. It refers to economic development in both member and non-member countries. Furthermore it refers to 'standards of living', 'sound economic growth', and 'sustainable economic growth'. Also, by browsing the activities and studies conducted by the OECD (see for instance www.oecd.org) it is evident that this organisation's interest reaches much further than mere economic expansion of its member countries. All aspects of social conditions (e.g. conditions for democracy, standards of living, health, the environment and education) are studied by the OECD, and not only in their member countries. It may be countered that such social indicators are studied because they are thought to be predictors of economic success (Neuman, 2003). Nevertheless, the OECD promotes such

²² The International Study Center, located at Boston College, runs the project and is also in charge of writing the international reports. Another US member in the consortium in charge of the study is the Educational Testing Service (ETS).

issues, and therefore, their ideological programme does not just promote economical success and liberal ideals for trade.

In summary, it may be stated that it is reasonable to expect that the organisations in charge (at different levels) and those who pay to some extent have an impact on aspects of studies like TIMSS and PISA. Indeed, it would be naïve to think that the wider context of these organisations would not influence the studies that they supervise. However, it seems to be difficult to (a) describe a very clear and coherent ideological perspective framing these organisations, and (b) to document how *particular* ideological factors (if such factors might be identified) have influenced *specific* aspects of the studies.

The study and analyses of politics and documents formulating policy and political intentions is beyond my expertise, which should be clear from the rudimentary commentary given in this section. The intention of this section has only been to present some observations indicating that very simple statements regarding the ideological basis of these organisations are prone to present only one side of the coin. This is not to say that we should not consider or discuss the ideology of LINCAS. There are many ideological issues regarding LINCAS that should be analysed and discussed, of which only a few are briefly mentioned in this thesis.

2.3.7 Different sampling designs

Table 2.2 summarises the main similarities and differences in methodology applied by TIMSS and PISA, some of which will be discussed in this section. TIMSS has chosen a grade-based sample, and furthermore they have included several populations (in 1995 there were three populations, and in 2003 there were two populations). For instance, population 2 is formally defined as the upper of the two adjacent grades with the most 13-year-olds at a specified date (Mullis *et al.*, 2001). In practice, however, given the variation in the average ages across the participating countries, it seems that the real definition differs somewhat from the formal definition. It seems that many countries have sampled a grade with a majority of 14-year-olds instead. The *de facto* definition of population 2 for TIMSS seems to be the 8th grade (see for instance the columns of ages and grades given in the table on p. 36 in Martin *et al.*, 2004b). As a consequence, age varies both within and across countries in TIMSS.

PISA on the other hand, has an age-based sample, defined in 2003 as those who were born in 1987 (OECD-PISA, 2003b). This means that PISA has older students, and the sample is more homogeneous in terms of age, but varies more in terms of grade level or years in school²³. Furthermore, both PISA and TIMSS have schools as their primary sampling unit, which means that the first step in the sampling is to select schools by drawing them from a list or database of all the schools in the country²⁴. PISA continues by drawing a random sample of students from the selected school, while TIMSS draws one or several classes from the

²³ This is conceived as problematic by the OECD. Several countries will from PISA 2006 also include a grade-based sample in order to study the age/grade context in some more detail.

²⁴ The sampling design for the selection of schools is quite complex. The sample is stratified and the schools are drawn with a probability proportional to the number of students at the school.

schools. Although this may seem a small difference involving only some technicalities regarding differences in the hierarchical levels of the data, it is clear when studying other aspects of the design that the different sampling designs reflect one of the most marked differences between the purposes of PISA and TIMSS.

	Similarities	Differences	
		TIMSS	PISA
Population and samples	<ul style="list-style-type: none"> Clearly defined populations and large, high quality samples (4000 – 5000 students in each country) Strict procedures for exclusion, defined minimum criteria for participation rates etc. 	<ul style="list-style-type: none"> Several populations (in 2003 two populations): The upper of the two adjacent grades with the most 9-year-olds (population 1) and the upper of the two adjacent grades with the most 13-year-olds (population 2) at the time of testing \Rightarrow age varies within and across countries. Class-based samples. 	<ul style="list-style-type: none"> One population: Students born in a particular year; the participating students are 16 years during the test-year \Rightarrow age relatively comparable, but grades vary within and across countries. School-based samples.
Design and instruments	<ul style="list-style-type: none"> Paper-and-pencil tests. Both constructed and selected response. Both include a student and school questionnaire. To some degree they have similar variables, e.g. demographic variables. 	<ul style="list-style-type: none"> Items more 'school-like'. Short or no stimulus material for most items. Includes a teacher questionnaire. Same instruments in all countries. Mainly reporting on single variables from questionnaire. A major focus on the class as a unit of analysis; questions related to instruction or teaching. 4-year cycle with 50/50 test time devoted to mathematics and science. 	<ul style="list-style-type: none"> Items organized in small clusters ('units') referring to a common stimulus material. The main instruments common in all countries, but some smaller instruments are optional. Mainly reporting on composites constructed from questionnaire. A major focus on the school context, socio-economic status, and learning strategies in student questionnaire. 3 years cycle. Major domain (2/3 of test time) circulates between reading, scientific and mathematical literacy.

Table 2.2: Short and simplified description of similarities and differences between PISA and TIMSS regarding sampling and design.

In general, the IEA studies, and in particular the TIMSS studies, have been framed within an elaborate model of curriculum at different levels in the educational system (Bos, 2002; Robitaille & Maxwell, 1996). This model goes from measurements of factors reflecting the intended curriculum (stated in the policy documents describing the school subject), via the implemented curriculum (measured by what is included in textbooks and what is taught in class) to the attained curriculum (measured by students' achievements on the cognitive tests). In this model the class unit is central in order to meaningfully represent the implemented curriculum. As a consequence TIMSS has sampled classes and they have accordingly also included questionnaires to teachers. Furthermore, as a consequence there are more items in the student questionnaire that intend to measure aspects of the teaching. In TIMSS 1995 the intended curriculum and the

implemented curriculum levels were even more explicitly targeted by the design through the extended curriculum analyses that were conducted (W. H. Schmidt et al., 2001; W. H. Schmidt et al., 1997a; W. H. Schmidt et al., 1997b). Furthermore, in-depth studies based on observations in classrooms were implemented in some of the participating countries (W. H. Schmidt *et al.*, 1996; Stigler & Hiebert, 1999). PISA on the other hand is school based, and consequently, the selection of background variables in the questionnaires reflects this perspective.

This has direct consequences for the types of inferences made from the two studies. Both studies can be used to make statements about the population of students, that is, the lowest level in the hierarchical design of the sample. However, in a multilevel modelling approach PISA targets issues related to schools, while TIMSS to a larger extent targets factors related to the classroom. Also, in principle, TIMSS can be used to target the school level since the classes are sampled from the schools. However, in general the differences in sampling designs make it impossible to compare this parameter in the two studies.

2.3.8 Complementary or incompatible?

From the sections above three arguments have been highlighted as to why it is difficult to, strictly speaking, directly compare results from TIMSS and PISA:

- I. PISA and TIMSS cannot be directly compared because the measured concepts of scientific competence, or achievement, are not the same in the two studies.
- II. PISA and TIMSS cannot be directly compared because the international context used as a reference is different in the two studies.
- III. PISA and TIMSS cannot be directly compared because they have different target populations and different sampling designs.

Therefore, it is important for those who read the results from surveys like TIMSS and PISA as evaluations of aspects of the respective countries' school systems, to be aware of how these two surveys differ. The main difference between the studies is that TIMSS sets out to communicate how successful the school system has been in implementing the policy defined by the science and mathematics curriculum, and furthermore to describe how the curricular intentions are mediated in the system through textbooks and the classroom context. PISA, on the other hand, may be used to evaluate how successful the curriculum and school system has been in fostering the competencies judged, by consensus among both experts and other representatives of the participating countries, to be central for citizens to master.

In December 2004 both projects published their main reports at about the same time, and in the newspapers and other public media the results were often referred to, in general terms, as international studies of maths or science performance, without any discussions of the differences between the studies. In countries participating in both studies this could lead to confusion if the two studies reported discrepant results for the same country. To a certain degree this happened for some countries in 2001, when the results from PISA 2000 were

published. UK, for instance, apparently scored relatively better in PISA 2000 than in TIMSS 1999, a matter commented on with concern by Prais (2003). He saw this discrepancy as peculiar since population 2 in TIMSS 1999 covered approximately the same cohort as PISA 2000 except that they were one year older when they participated in PISA. An argument with a similar logic has been raised by McGuinness (2004) regarding the reading results in PISA 2000. Both used the discrepancy between PISA and other international assessments to cast doubt on the validity of the interpretations made in PISA. Their arguments were in general not supported by evidence, and in his response Adams (2003) countered Prais's arguments mainly by referring to how the measures obtained in PISA could not be directly compared to those obtained in TIMSS in any simplistic manner, as is also obvious from the comparison given above. This specific example is backed up by personal experience in the way that results from TIMSS and PISA are talked about and discussed among academics, in the media or by policy makers. All in all, this highlights the importance of discussing the major differences of the two studies.

Even if the discussion so far might suggest the general conclusion that TIMSS and PISA are incompatible, it is still evident that the countries participating in both studies have a richer database from which to discuss school issues, and furthermore, I will claim that findings from both projects can be combined in fruitful ways if these differences are directly applied as an analytical resource. For instance, the two largely different perspectives on science as a subject in school provide a resource that can be used in analyses aiming to inform the issue about what kind of competency we would like our schools to foster.

Nevertheless, the issue of how to link TIMSS and PISA needs to be addressed. Allerup & Mejding (2003) have described an effort that was made in Denmark to link reading literacy in PISA 2000 with the assessment of reading literacy in an earlier IEA study conducted in 1991 (Elley, 1992). This was possible since an extra booklet with items from both studies was included in the Danish PISA survey. One lesson learned from this would be to include a similar design in countries participating in both TIMSS and PISA. Since the governments in these countries pay for the studies, they should be interested in making a link between the projects, and they should furthermore seek to find solutions for the problem of how this issue could be included as a regular part of the studies. This would necessitate some kind of agreement between the OECD and the IEA about linking future assessments. As already indicated, the governments have to some extent already signalled such a need by deciding that a thematic report on the comparison of PISA and TIMSS should be developed.

Some of the similarities and differences between TIMSS and PISA highlighted in Tables 2.1 and 2.2 will also be returned to in the following discussion. Please note once again that the intention of the comparison presented above has not been to rate one of the surveys as generally superior to the other. The overall message is that designing assessments for different purposes leads to different assessments. And more importantly, these differences highlight the different types of inferences that may validly be made from the two studies. It is,

for instance, reasonable to say that since PISA did not consider curriculum in the participating countries explicitly in the design, this assessment cannot be used to make all types of judgements about the merits of the educational system in a country. In a system, for instance, with a curriculum that emphasises the understanding of concepts and procedures that are fundamental in science, and with less focus on the application of science in authentic contexts, PISA cannot be used to monitor the degree to which students reach the standards defined in that country. In such a country the likelihood is that TIMSS provides data that are more aligned with the national standards. However, the description given by the international comparisons in PISA can, when put into the national normative context, be used to evaluate whether the curriculum in science adequately prepares young people for the future. This is, to a high degree, an issue where both science educators and the policy makers as stakeholders should have a mutual interest and much common ground.

2.4 LINCAS, policy makers and researchers in science education

Science education has many stakeholders. I will now turn to the main issue of the chapter: the link between two of the groups of people with a shared interest in education – researchers and policy makers. In particular, this section aims at describing how secondary analyses related to large-scale international comparative achievement studies in science may be a meeting place for researchers in science education and policy makers.

2.4.1 Three observations framing the discussion

Educational research (Tomlinson & Tuijnman, 1994; Tooley & Darby, 1998) including science education research (Jenkins, 2000; Millar, 2003), has to some extent not been considered as relevant for policy makers. It is also difficult to find evidence that the research in science education has had any lasting impact on practice leading to changes and improvements in how science is taught and learned in school (Lijnse, 2000). Of course this is a gross generalisation and should be moderated by the fact that many research projects have been done with designs implementing teaching, and these projects have, no doubt, lead to innovations in local curricula and the way science has been taught and learned in many schools. However, it is not easy to see how this typically small-scale research has been disseminated to a wider audience, leading to systemic change at a larger scale. Furthermore, it is not easy to see *how* this type of practice-oriented research could be disseminated to a wider audience in a way that allows for a generalisation of the findings that in most cases are constrained by specific and local factors such as the availability of resources (e.g. time or laboratory equipment).

This leads me to the first of three observations framing the forthcoming discussion about the link between research and the policy makers: all in all, it seems that science education research has been evaluated by many as not very policy relevant, although the community of science education researchers feel

that their research should be regarded as a central contribution to the development of national and regional policy.

On the other hand, large-scale international surveys such as TIMSS and PISA have received a lot of attention from policy makers. Results and analyses from these projects have been communicated, to a varying degree in the different countries, to decision makers, to teachers and to the wider audience, mainly through reports and oral presentations and discussions in a wide variety of arenas. The continued support of LINCAS, and the extensive use of data from LINCAS in national and international policy documents is the basis for the second simple observation relevant for my discussion about the link between researchers and the policy makers: LINCAS are regarded as very relevant by policy makers.

When LINCAS, such as TIMSS or PISA, are talked about among researchers in the field of science education (or education in general), they are frequently mentioned in a context of criticism (in the negative sense of the term). I have no intension to criticise other researchers for being critical. On the contrary, when I summarise the reasons why science educators should be engaged in secondary analyses of data from LINCAS later in this chapter, one of the arguments I focus upon is that these studies are so immensely influential at the policy level that constructive criticism is indeed needed. Nevertheless, it strikes me as odd that many researchers are not willing to engage more positively regarding findings from LINCAS, findings that in my opinion may be considered as relevant also for the community of researchers in science education. This may be related to the more fundamental issue raised by Black (2000) that researchers in science education pay too little attention to assessment in general²⁵.

Another related aspect is that although results from LINCAS are reported through many channels, dissemination targeted at the science education research community, through research conferences or journals has not been equally common. Some papers have been presented at science education conferences (e.g. Kjærnsli, 2003; Turmo, 2004), a few articles have been published (e.g. Harlen, 2001; Kjærnsli & Lie, 2004; Olsen *et al.*, 2001 (paper I in this thesis); E. C. Papanastasiou, 2003; Yip *et al.*, 2004) and books have been published with secondary analyses of data from TIMSS (Howie & Plomp, in press; C. Papanastasiou, 2004; Robitaille & Beaton, 2002; Shorrocks-Taylor & Jenkins, 2000).

My third observation may therefore be summarised as follows: all in all, it may be concluded that LINCAS have been considered as not very relevant in the science education community, although the researchers within these projects feel that their work should be regarded as a central contribution to the knowledge base shared by the science educational community.

The first and third observation presented above may be seen as very similar statements: the two groups, science education researchers and researchers in large-scale international comparative studies, have a joint communication

²⁵ Just to put things straight: this is not to say that Black supports LINCAS. Rather the opposite is true, as is evident in a paper he co-authored with Atkin arguing that TIMSS is misused by policy makers (Atkin & Black, 1997).

challenge. They both lack arenas to disseminate and discuss the significance of their research to groups outside their own community, and both groups have a common conviction that their research should be of interest to others. In the following, the fact that LINCAS are regarded as relevant by policy makers is applied in an attempt to suggest a solution to this challenge: LINCAS provide a context for communication between researchers from many fields and policy makers. In the following I will argue that LINCAS give opportunities for secondary research in science education. Through such research it could in principle be possible to integrate other research in science education with LINCAS. With any luck, one possible consequence might be that research in science education may be mediated into the process of policy making, as part of the information base for policy makers.

2.4.2 A shift in how research is financed and organised

From the above there should be no doubt that both PISA and TIMSS are regarded as policy-relevant studies. Furthermore, they are both very much in the public eye. I will suggest that these characteristics to some degree reflect important changes that have taken place for research in general during the last decades.

In the last decades we have experienced big and influential changes in the way research in general is funded and organised. These changes have been driven by both new knowledge of what characterises research as a socially embedded human activity and changes in research policy; however, they also derive from new problems arising in the disciplines themselves, pointing towards the need for large-scale multidisciplinary research projects (e.g. biotechnology). Instead of being publicly funded and with full autonomy, researchers are increasingly confronted with the reality that they have to apply for funding. In the guidelines formulated by the agencies issuing research grants, an increasing focus is given to the possible social implications of the research, and more of the government funding is allocated into larger research programmes with a defined agenda.

This has to some extent made the distinction between academic basic research, typically organised through universities, and applied research, often organised through research centres outside traditional academic institutions, less clear. Gibbons *et al.* (1994) have suggested that research has shifted modes, from 'Mode 1' (old) to 'Mode 2' (new). Ziman (1996) also talks about a qualitative shift in research in science, and labels the new as 'post academic science'. Funtowicz *et al.* (1990) use the label 'post-normal science' to describe the same trend. Although all of these labels are primarily derived from studies of how research in the natural sciences are governed and conducted, I will claim that these concepts are also relevant for describing research in the social sciences.

One of the characteristics of Mode 2 research is the way in which it is organised. The research groups typically consist of scholars with different backgrounds, and they are administered as temporary projects or networks working relatively independently of the larger structural frames (e.g. the university structure). Another element in this heterogeneous and ad hoc structure is the fact that the user is often also part of the project (Nowotny *et al.*, 2001).

This ensures that the original call for tender is followed up in the research activities. The positive evaluation of this shift in how research is organised would be that research has become more socially responsive and responsible. The negative view is that researchers may have problems in keeping their integrity, and furthermore, this could cause a shift in priorities which then do not reflect the internal priorities of the research community.

To some extent both PISA and TIMSS have some Mode 2 characteristics. The studies are divided into different projects involving researchers from different academic fields such as educational science, psychology, psychometrics, and sociology. Many of the organisations and persons involved got their job by bidding on a call for tender specifying the task. And most importantly, the key decisions are made by representatives from the ministries in the participating countries in a dialogue with the researchers involved. This ensures that the researchers working on the project and the users who have specified the assignment have a shared understanding of the joint task.

2.4.3 The role, function and position of the policy makers in PISA and TIMSS

Both PISA and TIMSS are therefore put into a policy context largely by the way they are organised and financed. Both are administered and implemented on a daily basis by a consortium of research organisations, with one dominant organisation²⁶. Furthermore, in both studies some expert groups (subject matter groups, questionnaire groups and technical advisory group) have central roles as consultants to the consortia. In TIMSS the person in charge at the national level is referred to as the National Research Coordinator (NRC) and in PISA the person with this function is referred to as the National Project Manager (NPM). In both studies there are frequent meetings between the consortia and those representing the national projects. Many of the key decisions are informed to a large extent by the feedback from the people in charge at the national level.

On a daily basis the OECD monitors PISA by a secretariat. However, the secretariat functions as a mediator of the decisions made by the PISA Governing Board (PGB)²⁷. This board consists of representatives from the central policy level in the participating countries, and the board meets regularly to monitor the progress of PISA, make key decisions, and otherwise instruct those in charge of implementing the studies. The frameworks for PISA are, for instance, mainly developed by groups of experts, but in the process leading to a final document there is ongoing negotiation with the PGB. All in all, it is to be expected that this ensures that PISA is aligned with the overall rationale for the OECD engagement in education. I have already suggested that the wider scope of the OECD, where education to a large extent is connected to concepts like ‘lifelong learning’ and ‘human capital’, has affected the content and shape of PISA, and this is one likely reason for some of the major differences between PISA and TIMSS.

²⁶ The dominant organisations in the PISA and TIMSS consortia are the Australian Council of Educational research (ACER) and TIMSS & PIRLS International Study Center at Boston College, respectively.

²⁷ Formerly known as the Board of Participating Countries (BPC).

The IEA also has a structure where the decisions are formally taken by their members. The members in the IEA are various institutions representing countries²⁸. There are mainly two types of member institutions; universities (or departments/institutes at universities) and ministries of education (or other institutions with authority at a national level). Representatives from these institutions gather once a year at the General Assembly. However, since the frequency of these assemblies is lower than the frequency with which the PGB has meetings, the General Assembly may be suspected of having less influence on the IEA studies than the PGB has on PISA. Reviewing minutes from these meetings largely confirms this. The PGB, for instance, formally decides on details in the assessments, such as the distribution of the testing time across categories in the framework. For TIMSS such decisions would probably be taken by the National Research Coordinators. IEA is, however, also dependent on producing surveys or other studies that the General Assembly finds relevant, and thus the Assembly has an impact on the studies that IEA conducts.

2.4.4 LINCAS as a link between policy makers and researchers?

Jenkins (2000) discusses some fundamental differences between educational research and policy making based on Loving & Cobern (2000) and Huberman (1994). In his discussion he suggests that the science education researcher and the policy maker not only have different agendas, they also live within different knowledge systems, and

The knowledge produced within one system and for the one set of purposes cannot normally be readily transferred to another. (Jenkins 2000, p. 18)

Jenkins does not provide a definition of the concept knowledge system and he does not identify more specific aspects of the two knowledge systems claimed to be very different. Furthermore, he does not come up with a solution for how the problem in the above quote may be amended. Nevertheless, his statement referred to above coincides with my own observation and captures in a very good way the experience-based and common sense notion that policy makers and researchers have different criteria for what constitutes relevant knowledge, mainly because the two systems utilise knowledge for different purposes.

Policy makers are, to a much larger extent than the researchers, confronted with decision making. This entails at least two characteristics of the knowledge seen as relevant for decision making. Firstly, decisions are bound by time. The pace of decision making is usually much faster than the timelines for most researchers, who would have to accept, for instance, that an article may take two years to be published after it has been submitted to a publisher. It is therefore likely that, due to the pressure to produce policy in a short time, the knowledge that may be digested and understood without occupying too much time is considered as more relevant by the policy maker. Secondly, knowledge that is likely to be true, rather like evidence that will ‘hold up in court’, is likely to be

²⁸ It would be more precise to label the members as representatives of educational systems since some nations have more than one educational system and also more than one member in the IEA, e.g. Flemish and French Belgium.

more appreciated when confronted with the realities of decision making. This provides a context for understanding that what we as researchers appreciate as ‘thick descriptions’, may be regarded by the policy makers as unnecessary complex, and difficult to digest and transform into policy making. On the other hand, knowledge obtained through PISA, for instance that teachers in the Norwegian classrooms are to a larger degree than in most other countries, confronted with readers at very different levels of proficiency (Lie *et al.*, 2001), may be seen as identifying a major problem that policy should address. Another part of the story is that, in order to identify the possible solution to the problem in this example, more detailed and ‘thicker’ descriptions of classroom processes are needed.

Through PISA, the OECD has established procedures and arenas for the dissemination of research to the policy makers. And vice versa, through the same arenas, the policy makers are able to communicate their needs for information on which to base their decisions. This is at least part of the solution for how it might be possible to get a good transfer of information back and forth between the two knowledge systems²⁹. This means that the overall aim of the PISA study is very much aligned with how policy makers define and justify the subject of science in school. This also means that the cognitive measures are contextualised by variables perceived to be of importance for the policy level. This includes variables that can be directly manipulated by the policy level (e.g. school size), the relationship between achievement and variables that the policy is aimed at strengthening or weakening (e.g. the relationship between achievement and socio-economic status), and variables used to control this information.

It has been argued that international organisations such as the OECD, the European Union, the World Bank and the UN, are major mediating channels for policy changes taking place all around the world (Drori, 2000; Goldstein, 2004a; Kellaghan & Greaney, 2001). This means that, irrespective of what one, as a science educator, may think should be the core of the school subject of science, the PISA definition of scientific literacy and its operationalisation through the cognitive items used are important to be aware of, in order to be informed about the decisions being taken regarding policy and curriculum in many countries in the years to come. In fact, the PISA assessment might be seen as one of the most influential operational definitions of scientific literacy (Millar, 2003). I would suggest that since PISA will influence decision making regarding the curriculum and teaching of science in many countries, researchers in science education should be highly motivated to participate, if joint arenas are set up by the national authorities to discuss the possible impact of the results in PISA. On the other hand, since science education research may be used to obtain rich and contextualised information about how learning processes actually may be implemented in science classrooms, the OECD should be motivated to engage science educators in research that would target possible ways of addressing

²⁹ There are other more direct ways to communicate research issues and findings to policy makers. Some researchers may for a period hold positions in a government institution where the policy is actually made. Some science educators are from time to time called upon by policy makers to give an account of the state of the art, or to participate in writing a green paper (or a white paper) with direct consequences for policy making.

issues from PISA through policy making. Furthermore, there is additional information in the data that is typically not analysed by the primary analysts working on the projects, and this information may inform the policy maker, it may inform science education research, and it may inform the future conduct of LINCAS in science.

All in all, PISA can be characterised as commissioned research with a high user influence. It is initiated and administered through the OECD, one of the most influential international organisations on educational policy, at least in the industrialised world. It has been developed and implemented by consensus among experts and representatives from the ministries in the participating countries. At the national level the project is organised in a great number of ways: in some countries the people involved work in government institutions directly under the ministry (e.g. in the USA, where the National Center for Education Statistics is responsible, with Westat as a major partner), while in others it is implemented, analysed and reported on by researchers in academic institutions (e.g. in Norway, where the people working with PISA are university-based researchers). It is to be expected that whoever is involved at the national level will influence how the project is presented, which analyses are done and what kind of publications are produced. Since the national steering of the study in many countries is done by national government agencies, it cannot in general be anticipated that the PISA data are utilised to their full potential to target research questions directly related to the science educational component.

Hopefully, the discussion above has provided adequate ground for my suggestion that science educators throughout the world, either those involved in the project or researchers independent of the project, should be motivated to be informed about and engage in the national debates about how to make use of the results reported by both PISA or TIMSS (even though the above discussion has been explicitly related to PISA, much of it is equally relevant for TIMSS). Moreover, as I will now turn to, I will also suggest that these studies provide valuable resources for researchers, and researchers in science education should also be motivated to explore the possibilities of using the data or documents from these studies in their own research.

2.5 Exploring the possibilities for secondary research

Large-scale international comparative studies in education, such as TIMSS and PISA, have a primary agenda. This is summarised by Schleicher (2000) in the three outcomes that PISA is designed to produce:

- *A basic profile of knowledge and skills among students at the end of compulsory schooling.*
- *Contextual indicators relating results to student and school characteristics.*
- *Trend indicators showing how results change over time.* (p. 65)

It is important to supplement this primary agenda by secondary research. This may range from theoretical contributions to secondary analysis of the data. A number of slightly different definitions of the term secondary analysis have been suggested in some of the literature on research designs in the social sciences.

They usually focus on the fact that secondary analyses are analyses of already existing data, conducted by researchers other than those who originally collected the data, and with a purpose that most likely was not included in the original design leading to the data collection (Bryman, 2004; Burton, 2000; Dale *et al.*, 1988; Gorard, 2003; Heaton, 1998; Kiecolt & Nathan, 1985; Neuman, 2003; Pole & Lampard, 2002; Reeve & Walberg, 1994; Rew *et al.*, 2000). The definition that is best suited for the discussion presented below is probably the one suggested by Bryman (2004):

Secondary analysis is the analysis of data by researchers who will probably not have been involved in the collection of those data for purposes that in all likelihood were not envisaged by those responsible for the data collection. (p. 201)

This definition also opens up the possibility that the original researchers may be involved in secondary analysis, and furthermore that the purpose of the secondary analysis may have been included in the original research design. The latter point is highly relevant for many of the large-scale official surveys of different aspect of social life, many of which may be considered as having multiple purposes (Burton, 2000; The BMS, 1994), and where the potential for secondary analysis by social scientists is an important part of the design. One example of this is the efforts made to make the data available for other researchers, with comprehensive documentation (see also section 2.5.1).

My definition of secondary research or secondary analysis is in keeping with the open-ended and probabilistic nature of Bryman's definition given above. However, I find it necessary to write 'data and documents' instead of only 'data'. I consider the term 'data' to be too narrow to cover all the types of secondary research that, as I suggest in the following discussion, researchers in science education should have an interest in pursuing. Or, to put it even more simply, for the purpose of my argument secondary analyses can be seen as all analyses that go beyond those typically reported in the international and national reports from LINCAS. This simple statement would also include analyses of resources other than the data: analyses of instruments and items, analyses of the theoretical framework and rationale underlying the studies, analyses of the consequences LINCAS have (or the consequences they should or should not have) for educational policy. However, since the three papers I have enclosed in this thesis relate to the secondary analysis of *data* from TIMSS and PISA, I will, in the following, emphasise more heavily secondary research based on the data from LINCAS.

In the remaining part of this section some more specific arguments as to why data (and documents) from LINCAS should be the subject of secondary research will be suggested. Some of these arguments will be very general arguments that are equally relevant for all kinds of secondary analyses, while others are more specifically related to the secondary analyses of the cognitive database in science in TIMSS and PISA. Parts of these arguments are to some degree repetitions of prior statements in this chapter and in chapter 1. Even if some secondary research issues are suggested, I do not intend to present an exhaustive list of possible secondary research questions. My only hope is to provide a rational basis for why secondary research relating to TIMSS and PISA

should be of interest to researchers in education in general and for researchers in science education in particular.

There are a number of perfectly sound reasons for why many researchers give priority to collecting their own data instead of analysing already collected data. The primary reason, as already examined in chapter 1, is that ‘the scientific approach’ to some extent may be pragmatically defined by a methodology starting with the posing of research questions and hypotheses. Data collected by others are collected with other specific questions in mind, and it may therefore be impossible, or at least very difficult, to use these data to analyse other issues. Secondly, there are often many technical obstacles in using data collected by others: they might not be publicly available; they may lack the documentation necessary to understand the data, e.g. a comprehensive codebook; or the data may require technical skills beyond those of most researchers. Thirdly, there may be ideological reasons for not wanting to base research on data collected by national or international organisations that are primarily collected for policy analyses. Some of these issues are also conditions that limit the potential for using data from LINCAS in secondary research.

2.5.1 Arguments for secondary analysis of LINCAS

It is my view that the benefits of secondary analysis relating to LINCAS outweigh the limiting conditions, and some of these arguments are briefly presented below. These arguments are collected under five different headings, but some of them consist of partly overlapping arguments. Furthermore, these arguments will also be supported in section 2.5.2, discussing some generic designs for how data from LINCAS may be used in research in science education. In addition to giving arguments I will refer to some examples of research that to some degree are justified by that argument.

The high quality of the data

There are a number of reasons for concluding that the datasets coming out of PISA or TIMSS have qualities that are not often seen in educational research. The primary reason for this claim is that the quality is documented. In the technical report for the PISA 2000 survey (Adams & Wu, 2002), all the procedures for the instrument development, sampling, marking and data adjudication are described in detail, and similar technical documentation of TIMSS is available (e.g. Martin & Kelly, 1997a, 1997b; Martin *et al.*, 2004a). By studying such reports, it is clear that the studies are based on:

- a very clearly defined population and adequate routines to sample this population in all countries;
- well-developed frameworks and instruments, including documentation of the quality of the translation into the different languages;
- well-developed and controlled routines for ensuring that the administration of the test was equal in all countries; and
- well-developed routines and quality monitoring of how student responses were scored, and how the data were entered and further processed.

To gather data with procedures like these is not usually possible in ordinary low-cost research.

Cost-efficient use of resources

Millions of dollars or euros have been spent on producing the high quality databases. Samples have been established, the instruments have been distributed to the students and back to the research centres in a way that ensures quality and comparability, and the data have been assembled and restructured through skilful work by experts to further secure the quality of the information available. Nevertheless, relatively little money is used in the analysis of the data. Most of the money has been spent on gathering the data. Evidently it would be good value for money to invest more in further analyses of the data. Relevant to this argument is the fact that the data from both TIMSS and PISA have been made publicly available (although not all the items are publicly available), and researchers interested in using the data can get access to them³⁰. Better even is to engage in a dialogue with the national centre. Through this contact it could be possible to get some advice and access to material that is not so easily available. For instance, many students (at least in Norway) are already using data from TIMSS and PISA as the basis for their Masters thesis, and several doctoral dissertations have been produced based on what could be labelled as secondary analysis of data from TIMSS and PISA (Angell, 1996; Isager, 1996; Kind, 1996; Turmo, 2003a).

The argument about making the most out of the investments made to collect and compile the databases is also an argument for advising governments that participate in LINCAS to make some funding available for secondary analysis, especially analysis related to the national context.

Re-analysis offers new interpretations

In many cases re-analysis is justified because since the time of the primary analysis new developments, either theoretical or methodological, may shed new light on the interpretations offered in the primary analysis.

In general, many of the large-scale data sets that are collected are important resources used by policy makers, and as such these data should be scrutinised from all possible perspectives. Pole & Lampard (2002) have argued that even if official statistics may be very influenced by certain ideologies, secondary analysis of the data can be used to document such a relationship. In this case it is the study itself that is the unit of analysis. This type of analysis would involve not only analyses of the data, but would also typically include analysis of the documents describing the design, the instruments used to collect the data, the reports, and other documents where the data are interpreted into the policy context.

Such arguments are equally relevant for secondary analysis related to LINCAS. The PISA/TIMSS data and documents need informed reviews from

³⁰ For access to the TIMSS data, see more on <http://timss.bc.edu/timss2003.html> and for PISA data, see <http://www.pisa.oecd.org>.

scholars in educational sciences who can frame the data and documents differently, and thus offer new interpretations. This is of particular importance since the policy impact of these studies is possibly much stronger than for most other educational research. I will suggest that this is extremely relevant for PISA. This study is likely to be widely used as a basis for curricular decisions in many countries with the aim of reforming science courses to promote scientific literacy.

In informal settings where I meet with other educational researchers fundamental criticism is often raised against TIMSS and/or PISA. Even if some references are given below to explicit statements of some of this criticism (mostly they relate to TIMSS or only parts of TIMSS), the main point here is that much of it consists of sentiments or expressions of contempt against LINCAS, and most of it is typically uttered in conversations and talks held in less formal settings to an extent that at least is surprising for me. Consequently, the few examples mentioned below are not exhaustive, and the volume of criticism against LINCAS is probably larger than the bibliographical references given below might suggest. Some of this criticism concerns ideological aspects of LINCAS; of these a few for which there exist bibliographical references are summarised below:

- Data from LINCAS are used invalidly by policy makers (Atkin & Black, 1997; Brown, 1998; Keitel & Kilpatrick, 1999);
- LINCAS (and testing in general) are based in a positivist epistemology (Romberg *et al.*, 1990);
- LINCAS may lead to globalisation and homogenisation of curricula (Goldstein, 2004a; Kellaghan & Greaney, 2001); and
- LINCAS are framed within a context that may reflect the agenda of the industrialised countries and the international organisations responsible for the assessments (Goldstein, 2004a; Orpwood, 2000; Reddy, 2005).

Other critical remarks are more specifically related to methodological issues, although several of the bibliographical examples mentioned below may equally reflect criticism that is in nature ideological:

- criticism of the notion that it is possible to develop an instrument that is equal across countries (Blum *et al.*, 2001; Bonnet, 2002; Freudenthal, 1975);
- the limited range of abilities typically tested in LINCAS (Atkin & Black, 1997; Sjøberg, 2005a; Wang, 2001);
- criticism of the use of multiple choice items (deLange, 1997; Harlow & Jones, 2004; Schoultz, 2000);
- criticism of poor sampling procedures or response rates (Bracey, 2000; Brown, 1998; Halliday & Riegelman, 2004; McGuinness, 2004; Prais, 2003); and
- criticism of the methods used to scale the data, and a lack of a longitudinal design (Goldstein, 1995, 2004b, 2004c).

It is valuable to have discussions of such issues, and the contributions made so far are fruitful starting points for academic debate, and as previously stated, the methodological development of LINCAS over the last decades may also be seen

as improvements driven by criticism of the methods used (Porter & Gamoran, 2002). Nevertheless, it is desirable that much of this criticism is further developed since some of the statements made are unclear, are based on a lack of understanding of LINCAS, and rest on assumptions that are not fully accounted for.

This implies that there is a two-way challenge regarding secondary analyses of data from LINCAS. Researchers working with LINCAS data should be encouraged to publish more in academic journals and at conferences since the official publications typically seen from these studies are international and national reports. Although these have been scrutinised by a great number of competent people, they have not undergone the anonymous peer-review process which is one of the most important characteristics of scientific discourse. Such products would also be informative for the rest of the community of researchers. And likewise, the criticism against LINCAS should be made available in a format that is based less on rhetoric and anecdotal evidence (although I appreciate rhetorical skills and texts with anecdotes, and value such elements of a text as important) and more on data and/or a comprehensive presentation of theoretical arguments. Furthermore, it would be fruitful if thematic issues in journals or edited books are devoted to addressing some of the recurring critical issues (as some of those mentioned above) by letting both sides challenge each other's views.

Utilising the information in the database

Large-scale surveys are expensive to administer. In order not to find at a later stage that some important variables were not included, those responsible for formulating the instruments will often include many more variables than it is possible for them to include in their primary analysis (The BMS, 1994). This implies that in several of the databases from such surveys there is an abundance of information that could be analysed, but not envisaged as a primary analytical focus in the original study.

The argument presented here is highly related to the main purpose for this thesis, as presented in chapter 1. In most analyses of the data from LINCAS, student achievement is represented by overall measures on one or a few scales. These scales are constructed by aggregating data from a large number of items. However, in the database detailed information exists relating to each of the items used to derive the scales – information that to some extent can be viewed as irrelevant for the primary agenda presented above.

The exploration of the item-specific information is, however, methodologically challenging; it may capitalise on chance and it can lead to an overload of information which cannot be reduced to general substantial findings (see also chapter 3 and paper II of this thesis for more detailed discussions). In addition, a major problem is the fact that even if the data are released, only a fraction of the actual science items have been released so far. However, there is every reason to expect that, following the release of the data from PISA 2006, when scientific literacy will be the major domain, a substantial number of items

will be released. Furthermore, a substantial number of items are already publicly available from the TIMSS studies.

The national or regional context

Since the results from the studies are mainly used to inform policy at the national level, it is necessary to have discussions on how the results may be used to evaluate the national school system. In order for LINCAS to provide an even better basis of information for this discussion, it may be necessary to develop a specific national design. This would ensure that one could obtain information seen as vital in the national context. Germany is the prime example of a country with their national extensions to the PISA study. In Germany the participating students respond to extra nationally developed instruments, and the country also has an extended sample in order to cover the educational system in each of the partially autonomous districts (Länder) (Stanat *et al.*, 2002).

Furthermore, it is not evident that the international comparative basis provided by all participating countries is the most informative standard with which to evaluate the national results. There are several reasons to suspect that the data can be even more meaningfully compared to neighbouring countries, or countries of the same kind: neighbouring countries are usually more similar to each other, and therefore more factors are ‘controlled for’ because they are more equal in these countries; the data are clustered so that countries within the same region have more similar profiles across items, which means that from a psychometrical perspective there is less error variance (Wolfe, 1999, see also paper III in this thesis for a more detailed discussion)); and it is easier for neighbouring countries to cooperate, for instance in the production of joint publications with a focus on factors that are considered particularly relevant for the region (e.g. as in Lie *et al.*, 2003; Vári, 1997). It is a matter of opinion whether this should be characterised as secondary research or not, but it is a design for analysing the data that is not implemented in most countries.

2.5.2 Targeting research questions in science education

The main reason why science educators could be motivated to invest their own time and resources on secondary analyses of the PISA or TIMSS data is that they may be used to address research questions of importance. I have presented some analyses in the papers in this thesis, but further examples are referred to in the following. In general, there are several designs that could be applied to secondary analysis of data from LINCAS: designs applying secondary data merely as a background; designs that target more specific research questions than those included in the typical national or international report; designs that target highly specific sub-samples of particular interest; and designs that combine data from several studies.

Using data, results, or interpretations as a background

Secondary analysis of already existing data, results or their interpretations may be included as a somewhat peripheral part of a research design; it may provide the background for generating hypotheses and research questions, or it may

provide data or findings with which to contrast or triangulate other data or findings. This type of research is not thought of as secondary analysis, but it is still one important application of the data from PISA and TIMSS for research in science education.

One example of this type of work, in a Norwegian context, is the research project entitled *PISA+*³¹. The researchers involved in the project will use transcripts of videotapes from classrooms, covering several hours of activities, as their primary data source. Therefore, it is clearly not secondary analysis of data from PISA. But as the title of the research project reflects it is triggered by some of the findings from PISA, in the Norwegian context, that to some degree resonated with findings from another Norwegian study based on classroom observations (Klette *et al.*, 2003). Science is one of the school subjects that will be studied as a part of this project.

Other types of research where the focus is on how phenomena change over time, or how one group of respondents compares to another group, may also use data or findings from LINCAS as a background. In some of these cases LINCAS can provide data that may be used as a baseline for comparison to which the researchers' own data may be related. For such a purpose it would, strictly speaking, be necessary to use identical instruments and similar routines for collecting and processing the data. One example related to LINCAS is the use of items from TIMSS 1995 in an evaluation of the science achievement in Norway before and after the curriculum reform in 1997 (Almendingen *et al.*, 2003). In such a design the data from LINCAS are still not explicitly the target for the analyses, but rather used as a background, and as such this type of study would probably not be thought of as secondary research related to LINCAS. However, as this example illustrates, it is not possible to draw a very sharp line between secondary analysis and analysis 'merely' using data or findings from LINCAS as a background.

In-depth analyses of some variables

Another way to use the data from LINCAS is to target highly specific research questions by in-depth analysis of the data. I have already argued that in the cognitive assessments in TIMSS and PISA, individual variables were included with the purpose of establishing reliable aggregates or composites, and as such these variables may be analysed in-depth in secondary data analysis.

One relevant example is the analysis by Turmo (2003b) of a few single items from the PISA 2000 study related to the environmental issue of depletion of the ozone layer. Another example is the study of physics items from TIMSS 1995 by Angell (1996) where the aim was to shed light on the issue of whether students' alternative conceptions reflect theory-like structures or if they are highly dependent on the situation or context provided by the items, and as such reflect rather intuitive ideas. Another example worth mentioning is the study of the relationship between the use of computers and scientific literacy in the USA reported by Papanastasiou *et al.* (2003), based on data from PISA 2000. This last

³¹ See <http://www.pfi.uio.no/forskning/forskningsprosjekter/pisa+/> for a description of the project in Norwegian.

study exemplifies that although specific issues have to some degree been briefly reported in the official reports, they may be the subject of closer attention in additional in-depth analysis.

In-depth analyses of a sub-sample

The third possible design for targeting research questions in science education is to analyse specific subgroups of the respondents in depth. Many datasets are so large that the researcher may extract a subset of respondents with similar characteristics.

With data from LINCAS one may, for instance, conduct an in-depth analyses of minority groups, as was done with data from TIMSS 1995 by Heesch *et al.* (1998, 2000), and with data from PISA 2000 by Hvistendahl & Roe (2004), although the latter was related to reading literacy. Using data describing students' backgrounds and attitudes it could be possible, for instance, to construct indicators identifying specific subgroups that some science educators have classified as holding world-views that may be incompatible with those inherent in science and school science itself (Aikenhead, 1996; Cobern, 1996; Kempa & Diaz, 1990; Kilbourn, 1994). The possibilities for doing so will be even higher with the data from PISA 2006 where many more variables relating to facets of students' motivation, attitudes and learning strategies are included.

Combining data from several studies

The last generic design that I would suggest for secondary analysis of data from LINCAS is to combine data from several studies. However, this may be technically challenging, and in some cases not feasible because different studies have different populations and different sample designs that make it impossible to find a common identifier or unit of analysis.

One successful example is the study reported by Kirkcaldy *et al.* (2004) on the relationship between health efficacy, educational attainment and well being. This study combined data from PISA with data provided by the World Health Organisation, the UN, and other sources. This study was not related specifically to science or scientific literacy, but with the large number of items that will be available from the PISA 2006 study, it may be possible to supplement this analysis by extracting a sub-test of the PISA material consisting of those items that relate to what could be labelled as health literacy. In general, health is one of the five areas of applications that have been defined in the new extended framework for scientific literacy in 2006 (OECD-PISA, 2004c).

Another candidate for combining data, perhaps the primary candidate, given the discussion in section 2.3, would be to find ways to combine the data in PISA and TIMSS. As addressed before this is not a straightforward task since the two studies are different in so many ways. However, it should in principle be possible to use data aggregated to countries to explore and describe typical features of students' achievements, attitudes, motivation, and background in different countries. As suggested in paper III, the consistency with which countries group together in several cluster analyses of different datasets indicates that more fundamental, or cultural, explanations might account for students'

relative strengths and weaknesses. Background data may be combined in original ways to construct indicators of fundamental societal, attitudinal or motivational traits hypothesised as explanations for the grouping of countries.

2.6 Closing remark

Until now the PISA and TIMSS assessments have not resulted in many publications in the science education journals or conferences. This is in my view the key to understanding why studies like TIMSS and PISA have not been widely acknowledged as contributors to the science education community. I have argued that national governments should address this issue by funding researchers who would like to utilise the databases from PISA and TIMSS, and I have also argued that the community of science education researchers should seize the opportunity to apply data from these studies in their own research. I have also suggested many arguments for why data from LINCAS may be a valuable resource for research, and following that I have tried to identify some generic designs, and to some extent given some specific examples of how such designs may be applied, particularly related to research in science education

Opportunities for engaging in secondary research concerning the scientific literacy dimension in PISA will be particularly present when the data, and a large amount of the test material from the survey in 2006, will be made publicly available by the end of 2007. With the publications of international and national reports at the same time, it may be expected that science education will be at the centre of the public school debate. It could be expected or maybe even required, that researchers in science education will engage in this public and academic debate with well-developed interpretations of the results, particularly so in the national context. As a consequence, we may look forward to a variety of contributions that, taken together, make up a critical examination of, on the one hand, what PISA (and TIMSS) has to offer, and on the other, what the limitations of these studies are.

2.7 References

- Adams, R. (2003). Response to 'Cautions on OECD's Recent Educational Survey (PISA)'. *Oxford Review of Education*, 29(3), 377-289.
- Adams, R., & Wu, M. (Eds.). (2002). *PISA 2000 Technical Report*. Paris: OECD Publications.
- Aikenhead, G. (1996). Border Crossing into the Subculture of Science. *Studies in Science Education*, 27, 1-52.
- Alexander, R., Broadfoot, P., & Phillips, D. (Eds.). (1999). *Learning From Comparing: new directions in comparative educational research. Volume 1: Contexts, Classrooms and Outcomes*. Oxford: Symposium Books.
- Alexander, R., Osborn, M., & Phillips, D. (Eds.). (2000). *Learning from comparing: new directions in comparative educational research - Volume 2: Policy, Professionals and Developments*. Oxford: Symposium Books.

- Allerup, P., & Mejdning, J. (2003). Reading Achievement in 1991 and 2000. In S. Lie, P. Linnakylä & A. Roe (Eds.), *Northern Lights on PISA* (pp. 133-145). Oslo: Department of Teacher Education and School Development, University of Oslo.
- Almendingen, S. B. M. F., Tveita, J., & Klepaker, T. (2003). *Tenke det, ønske det, ville det med, men gjøre det.? en evaluering av natur- og miljøfag etter Reform 97*. Nesna: Høgskolen i Nesna.
- Angell, C. (1996). *Elevers fysikkforståelse. En studie basert på utvalgte fysikkoppgaver i TIMSS*. Dr. Scient. thesis, Det matematisk-naturvitenskapelige fakultet, Universitetet i Oslo.
- Angell, C., Kjærnsli, M., & Lie, S. (1999). *Hva i all verden skjer i realfagene i videregående skole?* Oslo: Universitetsforlaget.
- Atkin, J. M., & Black, P. (1997). Policy Perils of International Comparisons: The TIMSS Case. *Phi Delta Kappan*, 79(1), 22-28.
- Beaton, A. E., Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., Smith, T. A., & Kelly, D. L. (1996). *Science Achievement in the Middle School Years*. Boston: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.
- Black, P. (2000). Policy, practice and research: the case of testing and assessment. In R. Millar, J. Leach & J. Osborne (Eds.), *Improving science education: the contribution of research* (pp. 327-346). Buckingham: Open University Press.
- Blum, A., Goldstein, H., & Guerin-Pace, F. (2001). International adult literacy survey (IALS): an analysis of international comparisons of adult literacy. *Assessment in Education*, 8(2), 225-246.
- Bonnet, G. (2002). Reflections in a Critical Eye [1]: on pitfalls of international assessment. *Assessment in Education*, 9(3), 387-399.
- Bos, K. T. (2002). *Benefits and Limitations of Large-Scale International Comparative Achievement Studies: The Case of IEA's TIMSS Study*. PhD-thesis, University of Twente.
- Bracey, G. W. (2000). The TIMSS "Final Year" Study and Report: A Critique. *Educational Researcher*, 29(4), 4-10.
- Brekke, G., Kobberstad, T., Lie, S., & Turmo, A. (1998). *Hva i all verden kan elevene i matematikk? Oppgaver med resultater og kommentarer*. Oslo: Universitetsforlaget.
- Brown, M. (1998). The Tyranny of the International Horse Race. In R. Slee, G. Weiner & S. Tomlinson (Eds.), *School Effectiveness for Whom? Challenges to the School Effectiveness and School Improvement Movements* (pp. 33-47). London: Falmer Press.
- Bruhn, J. (1995). Mathematics Education and Comparative Studies: Two Examples. In W. Bos & R. H. Lehmann (Eds.), *Reflections on*

- Educational Achievement. Papers in Honour of T. Neville Postlethwaite* (pp. 69-74). Münster: Waxmann Verlag GMBH.
- Bryman, A. (2004). *Social Research Methods* (2nd ed.). Oxford: University Press.
- Burton, D. (2000). Secondary Data Analysis. In D. Burton (Ed.), *Research Training for Social Scientists* (pp. 347-360). London: Sage Publications.
- Bybee, R. W. (2005). PISA: A Beneficial Perspective for United States Education. *Natural Selection* (Winter 2005).
- Coburn, W. W. (1996). Worldview Theory and Conceptual Change in Science Education. *Science Education*, 80(5), 579-610.
- Comber, L. C., & Keeves, J. P. (1973). *Science Education in Nineteen Countries*. Stockholm / New York: Almqvist & Wiksell / John Wiley & Sons.
- Dale, A. A., Arber, S., & Procter, M. (1988). *Doing secondary analysis*. London: Unwin Hyman.
- deLange, J. (1997). *Looking through the TIMSS mirror from a teaching angle*. Retrieved 12.04.2002, 2002, from <http://www.enc.org/print/topics/assessment/timss/additional/document.shtm?input=CDS-000158-cd158>
- Drori, G. S. (2000). Science Education and Economic Development: Trends, Relationships, and Research Agenda. *Studies in Science Education*, 35, 27-57.
- Elley, W. B. (1992). *How in the world do students read?* The Hague: International Association for the Evaluation of Educational Achievement.
- Ferrini-Mundy, J., & Schmidt, W. H. (2005). International Comparative Studies in Mathematics Education: Opportunities for Collaboration and Challenges for Researchers. *Journal for Research in Mathematics Education*, 36(3), 164-175.
- Foshay, W. A. (Ed.). (1962). *Educational achievements of 13-year-olds in twelve countries*. Hamburg: UNESCO Institute of Education.
- Freudenthal, H. (1975). Pupils' achievements internationally compared - The IEA. *Educational Studies in Mathematics*, 6, 127-186.
- Funtowicz, S. O., Leinfellner, W., & Ravetz, J. R. (1990). *Uncertainty and quality in science for policy*. Dordrecht: Kluwer Academic Publishers.
- Gibbons, M., Limoges, C., Nowotny, H., Schwartzman, S., Scott, P., & Trow, M. (1994). *The new production of knowledge: the dynamics of science and research in contemporary societies*. London: Sage.
- Gilmore, A. (2005). *The impact of PIRLS (2001) and TIMSS (2003) in low- and middle-income countries. An evaluation of the value of World Bank support for international surveys of reading literacy (PIRLS) and mathematics and science (TIMSS)*. Retrieved May 23rd, 2005, from http://www.iea.nl/iea/hq/fileadmin/user_upload/WB-report.pdf

- Goldstein, H. (1995). *Interpreting international comparisons of student achievement*. Paris: UNESCO Publishing.
- Goldstein, H. (2004a). Education for All: the globalization of learning targets. *Comparative Education*, 40(1), 7-14.
- Goldstein, H. (2004b). International comparative assessment: how far have we really come? *Assessment in Education*, 11(2), 227-234.
- Goldstein, H. (2004c). International comparisons of student attainment: some issues arising from the PISA study. *Assessment in Education*, 11(3), 319-330.
- Gorard, S. (2003). *Quantitative Methods in Social Science. The role of number made easy*. New York: Continuum.
- Halliday, W. G., & Riegelman, C. (2004). *Dealing with Data from Influential TIMSS and AAAS Studies*. Paper presented at the 7th Annual International Conference of the National Association for Research in Science Teaching (NARST), Vancouver.
- Harlen, W. (2001). The Assessment of Scientific Literacy in the OECD/PISA Project. *Studies in Science Education*, 36, 79-104.
- Harlow, A., & Jones, A. (2004). Why Students Answer TIMSS Science Test Items the Way They Do. *Research in Science Education*, 34(2), 221-238.
- Harmon, M., Smith, T. A., Martin, M. O., Kelly, D. L., Beaton, A. E., Mullis, I. V. S., Gonzalez, E. J., & Orpwood, G. (1997). *Performance Assessment in IEA's Third International Mathematics and Science Study*. Chestnut Hill: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.
- Heaton, J. (1998). Secondary analysis of qualitative data. *Social Research Update* (22).
- Heesch, E. J., Storaker, T., & Lie, S. (1998). *Språklige minoriteters prestasjoner i matematikk og naturfag*. Oslo: Institutt for lærerutdanning og skoleutvikling, Universitetet i Oslo.
- Heesch, E. J., Storaker, T., & Lie, S. (2000). *Språklige minoritetselever og realfag*. Oslo: Institutt for lærerutdanning og skoleutvikling, Universitetet i Oslo.
- Howie, S. J., & Plomp, T. (Eds.). (in press). *Contexts of learning mathematics and science: Lessons learned from TIMSS*. Lisse: Swets & Zeitlinger Publishers.
- Howson, G. (1999). The Value of Comparative Studies. In G. Kaiser, E. Luna & I. Huntley (Eds.), *International Comparisons in Mathematics Education* (pp. 165-188). London: Falmer Press.
- Huberman, M. (1994). The OERI/CERI Seminar on educational research and development: a synthesis and commentary. In T. M. Tomlinson & A. C. Tuijnman (Eds.), *Education research and reform: an international*

- perspective* (pp. 45-66). Washington, DC: OECD Centre for Educational Research and Innovation/US Department of Education.
- Husén, T. (1967). *International Study of Achievement in Mathematics*. Stockholm / New York: Almqvist og Wiksell / Wiley & Sons.
- Husén, T. (1973). Foreword. In L. C. Comber & J. P. Keeves (Eds.), *Science Achievement in Nineteen Countries* (pp. 13-24). Stockholm / New York: Almqvist & Wiksell / John Wiley & Sons.
- Husén, T., & Tuijnman, A. (1994). Monitoring Standards in Education: Why and How it Came About. In A. C. Tuijnman & T. N. Postlethwaite (Eds.), *Monitoring the standards of education. Papers in honor of John P. Keeves* (pp. 1-21). Oxford: Pergamon.
- Hvistendahl, R., & Roe, A. (2004). The Literacy Achievement of Norwegian Minority Students. *Scandinavian Journal of Educational Research*, 48(3), 307-324.
- IEA. (1988). *Science Achievement in Seventeen Countries: A Preliminary Report*. New York: Pergamon Press.
- Isager, O. A. (1996). *Den norske grunnskolens biologi i et historisk og komparativt perspektiv*. Dr. Scient thesis, Det matematisk-naturvitenskapelige fakultet, Universitetet i Oslo.
- Jenkins, E. W. (2000). Research in Science Education: Time for a Health Check? *Studies in Science Education*, 35, 1-25.
- Kaiser, G. (1999). International Comparisons in Mathematics Education Under the Perspective of Comparative Education. In G. Kaiser, E. Luna & I. Huntley (Eds.), *International Comparisons in Mathematics Education* (pp. 3-15). London: Falmer Press.
- Keeves, J. P. (Ed.). (1992). *The IEA Study of Science III: Changes in Science Education and Achievement: 1970 to 1984*. New York: Pergamon Press.
- Keitel, C., & Kilpatrick, J. (1999). The Rationality and Irrationality of International Comparative Studies. In G. Kaiser, E. Luna & I. Huntley (Eds.), *International Comparisons in Mathematics Education* (pp. 241-256). London: Falmer Press.
- Kellaghan, T., & Greaney, V. (2001). The globalisation of assessment in the 20th century. *Assessment in Education*, 8(1), 87-102.
- Kempa, R. F., & Diaz, M. M. (1990). Students' motivational traits and preferences for different instructional modes in science education. *International Journal of Science Education*, 12(2), 205-216.
- Kiecolt, K. J., & Nathan, L. E. (1985). *Secondary Analysis of Survey Data*. Beverly Hills: Sage Publications.
- Kilbourn, B. (1994). World views and curriculum. *Nordisk Pedagogik*, 14(1), 131-140.

- Kind, P. M. (1996). *Exploring Performance Assessment in Science*. Dr. Scient thesis, Det matematisk-naturvitenskapelige fakultet, Universitetet i Oslo.
- Kind, P. M., Kjærnsli, M., Lie, S., & Turmo, A. (1999). *Hva i all verden gjør elevene i realfag? Praktiske oppgaver i matematikk og naturfag*. Oslo: Institutt for lærerutdanning og skoleutvikling, Universitetet i Oslo.
- Kirkcaldy, B., Furnham, A., & Siefen, G. (2004). The Relationship Between Health Efficiency, Educational Attainment, and Well-Being Among 30 Nations. *European Psychologist*, 9(2), 107-119.
- Kjærnsli, M. (2003). *Achievement in Scientific Literacy in PISA: Conceptual understanding and process skills*. Paper presented at the 4th Biannual Conference of European Science Education Research Association (ESERA). Noordwijkerhout, The Netherlands.
- 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.
- Kjærnsli, M., Lie, S., Stokke, K. H., & Turmo, A. (1999). *Hva i all verden kan elevene i naturfag? Oppgaver med resultater og kommentarer*. Oslo: Universitetsforlaget.
- Klette, K., Aukrust, V. G., Hagtvet, B. E., & Hertzberg, F. (2003). *Synteserapport - klasserommets praksisformer etter Reform97*. Oslo: Norges forskningsråd.
- 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*. Oslo: Institutt for lærerutdanning og skoleutvikling, Universitetet i Oslo.
- 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*. Oslo: Department of Teacher Education and School Development, University of Oslo.
- Lijnse, P. (2000). Didactics of science: the forgotten dimension in science education research? In R. Millar, J. Leach & J. Osborne (Eds.), *Improving science education: the contribution of research* (pp. 308-326). Buckingham: Open University Press.

- Linn, R. L. (2000). The Measurement of Student Achievement in International Studies. In A. C. Porter & A. Gamoran (Eds.), *Methodological Advances in Cross-National Surveys of Educational Achievement* (pp. 27-57). Washington, DC: National Academy Press.
- Loving, C. C., & Cobern, W. W. (2000). Invoking Thomas Kuhn: What citation analysis reveals about science education. *Science & Education*, 9(1-2), 187-206.
- Martin, M. O., & Kelly, D. L. (Eds.). (1997a). *Technical Report Volume I: Design and Development*. Chestnut Hill: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.
- Martin, M. O., & Kelly, D. L. (Eds.). (1997b). *Technical Report Volume II: Implementation and Analysis*. Chestnut Hill: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.
- Martin, M. O., & Mullis, I. V. S. (2000). International Comparisons of Student Achievement: Perspectives from the TIMSS International Study Center. In D. Shorrocks-Taylor & E. W. Jenkins (Eds.), *Learning from Others* (pp. 29-47). Dordrecht: Kluwer Academic Publishers.
- Martin, M. O., Mullis, I. V. S., Beaton, A. E., Gonzalez, E. J., Smith, T. A., & Kelly, D. L. (1997). *Science Achievement in the Primary School Years*. Boston: Center for the Study of Testing, Evaluation and Educational Policy, Boston College.
- Martin, M. O., Mullis, I. V. S., & Chrostowski, S. J. (Eds.). (2004a). *TIMSS 2003 Technical Report*. Chestnut Hill: TIMSS & PIRLS International Study Center, Boston College.
- Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., & Chrostowski, S. J. (2004b). *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, Lynch School of Education, Boston College.
- Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., Gregory, K. D., Smith, T. A., Chrostowski, S. J., Garden, R. A., & O'Connor, K. M. (2000a). *TIMSS 1999 International Science Report. Findings from IEA's Repeat of the Third International Mathematics and Science Study at the Eight Grade*. Boston: International Study Center, Lynch School of Education, Boston College.
- Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., Smith, T. A., & Kelly, D. L. (1999). *School Contexts for Learning and Instruction in IEA's Third International Mathematics and Science Study*. Chestnut Hill: TIMSS International Study Center, Boston College.
- Martin, M. O., Mullis, I. V. S., Gregory, K. D., Hoyle, C., & Shen, C. (2000b). *Effective Schools in Science and Mathematics. IEA's Third International*

- Mathematics and Science Study*. Chestnut Hill: TIMSS International Study Center, Boston College.
- McGuinness, D. (2004). *Early Reading Instruction: What Science Really Tells Us about How to Teach Reading*. Cambridge: MIT Press.
- Mejding, J. (Ed.). (2004). *PISA 2003 - Danske unge i international sammenligning*. København: Danmarks Pædagogiske Universitet.
- Millar, R. (2003). Presidential Address: What can we reasonably expect of research in science education. In D. Psillos, P. Kariotoglou, V. Tselfes, E. Hatzikraniotis, G. Fassouloupoulos & M. Kallery (Eds.), *Science Education Research in the Knowledge-Based Society* (pp. 3-8). Dordrecht: Kluwer Academic Publishers.
- Mullis, I. V. S., Martin, M. O., Beaton, A. E., Gonzalez, E. J., Kelly, D. L., & Smith, T. A. (1998). *Mathematics and Science Achievement in the Final Year of Secondary School. IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.
- Mullis, I. V. S., Martin, M. O., Fierros, E. G., Goldberg, A. L., & Stemler, S. E. (2000a). *Gender Differences in Achievement. IEA's Third International Mathematics and Science Study*. Chestnut Hill: TIMSS International Study Center, Boston College.
- Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., & Chrostowski, S. J. (2004). *TIMSS 2003 International Mathematics Report. Findings From IEAs Trends in International Mathematics and Science Study at the Fourth and Eighth Grades*. Boston: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.
- Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Gregory, K. D., Garden, R. A., O'Connor, K. M., Chrostowski, S. J., & Smith, T. A. (2000b). *TIMSS International Mathematics Report. Findings from the IEA's Repeat of the Third International Mathematics and Science Study at the Eight Grade*. Boston: 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., Gonzalez, E. J., Chrostowski, S. J., & O'Connor, K. M. (2001). *TIMSS Assessment Frameworks and Specifications 2003*. Boston: International Study Center, Lynch School of Education, Boston College.
- National Center for Education Statistics (NCES). (2005). *Comparing NAEP, TIMSS, and PISA in Mathematics and Science*. Retrieved May 24th, 2005, from http://nces.ed.gov/timss/pdf/naep_timss_pisa_comp.pdf
- Neidorf, T. S., Binkley, M., Gattis, K., & Nohara, D. (forthcoming-a). *A Content Comparison of the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and Program for International Student Assessment (PISA) 2003 Mathematics*

- Assessments* (NCES 2005-112). Washington, DC: National Center for Education Statistics.
- Neidorf, T. S., Binkley, M., & Stephens, M. (forthcoming-b). *A Content Comparison of the National Assessment of Educational Progress (NAEP) 2000 and Trends in International Mathematics and Science Study (TIMSS) 2003 Science Assessments* (NCES 2005-106). Washington, DC: National Center for Education Statistics.
- Neuman, W. L. (2003). *Social Research Methods: Qualitative and Quantitative Approaches* (5th ed.). Boston: Allyn and Bacon.
- Nohara, D. (2001). *A Comparison of the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study Repeat (TIMSS-R), and the Programme for International Student Assessment (PISA)*. Working Paper No. 2001-07. Retrieved May 24th, 2005, from <http://nces.ed.gov/pubs2001/200107.pdf>
- Nowotny, H., Scott, P., & Gibbons, M. (2001). *Re-thinking science: knowledge and the public in an age of uncertainty*. Cambridge: Polity Press.
- OECD-PISA. (1999). *Measuring Student Knowledge and Skills*. Paris: OECD Publications.
- OECD-PISA. (2001). *Knowledge and Skills for Life. First results from PISA 2000*. Paris: OECD Publications.
- OECD-PISA. (2002a). *Reading for Change - Performance and Engagement Across Countries*. Paris: OECD Publications.
- OECD-PISA. (2002b). *Sample Tasks from the PISA 2000 Assessment: Reading, Mathematical and Scientific Literacy*. Paris: OECD Publications.
- OECD-PISA. (2003a). *Learners for Life: Student Approaches to Learning: Results from PISA 2000*. Paris: OECD Publications.
- OECD-PISA. (2003b). *The PISA 2003 Assessment Framework: Mathematics, Reading, Science and Problem Solving Knowledge and Skills*. Paris: OECD Publications.
- OECD-PISA. (2004a). *Learning for Tomorrow's World. First Results From PISA 2003*. Paris: OECD Publications.
- OECD-PISA. (2004b). *Problem Solving for Tomorrow's World – First Measures of Cross Curricular Competencies from PISA 2003*. Paris: OECD Publications.
- OECD-PISA. (2004c). *Scientific Literacy Framework*. Draft presented to the PISA National Project Managers in September, and approved by the PISA Governing Board, June 2004. [DOC: NPM(0409)5].
- OECD-PISA. (2005). *School factor related to quality and equity: results from PISA 2000*. Paris: OECD Publications.
- OECD. (1996). *Education at a Glance*. Paris: OECD Publications.

- OECD. (1997). *Education at a Glance*. Paris: OECD Publications.
- OECD. (1998). *Education at a Glance*. Paris: OECD Publications.
- OECD. (2001). *Education Policy Analysis 2001*. Paris: OECD Publications.
- OECD & UNESCO. (2003). *Literacy Skills for the World of Tomorrow*. Paris: OECD Publications & UNESCO Publishing.
- Olsen, R. V. (2004). *The OECD PISA assessment of scientific literacy: how can it contribute to science education research?* Paper presented at the 7th Annual International Conference of the National Association for Research in Science Teaching (NARST), Vancouver.
- Olsen, R. V., Lie, S., & Turmo, A. (2001). Learning about students' knowledge and thinking in science through large-scale quantitative studies. *European Journal of Psychology of Education*, 16(3), 403-420.
- Orpwood, G. (2000). Diversity of Purpose in International Assessments: Issues arising from the TIMSS test of Mathematics and Science. In D. Shorrocks-Taylor & E. W. Jenkins (Eds.), *Learning from Others: International Comparisons in Education* (pp. 49-62). Dordrecht: Kluwer Academic Publishers.
- Papanastasiou, C. (Ed.). (2004). *Proceedings of the IRC-2004: TIMSS* (Volume I and II). Nicosia: Cyprus University Press.
- Papanastasiou, E. C. (2003). Science Literacy by Technology by Country: USA, Finland and Mexico. Making sense of it all. *Research in Science & Technological Education*, 21(2), 241.
- Papanastasiou, E. C., Zembylas, M., & Vrasidas, C. (2003). Can Computer Use Hurt Science Achievement? The USA Results from PISA. *Journal of Science Education and Technology*, 12(3), 325-332.
- Pole, C., & Lampard, R. (2002). *Practical Social Investigation: Qualitative and Quantitative Methods in Social Research*. Essex: Pearson Education Limited.
- Porter, A. C., & Gamoran, A. (Eds.). (2002). *Methodological Advances in Cross-National Surveys of Educational Achievement*. Washington, DC: National Academy Press.
- Postlethwaite, T. N. (1988). Preface. In T. N. Postlethwaite & T. Husén (Eds.), *The Encyclopaedia of Comparative Education and National Systems of Education*. Oxford: Pergamon Press.
- Postlethwaite, T. N., & Wiley, D. E. (1992). *The IEA Study of Science II: Science Achievement in Twenty-Three Countries*. New York: Pergamon Press.
- Prais, S. J. (2003). Cautions on OECD's Recent Educational Survey (PISA). *Oxford Review of Education*, 29(2), 139-163.
- Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M., Pekrun, R., Rolff, H.-G., Rost, J., & Schiefele, U. (2004). *PISA 2003. Der*

Bildungsstand der Jugendlichen in Deutschland– Ergebnisse des zweiten internationalen Vergleichs. Münster: Waxmann.

- Reddy, V. (2005). Cross-national achievement studies: learning from South Africa's participation in the Trends in International Mathematics and Science Study (TIMSS). *Compare*, 35(1), 63-77.
- Reeve, R. A., & Walberg, H. J. (1994). Secondary Data Analysis. In T. Husén & T. N. Postlethwaite (Eds.), *The International Encyclopedia of Education* (2nd ed., pp. 5367-5373). Oxford: Pergamon.
- Rew, L., Koniak-Griffin, D., Lewis, M. A., Miles, M., & Ann, O. S. (2000). Secondary Data Analysis: New Perspective for Adolescent Research. *Nursing Outlook*, 48, 223-229.
- Roberts, D. A. (forthcoming). *Scientific literacy/science literacy* (fifth draft). To be published in NARST Handbook of Research in Science Education.
- Robinson, P. (1999). The Tyranny of League Tables: international comparisons of educational attainment and economic performance. In R. Alexander, P. Broadfoot & D. Phillips (Eds.), *Learning From Comparing: new directions in comparative educational research. Volume 1: Contexts, Classrooms and Outcomes* (pp. 217-235). Oxford: Symposium Books.
- Robitaille, D. F. (Ed.). (1997). *National Contexts for Mathematics and Science Education: an encyclopedia of the education systems participating in TIMSS*. Vancouver: Pacific Educational Press.
- Robitaille, D. F., & Beaton, A. E. (Eds.). (2002). *Secondary Analysis of the TIMSS Data*. Dordrecht: Kluwer Academic Publishers.
- Robitaille, D. F., & Garden, R. A. (Eds.). (1996). *Research Questions & Study Design* (Vol. 2). Vancouver: Pacific Educational Press.
- Robitaille, D. F., & Maxwell, B. (1996). The Conceptual Framework and Research Questions for TIMSS. In D. F. Robitaille & R. A. Garden (Eds.), *Research Questions & Study Design* (Vol. 2, pp. 34-43). Vancouver: Pacific Educational Press.
- Robitaille, D. F., Schmidt, W. H., Raizen, S., Mc Knight, C., Britton, E., & Nicol, C. (1993). *Curriculum Frameworks for Mathematics and Science*. Vancouver: Pacific Educational Press.
- Romberg, T. A., Zarinnia, E. A., & Collis, K. F. (1990). A New World View of Assessment in Mathematics. In K. G. (Ed.), *Assessing higher order thinking in mathematics* (pp. 21-38). Washington, DC: American Association for the Advancement of Science.
- Rosier, M. J., & Keeves, J. P. (Eds.). (1991). *The IEA Study of Science I: Science Education and Curricula in Twenty-Three Countries*. New York: Pergamon Press.
- Rychen, S., & Salganik, L. H. (Eds.). (2003). *Key Competencies for a Successful Life and a Well-Functioning Society*. Cambridge: Hogrefe & Huber.

- Schleicher, A. (2000). Monitoring Student Knowledge and Skills: The OECD Programme for International Student Assessment. In D. Shorrocks-Taylor & E. W. Jenkins (Eds.), *Learning from Others* (pp. 63-77). Dordrecht: Kluwer Academic Publishers.
- Schmidt, W. H., Jorde, D., Cogan, L. S., Barrier, E., Gonzalo, I., Moser, U., Shimizu, K., Sawada, T., Valverde, G. A., McKnight, C., Prawat, R. S., Wiley, D. E., Raizen, S. A., Britton, E. D., & Wolfe, R. G. (1996). *Characterizing Pedagogical Flow: An Investigation of Mathematics and Science Teaching in Six Countries*. Dordrecht: Kluwer Academic Publishers.
- Schmidt, W. H., McKnight, C., Houang, R. T., Wang, H. C., Wiley, D. E., Cogan, L. S., & Wolfe, R. G. (2001). *Why schools matter: a cross-national comparison of curriculum and learning*. San Francisco: Jossey-Bass.
- Schmidt, W. H., McKnight, C., Valverde, G. A., Houang, R. T., & Wiley, D. E. (1997a). *Many Visions, Many Aims. Volume 1: A Cross-National Investigation of Curricular Intentions in School Mathematics*. Dordrecht: Kluwer Academic Publishers.
- Schmidt, W. H., Raizen, S. A., Britton, E. D., Bianchi, L. J., & Wolfe, R. G. (1997b). *Many Visions, Many Aims. Volume 2: A Cross-National Investigation of Curricular Intentions in School Science*. Dordrecht: Kluwer Academic Publishers.
- Schoultz, J. (2000). *Att samtala om/i naturvetenskap. Kommunikation, kontext och artefact*. PhD-thesis, University of Linköping.
- Shorrocks-Taylor, D. (2000). International Comparisons of Pupils Performance: An Introduction and Discussion. In D. Shorrocks-Taylor & E. W. Jenkins (Eds.), *Learning from Others* (pp. 13-27). Dordrecht: Kluwer Academic Publishers.
- Shorrocks-Taylor, D., & Jenkins, E. W. (Eds.). (2000). *Learning From Others*. Dordrecht: Kluwer Academic Publishers.
- Sjøberg, S. (2005a). PISA, TIMSS og norske læreplaner. *Bedre skole* (1).
- Sjøberg, S. (2005b). TIMSS och PISA brickor i det politiska spelet. *Pedagogiska magasinet* (2), 10-19.
- Stanat, Artelt, Baumert, Klieme, Neubrand, Prenzel, Schiefele, Schneider, Schümer, Tillmann, & Weiß. (2002). *PISA 2000: Overview of the Study. Design, Method and Results*. Berlin: Max Planck Institute for Human Development.
- Steedman, H. (1999). Measuring the Quality of Educational Outputs: some unresolved problems. In R. Alexander, P. Broadfoot & D. Phillips (Eds.), *Learning From Comparing: new directions in comparative educational research. Volume 1: Contexts, Classrooms and Outcomes* (pp. 201-216). Oxford: Symposium Publishers.

- Stigler, J. W., Gonzales, P., Kawanaka, T., Knoll, S., & Serrano, A. (1999). *The TIMSS Videotape Classroom Study: Methods and findings from an exploratory research project on eight-grade mathematics instruction in Germany, Japan, and the United States*. Washington, DC: National Center for Education Statistics.
- Stigler, J. W., & Hiebert, J. (1999). *The teaching gap: best ideas from the world's teachers for improving education in the classroom*. New York: Free Press.
- The BMS. (1994). Correspondence Analysis: A history and French Sociological Perspective. In M. J. Greenacre & J. Blasius (Eds.), *Correspondence Analysis in the Social Sciences* (pp. 128-137). London: Academic Press.
- Tomlinson, T. M., & Tuijnman, A. C. (Eds.). (1994). *Education research and reform: an international perspective*. Washington, DC: OECD Centre for Educational Research and Innovation/US Department of Education.
- Tooley, J., & Darby, D. (1998). *Educational research: a critique*. London: Office for Standards in Education.
- Turmo, A. (2003a). *Naturfagdidaktikk og internasjonale studier. Store internasjonale studier som ramme for naturfagdidaktisk forskning: En drøfting med eksempler på hvordan data fra PISA 2000 kan belyse sider ved begrepet naturfaglig allmenndannelse*. Dr. Scient thesis, Det utdanningsvitenskapelige fakultetet, Universitetet i Oslo. Oslo: Unipub AS.
- Turmo, A. (2003b). *Understanding a newsletter article on ozone - a cross-national comparison of the scientific literacy of 15-year-olds in a specific context*. Paper presented at the 4th Biannual Conference of European Science Education Research Association (ESERA). Noordwijkerhout, The Netherlands.
- Turmo, A. (2004). *Testing the Outcomes of the Nordic Principle of Equity: the case of scientific literacy*. Paper presented at the 7th Annual International Conference of the National Association for Research in Science Teaching (NARST), Vancouver.
- Turmo, A., & Olsen, R. V. (2003). Naturfagdidaktisk forskning i et policy-perspektiv. Muligheter og utfordringer. In D. Jorde & B. Bungum (Eds.), *Naturfagdidaktikk: Perspektiver, forskning, utvikling* (pp. 441-462). Oslo: Gyldendal akademisk.
- Uljens in Mathiasson. (2005). PISA - en maktfaktor i skoldebatten. *Pedagogiska magasinet* (2), 30-36.
- UNESCO. (1990). *World Declaration on Education for All*. New York: UNESCO.
- UNESCO. (2002). *Education for All. An international strategy to operationalize the Dakar Framework for Action on Education for All (EFA)*. Retrieved May 21th, 2005, from http://www.unesco.org/education/efa/global_co/global_initiative/strategy_2002.pdf

- UNESCO. (2005). *Education for All Global Monitoring Report 2005 - The Quality Imperative*. Retrieved May 24th, 2005, from http://portal.unesco.org/education/en/ev.php-URL_ID=35939&URL_DO=DO_TOPIC&URL_SECTION=201.html
- UNESCO & OECD. (2003). *Financing Education - Investments and Returns: Analysis of the World Education Indicators*. Paris: OECD Publications & UNESCO Publishing.
- Vári, P. (Ed.). (1997). *Are We Similar in Math and Science? A Study of Grade 8 in Nine Central and Eastern European Countries*. Amsterdam: International Association for the Evaluation of Educational Achievement.
- Wang, J. (2001). TIMSS Primary and Middle School Data: Some Technical Concerns. *Educational Researcher*, 30(6), 17-21.
- Willms, J. D. (2003). *Student Engagement at School - A Sense of Belonging and Participation*. Paris: OECD Publications.
- 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* (pp. 225-240). London: Falmer Press.
- Yip, D. Y., Chiu, M. M., & Ho, E. S. C. (2004). Hong Kong Student Achievement in OECD-PISA Study: Gender Differences in Science Content, Literacy Skills, and Test Item Formats. *International Journal of Science & Math Education*, 2(1), 91.
- Ziman, J. (1996). Is science losing its objectivity? *Nature*, 382(August), 751-754.

3 Methods and methodological reflections

3.1 Introduction to the chapter

Each of the papers that comprise the major part of this thesis has, to a varying degree, sections presenting the methods used, and most of the papers also have parts which could be labelled as methodological deliberations or reflections. The latter is natural since all the papers in the thesis address the general methodological issue concerning the characteristics of the information contained in students' responses to cognitive items in large-scale comparative studies of students' achievement, and how this may be analysed.

However, the format of the papers or articles does not allow for elaborated discussions of the fundamental position from which they are written. This chapter therefore has three aims: (a) to present a general perspective on the analyses of data (section 3.2); (b) to characterise some generic properties of categorical data and specifically characterise the properties of the categorical data in the TIMSS and PISA cognitive data (section 3.3); and (c) to give additional information on some of the specific methods used in the papers (section 3.4).

The terms methodology, method and technique (with associated adjectives) are used throughout the text. The intention has been to use these terms consciously and consistently. *Methodology* is used to refer to general ideas, perspectives and to some extent ideological positions beyond the particular statistical methods. For instance the concept of quantitative methodology refers to the idea that concepts can be operationalised into variables or constructs with associated values. Other examples of methodological issues related to the use of quantitative methods are: general reflections on the role and status of statistical modelling; the underlying arguments for the use of a particular method; analysis and discussions about general characteristics of a data set; or prescriptions, descriptions and/or justifications for a specific research design. The term *method* is more specific. Examples of different quantitative methods are multiple regression analysis, correspondence analysis or homogeneity analysis, to name but a few. Even more specific is the term *technique*, which refers to a particular (mathematical) procedure used for a certain method, for instance 'multiple regression analysis with stepwise selection of variables' or 'hierarchical cluster analysis using correlations as the proximity measure and single linkage for clustering'. This chapter will mostly be about the methodological issues framing this thesis (sections 3.1-3.3). Also, short descriptions of the methods used are given (section 3.4); including condensed presentations of the techniques used, but without reference to the mathematical aspects.

A final term used throughout the text is *data analysis*. More will be said in the next section on this. Initially it is enough to say that I find the term data analysis useful because this term gives priority to the data themselves. In many situations one could use the term statistical analysis almost interchangeably. This would, however, through connotation give priority to the statistical models imposed on the data. The main purpose of data analysis, as I understand it, is to process the data and extract meaning. An alternative procedure is to give priority

to a theory or model and subsequently use data to evaluate the theoretical statements.

3.2 A fundamental issue regarding the use of quantitative methods: Data or model?

This section is written with a more personal voice, and targets a more general context, than the rest of the thesis. The reason for including this part in the thesis is that my own work to some degree has been affected by the reflections presented in the following. This section should therefore not be regarded as one in which I put forth and explain the methods used in great detail, but rather as a more general description of my own position regarding the use of relatively complex quantitative methods within research in the social sciences. This discussion is primarily related to the methods used in papers II and III.

3.2.1 Exploration and descriptive statistics versus confirmation and inferential statistics

There has been, and there still is, some debate within the community of researchers committed to the development and evaluation of how to analyse data, regarding the main purpose of data analysis. On the one hand, there are schools of methodology which start with the formulation of a model, whether this is a statistical model, e.g. the Rasch model for dichotomously scored items, or theoretical models for how variables, according to a hypothesis or theory, should be related to each other, the latter approximating the idea of a causal model, e.g. as in many applications of structural equation modelling. Building models is most often accompanied by strong assumptions, for instance that the variables used are normally distributed and/or that any relationships between variables are linear. Within this school of methodology data analysis, as a consequence, is concerned with testing whether the data themselves confirm the model.

On the other hand there has been a growing recognition of the opposite procedure, that is, to start with the data themselves, and in general, the term data analysis as presented here will refer to procedures that to some degree follow this principle. One of the most cited proponents for this approach is Benzécri, categorised by both Gifi (1990) and The BMS (1994) as a spokesman of the French data analytic school. Benzecri (1973) formulated five principles for data analysis, the second of which addresses most explicitly the relationship between data and model:

The model must follow the data, and not the other way around. This is another error in the application of mathematics to the human sciences: the abundance of models, which are built a priori and then confronted with the data by what one calls a 'test'. Often the 'test' is used to justify a model in which the number of parameters to be fitted is larger than the number of data points. And often it is used, on the contrary, to strongly reject as invalid the most judicial remarks of the experimenter. But what

we need is a rigorous method to extract structure, starting from data. (Benzécri (1973) cited in Gifi (1990, p. 25)³²)

This is an extreme formulation which exaggerates what happens when data are analysed according to models. I have, for instance, rarely come across a study where “the number of parameters to be fitted is larger than the number of data points”. The main criticism of this principle, that “the model must follow the data”, is that little attention will be paid to establish models or new theories. An example which could be used to counter this critique is Bourdieu’s research in sociology (e.g. Bourdieu, 1984). His famous, much cited and well-accepted theory regarding the relationship between different tastes (life styles) and some specified social situations or classifications is to a large degree based on analyses starting with the data themselves using correspondence analysis. This illustrates the obvious point that objective criteria for what is accepted methodology do not exist.

However, most data analysts are not searching for truth in an absolute sense or to establish theories of any general nature. The main purpose of using the data themselves as the starting point for analyses stems from the need to find the main patterns in the abundant amount of information available in typical questionnaire and test-type datasets. Data analysis as described by Tukey (1977) is very much concerned with finding novel ways to describe the data, in order to communicate the essence of the data better without reducing it to a few statistical parameters (Cohen, 1990).

There are several possible reasons for the growing use of these methods over the last decades. One suggestion is that many studies in the social and behavioural sciences are large-scale and cannot be replicated because of the costs. This forces the researcher to include an abundance of potentially relevant items, given the overall purpose of the study. This means that there is a need afterwards to sort out those items that are redundant or irrelevant (The BMS, 1994). A more technical reason could be that many data sets operate with variables at different measurement levels: for instance, ordinal Likert-scaled variables, and nominal variables grouping the respondents according to some criteria. In addition these variables are rarely linearly related. Traditional model-based statistics were developed primarily to deal with variables measured at the ratio or interval level and most techniques assume linearity and ideal distributional characteristics³³. The need for new techniques have therefore lead to the development of tools for excavating the main trends in the data, and many of these tools can deal with variables at different measurement levels and with none or few strong assumptions of the statistical properties of the variables. Yet

³² With no knowledge of French I must rely on the translations of others. My trust in this translation is increased by the fact that Hjelbrekke (1999) has given a direct translation into Norwegian, which captures the same essence. Also the first sentence is seen in numerous other sources in English (e.g. Greenacre & Blasius (1994) and The BMS (1994).

³³ Recent developments as a result of the increased processing power in ordinary computers, combined with the development of algorithms to estimate parameters have, during the last few decades, made it possible to include variables of almost any kind in model-based analyses. Examples are statistical computer packages such as M+ and MLwiN that allow you to estimate mixture models including both categorical and continuous variables.

another reason for the development of techniques such as correspondence analysis and cluster analysis could be that social and behavioural sciences are in general not characterised by having strong and clear theories with predictive power. Consequently, it may not be possible to develop questionnaires or tests resulting from a clear theoretical rationale from which predictions of the main characteristics of the data can be deduced. All in all, this suggests that the social sciences have a need for non-linear multivariate techniques which are able to “extract structure starting from data”. Lastly, a pragmatic reason for traditional inferential statistics being widely used is related to the problem addressed in section 3.2.2 below: the fact that these statistical methods are relatively easy to understand and learn to use. They are familiar to most social scientists through introductory courses, and some simpler versions are even implemented in high school curricula in some countries (e.g. hypothesis testing on the difference in means). Making strong assumptions leads to simpler mathematics.

The reasons given above could easily be perceived as giving a negative judgement about the use of inferential statistics. Let me therefore add to this picture that the choice of using inferential statistics is often based on sound judgements of the problem addressed. Social scientists have research questions related to the confirmation or rejection of some statement or model and they frequently need to find estimates of how likely it is that the results can be generalised to the population from which the sample is drawn (e.g. significance testing of a parameter). In a lot of research such questions reflect the main research problems. In the case of large-scale international comparative achievement studies in education, such as TIMSS and PISA, the primary purpose is to measure student achievement and indicators of how this achievement is associated with external factors such as students’ backgrounds, *for the participating school systems*. In other words, the issue of generalisation from the sample to the target population (the specified age or grade of a group of students in the system) is central to this purpose. In order to guide users of this information, it is of vital importance to include parameters such as levels of statistical significance, confidence intervals and standard errors of measurements.

Significance testing reminds us that one chief achievement of inferential statistics is to be able to generalise our findings from the sample used to the population in mind. However, inferential statistics cannot be used for all interesting research questions. Indeed, many leading scholars have clearly stated that the use of inferential statistics, and especially the focus on hypothesis testing, has been overemphasised to such a degree that it has diverted our attention from more meaningful questions (eg. Cohen, 1990; Meehl, 1978; Rennie, 1998; Rosnow & Rosenthal, 1989; Tukey, 1977). We should allow ourselves to give more attention to relatively open questions about what the data themselves contain in terms of meaningful structure. Or in the words of Tukey (1962):

The most important maxim for data analysis to heed, and one which many statisticians seem to have shunned, is this: ‘Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.’ Data analysis must progress by approximate

answers, at best, since its knowledge of what the problem really is will at best be approximate. (pp. 13-14, cited from Gifi (1990))

The research problem for this thesis is not primarily to find significant effects, or to confirm structures in the data proposed by theory. A pragmatic reason for significance not being that vital for the problem is that, with the data set offered by PISA, the number of respondents is so large that the effects of a size that captures our attention would be reported as significant, at least when analysing the data at the student level.

The reason for not using confirmatory analytical perspectives in this thesis is primarily the fact that the aim of the work presented in this dissertation has been to develop descriptions of the properties of the information contained in the data, and not for confirmation of some hypothesis which could be done by applying the framework of classical statistical inference. This thesis is therefore exploratory in spirit, and statistical techniques are used which are deemed appropriate for working with the stated problem. Additionally, in Paper II the aim is to describe properties of variables measured at the nominal level, and most traditional statistical techniques are not suitable for exploring such data.

In bringing to a close the discussion about these two different perspectives on data analysis, it should be mentioned that the trend today is for exploratory and confirmatory data analysis to converge. This is seen, for instance, in the major statistical software packages where methods originally developed for, and which are most used in, exploratory data analysis can also provide confidence intervals and significance levels for the parameters in the solution. Reading the presentation above, one could easily get the impression that exploratory data analysis can be regarded as an alternative to applying the confirmatory methods of inferential statistics. It was however not my intention to promote such a simplistic interpretation of this dichotomy. As so often is the case, describing different views, positions, or ideologies by dichotomies may be useful to present tensions or dilemmas one may be confronted to, but very often such dichotomous categories rather reflect the extremes on a continuous scale. The choice of method should be a pragmatic choice depending on the research problem or the purpose of the research.

3.2.2 Pure versus applied data analysis or statistics

I will assume that the great majority of educational researchers consider statistical methods as tools to be used in addressing substantial questions arising from the field of education. Therefore, educational researchers apply statistical methods that are traditionally included in their training. In my work with this thesis I had to get acquainted with methods beyond those taught in master's or even doctoral courses at my university. I was therefore forced to study literature related to these methods. The specialised literature on method that I was introduced to in the work with this thesis can roughly be put into two categories:

- Literature that is too simplistic, meaning that it is not followed by proper reflections on how to interpret results. Examples of such literature are typically manuals or other materials developed by the software producers focusing on the hands-on (and not the minds-on) aspect of the techniques. This

is not to say that this literature is not useful. I have had to make considerable use of this literature in order to actually perform the analyses presented in this thesis. However, this illustrates one of the main problems of being a user of statistical methods integrated in ordinary statistical software: they are deceptively easy to apply to a data set, but it is generally not easy to find guiding rules on how to interpret the solutions obtained.

- Literature that is far too technical/mathematical to be understood by most social scientists. This literature does give a detailed account of how the results are obtained. All possible parameters which could help the interpretation are often presented. Unfortunately though, it is presented in a techno-mathematical jargon from which it is difficult or impossible to extract substantiated guidelines for the way in which the results should or could be interpreted. I think I can place myself in the upper quartile of social scientists on a scale of mathematical background, but even though I have the advantage of a general background in mathematics, including for instance courses in linear algebra, I have serious problems in understanding much of the available literature. This literature is developed by academics whose main interest is in the methods themselves, and as most of us do, they tend to write for their own peers.

Skimming the content of the most cited journals in science education also tells us that very few people publish results that are based on the application of non-traditional statistical methods. Only occasionally can we find papers using methods that go beyond what is ordinarily taught in introductory courses for social scientists (one exemplary exception is Sadler (1998) using item response theory). To find articles describing the use of non-traditional statistical methods, we have to seek journals devoted to the methods themselves. Even if the journals have words like ‘applied’ in their title, much of the work is largely unavailable for the great majority of researchers within the field of education. These articles are probably peer reviewed by people with a main interest in the methods, and not primarily in the substantial questions that the methods are used to answer, or at least, it appears to me to be like this.

In conclusion, what I have tried to sketch here is the twofold challenge I have met in the work with my thesis, particularly related to papers II and III: (a) the challenge of reading the literature necessary to become qualified to apply non-traditional methods to analyse data; and (b) the challenge of writing papers/articles that are deemed appropriate by educational researchers refereeing manuscripts submitted to journals.

As a consequence of this experience, I decided that whenever describing the methods used, I would not focus on the mathematical/statistical aspects per se. Rather, my aim has been to communicate *what the method does to the data* in a qualitative way. In doing this, particular attention has been paid to recording the relevance for using the method. A further priority has been to provide the methods with the details necessary to critically examine whether the inferences made are based on solid and robust findings. Hopefully, a balance between these

two somewhat opposing emphases has been found. More importantly, in seeking to find this balance a sense of pragmatism has been necessary.

3.2.3 A pragmatic methodological position

In essence the deliberation above is given in order to substantiate my own pragmatic stance regarding methodology, or, as stated by Lederman (1992):

We must let the research questions direct the research approaches and data analysis procedures. (p. 1012, original emphasis)

My original reason for using methods like cluster analysis and homogeneity analysis was the purpose of my research. This clearly states that my main purpose is to extract meaning and structure from the abundance of information available in the cognitive items in PISA, including also analyses of nominal variables.

Before doing the actual data processing and analyses, I did not have any specific or strong hypothesis about what the main findings would be. I had a hope or a vision that patterns could be identified across items (paper II) or across countries (paper III) that could indicate some stable cognitive structures reflected in students' responses. I have disciplined myself to frame my work by some very limiting conditions, of which the most important is that in my research I will only use data already available in the PISA data set³⁴. From the beginning I realised that this could be risky, since I could end up with a description of data without being able to understand what the patterns revealed. Moreover, as this is a thesis in science education, I could end up with negative findings, meaning that I might not be able to describe patterns in the data with concepts and characteristics which would inform the community of science educators, a few of whom in the end will evaluate whether this is an acceptable thesis. In retrospect, this has to some degree happened.

A pragmatic methodological position means that the substantial questions must be given priority. It does not mean that anything goes; it is not equal to relativism. In spirit, this pragmatic view on the choice of method is also reflected in the interesting debate about the presentation of quantitative studies prior to and after the publication of the new Publication Manual of the American Psychological Association (APA, 2001). The main theme of these academic debates was the use of hypothesis testing associated with a focus on levels of significance. In this debate, it was reported that hypothesis testing with the purpose of avoiding type I errors follows a logic dictated by an arbitrary rule with no ontological basis, often formulated as a binary decisional in the form of requiring an alpha³⁵ of 0.05 or lower as the main criterion for evaluating the findings (Rosnow & Rosenthal, 1989). Not only is this rule of decision arbitrary; it is even logically false in the sense that, in theory, the actual effect size may be small, but almost never exactly zero (Cohen, 1990). This practice diminished the focus on the actual research questions which more often address the size and

³⁴ The exception being paper I which was the work originally initiating my project description, using data from a Nordic follow-up by TIMSS 1995.

³⁵ Alpha is the probability of committing a type I error. That is to erroneously conclude that the null-hypothesis should be rejected.

interpretation of the actual effects under study (the differences in the means, the correlations etc.) (F. L. Schmidt, 1996). Arguments and anecdotal evidence were presented to the effect that focusing on the alphas (the probabilities of having type I errors present) leads to the reporting of statistically significant effects of non-importance (that is when the sample sizes are very large), or alternatively, it could lead to the reporting of negative findings³⁶ where the effect sizes are actually moderate or high and carry substantial meaning (that is when the sample sizes are relatively small) (Cohen, 1990; Rosnow & Rosenthal, 1989); even worse, effects that might be of interest to peers might not be reported at all because they are statistically non-significant. The latter is associated with the lack of attention given to the statistical power of studies (Cohen, 1990; Sedlmaier & Gigerenzer, 1989). Very often the negative findings were due to the fact that the study did not have sufficient power to reveal the real effect. As a consequence the likelihood of committing type II errors³⁷ may in many cases be approximately ten times higher than that of committing type I errors (Cohen, 1994; Sedlmaier & Gigerenzer, 1989).

This debate is linked to the pragmatic position on methodology because what we would really like to report is the existence of meaningful effects. Most research questions are asking about the effects, and the significance level is answering the important, but still subordinate, question of whether the reported effect size can be trusted.

3.3 Marking rubrics and codes used for the items in PISA and TIMSS

Papers I and II represent two different approaches to studying students' responses to items by analysing the categorical information available. Before returning to the specific nature of the categorical information available in TIMSS and PISA, it is necessary first to say something about categories, their function and how they are formed, in general.

3.3.1 The nature of making categories

To develop categories is to classify. Classification means to group subjects or phenomena into objects, classes, clusters or categories. These objects are described so that subjects which are similar in some respect are represented by the same object, that is, criteria for judgement must be developed. Using the term object also hints at the fact that classifying has an objective. The objective helps define the criterion to use. In this respect, the original ontological entities, the subjects or the phenomena, are transformed according to some rule. Different rules can be used to classify the same phenomena, e.g. clothes can be classified according to their main colour, or they can be classified by their function: formal dress; protection against cold/heat; a symbol of group membership; a uniform;

³⁶ Negative findings refer here to results where effects are reported as non-significant.

³⁷ Type II errors refers to erroneously concluding that the nil-hypothesis (no effect) is "true", or in a less bombastic form, to conclude that there is not enough statistical evidence supporting the research or alternate hypothesis (that there is an effect).

training equipment, etc. None of the rules are truer than others; however, some may be more useful given a specific purpose. A classification therefore has to be judged according to its usefulness in context (Everitt, 1993).

Another aspect of classification schemes demonstrated by the simple example given above is the fact that they have some generic properties which can vary from one scheme to another. Some of the generic properties for systems of classification can be identified in the example on clothing given above. These and some other generic properties are identified below and some comments are offered on how these generic properties affect the analysis of the information contained in the categories.

- I. *Mutually exclusive categories or not.* The first rule in the example above (classification according to the main colour) contained a key to group the phenomena or subjects (the clothes) into mutually exclusive categories. As a consequence each subject will be found in only one of the categories. From an analytical perspective this means that the rule can be transformed into one single variable. In the second rule in the example on clothing we might agree that this classification key does not necessarily lead to mutually exclusive categories. It is obvious that most clothes may have multiple functions at the same time and also depending on the context in which they are used. This rule would therefore not be easily transformed into a variable
- II. *The degree to which information is reduced.* Classification is used to reduce or simplify the properties of the phenomenon, in order to be able to organise, analyse or retrieve information from a large data set. The number of classes can vary substantially. In the first rule above we can imagine versions with just a few categories (warm colours vs. cold colours), or quite a few (using a fine-grained palette of colours). From a measurement perspective the degree of reduction will affect the analytical potential of the associated variable, e.g. age as categorised by year or by larger intervals. Depending on the number of categories used, statistical parameters will change. In many cases the choice of how many categories that should be used, would be determined by arguments relating to feasibility. If a phenomenon is studied by coding of manual observations, a very fine-grained categorical system would be difficult to use and would require more time and thus it would imply higher costs than a less detailed system.
- III. *Complete system or not.* For most purposes a system is needed which can categorise all possible phenomena under study. Such a system could be characterised as complete or finite. In the example above, the second rule could be made finite by giving an exhaustive list of different functions of clothes; alternatively it could be done by giving an exhaustive list of main categories and lists of all subcategories contained in this; or it might be made finite by including a number of main categories and a last non-specific category labelled, for instance, 'other functions'. In most practical situations the latter solution is preferred, since a complete system would include too many categories and,

consequently, be useless in practice. From an analytical viewpoint the use of an 'other' category implies accepting a certain amount of missing information in the data. The 'other' category should therefore not be too large. An example of a system which is (certainly) not complete is this very list of generic properties of classification systems which is based on my own limited experience and imagination.

- IV. *The degree of precision in the criteria.* For scientific purposes classification systems have to be defined in terms that can be operationalised so that a single user of the system is able to use it consistently, and so that different users classify the same subject into the same category. From a measurement perspective, this is of course closely related to the concept of reliability. In the example with colours above, it is likely that the same user will be able to use the system in a consistent manner. Also, different users are likely to use it quite similarly. However, we can imagine examples of colours which would be classified differently by two different people, e.g. the colour turquoise could be called green by one person while another person perceives the colour as blue. Also, as the example of colour illustrates, the measurement conditions have to be specified, because the apparent colour is highly dependent on, for instance, the light source.
- V. *The degree of structure in the system.* Some categorical systems assume that the phenomenon has a structure or mirrors a structural model or theory used to describe the phenomenon. This implies that the categories can be ordered according to some ranking criteria, and/or they can be organised into a hierarchical system or a more complex network describing a relationship between the categories. From a measurement perspective some of these systems can be transformed into ordinal variables. The example with age groups given below illustrates a categorical system which represents an ordinal variable. Other examples of systems with hierarchical structure are obviously taxonomies for living organisms or systems used for organising archives, such as the Dewey decimal classification system for literature used in libraries.
- VI. *The degree to which the system is based on theory or empirical observation.* Some categorical systems may be based solely on theory. A good example from physics is the standard model for the elementary particles, a system based on theoretical consideration before the particles were actually observed. An example of an empirically based classification system is the classification of stars into red and blue giants, main sequence stars, white/brown/red dwarfs etc. This was a grouping resulting from empirical observations of temperature and light intensity for the stars. Of course, theoretical statements are used in any empirical observation, and probably vice versa. The main point here is that the relative importance of the two can vary.

Another important consequence of classification worth mentioning is that the process often implies naming of the subjects. Indeed the creation of language can be seen as a process of classification. Every noun in any language represents a

class of subjects (Everitt, 1993) and most verbs represent classes of actions. According to the purpose and criteria, a name carries meaning and connotations. The ‘same’ classification scheme can exist in different versions using different words. For instance a classification of age can be done by defining some intervals, such as <6-12 years>, <12-19 years>, ... <67 years or more>; alternatively the classes can be named as ‘young kids’, ‘youths’, ... ‘elderly’, or ‘kids in elementary school’, ‘kids in secondary school’, ... ‘retirees’. All these groupings might have the same or very similar criteria, but they communicate differently and the choice of name is therefore an important part of developing a classification.

3.3.2 The general nature of the codes and marking rubrics used for the items in PISA and TIMSS

I will now return to the nature of the categorical information in PISA and TIMSS. This section gives a general discussion, before the actual coding systems are presented in section 3.3.3.

Students’ responses vs. students’ knowledge and thinking

Students’ responses to the cognitive items used in PISA and TIMSS have, in the end, received a score evaluating the quality of the answer. Most items are scored dichotomously, but some are scored by multiple score points. When these score points are combined into a total score, each score point is assumed to contribute to the measure of the common trait measured by several items, and this assumption is thoroughly tested at several stages: first through the initial screening of the items (both substantial evaluation and empirical testing), and then in an extended field trial in all participating countries. The assumption is also tested in the final data. If this assumption is violated to some specified degree, the item will be removed from the pool at any of these stages. This is to underline the fact that the main purpose of the single items in PISA is to produce an overall reliable score.

However, as discussed in chapter 1, before the responses have been attributed to scores reflecting their appropriateness or correctness, they are punched into the data files as codes. These codes mostly represent categories that do not have a relational character beyond the fact that they are mutually exclusive, that is, one student can only be categorised by one of the codes for each item. Papers I and II in this thesis study such nominal variables, and, in particular, paper II aims at studying how the different categories used in two different items relate to each other. It is therefore necessary to devote some space to a reflection on the nature of these categories. Included in this reflection is the history behind the specific coding system used for marking many of the constructed response science items in PISA, the so-called double digit rubrics.

Central in this reflection on the categorical nature of the cognitive variables, and the nature of making the categories, is the concept of *information*. This is a very general term and may have different meanings in different contexts. Here the point is that a category carries information that describes qualities, meaning that its quantity is only part of the information (see also Figure

3.1). This information may be perceived as real in a referential sense; it refers to a real category in the category system used, and this category in turn refers to a real student response. However, to state that the category reflects students' thinking and knowledge (as is done several times in this thesis) is not so straightforward. It is important to add that all attempts to move from the student response into the students' minds involve *constructing* what is in the students' mind. Several studies have shown, for instance, that the choices that students make on selected response items, or the responses they construct, do not always give a good description of the students' knowledge and thinking (Clerk & Rutherford, 2000; Harlow & Jones, 2004; Schoultz, 2000). In general, when stating that the items can be used to say something about students' knowledge and thinking in science, this refers to the categories used to represent the students' responses. In light of the above discussion it is important to note that the categories are taken to be *indicators* of the students' thoughts and knowledge representations *in the context* of the item. However, as is also one of the main conclusions in paper I, to generalise from their responses to one single item to more general statements about their abilities or knowledge, is hardly possible.

The items used in the PISA cognitive test booklets and questionnaires have different formats. The items used in the test booklets can roughly be split into two types, selected response and constructed response (CR). As the terms indicate, the students select appropriate answers in the former type, while they have to write an answer themselves in the latter type of items.

Selected response items

There are two selected response formats used in PISA. Firstly, the well-known multiple choice (MC) format is used. In this format the student is asked a question and is given a number of alternatives, of which only one is rated as a correct answer, while the others are referred to as distracters. In PISA the majority of the MC items include four possible answers, one correct and three distracters (see for instance Figure 1.2 in chapter 1). The other selected response format used in PISA is the vector format. The term vector refers to the fact that in these items the student is asked to evaluate two or more statements, usually by the selection of one out of two given choices: *Yes/No*, *True/False*, etc. The response is therefore a vector of several responses.

The selected response items are not manually marked. They are punched directly into the data file. The data punched reflect *the actual choices* made by the students, not just whether they are correct or not. In the subsequent processing of the data these initial variables are transformed to variables containing score points. However, the presence of the actual choices in the original data file makes it possible to study the properties of each single distracter, or each single point in a vector item. These data are used, for instance, to evaluate the assumption that each single item contributes to the overall scale, that is, to evaluate the quality of the item from a test perspective. Simply stated, a high quality MC item will have a correct response chosen primarily by students who get high scores on the test, and consequently, the distracters in a good MC item will be preferred by students of lower abilities.

The fact that the actual choices of the students are present in the punched data makes the corresponding variable categorical or nominal. The categorical information represented in these variables is, in this thesis, considered as a good source of information that may be analysed in order to make more specific inferences about the characteristics of students' knowledge and thinking, in the kind of science they meet in the PISA items.

Constructed response items

The CR items can roughly be split into two types. The first type is called extended constructed response, a format where the student has to write an argument, a description, a conclusion etc. over one or more lines. Another type can be labelled as short constructed response. For these items the students respond by giving a single word or a single number. All science items with this format had to be manually marked, but for some mathematics items it was possible for the software used to automatically score the responses to some of these short constructed response items by the actual response given, that is, the puncher typed students' responses as they were. In the following, both these formats are referred to as CR.

Before returning to how these items were marked or coded it is useful to discuss more specifically the way in which different factors have affected the properties of the categorical systems used in assessment projects like PISA and TIMSS. This discussion will draw on the general descriptions of categorical systems developed in section 3.3.1. In general, relatively detailed marking guides are used in the manual marking of items in PISA and TIMSS. These guides consist of the *codes* used for each single item, and a description of characteristics for the responses that should be assigned the different codes, also referred to as *marking rubrics*. In developing these rubrics, assumptions are made and some practical conditions constrain how the system can be developed:

- A. It is assumed, and indeed demanded, that it is possible to group non-identical answers with common descriptors. This is, I guess, not always a very strong assumption. Even though two answers are visually different it is evident in many cases that these responses refer to the same quality. Imagine for instance responses to a question asking about what gas is needed as an input to photosynthesis in plants. The response "carbon dioxide" is in this context obviously equivalent to the response "CO₂". Of course, there are items where the hypothetical response universe is much larger than in this relatively simple case. Most science items in PISA, for instance, have a much larger universe of possible responses. In conclusion, therefore, the degree to which this demand can be met varies from item to item.
- B. Student responses will be coded by many markers in many countries. In relation to the generic properties of categories listed in section 3.3.1 above, the characteristics of the PISA marking rubrics must have the following properties: (I) It has to consist of mutually exclusive categories. (II) The marking has to be efficient, due to cost and time which limits the number of categories which can be used. (III) All responses have to be coded, which implies that the system must be complete or finite. For most items this means

that ‘other’ categories have to be used. (IV) The marking rubrics must be so precise that a quite diverse group of people can have a very similar understanding of them. Points I-III above also contribute to this. (V) It must be possible to distinguish the categories of good quality deserving credit from those of lower quality (see more on this subject in point C below). (VI) It is not easy to agree on how to mark a specific response if the marking rubrics are written with very general statements. The marking rubrics therefore have to be based on actual student responses with real examples to be operational and reliable. In PISA this is operationalised by including several specific examples for some of the codes to illustrate typical responses. Empirical testing is therefore essential in order to develop the categorical systems used. However, as I will return to below, parts of the information in the code reflect how the response should be scored, and this must to some degree be judged by theoretical considerations.

- C. The codes used for marking constructed responses in many studies, such as PISA, have to carry information about how the responses should be scored, at least if the item is part of a test with the aim of producing a test score. The quality intended to be reflected by a category has to be judged to be high/good or low/bad, or as this dichotomy is usually stated, as correct or incorrect. This is also a limitation that affects how the final categorical system will look³⁸. This will also impose a structure on the categorical system which will, to some degree, produce an ordered system. However, from a purely diagnostic perspective it would be more appropriate to make categories describing some characteristic in the response, regardless of its correctness or quality. In tests intended to produce a score this is of course not possible.

There is always a limited amount of time and money with which to make up the marking rubrics. The system used for the CR items in PISA is developed by the repeated reading of a selection of students’ responses; at least this was the way it was done for the development of science items for use in PISA 2006 by some of my colleagues and myself (PISA Norway, 2004). The development of the categories is done as an iterative process, where at the outset a rough expectation for a high quality answer is described. While reading real responses, distinctly different versions of this quality are discovered. Some initial criteria regarding these distinctions are established. If the criteria seem to be successful (i.e. they can be used to separate responses into groups) an elaborated description of the criteria is written down. By further reading of student responses, the *stability* of the criteria is verified, and when all responses seem to fit one of the established categories, the categorical system can be described as *saturated*.

³⁸ Even more importantly, this limits the universe of items that could be included in the test. If, for instance, an item intends to reveal some predefined structure in students’ thinking, the marking rubrics would not necessarily include judgements about the quality or correctness of the responses.

The reduction of information: a specific example

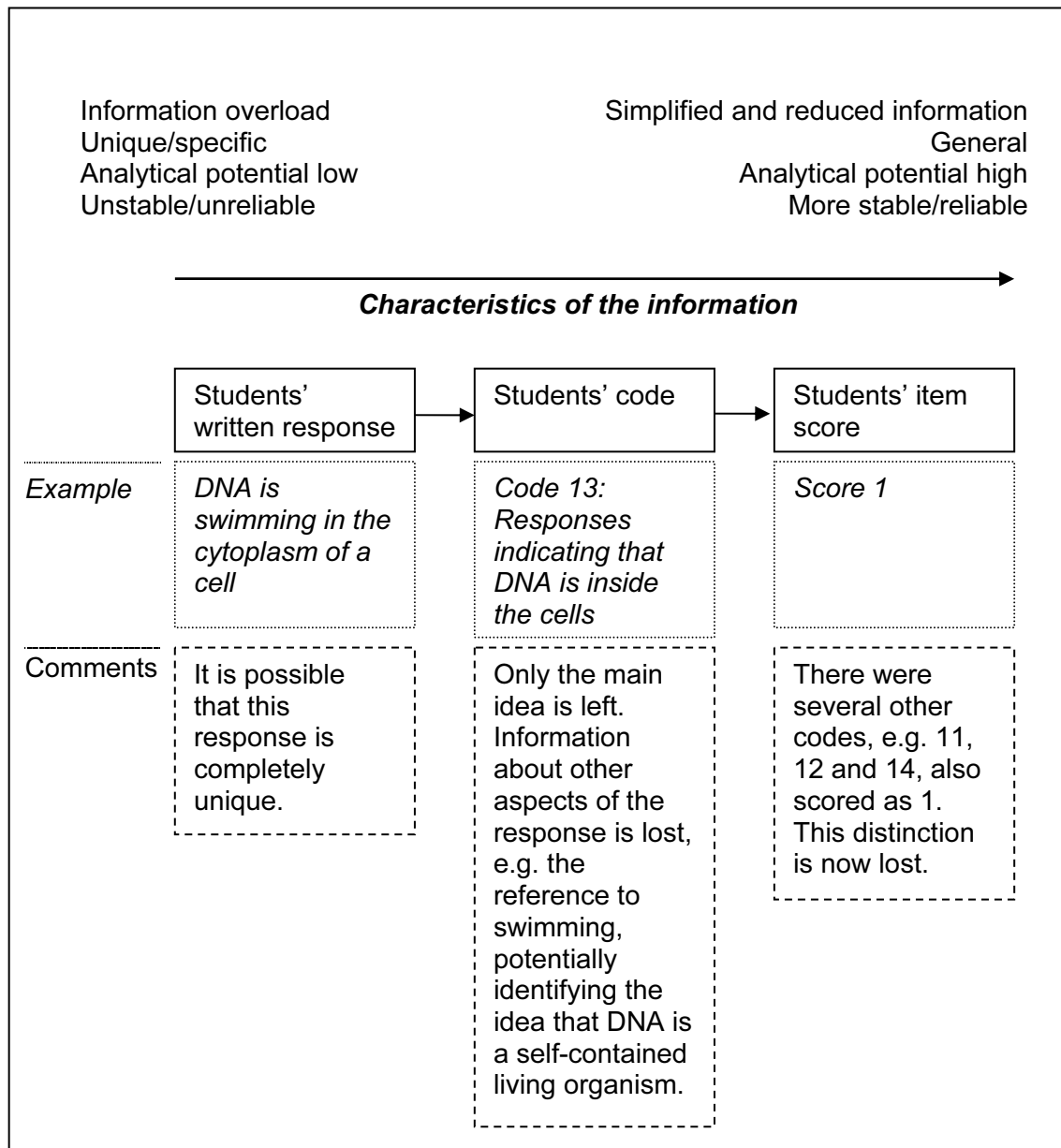


Figure 3.1: Illustration of how information is reduced when constructed responses are coded and subsequently scored.

Before giving a presentation of the marking systems and codes actually used in PISA (and TIMSS), we should first consider how information is reduced in the process, going from a specific student response, via coding of the response, to the scoring of the response. The example in Figure 3.1 will be used to discuss this process. In this figure a hypothetical example of a response to an artificial item is given. Although the hypothetical question asked is not stated, we can see from the student response that this item asks the student to localise DNA in the body. Furthermore, given that the response receives credit despite the fact that some formal errors are made (the response locates DNA inside the cell, but on the outside of the nucleus), this hypothetical item can also be seen as an example of

the use of partial credit scoring. In this case there are probably other codes referring to 'full credit' where the marking rubric would require that the DNA is located inside the nucleus of the cell. It will be clear to the reader familiar with the PISA marking procedures, that the progress from a response to a score for this particular hypothetical example could have been an example from PISA. For the reader who is not familiar with the marking and scoring of items in PISA and TIMSS, a more detailed presentation of the procedure is given in section 3.3.3, and a number of examples of items and marking rubrics from PISA is given in Appendix 1.

Figure 3.1 mainly illustrates that in the transformation of an actual student response, first into a code, and then into a score point, a reduction of the available information occurs. This is indeed intended, and necessary in order to produce data that can be analysed. This is shown by the arrow in Figure 3.1, which indicates how characteristics of the information change in the process of coding and scoring. It is simply not possible to analyse all the information contained in the specific responses given by several thousands of students on several items. This can be characterised as a problem of *information overload*. Each student response might, at least for items that require a relatively lengthy response, be *unique* for each individual student, and at the outset the information is therefore highly idiosyncratic of both the item and the individual students. All responses therefore include facets of information that in the end are considered irrelevant to the coding and the subsequent scoring.

Going even further, these responses are not just unique in the sense that each student gives slightly different answers. They are also unique, in the sense that in a hypothetical test-retest situation it is highly unlikely that the same student will write down exactly the same answer. To sum up, the information contained in one single response is unique for the student for that particular item responded to at a specific time or occasion. Nevertheless, in such a test-retest situation it is more likely that the same code will be used for the same student, and even more likely that the same score will be given to the same student. In this respect, the information is considered as becoming more *stable* or *reliable* as the students' written responses are transformed to more general codes, which in turn are transformed to even more general scores. The analytical potential is therefore considered as higher for the scored items.

On the other hand, and this is not represented in Figure 3.1, the analytical potential in the students' written responses may be perceived as higher than in the coded and scored information because some analysis may focus on other aspects of the responses. For instance, this could be a more qualitatively oriented secondary analysis, coding facets in the response that are considered irrelevant or redundant given the scoring criteria; and as such this information is consciously reduced or eliminated in the coding/scoring process. In secondary analyses of data from PISA, the researcher can choose what level of information is most suited for analysis according to the research question stated. Using, for instance, the information preserved in the nominal codes means using information that has been severely reduced as compared with the students' written responses.

Nevertheless, it contains a lot of detailed information compared to the scored variables.

3.3.3 The codes used in PISA and TIMSS– the double digits

After discussing the general aspects of the items in PISA and TIMSS, the different formats and how, in general, they are marked by using a detailed marking guide, we will now give a specific description of the marking rubrics used for CR items in PISA, starting with a historical account of the roots of the system, which is found in the TIMSS 1995 study.

The development of the ‘Viking’ rubrics in TIMSS

With TIMSS 1995 LINCAS started using constructed response items, or free-response items as TIMSS called them, to a greater extent than before. The reason for including such items was mainly to assess outcomes of mathematics and science education that were not possible to assess with MC items (Orpwood & Garden, 1998; Robitaille & Garden, 1996), but was also to make use of the added analytical potential provided by students’ responses in their own words (Lie *et al.*, 1996). However, this potential was also followed by some challenges, or, in the words of Taylor (1993, p. 1):

The inclusion of free-response items in the international item pool provides an opportunity to collect a rich source of data related not only to levels of student achievement but also to the method used by the students in approaching a problem, and to the misconceptions or error types which may be demonstrated by them. Inherent in the collection of these data, are issues of reliability and need for additional resources in the coding process. (cited from Lie *et al.*, 1996, p. 7-1)

The issues of reliability and constraints related to costs and time had to be addressed in the design of a coding system. Within these constraints and limitations, a generic categorical system for marking constructed responses was developed for TIMSS 1995.

The work began early in the 1990’s, with a quite complex system focusing on several facets in the student responses. For most facets the codes used were generic, thus using the same, or at least a very similar, set of codes for all items. As a consequence of several trials, this initial generic system was abandoned and a simpler system was developed, focusing on two facets only: the correctness of the answer (score) and another facet capturing the specific approach (strategy, response content or missing content, error type or misconception) taken by the student. This last facet emphasised specific aspects in the students’ responses.

As a consequence, the specific marking rubrics formulated were to a high degree idiosyncratic for each item, but the codes used followed a generic system that was the same across all CR items. This system was initiated and developed mainly by the Norwegian TIMSS team (Alseth *et al.*, 1993; Angell, 1995; Angell *et al.*, 1994; Angell & Kobberstad, 1993; Kobberstad *et al.*, 1994) and is therefore referred to in many TIMSS documents as the ‘Viking-rubrics’ (e.g. Orpwood & Garden, 1998). The final generic system is presented in Table 3.1.

For each item, specific marking rubrics were constructed according to this system. Each country could also include their own codes for specified types of

correct or incorrect responses. These codes were in the international data file recoded to 19 or 79 (unspecified responses). Some of the items had more than one score point, which meant that the system had to be extended by codes 20, 21,...,29 (and in some cases even 30, 31,...,39) following the same system. By using this system, the data carried both score information (first digit 3 corresponds to 3 score points, first digit 2 to 2 score points, first digit 1 to 1 score point and first digit 7 or 9 corresponds to 0 score points) and a code, the second digit, preserving qualitative characteristics of the response. The number of codes was kept fairly low, in order to be manageable. The rule adopted was that a code had to capture approximately 5% of all students in the international field trial to be worth having as a separate code (Lie *et al.*, 1996). The sequence of the codes within the rubrics for each item was usually ordered so that the most common correct response type came first in the group of correct responses, and similarly with the incorrect responses.

Code	Description
10	Correct response, type 1
11	Correct response, type 2
12	Correct response, type 3
13, 14, ...	Other specified correct responses
19	Unspecified correct response
70	Incorrect response, type 1
71	Incorrect response, type 2
72	Incorrect response, type 3
73, 74, ...	Other specified incorrect responses
76	Incorrect response, information in stem repeated
79	Unspecified incorrect response
90	Crossed out/erased, illegible, or impossible to interpret
99	Blank

Table 3.1: The generic system for marking free-response items in TIMSS 1995. Adapted from Lie *et al.* (1996, p. 7-7).

Double digit scoring of science items in PISA

In essence this system was adopted by PISA 2000 for the science items. This time the codes used were also heavily influenced by the initial work done by the Norwegian PISA team (Kjærnsli *et al.*, 1999b). In the final version there were some minor adjustments as compared to the TIMSS system. The PISA system is presented in Table 3.2. As for TIMSS above, some items in PISA were scored by several score points, and as such the system had to be extended with more codes.

Code	Description
01	No credit, type 1
02	No credit, type 2
03, 04, ...	No credit, type 3, 4, ...
11	Full credit, type 1
12	Full credit, type 2
13, 14, ...	Full credit, type 3, 4...
99	Blank

Table 3.2: *The generic system for marking constructed response items in PISA 2000.*

The codes for PISA were not labelled as ‘correct’ or ‘incorrect’. Instead they were labelled as ‘full credit’ and ‘no credit’ (and for items with more than one score point the label ‘partial credit’ was used). This reflects the fact that in many items the notion of correctness can be misleading. For instance, when asking the student to identify the evidence for some claim, students often gave statements with the scientific explanation for the phenomenon under study in the item/unit. This represents correct scientific knowledge, but is not necessarily a good response to the question asked (e.g. see the Semmelweis’ Diary unit, question 1, code 01 and 02 in Appendix 1).

Furthermore, some rubrics are left out in the PISA system as compared to the TIMSS system. Most important in this respect are the unspecified correct answers (or unspecified credited answers as they would have had to be labelled in PISA). The credited responses in PISA had to be captured by one of the specified codes in the marking rubrics, and codes referring to ‘unspecified credited responses’ were not included in the system. Through my own experience, as one of the markers in PISA, the lack of such a code was sometimes strongly felt, since it happened that responses that were obviously of a good quality had to be marked by a code referring to no credit.

The reason for not including such a code is not clearly documented, but it is most likely related to the aim of achieving consistent marking across countries. When browsing through the percentages of students in the respective countries receiving this code for TIMSS items, it is evident that countries to some extent have developed idiosyncratic notions on what an ‘unspecified correct response’ is (code 19 in Table 3.1). The proportions varied substantially between countries. The rubric for unspecified incorrect responses was kept in PISA, but not in a preset generic code as in TIMSS. The system adopted by PISA was that for most items, when all the specified non-credited responses were formulated, the next code at the ‘0X’ rubric level would be labelled ‘other non-credited responses’ (e.g. see the Semmelweis’ Diary unit, question 1, code ‘04’ in Appendix 1). In addition this category included any type of responses that could not be given credit, including responses completely off task as those coded by ‘90’ in TIMSS (see Table 3.1). The category ‘other non-credited responses’ in PISA is therefore

an aggregate of many different types of answers. In addition, the specific code 76 for those responses where only the stem was repeated was left out, probably because it was not overly used in TIMSS 1995. This code is also left out in the TIMSS 2003 marking rubrics.

In addition, both the TIMSS and the PISA data files contain some computer-constructed codes, such as code 97 given in PISA for students who were not administered this particular item. This is due to the rotation of booklets applied in both PISA and TIMSS. This implies that all students do not answer the same set of items. Another computer-generated category with associated code is the category 'not reached'. Basically, this is constructed by identifying a series of non-responses starting from the last item in the booklet and going backwards, the assumption being that items not responded to at the very end of the booklet represented items the student did not have time to read and respond to. In this way the extent to which the test was speeded³⁹ can be estimated. This code is also useful in the scaling of the item parameters: the assumption is that students who do not reach an item can be treated as if they had not been administered the item at all. These responses are therefore not included when computing the item difficulties. However, the responses coded as 'not reached' were scored with zero points and included in the estimation of the student scores.

Double digit coding and overall test quality

One of the reasons why double digit coding was well accepted by most people involved in the decision making in PISA and TIMSS might also be related to other advantages these codes gave. Many people involved in LINCAS are not primarily interested in the substantial diagnostic information represented by the double digits. However, they could easily accept this system, due to the positive impact these codes potentially have for producing a reliable overall score in the domains tested, which after all was the primary goal of both TIMSS and PISA⁴⁰.

The first argument is that the markers will, in any case, need a quite detailed list of response types, in order to score the item properly and to achieve an acceptable inter-rater reliability. Giving a specific code to the main response types was therefore not considered to introduce substantially more work. It could even be suspected that using such codes would enhance markers' attention to the criteria, and thus achieve even better inter-rater reliability. However, any firm evidence for this does not exist.

In addition, this would make it possible to keep more information for later analysis. This detailed information could, for instance, be used to identify whether all the responses receiving credit should really do so. When, for instance, more than one score point is used to score an item, this detailed

³⁹ For some readers this may seem to be an odd property of a test. That a test is speeded simply refers to tests where response time obviously is a limitation. Some tests are intentionally speeded because the ability of the respondents to complete within a time limit is part of the defined construct. Tests like TIMSS and PISA are not supposed to be speeded. Furthermore, in an international comparative assessment including reading of materials, we could also suspect that the test may be differently speeded in different languages. It is therefore important to document this property of the test.

⁴⁰ However, the diagnostic aspect was an important primary research goal for TIMSS, and the use of double digit codes could therefore be even better argued for within the TIMSS framework.

information can be used to see whether all types of answers really deserved 3 points, 2 points or 1 point. In other words, keeping this information makes it possible to produce an even more reliable overall score, by collapsing codes and recoding score points. In this way, the scoring may be optimised to make the items discriminate better. This is of particular importance for the field trials. As a consequence of the field trial several marking rubrics have been revised to improve the quality of the marking. Of course, any recoding of score points has to be done with care, and should in general be accompanied by substantial analysis of the actual response types to be collapsed. In general, such recoding was done with just a few items in PISA 2000 and 2003.

Sadly, there is now a marked trend for double digit codes to be gradually phased out in PISA. In PISA 2000 a total of 14 science items were manually marked and 12 of these were marked by double digit coding. The units publicly released after the 2000 survey (see Appendix 1) were replaced by new units in the survey administered in 2003, and all the new units had marking rubrics that only captured the score information; and as a consequence, only 8 of the 15 manually marked science items were coded by double digits in PISA 2003. In the field trial for PISA 2006 implemented in the spring of 2005, where science was the major domain, only 26 of the 95 manually marked items have coding systems with double digits, and furthermore most of the new items with a double digit system are very short systems, typically having only one code for the correct response and two types of non-credited responses. This means that the marking rubrics are gradually transformed into lists describing answers that should receive credit and similar descriptions of responses that are not credited. Or, referring to Figure 3.1, the intermediate code level is skipped, and markers could therefore rather be labelled as scorers.

Once more, the arguments behind this development are not very explicitly given in documents, but rather raised as concerns at different kinds of meetings. The main argument is probably that several countries have expressed their concern that using double digit coding costs more than using only scoring by a single digit. However, this argument lacks any substantial empirical documentation. Furthermore, it has been argued by several that since these codes are not used anyway to any extent, we do not need them. In addition, the marker consistency at the code level is very low for some items, which means that the information contained in the codes has a limited analytical potential.

I would claim, in light of the discussion in the next section, that this is a very unfortunate development, which would seriously reduce the potential for using PISA data in secondary analysis targeting science educational research questions. The most unfortunate aspect of this is that the marker, when confronted with a student response, in most cases has to categorise the response according to the typology of answers described in the lists defining the scoring criteria. In other words the information at the code level (see Figure 3.1) is there, for a split second at least, but in the very next moment when the score is written down, this information is lost.

CR items are typically included in order to have items that test students' ability to argue and reason, and to communicate their thinking. Also, with the

inclusion of constructed response items it is possible to have items with several acceptable response types, while in MC items only one acceptable response is usually included. An interesting paradox is that when double digit codes are not used in the marking of CR items, only the score information is kept, while in contrast, for MC items more information is preserved, since the choice of the student is kept as a code in the originally punched files. Consequently, CR items that were included to allow for a larger variety of student responses, are in the end the items providing least information about students' thinking and knowledge in science.

3.3.4 The use of double digit codes in analyses: some examples

Even though the diagnostic aspect was emphasised in TIMSS (e.g. Lie *et al.*, 1996), this was not followed up in the international reports. It has also only to a minor degree been followed up by others in secondary analysis based on the data. Some of the cases are, however, promising and illustrate that the information preserved in the double digits is potentially useful for analyses. In the following discussion, some examples of how this information has been used will be given.

Kind (1996) analysed the performance assessment units in TIMSS 1995. This was a study involving students in practical hands-on activities (Harmon *et al.*, 1997). They were given some simple equipment (for instance a tablet to be dissolved in water) and a set of written questions to answer. These questions were organised into units resembling PISA items in style. Students' written responses were analysed, using the system of double digit codes presented in Table 3.1. Using this information Kind (1996) was able to give a comprehensive and detailed report of aspects in the students' responses, item by item. He also investigated how items worked as measures of some general skills across items in different units, using the score information only. His finding was that the correlations between items within units were much higher than those between items across units intended to measure the same skill. In his conclusion, he addresses this finding by a set of remaining questions, and states that

All these questions point towards an analysis based on types of responses, i.e. discussing the responses at code-level, rather than score-level. (Kind, 1996, p. 251)

In other words, despite the finding that the general skills were not present as traits in the scores across units, such general skills might be studied by using the categorical information. However, neither Kind nor anyone else performed this analysis, so the issue remains open.

In a study of physics items from TIMSS 1995 (Mullis *et al.*, 1998) Angell (1996) utilised the double digit information to study whether students' misconceptions or alternative conceptions are consistent and theory-like, or whether they are fragmented and intuitive. Through the use of both MC items with diagnostic distracters and CR items with diagnostic information preserved by the double digit coding, he was able to conclude that students' conceptions in physics are more consistent with the view that they are fragmented and largely intuitive, heavily dependent on, for instance, the context. By the score information alone this analysis had not been possible.

Turmo (2003b) studied one of the released science units (four items) from PISA 2000 related to the depletion of the ozone layer. In this work he utilised the information at the code level, combined with actual student responses, to explicate what PISA assesses and how students are required to respond. Using the information at the code level as a communicative tool is probably the most widespread use of this information, occurring for instance in national reports and in a popular version of the PISA framework (OECD-PISA, 2002b). Furthermore, in his work Turmo (2003b) related this information to students' self-reported use of learning strategies, and he also studied the profiles across the codes for students in different predefined regions.

In secondary national reports aimed at teachers in Norway, the released items in science (Kjærnsli *et al.*, 1999a) and mathematics (Brekke *et al.*, 1998) from TIMSS 1995 were presented with results at code level. The results were accompanied by discussions focusing on the diagnostic aspects of the response types. The reports were organised thematically, and as such these publications could function as tools for teachers who would like to be better informed about some common elements in students' understanding of some of the concepts that are central in the curriculum.

The four examples given above are all Norwegian, and many more secondary studies from Norway related especially to TIMSS 1995 could be given. Common for all these examples is that the categorical information has been presented item-by-item, followed by more holistic views of the total information. The aim of paper II in this thesis was to use the categorical information more analytically, using tools for studying relationships between categories within and across items.

3.4 Analysis of nominal variables

This section is primarily written in order to give the background for understanding the most central aspects of the methods used in paper II. In revising the originally submitted article, it had to be shortened. The easiest way to make the text shorter was to assume that the reader was familiar with the methods used, and if not, that a brief text on the methods including references to further reading would be sufficient. I therefore find it necessary to include this section presenting the methods used in that paper in some more detail.

First in this section, there is a general introduction to how categorical variables may be treated. Second, a presentation of correspondence analysis (CA) is given. In the final version of paper II this analytical tool was not explicitly used, but there is a reference to the overall results obtained from CA in the paper. However, the more relevant reason for including a description of CA here, even though it is not extensively used in the empirical works included in this thesis, is that homogeneity analysis (HA), has been shown to be equivalent to multiple CA (Greenacre, 1993; Heiser, 1981), and therefore an introduction to CA is also an introduction to HA. In my view, understanding the aim of HA, as it has been used in my work, is better achieved by first taking a quick peek at the principles of CA. Having done that, this section presents HA, the method explicitly used in

paper II to study the association between categories within and across several nominal variables.

3.4.1 General introduction

In educational research most variables are categorical. Some of these may be more or less directly related to an assumed underlying continuum, while others are genuinely qualitative in the sense that the underlying construct is in itself discrete. Some of these discrete qualities are ordered, while some are not. The purpose of this thesis is, as stated before, to analyse information about qualities that by their very nature are discrete and not ordered in any sense; in other words, the variables measuring these qualities are nominal variables. In large-scale international surveys such as PISA, some examples of such variables are gender, country, school ID and the cognitive items coded by response type as presented in section 3.3 above.

In the standard presentation of, for instance, the PISA data, variables of this kind have been used as grouping variables, e.g. to compare differences between countries, between schools or between boys and girls. This is equally true for most educational research based on measurement.

In other kinds of research, such as sociology or biology, the constructs of interest are to a large degree measured by nominal variables. As a consequence there are many research questions in these disciplines related to the nature of the association between these nominal variables and the categories contained within them. Not surprisingly, methodological development to handle nominal variables is mostly implemented within these disciplines.

The simplest way to describe nominal data is by counting the number of respondents in the respective categories. This can give descriptions that by themselves are informative for some research questions. The analysis of single items is an example of such a procedure. By counting how many students circled each of the given alternative answers in a multiple choice question, a summary description of the total profile for the sample is given. Often we would like to relate this information to other variables. In projects like TIMSS and PISA, this has been done by using the assigned codes as a grouping variable. A comparison of the overall achievement in the tests for students in these groups is used in the test development procedure, as an indicator for how well the codes discriminate between students with different achievements. Such analyses are studies of the association between the single categories in a nominal variable and a continuous variable (the overall score). In paper II the focus is on the association between the codes used in two or more items, or in other words, the association between several nominal variables.

In the case of the association between two nominal variables, the most straightforward task is to construct a crosstabulation of the two variables. Such tables are also often referred to as contingency tables or correspondence tables. By the row and column totals only, it is possible to compute the expected count in each cell when assuming that there is no relationship between the variables. By summing up the squared differences between the actual count and the expected count, we get a measure of the total deviation from the expected. This is

the χ^2 -statistics used as a basis for numerous statistical measures of association or model fits. The measure of overall association is given by the ratio of χ^2 and the number of degrees of freedom (which equals $(n - 1) + (m - 1)$ where n and m are the number of categories in the respective variables). Some research questions seek to develop a more detailed description of this association. In this thesis, for instance, one of the underlying assumptions is that there exist patterns across items, patterns reflecting that students giving a specific type of response in one item will tend to also favour one or a few other response types in another item. If the number of categories for each variable is small it is relatively easy to describe this nature in a qualitative sense by simply inspecting the crosstable. However, when the number of categories becomes larger, or more variables are included in the analysis, it is difficult to extract the main trends in the material.

In a book about graphical representations Bertin (1981) addresses the issue of what characterises successful visualisations of a data set. The very general starting point for this reflection is that information is about relationships in the data. His concept of information can, however, be transferred to all kinds of multivariate analysis, not only those related to graphical representations of the data. Bertin refers to *three levels of information* in a multivariate set of data, each level targeting different types of questions. In the case of a crosstable, the first level of questions are related to the information in each cell: how one category in one variable is associated with a category in another variable. An example of a level 1 question regarding the analysis of how the codes on two items are related could be: “How many of the students coded 11 in question 1 are coded 99 in question 2?” At level 2, questions are formed that imply a reduction of both dimensions of the table; in other words, these questions are directed towards larger trends in the material. In the case of the items this could for instance be: “How many of the students receiving credit on question 1 omitted question 2, and how does this compare to the students who did not receive credit on question 1?” Finally, information at level 3 is related to the overall pattern in the table, e.g.: “What is the overall correspondence between students’ responses on question 1 and question 2?” By using CA and HA it is possible to study the type of information that Bertin (1981) labels as levels 2 and 3, while questions at level 1 are easily answered by the crosstabulation itself.

3.4.2 Correspondence analysis

Correspondence analysis (CA) is the statistical technique most commonly used to study patterns across nominal variables. The aim of CA is to transform a contingency table into a graphical representation. In doing so, the aim is to reduce the complexity of the original table, and to provide a representation of the data that facilitates interpretation (Clausen, 1998). The derived equations and the concepts used are therefore geometrical by nature. I have chosen to give a presentation of this method without including any mathematical description. The mathematical description of correspondence analysis is formulated in matrix notation which is largely unfamiliar to most educational researchers. A comprehensive source including a full account of the mathematics is given in Greenacre (1983), and a compact presentation of the mathematical formulation is

given in Blasius & Greenacre (1994). I have chosen instead to include a short qualitative description of the key ideas behind the technique, using an example of two items, questions 2 (Q2) and 4 (Q4), from the unit Semmelweis' Diary (see Appendix 1). The description of the method is in general influenced by the reading of a number of books introducing the method (Clausen, 1998; Greenacre, 1993; Hjellbrekke, 1999).

Table 3.3 shows the crosstable giving the distribution of the pairs of choices the students selected for these two MC items. All students administered this unit in all countries are included in this analysis ($N = 58\,025$)⁴¹ with equal weights. In this table we can see, for instance, that choosing A in Q2 and B in Q4 are the most common selections, and that the combination of these two responses is by far the most common response pair across the two items. In fact, this is not very surprising since these are the two correct responses. But, given that these selections of responses were the most common, it follows automatically that a combination of the two is also quite common. The question is then: is this combination more frequent than could be expected? This can be answered by using the data in the crosstable directly. Of those responding A in Q2, almost 70% also choose B in Q4. This is a much higher proportion than for those selecting any of the other options in Q2. This means that it is mainly the students choosing A in Q2 that contribute to the large proportion of B in Q4. It is therefore reasonable to say that these two categories are positively associated. An even more extreme positive association is found for the two categories representing the non-response. Of the students who did not respond to Q2, 86% also omitted Q4! For all other categories in Q2 less than 5% omitted Q4. In the correspondence analysis of this table, it will be quite clear that these two distributional characteristics of the table are the main trends in the material.

Semmelweis - Q2	Semmelweis - Q4					Active margin
	A	B	C	D	No response	
A	1790	24427	1972	5903	1055	35147
B	494	1684	827	1138	194	4337
C	388	1372	486	863	162	3271
D	638	4241	937	2125	382	8323
No response	84	561	120	218	5964	6947
Active margin	3394	32285	4342	10247	7757	58025

Table 3.3: Correspondence table for students' responses on two MC items from the unit 'Semmelweis' diary'.

The starting points of CA are the *row* and *column profiles*. These are simple transformations of the correspondence table presented in Table 3.3 where the

⁴¹ For those of you who have read somewhere else that almost 270 000 students participated in PISA 2000 this number might create confusion. However, since a number of booklets were used (9 booklets) with a rotation of items, not all students were administered science items. Furthermore, of those who did respond to science items, not all were administered each single item. The specific unit referred to here was included in 2 out of the 9 booklets.

entries in the cells are relative numbers showing the proportions of students in one category in the first variable across the categories in the second variable. This transformation for option A in Q2 is here a row profile, and this is a vector of the proportions of students selecting A in Q2 across the five options in Q4. The other profiles are similar vectors for the other categories. In the same way the column profiles for the response categories of the second variable (Q4) can be easily calculated. Similarly we can also calculate the average row and column profiles.

Since all these profiles are vectors, they can be represented by a point in space. Profiles that are similar will be points that are close to each other in this space. Furthermore, the points for profiles that resemble the average profile will be close to the point corresponding to the average profile. This point is called the centroid, and in the further processing of the data in CA, the coordinates for the other points are transformed so that the centroid is placed in the origin of the axis. This involves using the inverse of the elements in the average profiles as weights. The average column profile thus defines the weights to be used for the rows, and vice versa. The elements in the average profiles are often referred to as the masses of the associated points. As a consequence, small categories are weighted up and larger categories are weighted down. In effect, the distances between the points in this space are so called *chi-square* (χ^2) distances. This weighting has to be considered when interpreting the plot, which will be returned to shortly.

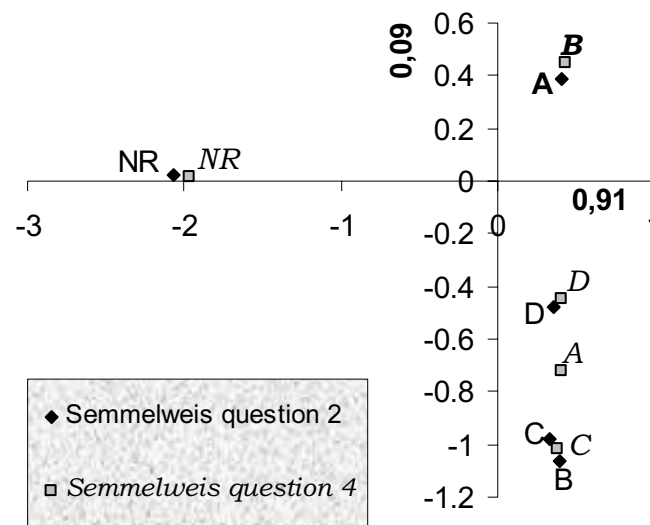


Figure 3.2: Correspondence analysis of Semmelweis' Diary question 2 versus question 4. The correct answers are boldfaced.

However, our ability to comprehend such spaces is for most of us limited to one- or two- (for some maybe three-) dimensional spaces. Correspondence analyses typically display the results as a projection of this multidimensional space onto a plane. Without going into detail, the procedure for identifying the best representation of the points in this space aims to identify the plane that best represents the distances between the points, and the points' distances to the

centroid. The dilemma of visualising multivariate data is to choose between simplicity on the one hand, and, representing the data accurately on the other (Rovan, 1994). Very often these projections are good in the sense that the points in the original dimensionality lie close to a plane, and also, very often the profiles' positions in the original space lie approximately along one single line intersecting this space. In the latter case the solution is one-dimensional.

The most important output when performing CA with the help of a computer is the plot of the solution, typically a scatterplot in two dimensions. One example is given in Figure 3.2, which is the plot for the crosstable given in Table 3.3. Such a plot consists of two point clouds, one cloud for each of the variables. In the interpretation of these clouds, one can assess the categories within one variable that have similar profiles. Such categories will be close to each other in the plot. Also, one can assess the categories from the two variables that are associated with each other. However, the distance between points in the plot referring to categories from different variables cannot be interpreted literally. Remember that the origin for the two clouds was decided by the average row and column profiles, respectively. In other words, the origins for the two clouds are not the same, even if they are placed in the same point in the plot.

Nevertheless, the relative positions that the points have within the two clouds can be compared and interpreted as an association. In Figure 4.2, for instance, we can clearly see that the two non-response (NR) categories are associated with each other. They are both separated from the other responses by the first dimension. The second dimension separates choice A in Q1 and choice B in Q4 from the remaining possibilities offered to the students. These two response categories are in fact the correct answers to these two questions. All in all then, CA presents a visualisation of the overall patterns in the responses across the two items: Primarily, the association between the two nominal variables is that non-response is a relatively consistent response across these two items. Secondly, students who respond correctly tend to do so consistently across the two items. However, there is no clear and distinct pattern in the wrong responses across the two items.

Together with the plot the statistical software usually gives out a wide array of parameters that are helpful in the interpretation of the plot. First of all, there is the *total inertia*⁴², a statistic derived from the χ^2 . This can be regarded as a measure of how much the data in the table deviate from the 'expected' or the nil hypothesis that there are no relationships between the categories of the two variables. This would imply that all the profiles for all the categories are equal, and as a consequence, they would be equal to the average profiles; hence all points would be lying in the origin of the plot. In other words, the inertia is a measure of the degree to which the points spreads out in the plane. It is therefore possible to say that the inertia is comparable to the concept of variance.

⁴² This is completely analogous to the concept of inertia in physics. It is found by multiplying the masses of the points in one of the clouds with the associated chi square distance (the distance from the origin in the plot), and then summing all these. The inertia will be the same independently of which cloud is used for the computation.

This measure can also be disaggregated in ways that are helpful for the interpretation of the plot. The total inertia can be decomposed to each of the dimensions or axes. Typically the axes on the plots are therefore denoted by either a percentage number referring to how much of the inertia is captured by each of the dimensions, or the absolute value of the inertia attributable to the axis is used as a label for the axis. In Figure 3.2 the percentages of the total inertia are used as the labels of the axis. This tells us that the first dimension alone accounts for 91% of the total inertia, and the second for the remaining 9%. In other words, this plot is an exact representation of the table analysed. These numbers also tell us that the pattern in these data could be represented quite well by one dimension only.

Also, two helpful diagnostic parameters for the points in the plot, *absolute contribution* and *relative contribution*, are usually included in the output from the statistical software. Absolute contribution tells us how much a point has contributed to the respective dimensions, while relative contribution is in many ways the opposite measure. This tells us how well the dimension explains the profile for the respective categories, that is, to what degree the projection of the point in the plot onto the axes represents the whole vector. By inspecting these parameters, it is possible to find which points are the defining points for the axes, and which points are well represented by the plot. These parameters for the CA presented in Figure 3.2 tell us clearly that the first dimension is defined by the non-respondents, and this axis separates the students who omit the item from those responding. Furthermore, in this case these parameters confirm that the second axis mainly separates the correct answers from the wrong answers. Also, outliers in the solutions can be detected by these parameters. Since correspondence analysis weights each profile, small categories can be very influential in the analyses, and in the worst case accidental fluctuations, which of course are more profound for categories with few students, can totally dominate the solution. In such a case the category can be defined as *supplementary*, meaning that it will not be used to develop the solution⁴³.

The interpretation given above can, in this simple example, subsequently be confirmed by inspecting the original contingency table. And indeed in this example we have already commented that the two main trends in the crosstable in Table 3.3 are that the correct answers are associated with each other, and the non-responses are associated. In other words, the plot itself does not give any new information: it is a description of the data themselves, and especially so in this case where the inertia of the axes summed up to nearly 100% of the total inertia. Therefore, small tables can often be analysed just as well by inspecting the table itself. However, when the number of categories is high and the tables correspondingly large, it is difficult to get a global view of the information. The plot is very useful for identifying *the most significant aspects* of the relationships between the variables.

⁴³ When a point is defined as supplementary, this means that the mass of the point is set to 0. Accordingly, points with zero mass have no inertia, and do not contribute to the construction of the dimensions. They can, however, receive relative contributions from the dimensions.

In order to develop a sound interpretation other parameters and other rules of thumb have been developed, e.g. for how many dimensions to include. These will not be treated explicitly here, but will be introduced as necessary when the results are presented. The method is most often used as an analytical technique to explore data. However, in the statistical software packages such as SPSS, one can also get significance levels and confidence intervals for most of the parameters, and this illustrates what was previously said about the convergence of the exploratory and confirmatory statistical analysis.

3.4.3 Homogeneity analysis

When analysing more than two nominal variables, HA has been shown to be mathematically equivalent to multiple CA (Greenacre, 1993; Heiser, 1981). Multiple CA is correspondence analysis of a table where each variable has been recoded with dummy variables identifying which category each respondent belongs to. HA of a set of categorical variables gives information of the same kind as CA, only for more than two variables analysed simultaneously. Therefore CA can be regarded as a special case of HA (Heiser & Meulman, 1994).

The essential mathematical problem which HA seeks to solve is in qualitative terms relatively easy to formulate, although the mathematical procedure in itself is not easy to follow. Each respondent has been coded into one of the categories of each of the nominal variables. The data matrix, with a row for each respondent and a column for each variable to be included in the analysis, can therefore be perceived as a set of non-metrical vectors giving (a) the response profile for the respondents (in HA the respondents or the cases are referred to as the objects), and (b) the profile for the variables over respondents. HA seeks to transform these non-metrical vectors to metrical vectors with minimal loss of information.

Skipping the details of how this is done, the mathematical procedure results in a numerical value which can substitute the vector for each respondent (the object scores). At the same time, a numerical value for each category within the variables (category quantifications) is found. This means that the object score for one particular respondent will be equal to the average value of the category quantifications in this respondent's profile, and conversely, the quantification for a specific category is equal to the average object scores for all the respondents placed in this category. Furthermore, the mathematical procedure used in HA seeks to find a solution for each respondent so that the category quantifications for the categories contained in the respondents' profiles, are as close to each other as possible.

This would result in numbers being close to each other for categories that are often combined in the response profiles, while categories that are seldom combined will receive quantifications that differs relatively more. At the same time this procedure ensures that respondents with similarities in the responses across the items receive object scores that are close to each other, and conversely, the respondents with largely different profiles will receive object scores that are far apart.

In the solution the object scores and the category quantifications are normalised so that the mean equals 0 and the standard deviation equals 1. After reaching one solution, a new solution can be sought, demanding of this new solution that it should not correlate with the first one. These solutions are often referred to as the first and second dimension. The maximum number of dimensions is one less than the number of categories in the variable with fewest categories.

There is no perfect solution for this problem, only an optimal solution. In practice, the problem is solved through an iterative computation method where a new and improved solution to the overall problem is found in several steps (a so-called alternating least squares algorithm (Gifi, 1990)). When the next solution is only incrementally better than the previous one, according to some criteria, the process has reached a final solution.

Although the essence of this method is concealed in heavy mathematical notations (e.g. as presented in Gifi, 1990), the main ideas were explained above. To illustrate the usefulness of this procedure, I will consider one hypothetical respondent with a specific response vector across four items. This specific respondent has received an object score, and at the same time all the four categories in this specific response vector have been assigned a number, the category quantifications. In this example I further assume that a two-dimensional solution has been developed. In the case where this hypothetical respondent has a response vector identical to a group of other respondents they will all have the same object scores and these object scores will most likely be close to each of the four category scores in this vector. In this case the analysis has revealed that there exists a subgroup of respondents with a characteristic response profile across the variables; in other words, there is a group which can be characterised as homogeneous, thereby the name 'homogeneity analysis'. A scatterplot of this two-dimensional solution, with both the category quantifications and the object scores plotted, would reveal that this group of respondents is placed in a cluster together with their response categories.

Imagining a hypothetical case where all the respondents belong to one of a very small number of unique profiles across the items, the solution (presented as a plot of objects scores, or as a plot of category scores) would consist of clearly separated clusters of categories or respondents. Usually, however, this is not the case. If the number of respondents is large compared to the number of categories, the number of different response vectors will probably approach the number of conceivable permutations of responses. However, some of the response vectors might be more dominating, and these would show up in a plot of the category quantifications. A category that separates students will have a relatively high numerical value (positive or negative), and the associated categories in the other variables in the set will be close to this category. Categories which do not separate students will have low numerical values. As a consequence, categories that are rare, will, by chance alone, relatively often receive high absolute scores or values along one or several of the dimensions, and the corresponding respondents classified with these categories will be seen as outliers in the distribution. And vice versa, categories used quite often will, as a statistical

necessity, most often be close to 0 since it is more likely that the respondents classified within this category have a repertoire of different profiles across the rest of the variables in the set. In the ideal data set for analysis, categories with very small or very high absolute frequencies should not be present.

In essence, the interpretation of the HA solution is similar to the analysis of CA solutions. Both methods have biplots or scatterplots representing the solution. Figure 3.3 is only included as an example of such a plot. This solution will therefore only be briefly discussed.

Figure 3.3 presents the solution of a HA of a unit titled *Ozone* including four items (see Appendix 1), one of the released units from PISA 2000. This particular unit has been presented and analysed in great detail by Turmo (2003a). One item (Q1) is a CR item with a relatively large number of categories, one is a MC item (Q2) and the two last items are CR items with only score points and non-response coded.

By plotting the quantifications for the categories in the variables in the set, the associations between categories are illustrated. In other words, categories that are related are close to each other in the plot. The solution presented in Figure 3.3 includes a total of 21 categories. The overall pattern is quite striking. The first dimension separates the students with no response on the items from those who actually did respond, more or less independently of the type or quality of the response. The second dimension is a more continuous dimension that obviously reflects the quality of the students' responses. All in all, this main pattern is very consistent with the finding in paper II, analysing a different unit. Also, the interpretation of this diagram is very similar to the interpretation of the CA presented in Figure 3.2. A similar plot could have been presented for the objects (in this case; the students).

The advantage of HA as compared to CA, besides the fact that more variables can be analysed simultaneously, is that the calculations are done on the raw data file, and not by analysing an aggregation of the data such as a contingency table, for instance. In effect, it has been stated that scores on the dimensions are computed for each of the respondents, the so-called *object scores*, and not only for the categories. Also I have, in a qualitative way, tried to show how these object scores are intimately linked to the scores for the categories: the average object score for all the respondents belonging to one specific category equals the score for this category in the dimension.

Having identified what the dimensions probably represent by examining the category scores, the object scores can be used as any ordinary scale, for instance as in Figure 2 in paper II, computing average object scores for countries. This feature of HA is also useful for purposes other than studying how categories are associated. For instance, it makes it possible to develop metrical scales for nominal or ordinal variables. This feature of HA is better known as optimal scaling.

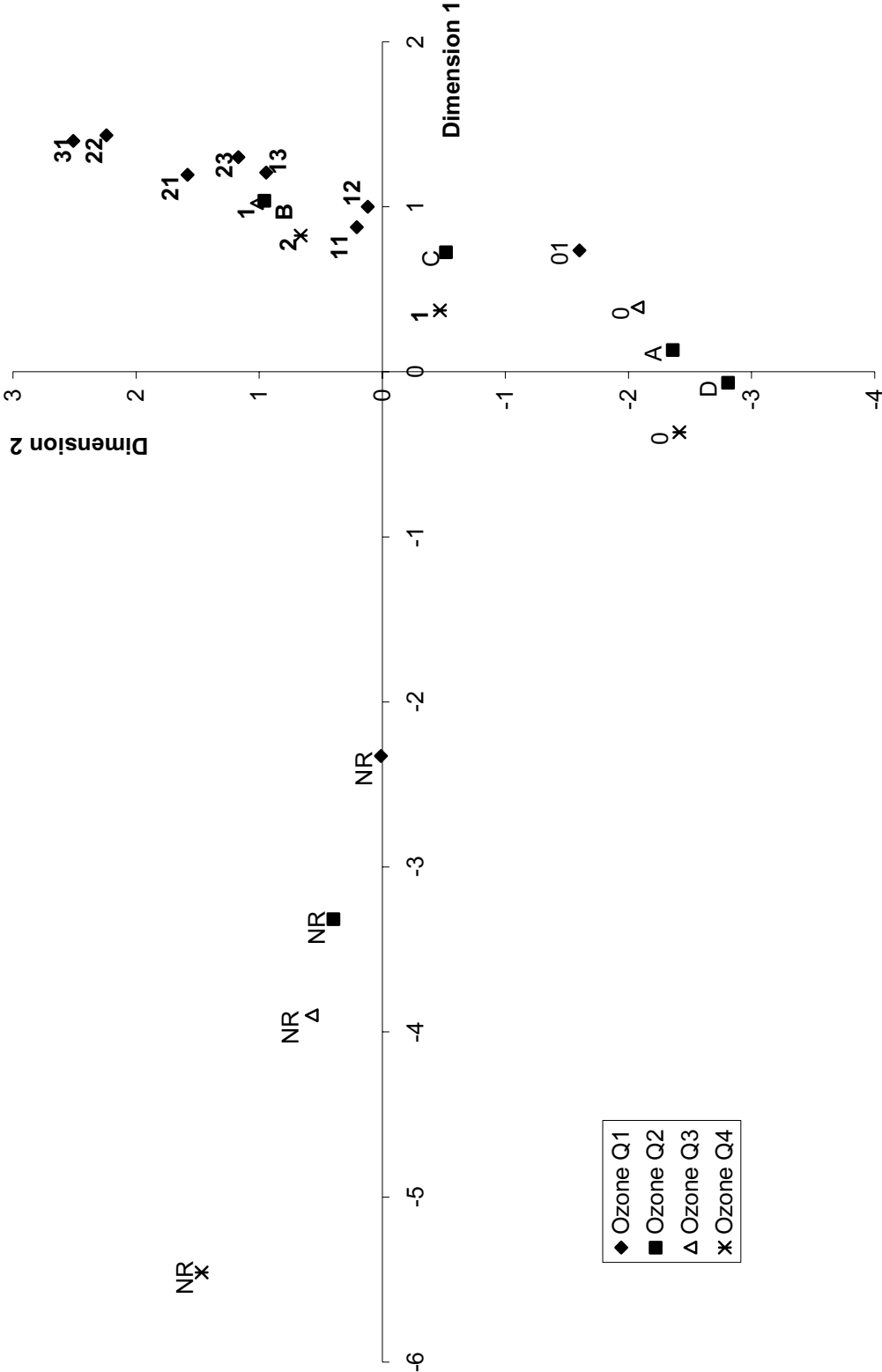


Figure 3.3: The plot of the solution obtained by homogeneity analysis of the unit of four items titled ‘Ozone’ (see Appendix 1). The boldfaced categories refer to credited responses, and NR refers to non-response.

I have previously mentioned that CA computes a number of parameters that are very helpful for interpreting what the solution means, and also parameters that tell us how well the solution represents the original information in the response profiles. One of the disadvantages of HA is that it is not supported by parameters such as absolute and relative contributions, as in CA. There is another set of parameters instead which to some degree is helpful in order to interpret the solution. In addition to the category quantifications (presented in Figure 3.3) and the object scores, the software developed at the University of Leiden implemented in SPSS Categories (Meulman & Heiser, 1999) gives some parameters which are helpful for obtaining a broader understanding of the solution.

First of all the HA solution is associated with an *eigenvalue* which can be interpreted very much in the same way as the eigenvalue in factor analysis or in CA. For the example with HA of the unit *Ozone* from PISA 2000, the eigenvalues for the two first dimensions are 0.59 and 0.35, respectively. The eigenvalue is a measure of how much of the categorical information is accounted for by each dimension, analogous to the concept of variance accounted for (a squared correlation). The higher the eigenvalue, the better the quantification is able to separate or discriminate between the respondents, or distinct groups of categories.

In addition a parameter called *discrimination measure* is produced which can be interpreted in the same way as the eigenvalues, only this time for each variable separately. To follow up the specific example, the discrimination measures for the first dimension are in the range 0.5-0.6 for all the four items, while they are lower and vary more across the four items for the second dimension. The values of the discrimination measures indicate the degree to which the dimensions separate the categories for that variable. The average discrimination measure across all variables for one dimension equals the eigenvalue for that dimension.

When including more than one dimension in the solution, the eigenvalues are computed for each dimension separately. Since each new dimension represents another quantification of the same data, the eigenvalues cannot be added to give an estimate of how much of the categorical information is accounted for by the dimensions combined.

Another disadvantage with HA, compared to CA, is that the method does not allow for supplementary points. This means that if there are categories or objects suspected to be outliers influencing the solution, they have to be excluded in the analysis. All in all this suggests that there are fewer criteria that are helpful for interpreting the solutions than in CA.

An important choice to be made, both in CA and HA, is how many dimensions to include in the solution. In the analysis presented in paper II, the purpose was to understand, or at least develop a description of, the association between a small set of nominal cognitive variables. For all the units analysed, it has been impossible to develop a meaningful interpretation of a third dimension. The subjective criterion of interpretability has therefore been used to delimit the analyses to two dimensions only.

Validation of the interpretation made of these dimensions is done by going back to the raw data to check whether there is support for the claims made. In general validation in this thesis is concerned to a high degree with looking at the same data with other tools, or slightly different versions of the tools, or by looking at different sub-samples within the total sample. In paper II the original solution for the whole international data set was triangulated using the Nordic data only, establishing the same interpretation. In multivariate data analysis this is often referred to as the stability (Gifi, 1990) of the solution: the fact that the interpretation is not an artefact of the method used to inspect the data or just an effect of the specific sample, but rather reflects real and overarching aspects of the data. The concept of stability is further elaborated in paper III. Given that the homogeneity analysis presented in paper II is based on the full international data file with close to 60 000 respondents it is very plausible that doing the same analysis on another set of data, collected within the same population using the same instruments or even slightly different instruments, would lead to a very similar interpretation.

Methods like correspondence analysis and homogeneity analysis are fairly standard in, for instance, sociology, biology and marketing research. This could be due to the fact that many studies in these fields are concerned with relating different groups or strata in the population being studied. It can be difficult to discover a useful key for such a grouping. The variables used for grouping are seldom scalable at the interval level. Through correspondence analysis or homogeneity analysis classification keys can be established. Although there is some use of these methods in educational science, I have not yet seen this method used to describe patterns in cognitive data. In this way the empirical work presented in paper II is an innovative analytical approach. Even if the aim for using this type of analysis to study patterns across items in a qualitative sense, to a large degree was not reached, it is hoped that this type of analyses can be used in future research on data that are more targeted towards this aim, an issue that is discussed in the concluding sections of paper II

3.5 References

- Alseth, B., Angell, C., Brekke, G., Kind, P. M., Kobberstad, T., Kjærnsli, M., & Lie, S. (1993). *The TIMSS Item Pilot 1993 - Norway: Comments to Free-Response Achievement Items. Report on Method or Approach and Error Type Listing*. Oslo: Department of Teacher Education and School Development, University of Oslo.
- Angell, C. (1995). *Codes for Population 3, Physics Specialists, Free Response Items*. Oslo: Department of Teacher Education and School Development, University of Oslo.
- Angell, C. (1996). *Elevers fysikkforståelse. En studie basert på utvalgte fysikkoppgaver i TIMSS*. Dr. Scient. thesis, Det matematisk-naturvitenskapelige fakultet, Universitetet i Oslo.

- Angell, C., Brekke, G., Gjørtz, T., Kjærnsli, M., Kobberstad, T., & Lie, S. (1994). *Experience with Coding Rubrics for Free Response Items*. Oslo: Department of Teacher Education and School Development, University of Oslo.
- Angell, C., & Kobberstad, T. (1993). *Coding Rubrics for Free-response Items*. Oslo: Department of Teacher Education and School Development, University of Oslo.
- APA. (2001). *Publication Manual of the American Psychological Association*. Washington, DC: American Psychological Association.
- Benzécri, J.-P. (1973). *L'analyse des données, volume II*. Paris: Dunod.
- Bertin, J. (1981). *Graphics and Graphic Information-Processing* (W. J. Berg & P. Scott, Trans.). Berlin/New York: Walter de Gruyter.
- Blasius, J., & Greenacre, M. J. (1994). Computation of Correspondence Analysis. In M. J. Greenacre & J. Blasius (Eds.), *Correspondence Analysis in the Social Sciences*. London: Academic Press Ltd.
- Bourdieu, P. (1984). *Distinction: a social critique of the judgement of taste*. London: Routledge & Kegan Paul.
- Brekke, G., Kobberstad, T., Lie, S., & Turmo, A. (1998). *Hva i all verden kan elevene i matematikk? Oppgaver med resultater og kommentarer*. Oslo: Universitetsforlaget.
- Clausen, S.-E. (1998). *Applied correspondence analysis: an introduction* (Vol. 07-121). Thousand Oaks: Sage.
- Clerk, D., & Rutherford, M. (2000). Language as a confounding variable in the diagnosis of misconceptions. *International Journal of Science Education*, 22(7), 703-717.
- Cohen, J. (1990). Things I Have Learned (So Far). *American Psychologist*, 45(12), 1304-1312.
- Cohen, J. (1994). The Earth is Round ($p < .05$). *American Psychologist*, 49(12), 997-1003.
- Everitt, B. S. (1993). *Cluster analysis* (3rd ed.). London: Edward Arnold.
- Gifi, A. (1990). *Nonlinear Multivariate Data Analysis*. New York: John Wiley & Sons.
- Greenacre, M. J. (1983). *Theory and applications of correspondence analysis*. London: Academic Press.
- Greenacre, M. J. (1993). *Correspondence analysis in practice*. London: Academic Press Limited.
- Greenacre, M. J. & Blasius, J. (Eds.). (1994). *Correspondence Analysis in the Social Sciences*. London: Academic Press Ltd.
- Harlow, A., & Jones, A. (2004). Why Students Answer TIMSS Science Test Items the Way They Do. *Research in Science Education*, 34(2), 221-238.

- Harmon, M., Smith, T. A., Martin, M. O., Kelly, D. L., Beaton, A. E., Mullis, I. V. S., Gonzalez, E. J., & Orpwood, G. (1997). *Performance Assessment in IEA's Third International Mathematics and Science Study*. Chestnut Hill: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.
- Heiser, W. J. (1981). *Unfolding Analysis of Proximity Data*. PhD-thesis, University of Leiden.
- Heiser, W. J., & Meulman, J. J. (1994). Exploring the Distribution of Variables and their Nonlinear Relationship. In M. J. Greenacre & J. Blasius (Eds.), *Correspondence Analysis in the Social Sciences* (pp. 179-209). London: Academic Press.
- Hjellbrekke, J. (1999). *Innføring i korrespondanseanalyse*. Bergen: Fagbokforlaget.
- Kind, P. M. (1996). *Exploring Performance Assessment in Science*. Dr. Scient. thesis, Det matematisk-naturvitenskapelige fakultet, Universitetet i Oslo.
- Kjærnsli, M., Lie, S., Stokke, K. H., & Turmo, A. (1999a). *Hva i all verden kan elevene i naturfag? Oppgaver med resultater og kommentarer*. Oslo: Universitetsforlaget.
- Kjærnsli, M., Lie, S., & Turmo, A. (1999b). *Two-digit codes for science and mathematics. Results from a Norwegian workshop*. Oslo: Department of Teacher Education and School Development, University of Oslo.
- Kobberstad, T., Lie, S., Brekke, G., & Kjærnsli, M. (1994). *Codes for Population 1 and 2 Free Response Items*. Oslo: Department of Teacher Education and School Development, University of Oslo.
- Lederman, N. G. (1992). You can't do it by arithmetic, you have to do it by algebra! *Journal of Research in Science Teaching*, 29, 1011-1014.
- Lie, S., Taylor, A., & Harmon, M. (1996). Scoring Techniques and Criteria. In M. O. Martin & D. L. Kelly (Eds.), *TIMSS Technical Report* (Vol. 1).
- Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46, 806-834.
- Meulman, J. J., & Heiser, W. J. (1999). *SPSS Categories 10.0*. Chicago (IL): SPSS Inc.
- Mullis, I. V. S., Martin, M. O., Beaton, A. E., Gonzalez, E. J., Kelly, D. L., & Smith, T. A. (1998). *Mathematics and Science Achievement in the Final Year of Secondary School. IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.
- OECD-PISA. (2002). *Sample Tasks from the PISA 2000 Assessment: Reading, Mathematical and Scientific Literacy*. Paris: OECD Publications.

- Orpwood, G., & Garden, R. A. (1998). *Assessing Mathematics and Science Literacy* (Vol. 4). Vancouver: Pacific Educational Press.
- PISA Norway. (2004). *Item submission bundle 1-5* (Report to PISA consortium). Oslo: Department of Teacher Education and School Development, University of Oslo.
- Rennie, L. J. (1998). Improving the Interpretation and Reporting of Quantitative Research. *Journal of Research in Science Teaching*, 35(3), 237-248.
- Robitaille, D. F., & Garden, R. A. (Eds.). (1996). *Research Questions & Study Design* (Vol. 2). Vancouver: Pacific Educational Press.
- Rosnow, R. L., & Rosenthal, R. (1989). Statistical Procedures and the Justification of Knowledge in Psychological Science. *American Psychologist*, 44(10), 1276-1284.
- Rovan, J. (1994). Visualizing Solutions in more than Two Dimensions. In M. J. Greenacre & J. Blasius (Eds.), *Correspondence Analysis in the Social Sciences* (pp. 210-229). London: Academic Press.
- Sadler, P. M. (1998). Psychometric Models of Student Conceptions in Science: Reconciling Qualitative Studies and Distractor-Driven Assessment Instruments. *Journal of Research in Science Teaching*, 35(3), 265-296.
- Schmidt, F. L. (1996). Statistical Significance Testing and Cumulative Knowledge in Psychology: Implications for Training of Researchers. *Psychological Methods*, 1(2), 115-129.
- Schultz, J. (2000). *Att samtala om/i naturvetenskap. Kommunikation, kontext och artefact*. PhD-thesis, University of Linköping.
- Sedlmaier, P., & Gigerenzer, G. (1989). Do Studies of Statistical Power Have an Effect on the Power of Studies? *Psychological Bulletin*, 105(2), 309-316.
- Taylor, A. (1993). *Coding Rubrics for Free-Response Items*. (Doc. Ref.: ICC648/NPC249). Discussion Paper for the TIMSS Project Coordinators Meeting, March 1993.
- The BMS. (1994). Correspondence Analysis: A history and French Sociological Perspective. In M. J. Greenacre & J. Blasius (Eds.), *Correspondence Analysis in the Social Sciences* (pp. 128-137). London: Academic Press.
- Tukey, J. W. (1962). The future of data analysis. *Annals of Mathematical Statistics*, 33, 1-67.
- Tukey, J. W. (1977). *Exploratory data analysis*. Reading: Addison-Wesley.
- Turmo, A. (2003a). *Naturfagdidaktikk og internasjonale studier. Store internasjonale studier som ramme for naturfagdidaktisk forskning: En drøfting med eksempler på hvordan data fra PISA 2000 kan belyse sider ved begrepet naturfaglig allmenndannelse*. Dr. Scient. thesis, Det utdanningsvitenskapelige fakultet, Universitetet i Oslo. Oslo: Unipub AS.

Turmo, A. (2003b). *Understanding a newsletter article on ozone - a cross-national comparison of the scientific literacy of 15-year-olds in a specific context*. Paper presented at the 4th ESERA conference "Research and the Quality of Science Education", Noordwijkerhout, The Netherlands.

4 Summary and discussion

In the following sections I will give short summaries of the chapters and papers contained in this thesis. These summaries emphasise some aspects of the chapters and papers that are seen as particularly relevant for the general aim of this thesis presented in section 1.1.5. However, as previously stated, it should be acknowledged that each of the papers should also be considered as separate contributions with their own purposes, goals and research questions, and some of these findings will therefore also be highlighted.

The three papers are placed in chronological order in the thesis, and this sequence reflects the process with which the data have been explored; paper II follows as a consequence of paper I, and the work documented in paper III is a consequence of some of the results in paper II. Thus, even if each paper can be seen as separate contributions, they have a history of development linking them together to some degree.

In addition to summarising and discussing the chapters and papers in the light of the overall aim of the thesis, I will in the following include some reflections of the possible implications that some of the findings may have for further analyses of data from LINCAS.

4.1 Summary of chapter 1

The thesis was introduced in this chapter by presenting the fundamental rationale for why analysis of items, either one-by-one, or by the study of profiles across a few items, is worthwhile. This rationale was based on a model of how items typically are correlated with each other and to the overall score in an achievement test such as those in TIMSS and PISA. It followed from this model that if we represent the total achievement measure by one overall latent factor, only a small fraction of the variance in the scored items is accounted for by a typical cognitive test score.

Furthermore, this argument was brought one step forward by also considering the categorical information in the codes initially used by the markers. Before the variables in the data file are scored, they are nominal variables with codes reflecting qualitative aspects of students' responses. Taken together with the theoretical model of the scored items, it was concluded that further analysis of the single items would be reasonable, and would involve the analysis of information beyond that contained in the overall score. All the empirical papers in the thesis are based on this rationale: to analyse the surplus information in the items.

The purpose of the thesis was then formulated as an exploration into the nature of this surplus information, and the potential of using this information to describe qualitative differences at the student or the country level. Furthermore, the underlying motivation for doing this was stated as a desire to inform the science education community about the potential for, and limitations of, using the data from LINCAS in secondary research. This latter issue was elaborated and discussed in the next chapter.

4.2 Summary of chapter 2

This chapter gave a broad presentation of LINCAS, their policy relevance, and their link, or lack thereof, to the field of science education research. The chapter consisted of several related elements that, taken together, addressed the issue of why and how researchers in science education could or should engage in analyses of LINCAS.

This was done by presenting the historical development of LINCAS, from the first IEA studies by the end of the 1950's to the contemporary studies PISA and TIMSS. I suggested that the development in this period reflects broader societal issues. Moreover, I suggested that the development illustrates a tension or dilemma that LINCAS have been confronted with from the very beginning: LINCAS was initially framed by the idea that international comparisons could be the basis of a powerful design for studying educational issues. Thus, the main idea driving the genesis of LINCAS (which I labelled *Purpose I*) was an ambition to utilise the international variation in the study of general educational issues. This research base has been maintained throughout the history of LINCAS. What made it possible to conduct the increasingly more expensive studies was the fact that policy makers evaluated the studies as providers of policy-relevant information. Over the years there has been a shift towards the purpose of finding evidence for effective policy at the system or national level (which I labelled *Purpose II*), and the discussion in this chapter demonstrates that this vision for LINCAS is very visible in the PISA study. It would be fair to say that my thesis aims to promote Purpose I, and, furthermore, it aims to promote the view that the tension that is often perceived between the two purposes is to some degree based on a lack of communication and interaction between the policy makers and the educational researchers.

The chapter then turned to a comparison between PISA and TIMSS. This is an issue that in itself is worthwhile because there are some indications that users of the information may be confused by discrepant results in the two surveys. However, by examining the differences between the studies, it is evident that the results should not be compared in a simplistic manner: they have different designs targeting different populations and different levels of the school systems, they have defined the achievement measures differently, and even if many countries participate in both studies the composition of the countries in the two studies is clearly not the same.

Chapter 2 continued by discussing how science education may be linked to the policy context by engaging in secondary analysis of data and documents from LINCAS. This was not to argue that all, or even most, of the research in science education should be linked to PISA or TIMSS. Nevertheless, a relatively comprehensive review of possibilities for secondary analysis related to LINCAS was presented in the chapter, and the increased potential for such analyses relating to scientific literacy in PISA after the 2006 study was emphasised.

4.3 Summary of chapter 3

Chapter 3 gave an overview of some methodological issues that have heavily influenced my work. It began by placing my work in a tradition that could best

be labelled as exploratory data analysis. The main idea of this tradition is that when confronted with a data set we should seek to develop a description of the overall structure in the data, the multivariate relationship, which is a challenging task since there is no general procedure to follow for finding such overall patterns in the data.

In addition the general issue of the nature of the information in the cognitive items in TIMSS/PISA was explored in this chapter. A novel innovation in TIMSS was the double digit codes and the associated marking rubrics used for the constructed response items. With TIMSS it was acknowledged that using only multiple choice items, which before TIMSS was commonplace in most large-scale assessments, would seriously limit the range of competencies activated by a test. By using open-ended questions, giving students the opportunity to construct their own responses, TIMSS had the ambition of developing descriptions of how students' represented and made use of concepts in science. The double digit codes were used to preserve that information. This was also the idea in the science assessment of PISA 2000, although the generic system was slightly modified. However, with PISA 2003, and with the items that have undergone field trials before PISA 2006, it is evident that the use of such coding is gradually disappearing. The reason for this change is not entirely clear, but it may be suggested that the codes have been of little use internationally. Nevertheless, constructed response items will still be used since they allow for the testing of competencies other than the selected response formats.

The paradoxical consequence of this is that from PISA 2006 more information about students' thinking and knowledge will be available from analysis of the multiple choice items than from students' own written accounts of their reasoning and thinking, since the former at least include a code reflecting the response selected by the students. The constructed response items that were originally introduced into these assessments as tools for making the students demonstrate their thinking and reasoning are, in the marking guides for PISA 2006, more or less directly reduced to a description of how to score the items. Even if the marking guide includes explicit descriptions of the criteria for scoring, for the great majority of items there are no longer separate codes for students with different types of responses. I will suggest that this development was perhaps inevitable given that these codes were not extensively used or reported on in the international reports. However, I regard this development as a decrease in the potential for communicating how students typically think and interact with the items in tests like PISA. Furthermore, this development can be viewed as unfortunate from the perspective that such data could possibly be an important resource for secondary analysis aimed at studying students' understanding of very specific scientific concepts or phenomena.

Figure 3.3 provided some bipolar characteristics of analyses of information at different levels in item processing from specific written responses, through the coded responses, and finally to the scored items. Information is continuously and consciously peeled off in this process. In the first process of coding, all aspects that are seen as irrelevant for the overall intention of the response are peeled off. This may, for instance, be information regarding errors

in spelling, errors in grammar, and other very specific elements in the response. However, it may also be information that reflects characteristic features of students' thinking and knowledge. The marking guide has to be understood similarly by all markers, in all countries, and thus, it is a necessary condition that the number of codes are limited, and that they reflect clearly identifiable features of students' responses. The codes therefore represent classes of typical responses that may be distinguished from each other. In the next process, when the items are scored, all aspects other than the overall quality or correctness of the item are peeled off. The score can therefore be considered as not representing aspects of the responses as such, but rather as representing aspects of the ability that students have used to create their responses. At least this is the idea. However, as demonstrated in Figure 1.1, the score information at the single item level is still highly specific for the item.

Furthermore, chapter 3 addressed more specifically the methods used in one of the papers: correspondence and homogeneity analysis. I have so far not seen any other analysis where these, or similar tools, are used to study the relationship between nominally measured cognitive variables. In that sense the work undertaken in this paper represents an innovative approach to the analysis of data from cognitive tests. The aim of this section in chapter 3 was to write about the methods at a level requiring very little mathematics. This was a conscious choice in order to make this part of the text available to a more diverse group of readers. One consequence of this would be that interesting aspects of the methods are not commented on. Furthermore, since the language of mathematics is a useful tool that allows for very precise and unequivocal communication, another unfortunate consequence may be that the text is ambiguous, thus allowing misunderstandings to develop. Nevertheless, writing for a wider audience has forced me to challenge my own understanding of the methods I have applied.

4.4 Summary of paper I

Olsen, Turmo & Lie (2001): Learning about students' knowledge and thinking in science through large-scale quantitative studies. *European Journal of Psychology of Education*, 16(3), pp. 403-420.

In this paper we used data collected by a group of Nordic colleagues who implemented an extra booklet in TIMSS 1995. In the extra booklet some of the official TIMSS items were slightly reworded, or the format was changed.

The paper relates to a long-standing research field in science education: the study of students' own mental representations of scientific phenomena, concepts, laws and theories. The paper included a discussion of how written assessment tasks may or may not give insight into these mental representations.

We established a theoretical framework of how distracters may have different functions. These functions can be used to predict what would happen to the difficulty of multiple choice items that are reworded into a constructed response item. Furthermore several of the identified functions of the distracters

may be used to explain some strengths and weaknesses in the use of multiple choice items to map students' mental representations:

- some distracters are merely treated as elements in a check list, while the actual item solution is based on a parallel construction of a response;
- some distracters may be eliminated for reasons other than those intended by the item writer: there may be flaws in the item, and the distracter may be inappropriate or irrelevant in the given context;
- some distracters may be formulated or structured in such a way that they, usually unintentionally, set up a cognitive trap for the student;
- some highly attractive distracters are, purposely or accidentally, not included;
- the response alternatives sometimes define the questions, and in other cases they are vital to clarify the question intent.

In a subsequent analysis of several items we demonstrated how these functions could contribute to the understanding of response profiles for some of the TIMSS 1995 science items. These descriptions of how distracters may function are useful tools or rules that item writers can use in several ways. They may be used to detect or understand items with problematic behaviour, or they may be used to reformulate or recycle some items that have been used previously. The item difficulty can, for instance, be increased or decreased consciously by applying one or several of these rules. The article can therefore be seen as giving some theoretical guidance for item writers.

One conclusion from this paper that relates more directly to the aim of this thesis is that the analysis of single items is the analysis of highly unreliable information. The article furthermore ended in a recommendation that one possible development would be to design clusters of items to be analysed 'holistically'. This could lead to a more reliable mapping of students' cognitive structures. However, the paper did not suggest how such a holistic analytical approach could be carried out. Thus, one of the aims of the work presented in paper II was to follow up this recommendation.

Nevertheless, the paper demonstrated that although single items may be considered as not very reliable measures, the presentation and discussion of single items are powerful tools for communicating what an assessment really is about. Furthermore, the analyses of single items can be used as very illustrative examples of typical ways in which students respond to tasks that are relatively common in school science. Moreover, in the case of PISA, the single items may be regarded as examples that at least resemble tasks that students are likely to be confronted with in the future. In that way, the descriptions that may be developed by single item analysis can prove to be valuable resources to be used in, for instance, teacher education. In conclusion, although the single item information is not very reliable when conceived of as a measures, it is still relevant for science education.

4.5 Summary of paper II

Olsen, R.V. (2004): The Search for Descriptions of Students' Thinking and Knowledge: exploring nominal cognitive variables

by correspondence and homogeneity analysis. *Scandinavian Journal of Educational Research*, 48(3), pp. 325-341.

Rather than constructing items intended to work together to map students' mental representations, this paper followed up the recommendation in paper I by exploring how a group of items may be studied 'holistically'.

The starting point was the decision to analyse items structured into 'units' in PISA 2000. These units of items related to the same stimulus material. The hypothesis was that by studying students' profiles across all the items within the same unit, it should in principle be possible to develop descriptions of different response profiles across the item set, and hopefully, such profiles could reflect students' own mental representation of the situation presented by the items. The unit was chosen as the key to grouping items since a preliminary analysis suggested that items within the same units were more tightly related than items across the units. In the model in Figure 1.1 this means that the residual variances of the items within the same unit are structurally related.

The paper therefore explored the nominal information in the PISA cognitive items by mainly using homogeneity analysis (HA). If there were distinct profiles across the items, as hypothesised, HA would show this as clusters of categories, and clusters of respondents. It turned out that in the analysed unit 'Semmelweis' Diary' (see Appendix 1), two powerful characteristics of the students' profiles across the four items dominated the solution. However, these characteristics, or dimensions in the solution, did not reflect distinct qualitative aspects of students' mental representations of the context. Instead, the profiles described the overall test behaviours: those not responding to one item also tend to omit or give no response to most other items, and those formulating or selecting an appropriate or correct answer in one item tend to do so to a greater extent for all items in the set. This is well known, and is the type of information that in general is captured in the overall scores in achievement tests like PISA and TIMSS. It is interesting to note, however, that at the student level these characteristics are not correlated; this indicates that some students receive low scores mainly because they do not respond, while others receive low scores because they consistently give responses of low quality.

The paper then explored the main characteristics when aggregated to country level. This revealed that these characteristics were moderately correlated at the country level. This means that in a country with a relatively high response rate, there is also a higher rate of responses of good quality, and vice versa; in a country where many students do not respond to the items, the responses tend to be of low quality. This indicates that students' overall scores are affected by two separate, albeit related, processes: to be motivated or confident enough to respond, and to be able to give a good response.

Some countries had a pattern that largely deviated from the overall pattern described by the overall correlation between the two dimensions. This indicated that the comparison of countries with very discrepant profiles, such as Denmark and the USA, could be invalid due to non-response bias. However, the subsequent detailed analysis demonstrated that while there was a higher degree of non-response in Denmark than in the USA, there was a correspondingly higher

rate of unspecified incorrect responses in the USA. This was taken to imply that the tendency for relatively more non-response in Denmark compared to the USA did not introduce a serious bias. The likely interpretation is that the students who did not respond in Denmark would, *if* they had responded, have given responses of low quality, or in other words, the non-response is probably an expression of a lack of the ability that the items measure. Other specific country profiles were also explored and contrasted with each other in the paper. These patterns of missing responses should be studied in further analyses aimed at identifying the possible sources of the missing responses, and thus aiming at reaching more solid conclusions about whether there is possible bias due to non-response in the tests.

Given that the HA could not be linked to specific mental representations of the concepts or the situations in the stimulus material linking the items, correspondence analysis (CA) between pairs of the items in the unit was conducted. The hope was to find patterns revealing mental models, or cognitive structures, activated consistently across the items. However, all the correspondence analyses gave solutions that were consistent with the interpretation of the HA solutions: (a) a first dimension separating those responding to both items from those not responding to both items; and (b) a second dimension separating responses of low *vs.* high quality. Moreover, the CA gave clues about the relative strengths of these dimensions, which is slightly more difficult to evaluate in HA. The consistent result was that in the CA 60–90% of the profile variation (the inertia) was accounted for by the first dimension (the tendency to respond or not), while the second dimension (the tendency to give responses of good quality), in most cases, could account for the rest of the variation. One example of correspondence analysis was further included as an example in Figure 3.2 in chapter 3.

In the HA I also tried to study the third dimension. The first two dimensions were interpreted as reflecting the overall ability or the test behaviour. The hope was that the information captured by this third dimension could reflect qualitatively different ways of representing the issue or phenomenon targeted by all the items in the unit. But the CA convincingly explained why the exploration of this dimension did not lead to any interpretable result: there is simply no significant variation to be studied in the third dimension. Most of the variation is already accounted for by the two major dimensions.

The pattern in the HA and CA analyses of other units and items was similar to the pattern described for the unit ‘Semmelweis’ Diary’ in paper II (the analysis of one more unit was included as an example in Figure 3.3 in chapter 3). Furthermore, the HA of the ‘Semmelweis’ Diary’ unit was repeated with the smaller sub-sample of the Nordic countries with the same overall pattern as the result. Taken together this indicates that the pattern is stable; it is not influenced by the specific item, and the pattern was the same for another sample.

It was suggested that one likely explanation as to why these analyses were unsuccessful in uncovering specific patterns of students’ knowledge and thinking was that the items were constructed to work this way, and that the items with such characteristics would have a higher chance of being kept in the final instruments. In many ways, even the categorical information preserved in the

codes used for marking was primarily related to the overall correctness or quality of the responses, and was not intended to describe very specific aspects of the responses. Furthermore, the distracters used in the multiple choice items (or in other words the ‘wrong’ responses) usually did not correspond to different qualitative ways of representing concepts or ideas relating to the situation in the unit. Instead the distracters primarily distinguished themselves from the correct response by being just that, distracters or wrong responses.

One recommendation would be to develop units of items that are deliberately constructed to capture a few central qualitative aspects across the items. I have attempted to work further on this challenge by developing a unit with such characteristics⁴⁴, but the sample size in my study was unfortunately far too low. A relatively large sample size would be needed to carry out HA analysis across, for instance, three items with a number of categories. Having, for instance, four categories in each of the three items would give 64 possible combinations of the categories. Since a minimum of ten respondents for the combination of interest would be needed, the sample should be at least 640 and probably even larger since all combinations will not be equally common.

4.6 Summary of paper III

Paper III: Olsen (2005): Item-by-country interactions in PISA 2003: Country-specific profiles of science achievement. *Not published*.

A condensed version with a more narrow focus has been published as Olsen (2005): An exploration of cluster structure in scientific literacy in PISA: Evidence for a Nordic dimension?, *NorDiNa*, 1(1), pp. 81-94.

As described above Paper II reported a study of the cognitive data of a set of items where students’ profiles in each country were aggregated to average country profiles. This did not produce any clear-cut evidence that countries differed along the major dimensions in the solution. Rather, the overall pattern was that the two main dimensions in the solution (the tendency to omit items and the tendency to provide accepted answers) were correlated when aggregating the data to countries. However, the HA in paper II did reveal that some countries had quite distinctly different profiles of these characteristics.

Paper III can be seen as a continuation of this: instead of studying the tendency to respond or not, paper III seeks to identify which items describe the typical strengths and weaknesses of each of the countries. Moreover, the paper seeks to establish which countries have similar achievement profiles. The departure point for the analysis presented in this paper was to calculate the p-value residual matrix for all countries across all science items in PISA 2003. The entries in this matrix are expressions of the item-specific strengths and weaknesses for each country, independent of the country’s overall achievement level in the test. Usually these are referred to as item-by-country interactions. A

⁴⁴ The unit was developed as part of PISA 2006 and it is for now kept confidential. It consisted of three items that challenged students’ conceptions of the fundamental building blocks of matter (atoms and molecules). Some very common and well-documented types of conceptions were targeted by distracters in all the items.

cell with a high value expresses the percentage point deviation from the expected value for this country on that specific item (positive or negative), given the overall achievement level of the country and the overall difficulty of the item. The sequence of residuals for one country may be labelled as the country's achievement profile.

A cluster analysis was conducted on this matrix, and some distinct clusters of countries were represented in the solution; these were labelled as 'East Asian countries', 'English-speaking countries', 'North-West European countries', 'South American countries', 'less developed countries' and 'East European countries'. The correlations between the countries' p-value residuals were analysed in this cluster analysis, and when these correlations were sorted according to the cluster analysis solution (Table 3) it was evident that correlations between countries within the same clusters were significantly higher than between countries from different clusters.

These correlations were further studied by analysing how some broad item descriptors correlated with, or could account for, the cluster profiles. A cluster profile was the aggregated p-value residuals for the countries in the cluster. This analysis gave several specific findings, and the clusters could thus to some degree be accounted for by broad descriptors of the items. The most successful descriptor in this respect was the separation of items requiring the use of the stimulus material from items that could be responded to without the stimulus material. This item descriptor particularly separated students from the North-West European countries from the East European countries; East European countries performed relatively better on items that were not dependent on the stimulus material, while the North-West European countries performed relatively better on items that were tightly related to information in the stimulus material. This characteristic demonstrated that the literacy aspect in the PISA science measure, when taken to mean literacy in its fundamental sense – the ability to interact with texts describing science-related phenomena – is an important characterisation of the competency measured in PISA. I suggested that this aspect of the concept should be treated in the framework in some more detail than has so far been the case.

Another distinct feature revealed by this analysis was that the group of English-speaking countries performed relatively better on items testing the students' mastery of scientific process skills. Furthermore, a specific Nordic perspective was included in paper III since prior analyses of other data sets have indicated the presence of a Nordic profile. However, the result from this analysis is that the Nordic countries are only moderately linked by these item residuals.

The residuals were also considered from a psychometrical perspective. In international assessments these residuals represent a source of measurement error that one seeks to minimise. This error could be labelled the 'standard error of international measurement'. The logic underlying this perspective is that an international test intends to measure the same trait in all countries, and thus, the items should be approximately equally difficult in each country, when controlling for the overall achievement level in each country. One way to look into this phenomenon would be to range all items, for each country separately, according

to the difficulty. Ideally, the sequence of the items should then be equal in all countries. The item-by-country interactions expressed in the p-value residuals is another way of studying more or less the same phenomenon. Non-zero p-value residuals reflect that the difficulties of the items vary from country to country. It is not easy to conclude from the paper that the residuals reported represent a major problem. However, it is evident that the problem is relatively larger for some countries than for others. It was recommended that countries could perform regional analyses where the emphasis is on comparisons with countries whose achievement profiles across the items are more similar. The country clusters reported in the paper represent possible regions for such analyses.

On the other hand, while the residuals would be regarded as expressions of an error component when seen from a psychometrical perspective, the item-by-country interactions could also be perceived as information that may give us an insight into how students' achievements vary across countries depending on the context, format, or issue addressed in an item. Natural phenomena and scientific concepts are embedded in the social and cultural contexts in which they are observed and used, for instance through language. The country-specific achievement profiles may reflect this, and international comparisons may be used to gain insight into this phenomenon. However, a clearer conceptual understanding of how the social and cultural context interferes with people's scientific world-views needs to be developed. In light of this discussion, the paper suggested that analyses aimed at utilising the differences in achievement profiles in order to study the regional differences could take advantage of the field trial data, since the residuals in general are larger in the field trials.

Paper III is a very long paper including much detail, and as such many very specific findings are reported in the paper that go beyond the overall purpose of this thesis. One example is the methodological contribution of this paper to finding ways of establishing the stability of the solution. This part demonstrates that even if the p-value residuals are perceived as item- and country-specific information, it is possible to use this information to establish firm and stable descriptions of distinct achievement profiles across the participating countries.

Some of the cluster-specific characteristics were mentioned in this summary merely as examples illustrating the more general point: that the p-value residuals are expressions of characteristics of countries beyond the overall achievement level. As such this analysis is also an analysis of the item-specific information not contained in the overall score. Paper III illustrates the fact that an international comparative perspective on qualitative aspects of students' competence in science may be studied in secondary analyses of the item data, and the potential to do so will be particularly high in PISA 2006 when scientific literacy is the major domain, and when even more countries will be participating. This would make it possible to study the finer details that can characterise relative strengths and weaknesses across countries. Such information could potentially inform us about how different languages, different everyday experiences, different curricula, or other differences in students' lives, affect their scientific understanding of their world.

4.7 Concluding summary

In the material presented above I have given a comprehensive rationale for why the study of information derived from the single items is worthwhile. I have reviewed the nature of this information, both theoretically and by empirical studies of the information, and I have furthermore used methods innovatively in this work. The overall aim has been to document the nature of the information in the single items, and it has been a central aim to learn how this source of information may be analysed, and for what purposes it might be analysed. I believe these aspects have already been thoroughly presented in the above. I will therefore, in this final summary, focus on the line of exploration presented in the three empirical papers, and furthermore, I would like to state where this exploration has, for now, ended.

Paper I demonstrated that the information derived from the single items is not very reliable when conceived of as a measure of students' knowledge and thinking in specific contexts, and paper II demonstrated that the item-specific information across a small set of items did not reflect student characteristics in terms of their mental representation of the concept or phenomenon addressed in the items. However, paper III has demonstrated that when the item-specific data is aggregated to countries it is evident that this information does not only reflect the overall ability measured in the test or the associated random errors or fluctuations due to the uniqueness of the item. Paper III establishes that the information in the item, beyond the overall difficulty of the item and the average achievement level in the country, is a source for describing qualitative differences between the participating countries. Furthermore, paper III sought to develop this description and concluded that with PISA 2006, where scientific literacy will become the major domain, it will be possible to develop richer descriptions of these differences, and furthermore, it may be possible to identify more clearly possible sources why some countries have achievement profiles that are closer to each other. Therefore there is good reason for being optimistic about the continuation of this exploration.

Paper I

Olsen, Turmo & Lie (2001): Learning about students' knowledge and thinking in science through large-scale quantitative studies, *European Journal of Psychology of Education*, 16(3), pp. 403-420.

Learning about students' knowledge and thinking in science through large-scale quantitative studies

Rolf V. Olsen

Are Turmo

Svein Lie

University of Oslo, Norway

The main issue addressed in this article is that there is much to learn about students' knowledge and thinking in science from large-scale international quantitative studies beyond overall score measures. Response patterns on individual or groups of items can give valuable diagnostic insight into students' conceptual understanding, but there is also a danger of drawing conclusions that may be too simple and nonvalid. We discuss how responses to multiple-choice items could be interpreted, and we also show how responses on constructed-response items can be systematised and analysed. Finally, we study, empirically, interactions between item characteristics and student responses. It is demonstrated that even small changes in the item wording and/or the item format may have a substantial influence on the response pattern. Therefore, we argue that interpretations of results from these kinds of studies should be based on a thorough analysis of the actual items used. We further argue that diagnostic information should be an integrated part of the international research aims of such large-scale studies. Examples of items and student responses presented are taken from The Third International Mathematics and Science Study (TIMSS).

Introduction

The Third International Mathematics and Science Study (TIMSS) is a large-scale comparative quantitative study involving 45 participating countries. The study was initiated and run by the International Association for the Evaluation of Educational Achievement (IEA). Nearly one million students from 15,000 schools participated in the tests taking place in 1995. The study focused on three different populations, 9-year-olds, 13-year-olds, and students in their final year of upper secondary education. This study could, therefore, be characterised as one of the largest and most ambitious studies ever within the field of education.

Over the last years, the Norwegian TIMSS research group has presented several publications on how students responded to the cognitive items. This gives a comprehensive documentation on students' knowledge and thinking in science in our country (Angell,

Kjærnsli, & Lie, 1999, 2000; Kjærnsli, Lie, Stokke, & Turmo, 1999). These publications especially emphasise how double-digit coding can be used to systematise students' responses on open-ended items. In this article we will reflect on how diagnostic analysis, in general, should take into account the item characteristics, such as wording and format, before valid conclusions can be drawn. The examples given in the article are used mainly to address this more general focus, and not to analyse students' knowledge and thinking in specific science domains.

Items used in international quantitative studies fall into two main format categories; multiple-choice items (MC) and constructed-response items (CR). Previous international studies such as the IEA Second International Science Study (SISS) have been criticised for their extensive use of MC items. As a consequence of this critique TIMSS included several CR items. The relative number of CR items is even higher in the ongoing OECD study Programme for International Student Assessment (PISA), which covers reading, mathematical and scientific literacy. Our focus in this article is *how* we can learn about students' knowledge and thinking in science through these large-scale international quantitative studies.

Theoretical considerations

A psychometrical vs. a diagnostic perspective

When constructing tests intended to measure a cognitive trait, i.e., students' scientific literacy, one is concerned with designing an instrument that measures the trait with high reliability and validity. Even though all the items, to some degree, measure the same trait, and thus contribute to a reliable overall score, there will be a major portion of item-specific variance. Let us take a simple example. For a typical test (e.g., TIMSS), simple isolated right-wrong items have a correlation (so-called *point-biserial*) with the overall test score of 0.30 to 0.40. In classical psychometrical terms this means that only 9 to 16 percent of the variance for an item can be seen as "true" variance, related to the common trait being measured, whereas the major portion of the variance is item-specific or "error" variance. Furthermore, the items correlate as low as 0.10 to 0.20 with each other, again a sign of very low (1-4%) common variance. This is illustrated in Figure 1. To obtain reasonable reliability, one therefore needs a test consisting of many such items. From a psychometrical perspective it is essential to have a good sample of items from the universe of possible items. On the other hand, from a science educator's perspective, the item-specific variance implies that each item is a universe in itself.

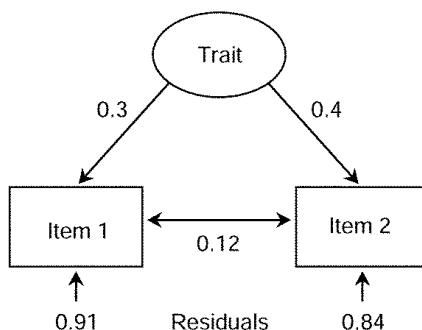


Figure 1. An illustrative example of typical inter-item and item-trait correlations. In this example the trait is represented as a latent variable in a one-factor solution. The standardised item-specific variances (the residuals) are calculated to be 0.91 and 0.84

Based on the above considerations, one could argue from a cost-efficiency perspective, that the major portion of information from a typical test is thrown away when only the overall score is analysed. From a science educator's perspective, there is much more information within reach by additional secondary analyses. The item-specific variance can, in this perspective, be viewed upon as relevant and highly interesting information, while from a test theoretical perspective the item-specific variance is simply regarded as "error" variance. This item specific information may be divided further into several components, for example, variance due to the topic addressed in the item, the item format, and contextual or situational factors specific for the item. So far, the argument has concentrated on the right-wrong dimension only. The argument is even stronger when one also takes into account the item specific information about *what* the student responses were, not only whether the responses were correct or not.

In this article we will focus on the diagnostic information that can be drawn from the distribution of responses to items in different formats. Although this article is written from a science educator's perspective, we will emphasise that this perspective is not necessarily in conflict with the psychometrical perspective. In our opinion it is possible to construct items, that together with all the other items form a reliable and valid test that measures a general trait (e.g., scientific literacy), and, at the same time, gives valuable diagnostic information. It is, however, important to point out that these diagnostic dimensions should not be regarded as traits. Nevertheless, we will argue that the study of such diagnostic dimensions should be included in the overall framework for international quantitative studies. This would ensure that the diagnostic perspective is pursued during the item construction process. As will be shown and discussed in this article, the structure of the items has a crucial impact on how much and what type of diagnostic information we obtain about students' knowledge and thinking in science.

The alternative framework paradigm in science education

Unfortunately, studies such as TIMSS have, to some extent, been ignored by the community of researchers in science education. The popular trend has, for some years, been to focus on qualitative studies.

A very important research aim in science education during the last decades has been to gain some insight into students' personal mental models of phenomena, scientific models, and theories that are part of science curricula. A rich source of evidence exists where students construct their own models that are in conflict with the scientifically accepted views. This research could be characterised as the alternative framework paradigm. A broad overview of the work that has been done, may be found in Pfundt and Duit's (1994) bibliography of research into students' alternative frameworks in this field. Also the review by Laws (1996) and the handbook edited by Fraser and Tobin (1998) are valuable sources for a more extensive overview of this research. We will argue that large-scale studies, such as TIMSS and PISA, can give valuable contributions to this field of research. A major strength of these studies is their international comparative perspective, which gives the opportunity to study similarities and differences in students' knowledge and thinking in science across the world. Some studies with similar aims have been presented (Angell et al., 2000; Smith, Martin, Mullis, & Kelly, 2000), but the rich data material has a great potential for more in-depth studies from this perspective. It is not within the scope of our article to address this comparative perspective in-depth, but we will present a brief example later.

Within this alternative framework paradigm a jungle of different terminology exists for students' own personal conception of scientific models, concepts, and theories. These terms reflect different dimensions of students' mental models and they also carry different connotations reflecting fundamentally different views on how learning occurs in science and how the knowledge is integrated within the students' cognitive structures. We find it relevant and necessary to address this issue further by giving a short review of the different terminology used within the field of science education.

The terms *alternative paradigms* and *alternative frameworks* (Driver & Easley, 1978) both give an immediate impression that these personal conceptions are part of larger, conceptual networks, something similar to Piaget's schemata. Also, phrases including the word *theory* give a similar focus. The same could be said about the term *children's science* (Gilbert, Osborne, & Fensham, 1982; Osborne & Freyberg, 1985). However, ample evidence exists that students' conceptions of science concepts and theories are not embedded in theory-like structures. Instead, their conceptions could be characterised as being highly situational (Hennessey, 1993), dependent on content, and in the end, they could be said to be fragmentary rather than part of a theory. Consider, for instance, the concept of force. In physics, students meet this concept through the laws of Newton, which give a definition of the term force. The same laws, and thereby the same universal force concept, apply to a wide range of phenomena explored in physics classrooms. However, studies have shown that students use different kinds of concepts for force in different contexts (e.g., Angell, 1996; Halloun & Hestenes, 1985; Kupier, 1994).

Terms like *facets* (Minstrell, 1992), *phenomenological primitives* (DiSessa, 1993), and *intuitive ideas* (Angell, 1996) do differ from one another, but they all give a "microscopic" view on knowledge as not being chunked up in "theories", but rather being piecemeal, fragmentary, and context-dependent.

These microscopic units of knowledge are, of course, connected into larger entities, otherwise thinking would be a random process. What could be observed among students is rather that students' thinking is very stable and cannot easily be changed by instruction (Duit & Treagust, 1995; Scott, 1992), but this stability is not due to a knowledge structure similar to overarching theories. We see these networks of knowledge as contextual and situational dynamic structures. This could explain why two slightly different items, that in a content analysis would be characterised as equivalent, are answered very differently by students.

In the Nordic and German research literature the term *everyday conception* (*hverdagsforestilling* [no], *Alltagsvorstellung* [ge]) is frequently used (e.g., Nielsen & Thomsen, 1983). Driver, Asoko, Leach, Mortimer, and Scott (1994) suggest the term "students' informal ideas" for this common sense knowledge. This term focuses on how these conceptions are formed, that is, they are conceptions formed in an everyday context, for instance, through the sensomotoric experiences being made in contact with the physical world. Consider, for example, how you actually can feel the "coldness" of an object. This everyday conception of coldness enables a person to act rationally in his or her daily life, for instance, by closing the open window that is perceived as being the source of coldness in a room. In a scientifically sound model of thermal phenomena, one will never meet a concept like coldness. Instead, we talk about heat (or energy) flowing in the opposite direction.

It is not our aim to supply this debate on the use of terminology with new arguments. We are, however, concerned with the possibility for research on students' conceptions from a variety of theoretical perspectives and for different purposes. We now know that interventions or specific teaching strategies, based on the knowledge of students' alternative conceptions, do not easily give the intended effect (Harlen, 1999). Students' conceptions seem to be almost immune towards extinction. One possible explanation for this is that these conceptions are based on everyday experiences and their validity is confirmed every time a phenomenon related to the concept is encountered.

What does it mean to "know" or "understand"?

The focus in this article raises a fundamental question: What does it mean to "know" or "understand" something? By raising this question, it is not our intention to focus on the nature of knowledge and understanding from an epistemological perspective. Our aim is to focus on how we may obtain access to students' knowledge by asking different types of questions. This brings the following questions into focus: Does "knowing" require the ability to produce a self-contained answer as in CR items? Or do persons "know" when they are able to choose the right answer in an MC item? Do they "know" if they are able to come up with the right answer in an interview situation, where interaction with the person asking the question is taking

place? Are all formats equally relevant in uncovering students' "knowledge"? Or do different formats give us access to different dimensions or aspects?

As mentioned earlier, TIMSS and other (particularly earlier) international studies (e.g., SISS) have been criticised because of the total dependence on MC items. This critique states that MC items do not give access to students' thinking and knowledge in science because the distractors are chosen for other reasons than the intended. MC items may be answered simply by guessing, or by strategies of elimination, where the students are not using their knowledge in science. Some studies have also shown that students' answers may be due to problems of interpreting what the question is about, often combined with some specific problems with a word or a phrase in the stem or in one or more of the distractors (Schultz, 2000). Clerk and Rutherford (2000) analysed interviews with students who had been exposed to a test with MC items. They found that distractors were chosen by students who did not hold the target misconception the item was intended to diagnose, and that language usage, for example, misinterpretation of the question text, was frequently the reason behind choices. However, all the items they studied had serious flaws and errors that could easily be corrected. There are still reasons to be somewhat cautious before claims on students' conceptions are made.

This critique is often followed by the more fundamental issue of a need for more qualitative research methods giving more direct access to students' reasoning patterns and problem solving strategies. In the final analysis, this fundamental critique could be said to reflect a different view of what constitutes knowledge in general. From a Vygotskian perspective, it could be argued that students' knowledge of a science topic is not related to how the student can answer questions in isolation. Knowledge is rather something one expresses in conversation with someone who has superior knowledge of the actual domain, for instance, a science teacher. A conversation such as this will give some evidence of how far a student can stretch under guidance, or in Vygotskian terms, it will give an insight into the student's zone of proximal development (Vygotsky, 1986). We believe that this perspective is important to bear in mind when interpreting results from written tests like TIMSS and PISA. We will return to this issue in more detail later.

Due to the many advantages of MC items (cost-efficient, high reliability), they will most likely continue to play an important role in large-scale assessments, regardless of the critique. When it comes to the usefulness of MC items to get insight into students' knowledge and thinking, we would like to support a position taken by Tamir (1990), who claims that although MC tests can be rightly criticised, their structure, when wisely used, makes them an excellent diagnostic tool for identifying students' conceptual understanding, especially when students are asked to comment on their choice. When interpreting responses to MC items, it is of crucial importance to take into careful consideration the role of the distractors, both their structure and content. In the next section, we will, therefore, discuss some features and the function of distractors in different cases, together with some consequences for identification of students' knowledge and thinking.

Interpreting responses on MC items

How should results from MC items be interpreted? How can results from this type of item give information about students' knowledge and thinking? When interpreting such results, it is important to take into consideration the function of the distractors: How do the students most likely make use of the distractors during the solution process? Our focus in the following is not to categorise items, but rather to discuss aspects that characterise the function of the individual response alternatives in the solution process. It should be emphasised that there will be individual differences between students when it comes to the tendency and ability to use the distractors actively in the process of producing a response.

Distractors as a check list. In some cases, all the alternatives appear to be plausible responses at first sight. A closer investigation, for example, a calculation, is necessary before one response can be selected. Therefore, it will not be helpful to use an elimination strategy to

reach the correct response. A typical situation is an item where students are asked to perform a calculation and then check if the answer is among the alternatives given. Other cases may be items where all the possible relevant answers are given as alternatives, so that there are no more relevant answers. It is a typical characteristic for this type of item that the MC format does not likely change the students' strategies, compared to a parallel CR version. Reasoning and calculations, based on information in the stem, are typically done the same way in both cases. The function of the distractors is reduced to a checklist for the established answer. The response patterns should, theoretically, not be very different in the two formats. However, students making an error in the calculations will not find their answer among the alternatives. These students will tend to recalculate. Theoretically, the p -value (the percentage of students answering the item correctly) for the item should, for that reason, be somewhat higher than for the CR version.

Distractors that can be eliminated. This category is for distractors that can easily be eliminated simply by being obviously incorrect, either due to silly or irrelevant content or by logical flaw. Because the students can reach the correct answer more easily than in a similar CR version, that is, by elimination, the MC version, most likely, will have a significantly higher p -value. In some cases, the stem asks for the correct or best explanation to a particular phenomenon. Some distractors may be eliminated because they represent incorrect science, whereas other distractors, albeit representing "correct" science, do not provide any plausible explanation to the phenomenon at hand. The latter case often represents a more demanding strategy for the students.

Response alternatives that help students to understand the question. Often, the response alternatives help students to understand the question intent, in particular when the stem includes difficult terminology. Compared to a parallel CR version, the MC version will have a higher percent of correct responses. An interesting hypothesis is that MC items, to some extent, reflect a Vygotskian perspective of knowledge. If a question is stated and the student is left alone to answer this question in his or her own words there is a possibility that some misunderstanding will occur. By giving response alternatives, the student is provided with additional information on how to interpret the question. In this sense, the distractors have the same function as a conversation partner, hindering some of the possible misinterpretations.

Missing distractors. In some cases, the students get help because one or more common misconceptions do not appear as a distractor. Again, in such cases, the MC version should get a much lower p -value when transformed into a CR version, where the actual misconception(s) readily will emerge.

Cognitive traps. It happens that distractors add difficulties compared to a parallel CR version. Therefore, based on this effect alone, the p -value should become lower for the MC version. Sometimes an alternative seems plausible when given as a distractor, but most students would never have come up with the alternative if the item had been in CR format. It could, therefore, be argued that the students have been tricked and led into giving an incorrect response. For this reason, we use the term "cognitive traps" for such distractors. Some students may be "trapped" into the tempting world of one of the distractors. At the same time, the combination of all response alternatives may still help many students to understand what the item is about and/or help them to reach the correct answer.

Response alternatives defining the question. In a special type of MC item the question consists of comparing the distractors. The correct answer cannot be reached without comparing the given alternatives. The stem may be formulated as "Which of these explanations is the best one?" etc. In these type of items, the response alternatives are often full and independent sentences or they are words that complete the stem into a full sentence. Obviously, such items cannot be given without the response alternatives, and thus have no parallel CR version.

Obviously, there are no sharp distinctions between all the “distractor aspects” presented here and we will again emphasise that they are not mutually exclusive. Individual distractors may well have more than one feature at the same time, and a particular MC item may have distractors with very different characteristics. In the empirical section, examples will be given that further specify and illustrate the aspects presented above.

Systematising responses on CR items

Both the TIMSS and PISA tests have made use of several CR items. A major motivation for including this type of item is to ensure higher test validity. It is obvious that, in many cases, it is more relevant to use the CR rather than the MC format. If one, for example, wants to assess students' ability to communicate their knowledge and understanding, as in PISA where communicating is seen as one important aspect related to scientific literacy, the CR format is clearly more adequate than the MC format. However, from a cost-efficiency perspective CR items represent a heavy burden. One may argue that, for CR items, it is even more important than for the MC counterparts to carry out diagnostic analyses in order to make use of the rich item-specific diagnostic information available.

By including CR items one obtains access to a lot of interesting information about students' knowledge and thinking. In ordinary score coding most of this information is lost. It is, therefore, necessary to apply a refined coding system to be able to make diagnostic quantitative analyses of responses to CR items. This coding system should encompass both the correctness dimension and the diagnostic aspect. In TIMSS this was provided by a two-digit system originally proposed by the Norwegian TIMSS team (Lie, Taylor, & Harmon, 1996). In PISA a similar development process has occurred.

The fundamental basis of coding CR items in both TIMSS and PISA was *simplicity, authentic student-response orientation and acceptable inter-marker reliability*. For many items the correctness, on the one hand, and method/error/type of explanation, on the other, are strongly interrelated. Instead of coding these two aspects separately, the idea behind the two-digit system is to apply only one two-digit variable that takes these issues into account. The codes follow a simple system. The basic idea is the following scheme (for a two-point TIMSS item):

- Codes 20-29: Correct response. Score=2.
- Codes 10-19: Partial response. Score=1.
- Codes 70-79: Incorrect response. Score=0.
- Codes 90 (off-task) and 99 (blank): Nonresponse. Score=0.

The first digit gives information about the score. The second digit indicates the method used, type of explanation/examples given, or error/misconception demonstrated. The score (dimension of correctness) is, thus, linked to the other integrated aspects in such a way that the data can be analysed both for correctness and for diagnostic information. It must be emphasised that the second digit is a pure categorical symbol with no general meaning across items, that is, it is not possible to use these digits for psychometrical analysis across items, simply because the second digit does not refer to any general trait and taken alone, it has no diagnostic meaning.

Empirical results and discussion

So far, we have presented important theoretical aspects and approaches related to the interpretation of results on both MC and CR items. In this empirical part of the article we will illustrate these theoretical perspectives by presenting some items and Norwegian results from the TIMSS study. All the results given are from the test of population 2 defined as the two adjacent grade levels with most students at the age of 13 at the time of testing. In Norway

these are grades 7 and 8, the last year of primary education and the first year of lower secondary education. The sample size in the TIMSS main test, in Norway, was 5,758 students.

Double-digit coding – One example

To illustrate how the double-digit coding system can be used to analyse students' response pattern on CR items we will present a TIMSS item dealing with the significance of the ozone layer:

Write down one reason why the ozone layer is important for all living things on Earth.

This item also exemplifies how the coding system may be used in international comparative analysis. As seen in Table 1, this item was coded by the use of double-digit codes falling in two main categories, correct and incorrect responses, in addition to the non-response category (here the codes 90 and 99 are collapsed into one category). Four codes for correct answers were used and six codes for no credit responses. In Table 1, we have given the Norwegian TIMSS results from grade 7 and 8 together with results from an additional Norwegian study at grade 9 in 1997 ($N=2,721$). From the table it is possible to study how the response pattern evolves from grade 7 to grade 9 in Norway. The international average for grade 8 is also given (students at the same age as Norwegian 8th-graders). It is, therefore, possible to analyse the Norwegian results in an international perspective.

Table 1

Distribution of student responses (%) on the “ozone layer” item

Code	Response characteristics	Grade			Int 8 th
		7 th	8 th	9 th	
	Correct response	53	70	65	54
10	Refers to protection against the UV radiation from the sun.	18	42	38	28
11	Refers to protection against dangerous or too strong radiation from the sun but does not mention UV. <i>Example: Because it keeps the sun's rays from being too strong.</i>	29	26	24	19
12	Mentions that the ozone layer protects humans so we do not get sunburned/skin cancer. NOTE: If UV is mentioned, code 10.	6	3	2	5
19	Other correct	0	0	1	4
	Incorrect response	39	25	26	33
70	Confuses the effect of the ozone layer with the greenhouse effect. <i>Example: It keeps the heat in.</i>	2	1	2	1
71	Confuses protection against heat. <i>Example: Everything will melt without it.</i>	7	5	6	6
72	Refers only vaguely to protection. <i>Examples: All living things will die without the ozone layer. It protects the Earth/us.</i>	11	7	6	7
73	Refers to or confuses oxygen, O ₂ with ozone, O ₃ <i>Example: It is needed for respiration.</i>	13	7	6	6
74	Sees the ozone layer as a barrier for the atmosphere. <i>Example: It keeps the air around the earth.</i>	3	1	1	2
79	Other incorrect	4	3	6	11
	Nonresponse	9	5	9	15

The results show that the Norwegian students could be characterised as knowing quite a lot about why the ozone layer is important. Among the Norwegian grade 8 students there is a significantly higher percentage of answers given full credit than the international average. As many as 42 percent of the Norwegian students refer explicitly to UV radiation (code 10). This is a very good result in an international perspective. Among Norwegian grade 7 pupils the type of radiation is not known to the same extent. It is also interesting to notice the higher percentage of correct answers for grade 8 compared to grade 9 in Norway. This has to be characterised as a surprising finding. One tentative explanation suggested refers to changes in media focus on this topic from 1995 to 1997, when the two tests were administered. Among the not credited answers are several interesting codes. Most of these responses are coded as 72 or 73. Answers coded as 72 refer to the fact that the ozone layer protects us (using only vague, general terms) while code 73 consists of responses where the students show that they are not aware of the difference between oxygen and ozone.

By this example we have shown some research ideas which can be pursued by the use of the double-digit coding system. Several other illustrative and more extended examples of how double-digit coding may be used to analyse students' responses are given in Angell et al. (2000).

The potential cultural bias of MC items

The Nordic countries have been particularly interested in finding out to what extent country performances on tests such as TIMSS partially can be a function of item formats. Initially, this interest was driven by a hypothesis that cultural differences exist because of the various assessment traditions in the participating countries. In the Nordic countries students are not used to being assessed by multiple-choice items while US students indeed are familiar with this test format. However, a review of the science and mathematics items in population 2 in TIMSS concluded that differences between the Nordic countries and USA could not, to any significant extent, be explained by item formats. For the science items US students performed relatively better on MC items which could be answered by an elimination strategy, while Nordic students, and Norwegian students in particular, performed well on CR science items as compared to both US and international averages. The same trend did not emerge from the analysis of the mathematics items, leading to the overall conclusion that there seems to be no major cultural bias due to item format per se (Lie, Kjærnsli, & Brekke, 1997). The Nordic co-operation turned, however, into a more fundamental and interesting reflection on how different item formats measure students' knowledge or different aspects of this knowledge (Gisselberg, Kjærnsli, Lie, & Weng, 1996; Thorseng, 1997).

The influence of the item format and distractor characteristics

In the above-mentioned Nordic study (ibid.), some of the items from the TIMSS main test were given in two additional versions. MC items in TIMSS were reformulated into CR and vice versa. Different versions of MC items were also given to students. As a part of this co-operation, two additional studies were implemented in Norway. The sample sizes in these studies were 929 (this study is from now on referred to as version 2) and 220 (hereafter referred to as version 3) respectively. Both the original study and the version 2 study had probabilistic samples, which were representative for the population of all Norwegian 13-year-olds. The version 3 study was not a probabilistic sample. Inferences from this last study should, therefore, be drawn with caution. In the following, we will use some of these data to discuss the influence of item format and to illustrate the different distractor characteristics presented earlier.

Pulse and breathing rate. The first item presented is about pulse and breathing rate. The item is chosen to illustrate how distractors can be eliminated. The TIMSS version of the item is an MC item with five alternatives. The stem is formulated as follows:

Immediately before and after running a 50 meter race, your pulse and breathing are taken. What changes would you expect to find?

In TIMSS the response alternatives were given as:

- A. no change in pulse but decrease in breathing rate
- B. an increase in pulse but no change in breathing rate
- C. an increase in pulse and breathing rate
- D. a decrease in pulse and breathing rate
- E. no change in either

The alternatives in this item all consist of two knowledge elements, knowledge about change in pulse and change in breathing rate. This obviously makes it more difficult to interpret the results. The MC version of the item is a typical example of an item where an elimination strategy is relevant. If the student knows that you breathe faster after running, he or she could logically eliminate all the four distractors. The student actually only has to know that the breathing rate increases after running a race. If the student combines this everyday knowledge with an ability to use an elimination strategy, he or she will arrive at the correct answer. If the student only knows that the pulse increases, he or she can eliminate three of the distractors. The item was also given in CR format by removing the alternatives. The results for the MC version (TIMSS) and the CR version (version 2) are given in Table 2. We have applied the two-digit coding system presented earlier.

Table 2

Distribution of student responses (%) on two different versions of the "pulse and breathing" item

	Responses	TIMSS	Version 2
	Correct response	93	27
20	An increase in pulse and breathing rate	93	18
29	Other correct answers (including both pulse and breathing rate)		9
	Partially correct response		66
10	An increase in/stronger pulse, breathing rate not mentioned		56
11	"It goes faster" or similar type of response		4
12	An increase in breathing rate or stronger breathing rate, pulse not mentioned		5
19	Other partial correct		1
	Incorrect response	9	4
70	An increase in pulse, but no change in breathing rate	4	0
71	No change in pulse but a decrease in breathing rate	1	0
72	A decrease in pulse and breathing rate	2	1
73	No change in either	2	0
79	Other incorrect		3
	Nonresponse	0	3

The results in Table 2 show that almost all students answer the item correctly in the MC version (93%). The response pattern for the CR version is very different. The percentage of full credited answers decreases dramatically to 27 percent. Most students only refer to an increase in pulse (code 10).

In version 3 of the item, the first sentence was the same, and two separate questions were formulated in order to separate the two knowledge elements more explicitly:

Immediately before and after running a 50 meter race, your pulse and breathing are taken.

- a) What changes would you expect to find in pulse?
- b) What changes would you expect to find in breathing rate?

Table 3

Distribution of student responses (%) on part A of version 3 of the "pulse and breathing" item

Responses		Version 3 (part a)
Correct response		87
10	An increase in pulse	87
Incorrect response		7
70	No change in pulse	0
71	A decrease in pulse	1
72	An increase in breathing rate	3
79	Other incorrect	3
Nonresponse		6

Table 4

Distribution of student responses (%) on part B of version 3 of the "pulse and breathing" item

Responses		Version 3 (part b)
Correct response		46
10	An increase in breathing rate	32
11	Stronger/more heavy breathing	14
No credit		21
70	No change in breathing rate	1
71	A decrease in breathing rate	1
72	An increase in pulse/heart rate	13
79	Other incorrect	6
Nonresponse		32

The results presented in Tables 3 and 4 show that almost all the students answer the first part correctly (87%). The percentage of full credited answers on part two is much lower (46%). The previous version of the item (version 2), and even more explicitly version 3, show that students seem to be very aware of the fact that pulse increases, but the increase in breathing rate is not equally familiar. This is clear from the fact that the percentage of correct answers decreases from part a to part b in version 3 and from the high percentage of students not responding to part b. It could, however, also be that students did not see part b as a new question. If so, this could explain the large increase in nonresponse from part a to part b. This is plausible because the Norwegian terms for breathing and pulse are not very different from each other as discussed below.

We would also like to comment on how items may be used to evaluate differences between countries. Overall, students in the Nordic countries gave very similar answers. When compared to other countries, or the international average, this gives a distinct Nordic profile. However, on these items the distribution of answers was different in Norway compared to Sweden and Denmark (Gisselberg et al., 1996). Swedish and Danish students were more familiar with the phenomenon of an increase in breathing rate than with the increase in pulse, which also is what could be expected from the Norwegian students. However, there are major differences in how the three languages translate the term "breathing". In Norwegian the word is "*pust*", not so different from the word "*puls*". It may be that Norwegian students are confusing these similar looking terms.

Our discussion of the results on the different versions of the item shows that interpretations of the results have to be carried out by thoroughly analysing the item format

and wording. It would have been a too hasty conclusion from the TIMSS results to state that 93 percent of Norwegian students “know” that both the pulse and the breathing rate increases after running. This high p -value could equally well be due to the item format, which gives students the possibility to use an elimination strategy. The alternative versions of the item give us additional information on how to interpret the results. Students seem to be more familiar with the phenomenon of an increased pulse rate than that of an increased breathing rate during running. But, as discussed above, this may partially be seen as a consequence of the wording.

From a science educator’s point of view, we will argue that MC items such as this do not give us much information on students’ knowledge and thinking. It is quite possible that even students with everyday conceptions, not consistent with the scientific model of respiration and blood circulation, can give a correct answer. The item does not test students’ knowledge in science, but rather common sense knowledge. From our perspective, it could, therefore, have been relevant to use structured clusters of items about the topic with this item included. Then, it would be possible to get a more in-depth understanding of what the students actually know and think about science phenomena, concepts and theories, and also check whether students are able to switch from everyday concepts used in everyday contexts to more scientific models used in more formal settings.

Greenhouse. The next item illustrates how response alternatives help the student to understand the question. The original item stem was formulated like this:

The burning of fossil fuels has increased the carbon dioxide content of the atmosphere. What is a possible effect that the increased amount of carbon dioxide is likely to have on our planet?

The original TIMSS version had these response alternatives:

- A. warmer climate
- B. cooler climate
- C. lower relative humidity
- D. more ozone in the atmosphere

This is a good example of an item where the alternatives give explanatory cues, that is, help the students to understand what the item is about and what kind of answer is expected. If this item is given without alternatives, it could be interpreted in different ways. It is not easy for a student to understand what type of answer one expects. Should the answer be a word or a complete sentence? Should the student refer to a primary (e.g., warmer climate) or a secondary effect (e.g., rising of sea level)? The MC format will help to eliminate these ambiguities. The analysis of the CR version (version 2, stem only) gave clues on new relevant distractors which were tested in a subsequent MC version (version 3). In this version distractor D was removed and replaced by two new ones:

- D. destruction of the ozone layer
- E. more difficult to breathe

As expected (see Table 5), the p -value decreases dramatically in the CR version from 61 percent in the TIMSS version to 14 percent. The great variation in the responses on the CR version verifies that this version is difficult for the students to interpret. Thirdly, it is interesting to notice that none of the constructed responses are compatible to the distractors in the original version. This implies that the distractors in the original TIMSS version have no diagnostic value.

The CR version (version 2) also reveals an interesting and well-documented misconception, that the destruction of the ozone layer and the increased greenhouse effect are being confused or even are regarded as the same phenomenon (e.g., Boyes & Stanisstreet, 1993, 1994). This misconception is amplified when explicitly given as a distractor (D in version 3). This could

illustrate a more general issue: When common misconceptions are included as distractors, they tend to function to a certain degree as cognitive traps, as discussed earlier. Stating that 48 percent of the students “have” this misconception would be to jump to conclusions. One alternative interpretation is that some of these students do not retrieve all the relevant information from the stem. Instead, they may simply recognise that the task is to identify an important environmental problem. Therefore, it could be argued that this distractor is a cognitive trap.

Table 5

Distribution of student responses (%) on three versions of the “greenhouse” item

	Responses	TIMSS	Version 2	Version 3
	Correct response	61	14	25
10	A warmer climate	61	9	25
11	Greenhouse effect		5	
	Incorrect response	37	55	66
70	A cooler climate	8	0	2
71	Lower relative humidity	12	0	3
72	More ozone in the atmosphere	17	0	
73	Destruction of the ozone layer		17	48
74	Change in amount of oxygen/more difficult to breathe		7	13
75	Pollution		6	
76	Changes in the atmosphere		5	
77	Plants/animals die or get injured		8	
78	Increased plant growth		3	
79	Other incorrect answers		9	
	Nonresponse	3	30	9

Mammal. The next item has a complex stem including what can be regarded as irrelevant information:

A small animal called the duckbilled platypus lives in Australia. Which characteristic of this animal shows that it is a mammal?

The response alternatives in the original version were:

- A. it eats other animals
- B. it feeds its young milk
- C. it makes a nest and lays eggs
- D. it has webbed feet

It could be argued that distinguishing relevant from irrelevant information is an important process skill in science. However, in our analysis we are concerned with how this irrelevant information, in combination with corresponding distractors, misleads the students. This is another good example of what we have called a cognitive trap. In this item, the irrelevant information is the first sentence about the duckbilled platypus. This information, in combination with either distractor C or D, sets up the trap. The term “duckbilled” may trigger the image of a duck, which of course, has webbed feet, makes a nest, and lays eggs! For students in countries (Australia) where the platypus is well known, the item may work rather differently. They may well know that all the alternatives are correct statements about the animal, whereas only alternative B answers the question at hand.

Table 6

Distribution of student responses (%) on three versions of the “mammal”-item

	Responses	TIMSS	Version 2	Version 3
	Correct response	59	14	30
20	It feeds its young milk	59	4	24
21	The females have breasts		5	6
29	Other correct answers		5	0
	Partially correct response		43	44
10	They give birth to living children		34	41
11	It does not lay eggs		8	3
19	Other partially correct answers		1	0
	Incorrect response	40	32	16
70	It eats other animals	8	0	0
71	It makes a nest and lays eggs	13	3	1
72	It has webbed feet	19	1	0
79	Other incorrect answers		28	15
	Nonresponse	2	15	12

The results in Table 6 indicate that the two distractors, characterised above, as cognitive traps, seem to work as such. In version 2, the same item was given as a CR item. None of the distractors in the MC version appeared frequently in these student constructed responses. It is particularly interesting to notice that 34 percent of the students give the answer “They give birth to living children” in the CR item (version 2). This could be an example of what we characterised as a missing distractor in the TIMSS original version, partly explaining the high *p*-value for this version, despite the two cognitive traps described earlier. This response was even more common in the third version (41%), which was formulated like this:

What is the difference between mammals and other animals?

The *p*-value for this version was much higher than for version 2, but significantly lower than in the TIMSS version (see Table 6). When comparing versions 2 and 3, we can observe that the fully correct responses increase from 14 to 30 percent and that the amount of other incorrect responses decreases from 28 to 15 percent. The only difference between these two versions is the removal of the first sentence giving task irrelevant information. If the intended task is to distinguish irrelevant from relevant information, it could be argued that the first sentence and the two corresponding distractors should be included. But, if the intention is to find out if students know what defines mammals then the item should be reformulated.

Earth’s surface. We will close this section by showing an example (without results) of an item type that we find particularly useful when assessing students’ knowledge and thinking in science. This is an example of an item where the response alternatives define the question. This class of items cannot, because the stem in itself is not meaningful, be reformulated into a parallel CR version.

Which best describes the surface of the Earth over billions of years?

- A. A flat surface is gradually pushed up into higher and higher mountains until the Earth is covered with mountains.
- B. High mountains gradually wear down until most of the Earth is at sea level.
- C. High mountains gradually wear down as new mountains are continuously being formed, over and over again.
- D. High mountains and flat plains stay side by side for billions of years with little change.

These type of items initiates a cognitive process where all the alternatives have to be actively evaluated. The student cannot, without relevant knowledge, simply eliminate any of the alternatives, because they are all plausible statements. In general, items asking the students to find "the best" alternative can include several "true" statements, even as distractors.

Summary and recommendations

Within the research community in science education, there has been a tendency to show little interest in the data from studies such as TIMSS. Qualitative and smaller scaled studies have been the popular trend. In this article, we have argued that data from both MC items and CR items, in international comparative studies, can give valuable insight into students' knowledge and thinking in science. However, interpretations of results from these kind of studies must be based on thorough analysis of the actual items used. We have demonstrated that even small changes in the item wording and/or the format can have large influences on the response pattern. This is a major challenge when drawing diagnostic interpretations from international comparative studies such as TIMSS and PISA. We have demonstrated that there is no systematic format effect. Rather, we have demonstrated that there is a complex interaction between item characteristics and students' responses. However, we have argued that these effects could be better understood by using a typology for different types of distractors. Furthermore, we have illustrated how an additional small scale study can give supplementary diagnostic information.

We have also argued that responses to CR items should be systematised in more detail than just deciding if the response is worthy of one or more score points or not. Double-digit coding has been presented as one very useful approach.

Based on our theoretical discussions and the empirical findings, we wish to give some recommendations for future large-scale international studies. Firstly, our discussion has shown that proper item construction is vital for diagnostic potential. We will, therefore, propose that the diagnostic perspective is emphasised when items are developed. In this article we have been particularly interested in discussing how the distractors that are used influence the inferences that can be made about student knowledge and thinking. We have argued that multiple-choice items, in addition to contributing to a reliable score, may also, when carefully constructed and interpreted, give valuable diagnostic information.

Secondly, we will suggest that clusters (or at least pairs) of items should be designed together to evaluate a common specific diagnostic dimension. This could be implemented in the test design of large-scale studies by giving different versions of the same question, as demonstrated, in separate test booklets. Such clusters could be linked to trace specific thinking patterns and problem solving processes.

Finally, we are not advocates for including large numbers of constructed-response items in large-scale assessments if the responses are not used beyond contributing to overall test scores. We have emphasised the importance of a stronger focus on the diagnostic information potential for the items used. In our view the diagnostic perspective should be regarded as an integrated part of the research aims for future international large-scale assessments.

References

- Angell, C. (1996). *Elevers fysikkforståelse. En studie basert på utvalgte fysikkoppgaver i TIMSS* [Students' understanding of physics. A study based on selected physics items in TIMSS]. PhD Thesis, University of Oslo.
- Angell, C., Kjærnsli, M., & Lie, S. (1999). *Hva i all verden skjer i realfagene i videregående skole?* National TIMSS report population 3. Oslo: Universitetsforlaget.
- Angell, C., Kjaernsli, M., & Lie, S. (2000). Exploring students responses on free-response science items in TIMSS. In D. Shorrocks-Taylor & E.W. Jenkins (Eds.), *Learning from others* (pp. 159-187). Dordrecht: Kluwer.

- Boyes, E., & Stanisstreet, M. (1993). The "greenhouse effect": Children's perception of causes, consequences and cures. *International Journal of Science Education*, 15, 531-552.
- Boyes, E., & Stanisstreet, M. (1994). The ideas of secondary school children concerning ozone layer damage. *Global Environmental Change*, 4, 311-324.
- Clerk, D., & Rutherford, M. (2000). Language as confounding variable in the diagnosis of misconceptions. *International Journal of Science Education*, 22, 703-717.
- DiSessa, A.A. (1993). Toward an epistemology of physics. *Cognition and Instruction*, 10, 105-225.
- Driver, R., & Easley, J. (1978). Pupils and paradigms: A review of literature related to concept development in adolescent science students. *Studies in Science Education*, 5, 61-83.
- Driver, R., Asoko, H., Leach, J., Mortimer, E., & Scott, P. (1994). Constructing scientific knowledge in the classroom. *Educational Researcher*, 23, 5-12.
- Duit, R., & Treagust, D.F. (1995). Students' conceptions and constructivist teaching approaches. In B.J. Fraser & H.J. Walberg (Eds.), *Improving science education*. Chicago, IL: The University of Chicago Press.
- Fraser, B.J., & Tobin, K.G. (1998). *International handbook of science education* (2 vols.). Dordrecht: Kluwer.
- Gilbert, J.K., Osborne, J., & Fensham, P.J. (1982). Children's science and its consequences for teaching. *Science Education*, 66, 623-633.
- Gisselberg, K., Kjærnsli, M., Lie, S., & Weng, P. (1996). *Preliminary notes from a Nordic study on item formats in TIMSS*. Unpublished manuscript.
- Halloun, I.A., & Hestenes, D. (1985). Common sense concepts about motion. *American Journal of Physics*, 53, 1056-1065.
- Harlen, W. (1999). *Effective teaching of science. A review of research*. Edinburgh, UK: Scottish Council for Research in Education.
- Hennessey, S. (1993). Situated cognition and cognitive apprenticeship: Implications for classroom learning. *Studies in Science Education*, 22, 1-41.
- Kjærnsli, M., Lie, S., Stokke, K.H., & Turmo, A. (1999). *Hva i all verden kan elevene i naturfag?* [What do students know in science?]. Oslo: Universitetsforlaget.
- Kupier, J. (1994). Students ideas of science concepts: Alternative framework? *International Journal of Science Education*, 16, 279-292.
- Laws, P.M. (1996). Undergraduate science education: A review of research. *Studies in Science Education*, 28, 1-85.
- Lie, S., Kjærnsli, M., & Brekke, G. (1997). *Hva i all verden skjer i realfagene?* National TIMSS report population 2. Oslo: Universitetsforlaget.
- Lie, S., Taylor, A., & Harmon, M. (1996). Scoring techniques and criteria. In M.O. Martin & D.L. Kelly (Eds.), *Third international mathematics and science study. Technical report: Vol. 1. Design and development* (chap. 7, pp. 1-16). Chestnut Hill, MA: Boston College.
- Minstrell, J. (1992). Facets of students' knowledge and relevant instruction. In R. Duit, F. Goldberg, & J. Niedderer (Eds.), *Research in physics learning: Theoretical issues and empirical studies. Proceedings of an international workshop held at the University of Bremen* (pp. 110-128). Kiel: Institute for Science Education, University of Kiel.
- Nielsen, H., & Thomsen, P.V. (1983). *Hverdagsforestillinger om fysik* [Everyday conceptions in physics]. Århus: University of Århus.
- Osborne, R., & Freyberg, P. (1985). *Learning in science. The implication of children's science*. Auckland: Heineman.
- Pfundt, H., & Duit, R. (1994). *Bibliography. Students' alternative frameworks and science education* (4th ed.). Kiel: Institute for Science Education, University of Kiel.
- Schoultz, J. (2000). Conceptual knowledge in talk and text: What does it take to understand a science question? In *Att samtala om/i naturvetenskap. Kommunikation, kontext och artefakt*. PhD thesis (partly in English), University of Linköping.
- Scott, P.H. (1992). Pathways in learning science: A case study of the development of one student's ideas relating to the structure of matter. In R. Duit, F. Goldberg, & J. Niedderer (Eds.), *Research in physics learning: Theoretical issues*

- and empirical studies. *Proceedings of an international workshop held at the University of Bremen* (pp. 203-224). Kiel: Institute for Science Education, University of Kiel.
- Smith, T.A., Martin, M.O., Mullis, I.V.S., & Kelly, D.L. (2000). *Profiles of student achievement in science at the TIMSS international benchmarks: U.S. performance and standards in an international context*. Chestnut Hill, MA: Boston College.
- Tamir, P. (1990). Justifying the selection of answers in multiple choice items. *International Journal of Science Education*, 12, 563-573.
- Thorseng, H. (1997). *Spørsmål og svar i naturfag* [Questions and responses in science]. MSc Thesis, University of Oslo.
- Vygotsky, L. (1986). *Thought and language*. Cambridge MA: MIT Press.

L'article met à jour le fait que les études quantitatives internationales à grande échelle nous disent beaucoup, au-delà des mesures générales de scores, sur les connaissances et les pensées des élèves par rapport aux sciences. Les patterns des réponses données à des items individuels ou bien à des groupes d'items peuvent nous éclairer sur la compréhension conceptuelle des élèves, mais risquent de conduire à des conclusions trop simples et non-valides. On discute les moyens de systématiser et d'analyser les réponses à des items à choix multiple. Enfin, on étudie empiriquement les interactions entre certains caractéristiques des items et les réponses des élèves. On montre qu'il suffit d'une petite modification dans la manière de formuler et/ou formater l'item pour produire un effet substantiel dans le pattern de la réponse. Il s'ensuit que l'interprétation des résultats provenant de ce type d'études doit être basée sur une analyse approfondie des items administrés. Il s'ensuit également que l'information diagnostique peut être une partie constitutive des objectifs de recherche internationaux qu'on cherche à réaliser par ce type d'études à grande échelle. Les exemples présentés d'items et de réponses des élèves sont extraits de la Third International Mathematics and Science Study (TIMSS).

Key words: Assessment, Item format, Quantitative methodology, Science education, TIMSS.

Received: December 2000

Rolf V. Olsen. Department of Teacher Education and School Development, University of Oslo, 1099 Blindern, N-0317 Oslo, Norway. E-mail: rolfvo@ils.uio.no

Current theme of research:

Deriving diagnostic information on students' thinking and knowledge in science from data collected in large-scale international comparative research projects.

Most relevant publications in the field of Psychology of Education:

Olsen, R.V. (in press). Introducing quantum mechanics in the upper secondary school: A study in Norway. *International Journal of Science Education*.

Are Turmo. Department of Teacher Education and School Development, University of Oslo, 1099 Blindern, N-0317 Oslo, Norway. E-mail: are.turmo@ils.uio.no

Current theme of research:

Students' knowledge and thinking in science and mathematics in an international comparative perspective.

Most relevant publications in the field of Psychology of Education:

Brekke, G., Kobberstad, T., Lie, S., & Turmo, A. (1999). *Hva i all verden kan elevene i matematikk?* [What do students know in mathematics?]. Oslo: Universitetsforlaget.

Kjærnsli, M., Lie, S., Stokke, K.H., & Turmo, A. (1999). *Hva i all verden kan elevene i naturfag?* [What do students know in science?]. Oslo: Universitetsforlaget.

Svein Lie. Department of Teacher Education and School Development, University of Oslo, 1099 Blindern, N-0317 Oslo, Norway. E-mail: svein.lie@ils.uio.no

Current theme of research:

Students' knowledge and thinking in science and mathematics in an international comparative perspective.

Most relevant publications in the field of Psychology of Education:

Angell, C., Kjærnsli, M., & Lie, S. (2000). Exploring students responses on free-response science items in TIMSS. In D. Shorrocks-Taylor & E.W. Jenkins (Eds.), *Learning from others. International comparison in education* (pp. 159-187). Dordrecht: Kluwer.

Brekke, G., Kobberstad, T., Lie, S., & Turmo, A. (1999). *Hva i all verden kan elevene i matematikk?* [What do students know in mathematics?]. Oslo: Universitetsforlaget.

Kjærnsli, M., Lie, S., Stokke, K.H., & Turmo, A. (1999). *Hva i all verden kan elevene i naturfag?* [What do students know in science?]. Oslo: Universitetsforlaget.

Lie, S., Taylor, A., & Harmon, M. (1996). Scoring techniques and criteria. In M.O. Martin & D. Kelly (Eds.), *Third International Mathematics and Science Study, Technical Report: Vol. 1: Design and Development*. Boston College.

Sjøberg, S., & Lie, S. (1981). *Ideas about force and movement among Norwegian pupils and students* (Report 81-11). University of Oslo.

Paper II

Olsen (2004): The Search for Descriptions of Students' Thinking and Knowledge: exploring nominal cognitive variables by correspondence and homogeneity analysis, *Scandinavian Journal of Educational Research*, 48(3), pp. 325-341.

The Search for Descriptions of Students' Thinking and Knowledge: exploring nominal cognitive variables by correspondence and homogeneity analysis

ROLF VEGAR OLSEN

*Department of Teacher Education and School Development, University of Oslo,
PO Box 1099, Blindern, N-0316 Oslo, Norway*

ABSTRACT *In the Programme for International Student Assessment (PISA) the items are organised in small clusters relating to the same stimulus material (called 'units'). Homogeneity analysis (HA) is used to develop a detailed description of the relationship between all the items in one unit, using the categorical information available in the PISA data. The main findings of the analyses presented are the following: (a) non-respondents are separated from those responding; (b) student patterns across the nominal variables cannot be used to develop a detailed insight into how students' thinking and knowledge is organised beyond the fact that good students perform well on most items; and (c) the profiles for countries are described by the same dimensions, indicating that the relative success of the countries can to a certain degree be explained differentially by these dimensions. From these findings implications for future assessments are suggested.*

Key words: *categorical data analysis; PISA; international assessment; diagnostic assessment*

INTRODUCTION

Students' responses to open-ended (OE) items on cognitive tests are typically scored with points reflecting whether the quality intended to be measured by the item is present or not. The scoring has to be done according to clearly formulated criteria. As a consequence, detailed scoring rubrics are developed. This is the case in PISA, where detailed marking guides are used to classify students' responses to OE items into different categories. The coding scheme used for most of the PISA science items is very similar to the one introduced in the Third International Mathematics and Science Study (TIMSS) (Lie *et al.*, 1996). The codes used capture not

only the score on the item (i.e. to what degree the answer can be classified as correct) but also information on the characteristic features of the answer beyond the right–wrong dimension.

This means that in the PISA 2000 data file most science items were entered as variables at the nominal level. This was also true for multiple-choice (MC) items, for which the data entry was a number corresponding to the alternative chosen by the student. However, in the further processing of the data each item was scored so that a total test score value could be computed for each student. This processing can be viewed as a reduction of highly detailed data into one number. Of course, this is exactly what tests in general are designed to do. However, from the perspective of science, mathematics or reading education the nominal variables contain valuable information. This article presents an exploration of the multivariate characteristics of a small set of items from the PISA 2000 cognitive instrument.

The items selected for the analyses are taken from the PISA domain of scientific literacy. Some inferences made from the results will therefore relate to research in science education, in particular. However, attention will be focused on more general issues related to testing in surveys such as PISA.

A THEORETICAL RATIONALE FOR ANALYSIS AT AN INTERMEDIATE LEVEL

The analyses of PISA 2000 data in the international report (Organisation for Economic Co-operation and Development, 2001) focused mainly on test scores. Some single items were presented, but the focus was chiefly on how the items relate to the overall trait being measured by the test. On the other hand, in the Norwegian national report (Lie *et al.*, 2001), quite a few pages were devoted to the results for single items. In this section I will comment on these two levels of analysis, that is, the test level and the item level, and, based on theoretical considerations only, I will suggest that analysis at an intermediate level can shed further light on student characteristics by broadening our understanding of what the individual items measure.

The Test Score Level

Assessments such as PISA are developed within the framework of test theory. Test theory is a collection of methods (and a rationale for these) for developing and analysing measures of psychological entities, in the literature often referred to as traits. However, the measures or scales produced and the theoretical underpinnings of these scales do not help us to interpret what students actually *think*, *know* or *can do* within the domain measured. Even though all the items measure, to a certain degree, the same trait, and thus contribute to a reliable overall score, there will be a major portion of item-specific variance. Olsen *et al.* (2001) have argued that a typical item in tests such as PISA has an item-specific variance component of 90%. This means that the scale developed accounts for only 10% of the pooled variance in students' answers.

The Item Level

Based on the above considerations, one can argue that from a cost-efficient perspective a major portion of information from a typical test is thrown away when only the overall score is analysed. From a science educator's perspective, much more information can be obtained through additional secondary analyses. Item-specific variance can, in this perspective, be viewed as relevant and highly interesting information, whereas from a test theoretical perspective it is simply regarded as 'error' variance.

However, there are also some limiting factors to be considered. Item-specific information may be divided further into several components, such as variance due to the topic addressed in the item, the item format, where the item is placed in the test booklet, and contextual or situational factors specific for the item. This suggests that when analysing single items a major problem arises from all these uncontrolled facets. If one considers a single item a test, these single items are highly unreliable, unless these facets are controlled for.

Another major problem with single-item analyses is their disconnectedness. We can very well perform carefully designed single-item analyses; however, all the analyses will be separate analyses. This leads to a relatively fragmented picture of very specific pieces of knowledge in specific contexts, and we will soon be overloaded with information.

The Intermediate Level

To recapitulate, the total test scores for students, schools or countries participating in PISA provide highly reliable information on very well-described concepts. Also, the procedure used to operationalise these concepts ensures that these are valid measures of the concept (Adams & Wu, 2002). However, no detailed insight into students' thinking and knowledge is available through these scores alone. On the other hand, detailed information is available through analysis of single items, but these analyses are not very reliable and it is difficult to condense or synthesise this information.

It is reasonable to assume that it is possible to analyse data at an intermediate level, avoiding some of the problems described above. It should, in principle, be possible to maintain some of the item-specific information of nominal character in analyses at an intermediate level, where groups of items (two or more) are studied. Such analyses could potentially yield information on students' thinking and knowledge beyond the very limited context of one specific item.

Purpose and Goal

So far it has been argued that analyses of test scores or single items give valuable but limited information about students' knowledge and thinking in science. It is, however, not entirely clear how one should proceed in order to analyse data at an intermediate level. The purpose of this article is therefore to establish whether there

are patterns in students' thinking and knowledge across items such as those used on the PISA test. Since this is an explorative work in its initial stages and since so little relevant research is available to inform the exploration, it is difficult to state exact and precise research questions. However, the exploration is informed by the goal, and this can be formulated in terms of two general questions:

1. What is the nature of the information across items coded nominally?
2. Which recommendations can be made for the future development of PISA based on the answer to the first question?

The first question directs the analyses to be performed and contains the motivation for the explorative nature of the study, while the second question addresses the ultimate goal of this exploration. It is also important to note that the study presented is first and foremost methodological in the sense that the ultimate goal is to inform the methodology used in PISA and not to give a detailed description of the data as such. The items in these analyses are therefore used mainly to exemplify the exploration and not primarily to give a substantial analysis of students' responses to these items. Also, the latter will be addressed to a certain degree, but the prime purpose is to illustrate more general issues regarding the items used in PISA as such.

METHOD

The procedure suggested by the stated questions is first to choose relevant subsets of items for analysis and then to analyse these subsets in a conjoint manner by multivariate techniques to explore the nominal data.

The Unit Semmelweis' Diary

In PISA the items are clustered in *units*. These units differ in size, but typically they consist of two to five items. All the items within one unit relate to some common stimulus material, typically a text or a graphical presentation. The diversity of these stimulus materials is large, but they all share the common feature that they strive to be authentic; that is, they are based on actual materials from sources students would be likely to encounter in their daily life or future daily life as adults. The items within each unit are locally independent in the sense that every item can be answered in isolation from the rest of the items. It is the unit that has been chosen as the intermediate level for the analyses presented below. This choice was done not only because units are 'physically' clustered, but also because an initial exploration of the data clearly suggested that they are clusters also in an empirical sense.

The unit analysed in this article consists of excerpts from the diary of Ignaz Semmelweis, a medical doctor at an Austrian hospital in the middle of the nineteenth century. The excerpt is accompanied by a diagram showing that the occurrence of puerperal fever is substantially higher in one of the wards of his hospital as compared with another. The scientific context of the unit could be labelled as the relationship between personal hygiene and the spread of disease through micro-organisms. The structure of this unit, the types of questions asked and the assumed

competencies involved in solving these questions are good exemplars of the PISA scientific literacy framework (see Kjærnsli & Lie, 2004). The stimulus material is based on authentic sources and relates to a context of personal and social relevance for the students. In two of the four questions the task is to process and evaluate the information given, while the other two questions draw heavily on scientific knowledge not presented in the stimulus. Two of the questions MC items and the remaining two OE items. Only one of the questions will be referred to in a substantial sense in this article. This question asks the student to explain why the washing of sheets at high temperature is important. The entire unit with marking rubrics has been released and is accessible from a number of printed and electronic sources, including OECD (2002).

Homogeneity and Correspondence Analysis of the Categorical Variables

Although it is to be anticipated that most readers will be unfamiliar with multivariate techniques for exploring the relationship between variables at the nominal level, the format of this article does not allow a detailed account of these techniques. A useful resource for a detailed insight into correspondence analysis (CA) is offered by Greenacre (1983); for homogeneity analysis (HA) I refer to Gifi (1990). Other, more popular, accounts are given in, for instance, Greenacre (1993), Clausen (1998) and Hjellbrekke (1999).

CA is the statistical technique most commonly used to study patterns or associations among nominal variables. HA has been shown to be mathematically equivalent to multiple CA (Greenacre, 1993). Ordinary CA transforms a contingency table into a graphical representation, and multiple CA, or as in this case, HA does the same for a multiway table. The plots are two-dimensional representations of a multidimensional space. Without going into detail, each point in this multidimensional space is an exact representation of the *profile* of each category in the variables analysed. The profile is the distribution of the weighted chi squares for a category in one variable across all the categories in the other variable. The plots are followed by diagnostic parameters provided in order to help the interpretation of the plots. Unfortunately, CA and HA are not accompanied by parameters of the same kind. A two-dimensional plot is, of course, an approximation of this multidimensional space. However, very often a two-dimensional solution is a very good representation of the full information in the contingency table.

RESULTS AND DISCUSSION

There were 35 science items altogether in PISA 2000, and due to the test design this gave 514 pairs of items responded to by the same students. Correlation analysis of the scored items yielded a mean correlation coefficient (r) of 0.24 between the items within a unit, and $r = 0.17$ between the items from different units (statistically significant with $t = 4.1$). To explore this further, a multiple regression with each item pair as a case and certain bivariate characteristics as variables was conducted. With the Pearson correlation coefficient as the dependent variable and the other variables describing certain important similarities and differences between the items

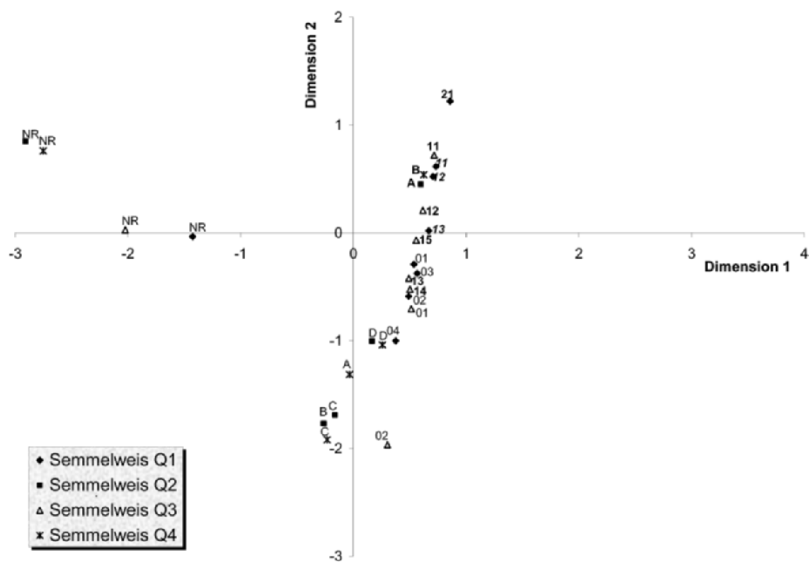


FIG. 1. Homogeneity analysis of the four items in the unit Semmelweis' Diary. Plot of category scores in a two-dimensional solution. NR refers to non-response; A, B, C, and D refer to student choices in MC items, and two-digit numbers refer to marking rubrics for OE items.

as the independent variables, the total R^2 was 0.06. The variable Unit, a dummy variable identifying items belonging to the same unit, had a positive and statistically significant beta ($\beta = 0.18$, $t = 4.2$). This was the only independent variable in the model that had explanatory power. All in all, these analyses justify the use of units as clusters in analyses at the intermediate level.

Homogeneity Analysis of the Unit Semmelweis' Diary

The results of the homogeneity analysis of the unit Semmelweis' Diary are presented in Figure 1 and Table I.

Exploring Table I first, we can note that the first dimension has an eigenvalue of 0.7, while the other dimension has an eigenvalue of 0.4. This means that Dimension 1, as is always the case, is the one with more explanatory power; that is,

TABLE I. Discrimination measures for the homogeneity analysis of the unit Semmelweis' Diary

Question	Dimension	
	1	2
1	0.52	0.32
2	0.74	0.45
3	0.72	0.39
4	0.76	0.49
Average (eigenvalue)	0.69	0.41

this dimension separates the categories better than does Dimension 2. This separation is stronger for the last three items, as shown by the discrimination measures.

Looking at the plot of the solution in Figure 1, we can identify the categories for the responses to the OE items by the fact that they are coded with double digits, while the MC items are coded with a single letter. The main feature in the data as captured by the first axis is the separation of the non-response categories from the categories used for students who gave responses. The second axis separates the categories for those students who responded to the items. The separation is between credited answers and non-credited answers. In the Figure the codes for the credited answers are in bold type, and the partial credit answers are, in addition, italicised.

All in all, this suggests that the first dimension could be labelled as the ability or willingness to respond. The underlying characteristic captured by this dimension is more or less discontinuous, meaning that this dimension separates students roughly into two groups on the basis of their willingness to respond. One of the groups responded to almost all of the items, while the other did not respond to any of the items. The second axis seems to represent the ability of students willing to respond to consistently give answers of good quality. This axis does not separate students into clear groups. Rather, the characteristic described by this axis seems to constitute a continuous ability scale.

The interpretation of the axes is further strengthened by a more thorough inspection of the distribution of the categories. Supporting our interpretation of the first axis is the fact that the MC items are separated from the OE items. The codes referring to students omitting the MC items are further to the left on the plot. What this implies is that students not responding to these items were students consistently not responding throughout the entire set of items; that is, these students have also to a large degree not responded to the OE items. This seems reasonable based on common sense alone. It was also verified by a closer inspection of the data. Of those not responding to the first MC item, approximately 90% also left one or more of the other items unanswered.

There are four important details supporting the interpretation of the second axis. Firstly, the answers to Question 1 (Q1) receiving full credit, coded as 21, are at the very top, well above any of the other categories for credited answers. Secondly, the codes referring to 'other incorrect answers', Codes 04 and 02 for Questions 1 and 3 respectively, are at the bottom of the figure, well below the other specific categories for incorrect answers to these OE items. The answers not fitting in any of the specific categories for incorrect answers included statements of a nonsensical character; that is, irrelevant drawings, crosses, or sentences such as 'I don't know'. Many of these answers could be characterised as completely off-task. It is reasonable to believe that these utterances were expressions of frustration, meaning in fact that the student did not understand the text, the question or the combination of these. Thirdly, inspecting the five credited answers to the third question (Q3), we can see that they span a large section of the axis. The one on the top coded as 11 refers to answers explicitly mentioning that washing sheets *kills bacteria*. Code 12, also high up on this axis, is used for answers referring to the *killing of germs, micro-organisms or viruses*. Code 15 is used for answers referring to the fact

that washing *sterilises* sheets. Codes 13 and 14 in the lower half are used for answers stating that after washing, the bacteria, viruses or germs *are gone*. In other words, even though these responses yielded the same credit, there are qualitative differences between these five response categories in terms of the degree to which they use appropriate scientific terminology. This was also verified empirically by the average science scores of students falling in these five categories. The progression along the second axis is perfectly aligned with the corresponding student groups' progression on the science score scale. The average science scores corresponding to the responses coded as 11, 12, 13, 14 and 15 were 551, 534, 504, 504 and 525 respectively. The fourth detail lending support to the interpretation of this axis as an ability scale is the position of the distractors in the MC items at the bottom of Figure 1. With the interpretation of this axis as the tendency to consistently give answers of overall high quality across a set of items, this would mean that students choosing one of the wrong answers to these items are those most consistently not credited in this unit. This is very much in line with the above interpretation, since both of the MC items were relatively easy items, meaning that students not credited were those furthest down on the ability scale.

All in all, this suggests that the axes can be reasonably interpreted as *the willingness to consistently respond to the items in a set* (1st axis) and *the ability or tendency to consistently give answers of high quality throughout a set of items* (2nd axis).

Comparing Countries with HA

The scores on each of the dimensions presented above were also calculated for each of the students (object scores). Figure 2 shows the average object scores for all the countries participating in PISA.

First of all, it is worth noting that the values for these average scores are much lower than those for the categories. The obvious reason for this is that the profiles of the categories described in Figure 1 were the averages for students giving the same type of answer to one of the questions in the unit. However, within each country there are students giving all types of answers to the same item. The average for a country would therefore be the mean for a more heterogeneous group. Despite this, the countries' spread along these dimensions seems to be meaningful. The markers used for each of the countries on this plot group the countries according to their overall science scores. The group of countries with science scores significantly above the OECD mean can be found in the upper right quadrant, while the countries with their scores significantly below the OECD mean are mainly placed in the bottom left corner. The countries scoring at the OECD mean are scattered around the centre. If we imagine a line with an axis placed along the diagonal from the bottom left to the upper right and projecting the points onto this line, this would be a one-dimensional solution with the countries ranked very much according to their overall science scores in PISA. Although the object scores are uncorrelated, the average object scores of the countries for these dimensions are correlated with $r = 0.34$ ($p = 0.06$).

It is interesting to consider some of the points in more detail. Certain countries do not follow the overall pattern. In the following, some pairs of countries with

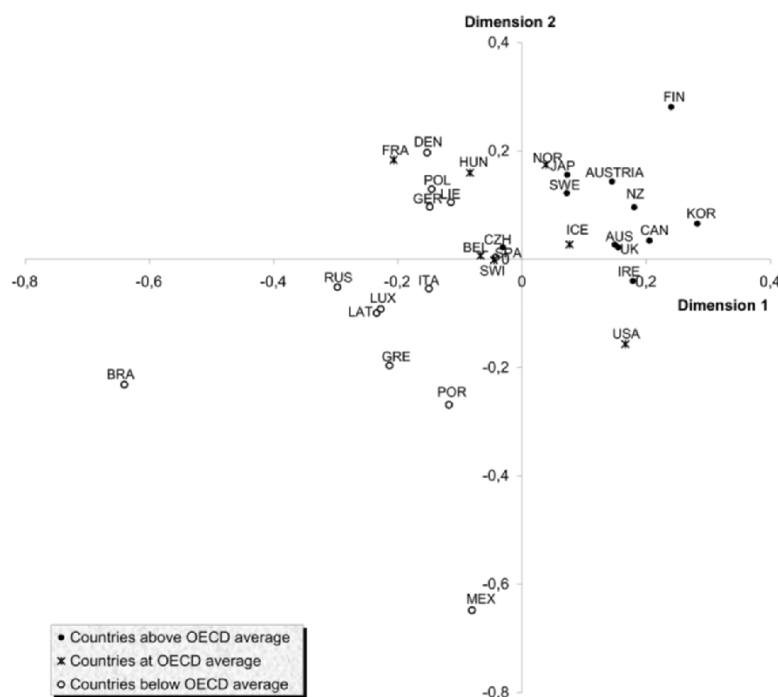


FIG. 2. Homogeneity analysis of the four items in the unit Semmelweis' Diary. Average object scores for the countries participating in PISA, divided into three groups based on whether their overall science score was above, at or below the OECD mean.

extreme values will be contrasted to illustrate the spectrum. Denmark and the USA make up one such pair. Both belong to the countries scoring little below the OECD average, and the difference between their scores is not statistically significant. The total score for this unit only was also calculated, with the result that Denmark and the USA showed the same overall ability. However, in Figure 2, Denmark and the USA have entirely different positions relative to both axes, both countries being 'outliers' in relation to the imagined regression line through the points. Denmark is in the upper left quadrant. Given the interpretation of the axis presented earlier, this would imply that Denmark had a relatively large number of non-respondents, but those responding fairly often received credit. On the plot the USA is placed diagonally opposite Denmark. According to the established interpretation of the axes, this indicates that American students had a high willingness to respond, but of those responding, relatively few received credit.

This contrast between Denmark and the USA indicates, first of all, differential willingness to respond. If this affects the countries' scores, it could be regarded as a bias in the scores. It is at least reasonable to expect that it did affect the scores on the MC items, because a higher willingness to respond will automatically, thanks to guessing alone, lead to a higher number of correct answers. Another possible reason that could be hypothesised to explain this phenomenon is that the test was speeded in Denmark but not in the USA. The PISA test is not intended to be speeded, so if this hypothesis is verified it would mean that the scores were biased due to differential speediness across countries.

TABLE II. Proportion of some response features in selected countries

	Item	Non-responses	Not reached	Credited responses	Credited responses of those responding	Unspecified non-credited responses
Denmark	Q1	39	3	24	40	19
	Q2	15	6	62	73	
	Q3	29	6	59	84	9
	Q4	19	7	63	78	
USA	Q1	19	4	25	30	26
	Q2	9	5	62	68	
	Q3	19	6	57	70	20
	Q4	10	6	54	60	
Brazil	Q1	56	17	5	11	18
	Q2	34	21	30	46	
	Q3	52	24	28	60	18
	Q4	38	25	25	40	
Mexico	Q1	24	4	9	11	30
	Q2	12	6	36	41	
	Q3	21	6	45	57	29
	Q4	15	7	29	34	
Korea	Q1	16	0	42	50	25
	Q2	4	1	80	83	
	Q3	12	1	78	88	9
	Q4	4	1	64	67	
Finland	Q1	19	2	35	43	19
	Q2	5	3	73	77	
	Q3	11	3	83	93	6
	Q4	7	3	76	82	

Note: The second last column is proportionate to those responding to the item, while the rest of the columns are percentages relative to all students with this unit in their booklet.

Another noteworthy feature of this description is the fact that both Brazil and Mexico were among the lowest scoring countries in PISA 2000. On this plot we can see that there were different reasons for these two countries' relatively low science scores. Brazil is extreme in the sense that it had relatively many non-respondents, while Mexico is extreme in the sense that it had quite a few students answering the items, but they seemed to give answers that did not receive credits.

Korea and Finland is another pair of countries worth noticing. They were two of the top-ranking countries on the science test; Korea, however, scoring significantly higher than Finland. In the subscore generated for this unit they were also the two top countries, scoring more than 0.5 standard deviations above the average. This is reflected in Finland's position in Figure 2. Korea's position on the second axis, however, is much lower than expected. Its position cannot therefore be readily explained by the interpretation of the axis given. It is an anomaly, which will be discussed later on the basis of correspondence analysis.

The data given in Table II display some key numbers for the countries discussed above. With this table the lines of interpretation presented above can be

studied in more detail. The Table provides information on the proportion of students not responding; the proportion of students who did not reach the item; the proportion of credited responses, both in total and as the proportion of those responding; and the proportion of students who gave non-credited responses of a non-specific character, as presented earlier.

Comparing first Denmark and the USA, we notice that the interpretation given above is confirmed by these data. Denmark had a fairly large proportion of students not responding, while US students were much more willing to respond. The proportion of students receiving credit was very much the same in the two countries, but isolating those responding, as is done in the second last column of Table II, we can see that those responding in Denmark received credit more often than those responding in the USA. However, as can be seen from the last column, of those responding in the USA, relatively many gave answers in the unspecific non-credited category. This was likely due to the relatively large number of irrelevant and nonsensical answers in the USA. One possible interpretation of this is that in Denmark students who did not have a good answer to offer did not respond at all, whereas in the USA they were more inclined to express this in words, thereby contributing to different proportions of responses in these two countries. Upon scoring, non-responses and incorrect responses were treated equally; that is, both were given no credit. This suggests that the hypothesis of bias due to differential willingness to respond should be rejected.

In PISA an attempt was made to estimate the speediness of the test by categorising the answers not responded to at the end of the booklets as 'not reached'. As can be seen from Table II, there were no clear differences between the USA and Denmark regarding this category, suggesting that there was no bias between these countries as to how much time students devoted to the test. This means that the relatively high number of non-respondents in Denmark was rather an effect of the students' relative unwillingness to respond; it was not due to a bias caused by the test being differently speeded in the USA and Denmark.

In contrast, the test does seem to have been speeded in Brazil. This, however, could be an artefact of the procedure of categorising items as not reached, as briefly described above. In Brazil there are in general a lot of students who, throughout the booklets, tend to leave items unanswered, irrespective of whether the item is administered first or last within the time available. It is therefore also likely that several of the items administered last are in fact reached, but not responded to. All in all, the analysis of the positions of Brazil versus Mexico in the HA presented above is supported by the data in Table II.

The previous comparison between Korea and Finland cannot be fully understood by inspecting Table II. What we can see in the Table is that both countries had relatively high proportions of credited answers. On the first two items Korea did slightly better than Finland, while the opposite was true for the last two items. We can, however, notice that Korea had relatively many students giving unspecific incorrect answers to the first question. Since this code receives a high negative value

on the second axis, this could be one reason for Korea's position in Figure 2. Another possible reason not found in the results presented so far is revealed by a more detailed inspection of the profiles of these two countries for Question 3. This was done by correspondence analysis, with the item (Q3) as one variable and the country as the other variable. The main result from this analysis was that Korean students relatively often used the more vague concepts of micro-organisms and germs (Codes 12 and 14), while Finnish students preferred to use the term killing of bacteria (Code 11). In the Korean language there is an overarching term, *SaeKyun*, referring to either a germ, a microbe, a bacterium, etc. (Y. Kwak, personal communication, April 2003). This might be hypothesised to be the reason why relatively many students in this country gave responses that were coded as 12 and 14. These codes received relatively low values on the second dimension in the HA. Hence, even though these codes denote correct answers, they tend to separate a different profile across the items in this unit as compared to the rest of the codes for credited responses.

The same HA was also performed in the Nordic countries alone, excluding all the other countries. The solution was very much the same as that for all participating countries. The dimensions were the same and the relative positions of the countries were similar. The object scores for the dimensions in the overall analysis were also highly correlated to those in the Nordic analysis ($r = 0.97$ for Dimension 1 and $r = 0.89$ for Dimension 2). All in all, this implies that a purely Nordic perspective is also available through the analysis presented in Figure 2.

Correspondence Analysis of Pairs of Items

It is evident that the homogeneity analysis above gives information similar to that which is typically present in analyses of items scored along a correct–incorrect dimension. The analysis, however, does not reveal patterns of associations between categories that could serve as evidence for some specific conclusions about the degree to which students' knowledge and thinking is conceptually consistent across items. Analysing two items might give information of this kind. Correspondence analyses of all pairs of items in this unit were therefore performed.

In general, when performing CA on pairs of items from the PISA item pool, we get the same kind of information as in the HA presented in Figure 1 above. The first and dominating dimension, accounting for 60–90% of the inertia, is the dimension separating non-responses from responses. The second dimension, accounting for 10–30% of the inertia, separates the categories of good responses from those of poor answers. The first two dimensions together accounted for more than 90% of the inertia in all analyses. Beyond this information, the analyses did not suggest any major consistencies in students' responses; for example, that one particular non-credit code on one item would be strongly associated with another non-credit code in another item. Had the latter been the case, we could inspect the categories in order to find reasonable interpretations based on the rich research literature on students' cognitive structures.

CONCLUSIONS AND RECOMMENDATIONS FOR FURTHER WORK

This article set out to discuss two related questions. The results presented address the first of these questions, dealing with the main patterns in students' responses across a set of items contained within a unit. Since this question inquired about such patterns in nominal information, the hope was to be able to give a substantial description of students' thinking and knowledge beyond what is available through analyses of items scored as right or wrong.

Summary of the HA Analysis

Analyses of the type presented above were done on a number of units and a number of items. The HA of similarly constructed units gave similar results to the one presented in this article. This was also the case for the CA on pairs of cognitive items. In other words, no clear qualitative patterns describing types of thinking or structures in knowledge were discovered in the PISA units taken as a whole. In the HA and also in the CA on pairs of items the first axis repeatedly captured information on the consistency with which students were willing to respond, and the second axis repeatedly represented the ability to consistently provide answers of good quality. This information was similar to that from analyses of the scaled versions of these items. The ability to consistently give answers of good quality is, for instance, what is measured by point biserials or other indices of discrimination. The major conclusion to be drawn is therefore that if we were to produce small subtests with the length of about 3–5 items, the items belonging to the same unit would be among the most optimal 'testlets' available. This is because these items can be clustered and aggregated based on theory alone and also because items belonging to the same unit have proven to be the dominating bivariate characteristic accounting for the variability of correlation coefficients. The unit discussed in this article had inter-item correlations in the range 0.3–0.4, and this particular testlet had a Cronbach alpha approximating 0.7. Including a few more items relating to the text and conceptually to the subject of hygiene, we would have a test of reasonable overall test quality measuring what might be labelled students' scientific literacy in hygiene.

Although the analyses did not give the kind of description initially sought, it is noteworthy that the dominant information at the nominal level is very much like the information available from the scored items. I would like to suggest some reasons for this. First, the nominal information present in the OE items could be characterised as quasi-nominal. The codes refer to both the score and the type of the answer. In other words, the codes show a progression from poor to good answers. This progression is particularly manifest in items with more than one score point, such as Item 1 in the Semmelweis unit; in addition to this, however, we saw in the HA that codes at the same score level were rank ordered so that the progression in quality among these equally credited responses was also captured. Second, the way items are selected and marking rubrics constructed contributes greatly to the above effect. All codes used are checked in trials to see whether they differentiate well between students at different levels of ability as measured by the overarching concept

measured by the test. All in all, the formulations behind the double-digit codes used in the marking guide for OE items mainly serve the function of scoring criteria, and we should not be surprised to find that the scoring aspect is also reflected in the analyses of the nominally coded data.

Even though the HA suggests that the nominal data can be reduced to one dimension only—a dimension going diagonally from the bottom left corner towards the upper right corner in Figure 2—yielding more or less the same information as the scored items, the presence of two dimensions gives a more detailed insight into country-specific features. Among the country profiles studied, we saw that two countries with medium ranks on the PISA test, the USA and Denmark, had distinctly different response profiles. The main difference was that US students were much more willing to respond. The same distinction was revealed for Brazil and Mexico, with Mexican students more willing to respond. The results presented do not suggest that such differences across countries could be explained by some bias—due either to differential willingness to respond or to differential speediness across countries. The results might be related to the fact that students living within different cultures and participating in different school systems have very different experiences with regard to testing as such. This phenomenon might also be explained by some background factors, as in the case of abilities. In PISA there are, for instance, many constructs related to motivation and self-efficacy. Matters and Burnett (2003) maintain that the most important predictor of the number of omitted items in a high-stakes test is students' academic self-concept, with gender as a mediating variable. It is therefore recommended that a separate study be conducted in order to better understand this phenomenon. In the case of PISA, the stakes are not high for students, and it would thus be easy to hypothesise that students' internal motivation would be another important predictor. It is consequently reasonable to suggest that the tendency to omit items in PISA is not directly attributable to a bias between countries, but rather that it is an indirect effect mediated through different perceptions of, for instance, the stakes, motivation and self-concept.

Establishing the dimensions as 'the willingness to respond' and 'the ability to give answers of high quality' could not explain every detail on the plot. One specific anomaly was the case of Finland versus Korea. To explain this phenomenon, a more thorough inspection of the profiles had to be done, revealing that Korean students in this specific context used more vague concepts ('germs'), while Finnish students relatively often referred to the more scientific concept of 'bacteria'.

Consequences for Item Development

In PISA, as in most test development processes, the marking guides are made for each item in isolation and in relation to the items' test properties. This means that beyond the right–wrong aspect, the codes used in marking each item reflect something very specific for that particular item only. It is not easy to see how one code used in one item captures similar information on students' thinking to another code in another item. In future item development, the inclusion of clusters of items designed to map students' cognitive representation of some clearly defined scientific

concept or process could be considered. These could be within one unit relating to a common stimulus, as is done in PISA, or the items could be related to different stimuli. We could, for instance, think of two units related to the concept of energy, both including a very similar item which, however, would be tightly related to its actual stimulus so that students would not be able to spot very easily that the items are conceptually equivalent. In this way we could see, for example, how consistently students talk about energy being used or energy being transformed into new forms and so forth.

Of course, we do not need large-scale assessments such as PISA in order to do research with the aim of diagnosing students' knowledge and thinking. A lot of smaller-scale research within science education addresses this, predominantly through qualitative methods. Nonetheless, including this aspect in PISA would allow us to generalise our findings to a clearly defined population, something that is difficult to achieve in independent low-cost research. What is even more important, however, is that in this way it would be possible to map how these cognitive patterns vary across countries, languages or school systems and cultures.

When selecting the final pool of items for a test, there are some constraints that could be used to argue against the inclusion of diagnostic clusters of items. Test time is limited, but the design in PISA, where test booklets are rotated, has made this limitation less severe. There is a trade-off between what Keeves and Masters (1999) have called the *fidelity* and the *bandwidth* of a test. On the one hand, the defined concept or trait has to be measured with high precision and the concept delimited from the universe of concepts (fidelity); on the other hand, the whole part of the universe has to be spanned by the concept (bandwidth). Arguing for diagnostic assessment is to argue for high-fidelity measurement of a minor subconcept and thereby possibly reducing the ability to reach the desired bandwidth in the measurement. Had we, however, let us say, two subconcepts measured by five items, neither of these would influence the bandwidth in any major way if the items used had satisfactory test properties. That is, they must be scorable and they should discriminate positively. Good diagnostic items should have many students falling into the hypothesised diagnoses. Very often we are mainly interested in misconceptions or alternative conceptions, in other words, conceptions that in most cases are formally wrong and should not receive credit. This implies that a substantial number of students should fall into these corresponding non-credited categories. This, again, could lead to items with relatively low p values, which is not optimal for a test. Also, some of the well-known misconceptions are conceptions shared by students at all ability levels. The discrimination indices for many good diagnostic items could therefore be expected to be relatively low. Low p values and low discrimination could at least be hypothesised to turn up for some items with very good diagnostic qualities. The combination of these two properties, low p values and low and possibly negative discrimination, is in conflict with the traditional requirements set for items in a test designed to measure one or a few overall traits. This potential conflict between items' test and diagnostic properties has to be resolved in a compromise between the diagnostic versus the measurement perspective. The compromise could be that low p values are acceptable since the whole ability range

has to be measured in the test, and that the discrimination must be at least positive, but not necessarily very high. A natural extension of the work presented in this article is therefore to make an attempt to produce clusters of appropriate test quality combined with the potential to be used in diagnostic analysis.

Accordingly, a diagnostic perspective—instead of being something that is largely regarded, as is the case today, as secondary analysis beyond what the studies are primarily aimed at—should be explicitly included in the framework, and it should become, from the very outset, an integral part of the research design. The diagnostic perspective has to be incorporated into the selection and description of the domain to be tested, item construction (the selection of relevant stimuli, the writing of questions and distractors, and the development of marking guides) and item selection, if it is to become an integral perspective of the test: (a) the framework should, based on previous research, make explicit the assumptions about how students' knowledge within a specific domain to be diagnostically tested can be typically structured and developed. These assumptions will provide valuable guidelines for the next design steps; (b) item writing will be the more targeted and easier, the more explicit the assumptions. However, it is never possible to foresee how an item will work in the end. Items should be written and field-trialled in different formats (Olsen *et al.*, 2001) and with systematically varied problem characteristics (Nichols, 1994) to see how this affects students' responses.

Although diagnostic testing has, for many decades, been one of the core elements within both mathematics and science education, there is little evidence that this research and knowledge base has had any major impact on practice (Jenkins, 2000). On the other hand, large-scale international comparative studies do have a great impact on curriculum design and evaluation in many of the participating countries, and including a diagnostic perspective in these studies may be of vital importance if we want to promote the use of diagnostic assessment in the classroom.

REFERENCES

- ADAMS, R. & WU, M. (Eds.). (2002). *PISA 2000 technical report*. Paris: OECD.
- CLAUSEN, S.E. (1998). *Applied correspondence analysis: An introduction*. Thousand Oaks, CA: Sage.
- GIFI, A. (1990). *Nonlinear multivariate data analysis*. New York: Wiley & Sons.
- GREENACRE, M.J. (1983). *Theory and applications of correspondence analysis*. London: Academic Press.
- GREENACRE, M.J. (1993). *Correspondence analysis in practice*. London: Academic Press.
- HJELLBREKKE, J. (1999). *Innføring i korrespondanseanalyse* [Introduction to correspondence analysis; in Norwegian]. Bergen, Norway: Fagbokforlaget.
- JENKINS, E.W. (2000). Research in science education: Time for a health check? *Studies in Science Education*, 35, 1–25.
- KEEVES, J.P. & MASTERS, G.N. (1999). Introduction. In J.P. KEEVES & G.N. MASTERS (Eds.), *Advances in measurement in educational research and assessment* (pp. 1–19). Oxford, UK: Pergamon.
- KJÆRNSLI, M. & LIE, S. (2004). PISA and scientific literacy: similarities and differences between the Nordic countries. *Scandinavian Journal of Educational Research*, 48, 371–386.
- 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* [Well prepared for the future? Norwegian 15-year-olds' competence in reading, science and mathematics in an international perspective?; in Norwegian]. Oslo, Norway: Universitetet i Oslo, Institutt for lærerutdanning og skoleutvikling.
- LIE, S., TAYLOR, A., & HARMON, M. (1996). Scoring techniques and criteria. In M.O. MARTIN & D.L. KELLY (Eds.), *TIMSS technical report (Vol. 1)* (pp. 7-1–7-16). Chestnut Hill, MA: Boston College.

- MATTERS, G. & BURNETT, P.C. (2003). Psychological predictors of the propensity to omit short-response items on a high-stakes achievement test, *Educational and Psychological Measurement*, 63, 239–256.
- NICHOLS, P.D. (1994). A framework for developing cognitively diagnostic assessments. *Review of Educational Research*, 64, 575–603.
- OLSEN, R.V., TURMO, A., & LIE, S. (2001). Learning about students' knowledge and thinking in science through large-scale quantitative studies. *European Journal of Psychology of Education*, 16, 403–420.
- ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT. (2001). *Knowledge and skills for life. First results from PISA 2000*. Paris: Author.
- ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT. (2002). *Sample tasks from the PISA 2000 assessment: Reading, mathematical and scientific literacy*. Paris: Author.

Paper III

Olsen (2005): Item-by-country interactions in PISA 2003: Country-specific profiles of science achievement. *Not published*.

A condensed version with a more narrow focus has been published as Olsen (2005): An exploration of cluster structure in scientific literacy in PISA: Evidence for a Nordic dimension?, *NorDiNa*, 1(1), pp. 81-94.

Item-by-country interactions in PISA 2003: Country-specific profiles of science achievement¹

By Rolf V. Olsen, Department of Teacher Education and School Development, University of Oslo

ABSTRACT

The cognitive items covering the domain of scientific literacy in the Programme for International Student Assessment (PISA) are explored through a cluster analysis of the item p-value residuals. Such residuals are often referred to as item-by-country interactions. The analysis clearly indicates distinct clusters of countries with similar profiles. The most stable country clusters have been labelled 'English-speaking countries', 'East Asian countries', 'German-speaking countries' and 'South American countries'. A more detailed inspection is done of the profiles for the Nordic countries, and they are shown to be members of a larger group of countries which is labelled North-West European countries. Some detailed features of the profiles are described using item characteristics such as the categories used in the operational definition of scientific literacy given in the framework. In projects like TIMSS and PISA efforts are made to minimise such interactions. 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.

Introduction

In this article country-specific strengths and weaknesses across cognitive items in scientific literacy² from the study Programme for International Student Assessment implemented in 2003 (PISA 2003) are explored. The study is organised through the Organisation for Economic Co-operation and Development (OECD). 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

¹ A short version of this paper has been published in NorDiNa (R. V. Olsen, 2005)

² Throughout the article the more convenient and less accurate term 'science' is also used when referring to scientific literacy

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 Finland in particular that did not cluster together with the other Nordic countries, followed by Denmark, which also had a profile that to some degree was drawn away from the Nordic cluster.

In the above-mentioned analyses of data from PISA 2000 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 very distinct pairs of countries (France & Belgium French and the Netherlands & Belgium Flemish) were present. Lastly, in TIMSS 1995 there was a cluster of less developed 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 three interrelated questions:

- I. What groups or clusters of countries are indicated by the cognitive items in the science domain of PISA 2003?
- II. To what degree does the cognitive data in the science domain of PISA 2003 suggest that there is a common Nordic profile?
- III. 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 cannot reach any solid conclusions regarding these questions. However, the analysis will point forwards 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 interactions. 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 minimise such effects by avoiding items with high item-by-country interactions in the final test instruments (Adams & Wu, 2002, pp. 25-26 and 102-105). 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 has a cyclic design and is repeated every three years, and three different domains are given different weights in the test material each time: reading literacy, mathematics literacy and scientific literacy. The study has so far been implemented twice, in 2000 and in 2003. Scientific literacy was a minor domain in both studies. When the study is conducted next, in 2006, science will be the major domain, occupying about two-thirds of the testing time.

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. in the Nordic countries' reports Björnsson *et al.*, 2004; Kjærnsli *et al.*, 2004; Kupari *et al.*, 2004; Mejding, 2004; Skolverket, 2004).

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

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 (p. 133).

This definition is further developed and operationalised throughout the document. It ends with descriptors of three main dimensions that the items should cover:

- A. The *content* dimension identifies several areas within science that are seen as particularly relevant given the overall definition.
- B. The *competency* dimension 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 (Kjærnsli, 2003). 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'.

Categorising and describing the items

All items are categorised within the three framework categories A, B and C listed above. When analysing 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 A is that this dimension has not been equally important in the item development. There are two possible reasons for this. First of all this dimension is described through examples only, so even though the number of examples is quite high, it is nevertheless not a complete description. No content *per se* is excluded by this dimension. Secondly, dimension C gives a description of ‘areas of applications’ that are suitable for PISA science items. Such areas of application roughly correspond to broadly defined thematic content, and as such also give an outline of what content is considered appropriate. This dimension has been more important for item developers.

Since science was a minor domain in 2003, only 34 items were available for analysis. Consequently, it is important to use descriptors of a general character not splitting the items into more than two groups. In the analysis presented below five item descriptors will be used:

- I. *Competency*: Item analyses from PISA clearly demonstrate that countries perform differently on items testing mainly factual knowledge or understanding of concepts (competency I) and items testing the mastery of some fundamental scientific processes (competencies II and III) (Kjærnsli et al., 2004; Lie *et al.*, 2001). The variable ‘Competency’ in Table 1 is coded 1 for items in Competency I, and 2 for items in competency II or III.
- II. *Context*: 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 different countries. The framework operates with three *situations* or *contexts* describing the areas of application. Using this as a key would result in too few items in each category to see any stable profiles. The situations from the framework, after an initial screening of the data, have 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’.
- III. *Format*: Previous research is ambiguous regarding the differential effects of item response format across countries. This will be explored through the dichotomy given by constructed response items (coded 1) vs. selected response items (coded 2).
- IV. *Textdist*: 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 can 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).
- V. *p-value*: In addition the difficulty of an item, in terms of the percentage of correct responses averaged over all countries, 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 differs

from the other characteristics mentioned above in that it is a continuous variable.

Descriptor	Code	Label	Number of items
Competency	1	Conceptual understanding	16
	2	Process skills	18
Context	1	Life and health	12
	2	Physical world	22
Format	1	Constructed response	14
	2	Selected response	20
Textdist	1	Stimuli independent	21
	2	Stimuli dependent	13

Table 1: Distribution of item descriptors across the science items in PISA 2003

Table 1 summarises the distribution of the 34 science items across the item descriptors I-IV. The *p*-value is not categorical and hence the distribution of this variable is not presented in Table 1. The *p*-value mean is 0.48 with a standard deviation of 0.16. The vast majority of the items are therefore of medium difficulty in the range 0.3-0.7. It is furthermore important to note that Table 1 summarises the distribution of the number of items, and not the number of score points. The distribution across the two formats is, for instance, more evenly distributed across score points than across the number of items as shown in Table 1. The table therefore gives the wrong impression that the PISA science test score is mainly based on multiple choice and other forms of selected response items. However, in the analyses presented in this article the item is the unit of analyses.

Although these item descriptors can be seen as mapping substantially different kinds of characteristics of items, they are not empirically unrelated. *Textdist* is positively correlated with *Competency* ($r \approx 0.5$), meaning that the successful solution of items testing students' understanding of scientific processes to a larger degree requires that the students make use of the textual stimuli given. Also, *Textdist* is negatively correlated with *Context* ($r \approx -0.4$), implying that the items targeting issues related to life and health are more dependent upon the text. Furthermore, the response format (*Format*) is positively correlated with the overall international difficulty (*p-value*) ($r \approx 0.4$), meaning that items with a selected response are easier than items with a constructed response. When using these item descriptors as explanatory variables interpretations should be made paying attention also to the dependencies between them.

Method

The residual matrix

The data input for the analyses presented in this article is a matrix of *p*-values, the percentages of students credited with a score point, for each science item in the PISA 2003 cognitive test for each of the participating countries. For most items the scoring is done by a single score point. Some items, however, have two score points separating answers deserving full credit (2 points) from responses

given partial credit (1 point). For these items, the p-value has been calculated by weighting the partial credit score point by a factor of 0.5. The number of items and countries is 34 and 41 respectively.

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 more difficult 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 main information contained in the p-values is the overall level of achievement for the countries, and the overall level of difficulty of the item.

The p-value matrix is therefore transformed to cancel out these general effects. This is done by first calculating the grand mean (\bar{p}) which is the average p-value over all items and all countries. The average performance for a country across all items (p_c) can be expressed as a deviance from this grand mean ($\Delta p_c = p_c - \bar{p}$), shown in the column labelled as country residual in Table 2. In the same way the average difficulty for an item across countries (p_i) can be expressed as a deviance from the same grand mean ($\Delta p_i = p_i - \bar{p}$), shown in the bottom row in Table 2. The item-by-country interaction or the p-value residual (p_{res}) is then computed as:

$$p_{res} = p_{ci} - \bar{p} - (\Delta p_c + \Delta p_i),$$

where p_{ci} is the actual p-value for country c on item i .

These values are shown in Table 2. In general then, the original p-value for a country can be reproduced by adding the grand mean, the country residual and the item residual to the value in each of the cells in the table. Furthermore, Table 2 shows the standard error of international measurement (Wolfe, 1999) which will be returned to shortly.

In other words, the residuals represent the achievement for a country on a specific item, beyond what can be expected from the item and country averages³ alone. If the p-value for a particular country on a specific item is as expected from the overall difficulty of the item and the overall performance of the country, the residual is 0. On items with positive/negative residuals the interpretation is that for this item the country is performing better/worse than expected. This transformed matrix can therefore be considered as giving the profiles for the

³ It should be noted that the achievement scores reported in PISA reports (eg. OECD-PISA, 2001, 2004) are not based on p-value metrics. Instead, psychometrically advanced models have been used. In this metric the difficulties for each item is computed by item response theory, or more specifically, by a so-called Rasch model. In this model the likelihood of receiving a score point is modelled as a logistic function of students' ability. The difficulty of an item is commonly represented by the ability level of a student who has a 50% likelihood of receiving this score point. To check whether this metric is comparable with the p-value metric used in the analyses here, the average p-value for each country has been correlated with the scores for the scale used in PISA and the scales are indeed highly correlated, $r = 0.97$, suggesting that the p-value metric used in this article does not introduce any major errors. Some of the reasons why the correspondence between the two metrics is not perfect could be suggested. In the Rasch model each score point is treated as an item while in the p-value metric items with several score points were transformed as described above; in the Rasch model the item parameters are expressed on a log-linear scale; items are rotated in booklets and all items do not appear equally often in the test material and this is not taken into account in the p-value metric; in the Rasch model used to scale the PISA data only the OECD countries were included, while all countries were included in the p-value metric.

country-specific patterns across items. A relatively low-performing country can in theory have a profile very similar to a country with higher overall performance since this effect is cancelled out by this transformation. However, this is not completely correct. The problem with the p-value metric is the upper and lower limits of 1 and 0, respectively. This creates what is often referred to as floor and ceiling effects. For very easy items, it can for instance be expected that all countries will have fairly high p-values. Thus, it is very likely that high-performing countries will end up with negative residuals, and similarly, for these items it is more likely that low-performing countries will end up with positive residuals. As a consequence the overall performance will influence the residuals in a systematic way. This problem can be avoided by using a transformation of the p-values, for instance the logistic transformation. Using the logistic function the p-value metric bounded by a lower and upper limit will be transformed into a metric with no upper or lower bounds. Nevertheless, in the results presented below, the p-value residuals have been used since this metric is more intuitive, and in general the vast majority of the items (29 of 34) have p-values in the range 0.3-0.7. There is, therefore, little reason to believe that these effects will have a major influence on the solution. Nevertheless, all analyses have in addition been done on the logistically transformed data as part of the procedure for checking the stability of the proposed solution, as will be returned to shortly.

From a measurement perspective these residuals should be as low as possible since the test intends to measure a trait that is independent of the actual items used. If this is not the case it is reasonable to question whether the produced test score is reliable, and in the end, whether the test is a valid measure of this trait. The error introduced by the item-by-country interactions to the measurement can be represented as the standard error of international measurement (Wolfe, 1999), SEI. For a particular country this is found by

$SEI_c = \frac{\sigma_{r,c}}{\sqrt{N}}$, where $\sigma_{r,c}$ is the standard deviation of the residuals for country c and N is the number of items.

Country	Standard error of international measurement	Country residual
Hong Kong	11	8
Macao	3	10
Japan	8	6
Korea	1	9
Ireland	5	17
UK	10	10
Australia	7	1.1
New Zealand	7	6
Canada	11	8
USA	5	10
Switzerland	-1	1.2
Liechtenstein	-2	5
Germany	-5	7
Austria	-5	1.5
Luxembourg	-6	0
Iceland	3	1.2
Finland	3	0.8
Denmark	7	1.1
Norway	6	1.3
Belgium	4	1.2
France	10	1
Netherlands	-5	0.8
Sweden	6	6
Mexico	-19	1.1
Brazil	-7	7
Uruguay	1	1.7
Portugal	8	1.4
Tunisia	4	1.1
Italy	-5	1.2
Spain	5	1.5
Hungary	-2	1.0
Poland	-12	3
Turkey	-4	1.3
Indonesia	-6	0
Thailand	-7	1.5
Latvia	-3	2.3
Russia	-4	1.9
Czech Rep.	-8	0
Slovak Rep.	-3	1.1
Serbia	-15	1.5
Greece	-6	1.3
Item residual	3	1.9
S327Q01T	11	8
S326Q04T	3	10
S326Q03	8	6
S326Q02	1	9
S326Q01	5	17
S304Q03b	1	10
S304Q03a	5	1.1
S304Q02	10	6
S304Q01	11	9
S269Q04T	29	10
S269Q03T	9	13
S269Q01	7	30
S268Q06	1	12
S268Q02T	-9	29
S268Q01	-5	10
S256Q01	-3	15
S252Q03T	-3	12
S252Q02	0	4
S252Q01	5	6
S213Q02	1	7
S213Q01T	15	18
S133Q04T	-21	13
S133Q03	-18	9
S133Q01	-13	-2
S131Q04T	6	-4
S131Q02T	-5	5
S129Q02T	-9	-4
S129Q01	-17	11
S128Q03T	-4	-18
S128Q02	0	-13
S128Q01	0	-4
S114Q05T	-5	2
S114Q04T	3	-15
S114Q03T	1	5

Table 2: Item-by-country interactions expressed as *p*-value residuals. Countries sorted as in the dendrogram in Figure 1. Countries in the established clusters (see later in the paper) are shaded.

As a consequence large-scale international comparative assessment studies have put a lot of resources into item development to minimise this error component. In the case of PISA the items are developed by people in different countries. These items are in turn judged by experts in each country in order to identify items which can be suspected of being biased. However, this alone is no guarantee for success, so a large-scale field trial is administrated one year prior to the main study, giving empirical evidence about how the items work across countries. Items with large item-by-country interactions are not included in the main study. And, as a final check, item-by-country interactions are estimated after the main study. In PISA 2003 three interactions were judged to be too high, and consequently the science scores were produced by omitting one particular item for each of these three countries⁴. In the residual matrix (Table 2), the cells representing these interactions have been replaced by the expected value 0 for the particular item for each of the three countries. Since this is only 3 out of a total of 34×41 entries in the analysed matrix, it is reasonable to expect that these replacements will not influence the analysis

The Nordic river

The ‘Nordic river’ is a label for a type of diagram that was developed originally for the Norwegian TIMSS 1995 report by Algirdas Zabulionis (Lie *et al.*, 1997) and it was also used in the Norwegian PISA 2003 report (Kjærnsli *et al.*, 2004). The diagram uses the percentage correct metric (p-values) for items. With this diagram the aim has been to *visualise* the Norwegian profile across items as compared to the Nordic cluster of countries and as compared to the overall international profile represented by the mean and the international maximum and minimum p-values. In the results presented below this type of diagram will be used to give an initial description of an *a priori* given Nordic cluster. The diagram presented in Figure 1 is based on the procedures established in the reports referred to above, but it has been slightly modified and instead of using the matrix of p-values, the matrix of residuals is used. All in all, the simple graphical tool used to construct Figure 1 can be characterised as being in accordance with the ultimate aim of the use of graphics in multivariate data analysis which is to represent *all* the data so that the main characteristics of the information is visualised more clearly (Bertin, 1981; Tukey, 1977).

Cluster analysis⁵

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

⁴ This is possible because item response theory (IRT) is used to develop the scales. One of the benefits of applying IRT in scaling is that the parameters developed for students and individual items are independent (Keeves & Masters, 1999).

⁵ This section is in general heavily influenced by two primary sources on cluster analysis. The most comprehensive and recent source is the book by Everitt, Landau & Leese (2001), which is a thorough update and revision of Everitt (1993). The manual for the SPSS statistical software package (which is used in the analysis presented here) is a very good starting point (Norusis, 1988; SPSS, 2003).

referred to as *internal cohesion* and *external isolation*, respectively. In many ways this general aim of grouping objects with similar characteristics is common for many methods of multivariate analysis (e.g. such as factor analysis or homogeneity analysis). What all variations of cluster analyses have in common is that they have as the main input some matrix of c cases across i variables. The aim is to find a cluster structure in this data matrix. This is done by first defining a *measure of proximity*, either indicated by a measure of *distance* or a measure of *similarity* between all the objects. This produces a matrix with $\frac{n(n-1)}{2}$ proximity

measures, where n is the number of objects. The most common distance measures is the (ordinary or squared) *Euclidian distance* (calculated from the sum of squared differences between two objects). Another measure used by, for instance, Lie & Roe (2003) is the *Manhattan* or *city block distance* (the sum of the absolute differences between two objects). Alternatively, a similarity measure such as ordinary *Pearson product moment correlation coefficient* (r) can be used, as is done in the analysis presented here. The reason for choosing this proximity measure is primarily based on three pragmatic reasons: it is the most familiar of the proximity measures; it is consistent with the use of correlation coefficients throughout the article; and it is the proximity measure that with the data at hand, gave the most distinct cluster structure. In addition, the choice of correlation coefficient as the measure of proximity makes cluster analysis similar to a factor analysis, but unlike factor analysis, this cluster analysis distinguishes between negative and positive loadings (Norusis, 1988; Zabulionis, 2001).

The results presented in this article are based on *agglomerative hierarchical clustering*. The overall aim of this type of cluster analysis is to show how the objects can be merged in successive steps. In other words starting with n objects, they are merged in $n - 1$ steps. As a result the n clusters in the beginning of the process (the objects) ends up in one overall cluster containing all the objects. The difficult part is then to evaluate at what stage in this process there is a solution with groupings of the objects that seems to capture a meaningful clustering structure of the data.

The starting point of the procedure is to examine all the proximity measures. The first cluster to appear is the pair of objects closest to each other, or using similarity measures, the pair of objects most similar to each other. At each stage following this, objects will be merged together based on the proximity measure used. The proximity measures (distances or similarities) between two objects have been treated above. However, in hierarchical analyses, after the first step, we cannot continue by simply using the original proximity measures between the single objects. In the first step a group, a pair of objects, has been formed, and this group should now be included in the analyses as a new composite object. Thus, the proximity between this pair and the rest of the objects must be represented somehow. Furthermore, as the process continues larger groups are formed and the proximities between such groups also have to be represented somehow. In so-called *single linkage* the proximity between two groups is represented by the minimum distance between pairs of objects, one in one group and one in the other. This method is therefore also referred to as

nearest neighbour. *Complete linkage* is similar, but in this method the maximum distance between pairs of objects is used. Accordingly, this clustering method is often referred to as *furthest neighbour*. Alternatively, one can use a parameter representing the average distances between all pairs consisting of an object in each of the groups. This is done in *average linkage* which is used in the analysis presented here⁶. This choice is also based on pragmatic reasons. Single or complete linkage represents the distance to a group by one single measure, while in average linkage all pairs of distances between objects in two different groups are used to evaluate the cluster structure. 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 (see below).

The result of a cluster analysis is commonly presented by a dendrogram (as seen in Figure 3). Dendrograms are line diagrams representing the hierarchical structure in the data, and they should be read from the left to the right. They illustrate when and how, in the stepwise procedure from n single objects to one single metacluster, objects merge to form the clusters. On the left all objects are separated and then proceed with lines showing the clustering. The points where lines meet, that is, where objects or groups of objects are merged, are referred to as *nodes*. In SPSS, which is used in the analyses presented below, the objects are sorted from top to bottom so that the objects merged together follow underneath each other in a sequence that allows the diagram to be drawn without lines crossing each other. This enhances the readability of the diagram. Also, the dendrograms are shown with a standardized metric for the distances in a range from 0 to 25. In this metric the ratios of the distances are preserved, whether they originally were Euclidian distances or a measure of similarity such as correlations (Norusis, 1988, p B-78). Thus, this metric facilitates comparisons between solutions using different proximity measures.

Agglomerative hierarchical cluster analysis is deceptively easy to do since it is integrated in most statistical software packages. There are two choices which have to be made, choices not done by the software. First, one has to choose which proximity measure and clustering method to use. In general, making a different choice might produce a different solution. Everitt *et al.* (2001) conclude their review of empirical studies of the appropriateness of different proximity measures and clustering methods by stating that: “*What is most clear is that no one method can be recommended above all others...*” (p. 67). It is therefore evident that performing and reporting cluster analyses needs to be followed by a way of documenting to what degree the solution represents the data in a robust manner, and therefore whether the solution and its interpretation are valid given the questions studied. Additionally, when interpreting and presenting the

⁶ There are several other clustering methods using proximity measures between two groups that in various ways represent the centre of the group: centroid linkage, median linkage or Ward’s method. Common to all these three methods is that they require the use of distance measures. Since the proximity measure used here is a similarity measure (correlation), these methods are not appropriate to this study.

analyses the decision on how many clusters to report has to be made. Reading the diagram from the left to the right, when should you stop? If there is an interesting clustering pattern in the data this will obviously lie somewhere in between the two extremes. One can do this by drawing a vertical line that intersects the diagram so that the nodes closest to the left of the line represent the clusters perceived to be the solution to report. Explicit procedures have been suggested for deciding where to put such a line, but Everitt *et al.* (2001) conclude that there is no consensus about which rule to apply and they cite Baxter (1994):

...informal and subjective criteria, based on subject expertise, are likely to remain the most common approach. In published studies practice could be improved by making such criteria more explicit than is sometimes the case. (Everitt et al. 2001, p. 77).

In the following, the advice in the last part of the quote has been followed in developing procedures used to identify a stable and robust solution.

Stability and validity

The analysis presented in this article has a small number of cases (41 countries), and also a limited number of variables (34 items) since scientific literacy was a minor domain in PISA 2003. Fortunately, the data used, in the form of the residual matrix, consist of aggregated data for large groups. This ensures that the data to a large degree represent a very stable input for the analysis. If, instead, the analysis had been performed on some other dataset where the cases had been responses from individuals, it could be expected that a large amount of random data would be present. In this sense the data used in this analysis can be assumed to be fairly stable.

Nevertheless, the proximity measures used are correlations between the p-value residuals for countries and there are two sources for concerns regarding the stability of these measures. Firstly, the p-values of the original p-value matrix were reduced to four components in the process of computing the p-value residual matrix presented in Table 2. These components were in order of decreasing stability: (a) the overall international grand mean (mean value for the whole p-value matrix); (b) the country residuals (the mean value for the rows in the p-value matrix); (c) the item residuals (mean value for the columns in the p-value matrix); and (d) the item-by-country interaction or the p-value residual. If there is noise or there are random errors in the p-values, this will be found in the residuals, and seen from a measurement perspective the single items are by themselves regarded as imperfect measures. They can, however, when taken together as in test scores, work as reliable measurements that may lead to valid inferences. From this perspective the residuals are nothing but noise. One of the aims of this paper is to establish the opposite: that the p-value residuals have properties that are not characteristic of noise. These residuals are, for instance, correlated with each other in a systematic manner, and they are correlated with external variables such as some broad descriptors of the items. According to Bertin (1981) the main characteristic of *information* in a data matrix is the presence of empirical relationship between the variables or between the cases. Using this rather vague notion of information, the fact that there are relationships

in the p-value residual matrix, and furthermore that these relationships can be described and understood in qualitative terms, justifies the assumption that the p-value residuals can be perceived as *information*. However, it has to be accepted that this is the most random component in the data, and as such the residuals are not as stable as, for instance, the p-values themselves.

The second source of concern is that the correlations are measures comparing the pairs of countries over 34 items. In general, there is no clear advice in the literature regarding the ratio of variables to cases in a hierarchical cluster analysis. It is, however, quite obvious that these correlation coefficients will be more and more stable as the number of items increases. In multiple regression analyses and factor analysis, two other statistical methods used to study the relationship between multiple variables, it is generally recommended that the number of cases should be about ten times higher than the number of variables (Tabachnick & Fidell, 2001). It should therefore be reiterated that the analysis presented here is to be regarded as a feasibility study of what is possible with the data collected in PISA 2006 when science becomes the major domain with at least three times as many items.

It is important to note that the analysis presented is not meant to be used for generalisations beyond the cases present, and thus the paradigm of inferential statistics does not apply. In conclusion it has to be demonstrated that the solution and its interpretation are valid for the actual cases included, but that whether or not any specific number reported is statistically significant is in itself largely irrelevant. This is to demonstrate that the solution is *robust*, and in the end that the interpretations of it are *valid*.

There is no clear advice in the literature regarding 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 sometimes 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 discussions above about whether the residuals are to be perceived as noise or as information were related to the stability of the data. In the end what is important is the stability of the solution and the two first notions of stability – the stability of the data and of the statistical techniques used – are subordinate to the ultimate question about the stability of the solution presented. 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 labelled 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. (Gifi, 1990, p. 38)

A solution should not simply be an artefact 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, p 37). A robust solution is therefore obtained if applying different methods gives similar solutions or if removing parts of the dataset does not alter the solution substantially. The fact remains that in general there are no clear guidelines to inform us about which measure or methods to use in a cluster analysis, and there is no clear advice on how to interpret and validate the obtained solution. Nevertheless,

Simply applying a particular method of cluster analysis to a data set and accepting the solution at face value is in general not adequate. (Everitt *et al.*, 2001, p. 196)

To have stable data, and to apply methods that are robust, are of detrimental importance in order to make inferences that are likely to have a satisfactory degree of validity. Four guiding rules and procedures have been set up and followed in order to evaluate whether or not the solution is likely to represent a real clustering in the data. This procedure is mainly based on what is possible to do with the software used, and the data at hand:

- I. *Internal stability or stability under the selection of data*: The interpretation is based on the dendrograms. Before a group of cases (in this case countries) is to be regarded as a cluster four criteria would have to apply:
 - i. External isolation: The node representing the merging point for the cases interpreted as a cluster must be relatively distant from the next node in the hierarchy. If this distance is relatively large it means that the cluster is separated from the rest of the countries, and therefore removing one or more of the other countries outside the group will have a very small or no effect on this cluster.
 - ii. Internal cohesion: The residuals for the countries forming a cluster should be positively correlated. Average correlations between the countries within a cluster will be reported as indicators of internal cohesion. Furthermore, to judge how meaningful it is to aggregate the residuals for these clusters coefficient alpha is computed. Included in this analysis are parameters showing what happens to the coefficient alpha if one of the countries within the clusters are deleted.
 - iii. A cluster should consist of more than two countries to make up a meaningful aggregate. Removing one of the countries in such a cluster would of course result in the total disappearance of this cluster.
 - iv. The clusters should be of approximately the same size. This would help when comparing parameters for the clusters, such as averages or coefficient alphas.
- II. *External stability*: The clusters obtained in this analysis are compared to similar studies undertaken on similar kinds of datasets.
- III. *Stability under selection of technique*: The stability of the method is studied by carrying out the analysis with other proximity measures and other clustering methods. Everitt *et al.* (2001, p. 177) suggest that widely different solutions might be taken as evidence against any clear-cut cluster

structure. Furthermore, in order to study whether the floor and ceiling effects occurring with the p-value metric have had a major influence on the solution, the logistically transformed residual matrix is analysed.

- IV. *Face value*: The clusters have to make sense in some way. This includes being able to conceptualise the clusters and provide a descriptive label reflecting a unifying property of the countries included in the cluster. This criterion is of course highly subjective and the ability to name composite entities is in general dependent on the analyst's perspectives and knowledge.

Within the established clusters it can be expected that some of the cases included are more stable representatives of the group than others. It might be that one or more cases coming into the clusters at a very late step really is/are more similar to another case or another group of cases 'hidden' within or across other clusters. This might happen because the clustering method used – average linkage – is based on average proximity measures. The matrix of correlations between countries' residuals is used as the proximity matrix in the cluster analysis, and a closer inspection of these coefficients can give further insights into the internal cohesion of the clusters, and whether there are structures in the data hidden by the analysis.

Results

The Nordic river

The actual residual matrix for all countries across all items is not included here. Instead a summary of these residuals is presented in Figure 1. This type of diagram is referred to as 'the Nordic river' due to the shape of the shaded area visualising the range of variation in the residual values for the Nordic countries.

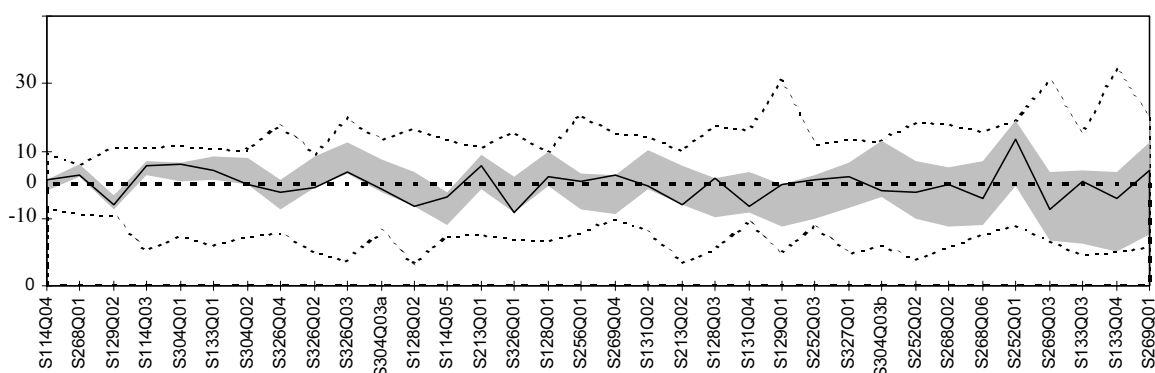


Figure 1: *The Nordic river. The dotted lines at the top and bottom represent the maximum positive and negative residuals across all participating countries. The dotted line in the middle represents the international average profile with residuals equal to 0 for all items, and the shaded area represents the range of residual values in the Nordic countries. The Norwegian profile is illustrated by the thin solid line within the Nordic profile.*

In Figure 1 the codes on the horizontal axis are the item labels for the 34 science items. The data are sorted from left to right with increasing Nordic range. Displaying this figure is an attempt to visualise a Nordic cluster before any such

clustering structure is established. This figure illustrates the characteristics of the Nordic profile, as compared to the overall international range and mean⁷. The Nordic river has at least two fundamental properties carrying different types of information. Firstly, the width of the river, or in other words the spread or distribution between the residuals in the Nordic countries, varies across items. This can be estimated by the range in p-values (as in Figure 1) or the standard deviation from the Nordic mean for all items. A relatively small range or standard deviation indicates an item where the residuals of the Nordic countries are very similar to each other, or in other words, it indicates a Nordic unity. Secondly, the mean Nordic residuals vary across items, and this indicates the relative strengths and weaknesses of the Nordic countries. Items where both these properties are distinct could be perceived as extremely characteristic of the Nordic countries.

One example is the publicly released item S129Q02, third item from the left in Figure 1. This is the second item in a unit labelled Daylight (see Appendix 1). This is the most difficult item in the pool with the overall international p-value of 0.17. The item is about modelling how the Earth is oriented relative to the Sun's rays, by indicating on a drawing the North and South poles, the axis between them, and the Equator. This item therefore requires factual knowledge in a physical context, it requires that students construct their own responses, and it is relatively independent of the stimulus material given although the stimulus material contains information about the tilt of the Earth's axis.

In the Figure light rays from the Sun are shown shining on the Earth.

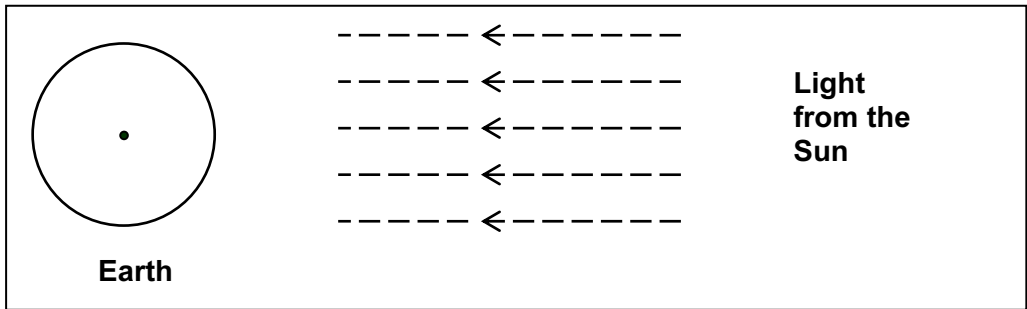


Figure: light rays from Sun

Suppose it is the shortest day in Melbourne.

Show the Earth's axis, the Northern Hemisphere, the Southern Hemisphere and the Equator on the Figure. Label all parts of your answer.

Figure 2: Item S129Q02 from the unit Daylight

The Nordic countries on average perform 6 percentage points below what could be expected for this item. Fundamentally, this item is about having a robust mental model of the Earth in the Solar System. It is possible to imagine that even

⁷ The Norwegian profile is only included as an example and will therefore not be discussed, but it illustrates that the Norwegian profile to a certain degree coincides with the overall pattern of the Nordic profile, that is, the overall pattern of local minima and maxima occurs for roughly the same items.

students who in another context might have formulated acceptable answers to separate questions relating to each of the isolated pieces of factual knowledge involved in the item (the inclination of the Earth's axis, and Equator as the line defining the Northern and Southern Hemispheres), would not necessarily be able to express the more comprehensive conceptual understanding involved in this item. The notion of such robust models of factual knowledge is not very prominent in the Norwegian science curriculum (KUF, 1996). Typically the specific aims of the Norwegian curriculum begin with formulations on the lowest cognitive levels, such as "students should become familiar with" or "be introduced to" some concepts, phenomena or objects. Whether this is a description that would also hold for the other Nordic countries requires further study, but the main issue addressed by presenting this single item is to illustrate that at the level of the single items data are highly specific and therefore single item analyses cannot be used to generalise beyond the item itself (R. V. Olsen, 2004; R. V. Olsen *et al.*, 2001). However, as the discussion above suggests, having several items requiring, for instance, that the students demonstrate their conceptualisation of robust mental models of scientific phenomena, would make it possible to study whether there is a pattern across these items that is distinctive for specific countries or clusters of countries. In other words, when the number of items increases in PISA 2006, not only will the proximity measures used in the cluster analysis be more robust, the possibility of generalising from item characteristics will be improved substantially.

Returning to Figure 1, The Nordic countries have a relatively small range as compared to the total international range. This is of course mainly due to the fact that the range always increases as the group size increases. However, the ratio of the ranges varies across items, e.g. items S129Q02 (the item presented in Figure 2) and S326Q03 have particularly narrow Nordic ranges as compared to the international range. In effect, the international range and the Nordic range (or the corresponding standard deviations) are moderately correlated ($r \approx 0.6$), which in this context is taken as evidence for that the variation in the Nordic range to some degree deviates from the international.

An extremely distinct Nordic profile would manifest itself in this type of diagram as a very narrow 'river' with high and low average residuals. This diagram is not such an extreme, but it does indicate that there is a Nordic cluster distinctly separable from the overall international profiles. In other words with the operationalisation inherent in this representation of what constitutes a cluster, it is indicated that there is a Nordic profile with some degree of internal cohesion and external isolation. On the other hand, this diagram also illustrates that there are distinct differences between the Nordic countries across items.

The type of diagram presented in Figure 1 is a helpful visualisation for evaluating *a priori* given clusters of countries. The shortcomings of this type of diagram in order to detect cluster structure is that we cannot rule out that one or more Nordic countries really has more in common with some other countries. This diagram forces or imposes a cluster structure on the data that we might reasonably expect to be present. It might also be that in a wider context, the Nordic cluster is not very prominent as compared to other clusters. In the 'Nordic

river' diagram the external isolation of the Nordic countries is only established by comparison with the extreme and average international profiles. Therefore, this tentative finding will be re-evaluated in relation to the cluster analysis below.

The cluster analysis

Identifying the main clusters

Figure 3 shows the dendrogram representing the solution of the cluster analysis. In this figure some groupings of countries have been marked by solid 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). In addition dotted lines have been used for subclusters within larger clusters (e.g. Hong Kong and Macao within the East Asian cluster) and for possible extensions to larger clusters (e.g. Tunisia to the South American cluster). According to the criteria previously presented, six groups of countries remain as distinct and possibly meaningful clusters:

- I. 'East Asian countries' (short label 'EastAsia'): Hong Kong, Macao, Japan and Korea
- II. 'English-speaking countries' (short label 'English'): Ireland, UK, Australia, New Zealand, Canada and the USA.
- III. 'North-West European countries' (short label 'NorthEur'): Switzerland, Liechtenstein, Germany, Austria, Luxembourg, Iceland, Finland, Denmark, Norway, Belgium, France, the Netherlands and Sweden.
- IV. 'South American countries + Portugal' (short label 'SouthAm'): Mexico, Brazil, Uruguay and Portugal.
- V. Less developed countries (short label 'LessDev'): Turkey, Indonesia and Thailand.
- VI. 'East European countries' (short label 'EastEur'): Latvia, Russia, the Czech republic, the Slovak republic and Serbia & Montenegro.

Although the East European countries are not externally very well isolated, the internal coherence is very distinct and as a result of the stability analyses, which will be returned to, this seems to be a meaningful cluster of countries.

The cluster of North-West European countries

Figure 3 gives little support for the hypothesis of a distinct Nordic cluster. Instead the Nordic countries are merged into the largest group of countries. This is a cluster of countries sharing many characteristics. It is a cluster of neighbouring countries; it is to some degree a linguistic cluster; it is a cluster of countries with a common political, socioeconomic and historical identity. As will be returned to later, all these underlying characteristics may influence school policy in general and in effect they might even influence science curricula. The least speculative of these characteristics is the geographical unity of the countries and this has therefore been chosen as the basis for labelling this cluster.

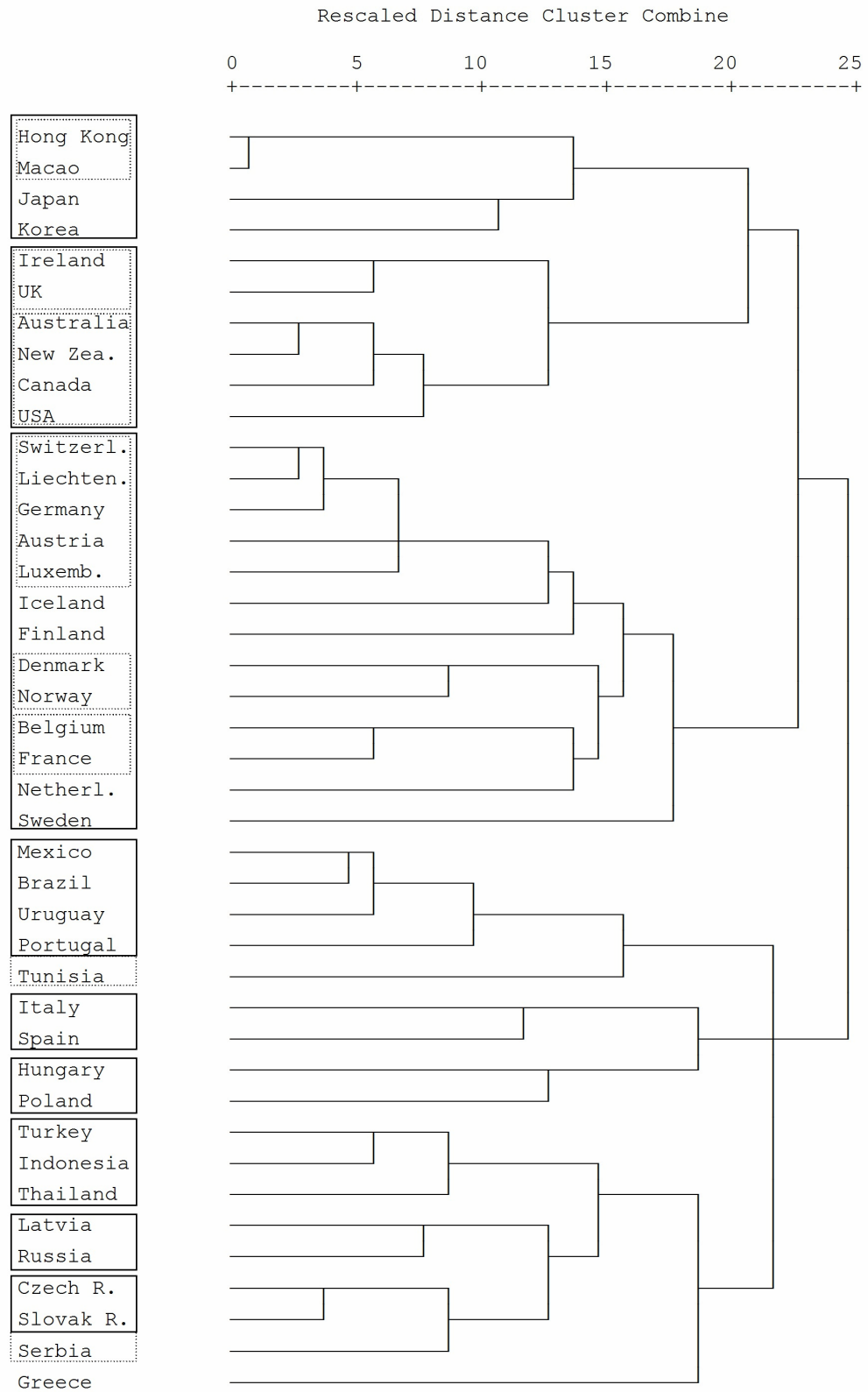


Figure 3: The dendrogram for country clustering. The groups defined as those with high degree of external isolation are framed.

[illegible]

Table 3: Matrix of correlations between all countries' residuals. The shaded triangles show the correlations within the four clusters (see text). In addition the subgroup of German-speaking countries within cluster 3 is marked. All positive correlations significant at the 0.05 level are boldfaced, and all negative correlations significant at the same level are boldfaced and in italics.

In such a large cluster it cannot be expected that all pairs of countries are similar. Sweden is for instance included in the group at a very late stage, that is, at a large distance from the rest of the cluster. However, all the countries share similarities with the average profile for the countries within the group. The average correlation coefficient is 0.32 and the coefficient alpha is as high as 0.86 (see Table 4), both taken as indicators of moderate internal coherence, although the magnitude of the coefficient alpha in this case is also due to the higher number of countries in this group as compared to the other groups. 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 small negative correlations between countries within this group as shown in Table 3. However, this cluster is not in accordance with the aim of reaching a final solution with clusters of approximately the same size.

Within this group there is one distinct subgroup which could also be perceived as a cluster by itself, the 'German-speaking countries' (short label 'German'), which has a high degree of internal cohesion. Table 3 shows that many of the countries' residuals in group III are relatively highly 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 the cluster. In the larger cluster it seems as though this subgroup is a 'centre of gravitation' attracting the other countries. The substantial nature of the average internal cohesion in the cluster of North-West European countries can in other words be an expression of this moderate to strong relationship with the German-speaking countries. The country standing out as the main mediator of this effect is Switzerland. All countries within the larger group 'NorthEur' have relatively high and positive correlations with this country.

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'). This is based on: (a) the Nordic river in Figure 1 documenting that for some items there is a Nordic unity; (b) prior studies documenting a Nordic unity across cognitive items (Angell *et al.*, in press; Grønmo *et al.*, 2004b; Kjærnsli & Lie, 2004; Lie & Roe, 2003; Zabulionis, 2001); and (c) the existing priority given to the study of international comparative data from a Nordic perspective (Lie & Linnakylä, 2004; Lie *et al.*, 2003). However, the cluster analysis has redirected this exploration of a common Nordic profile in science achievement to also include the study of the differences between the Nordic countries.

The coefficient alpha and the average correlation given in Table 4 strengthen the findings from the cluster analysis that the hypothesised Nordic cluster is not a very distinct cluster of countries ($r_{\text{avg}} = 0.24$, $\alpha = 0.59$). However, although these measures of similarities (internal cohesion) within the Nordic cluster are low as compared to other groups of countries, several Nordic countries have moderately positively correlated residuals, and this will be returned to shortly.

Throughout the article the cluster of German-speaking countries (cluster IIIa) and the Nordic cluster (cluster IIIb) will be included. It is natural to include the German-speaking group of countries 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 one of the objects of study for this article.

The other clusters

The other clusters will not be discussed in the same detail.

For the cluster of South-East Asian countries it should be noted that the internal cohesion is special since the two ‘countries’ 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, an interpretation that is highlighted in the short name ‘SouthAm’. This is in part also based on the fact that Portugal is the last country that comes into this cluster (see Figure 3). In PISA 2006 more countries from South America will participate (Argentina, Columbia and Chile) and so it will be possible to study the hypothesis that this cluster is mainly related to this geographical component in more detail. Furthermore, the cluster analysis indicates that Tunisia is also grouped into this cluster at even larger distances. Including this country will, however, lead to a noticeable decrease of the coefficient alpha for this group of countries. Primarily based on the criteria of ‘face value’, that it should be possible to conceptualise the cluster of countries with a representative label, and supported by the decrease in coefficient alpha, Tunisia is not included in this cluster.

The fifth group is even more problematic to label. Turkey, Indonesia and Thailand do not share any geographical or linguistic characteristics. Nevertheless, the group is found meaningful through the label ‘less developed countries’. In general most of the countries participating in PISA are OECD countries with relatively strong economies and well-developed democratic institutions. Even if Turkey is a member of the OECD and the country has ambitions of becoming a member of the European Union, it is not a typical representative of either Europe or the OECD. There are other countries included in the analysis that could also be labelled as less developed, e.g. some of the

countries in the South American cluster or the East European cluster, and also Tunisia. At longer distances all these countries are merged. In the very last step in the cluster analysis two larger groups are merged. These two groups are distinctly different in the level of economical development. In the upper half of the diagram there is a metacluster of rich and highly developed countries (EastAsia, English and NorthEur), while the lower half mainly includes less developed countries (SouthAm, LessDev and EastEur). Group V is therefore included as a cluster with a distinct external isolation, and as an example of a structure illustrating that clusters might be related to factors other than linguistic, geographical or historical identities. This group illustrates a structure that possibly could be related to social, economical or political factors. Furthermore, this cluster is kept since this is a feasibility study of what it might be possible to do with the data from PISA 2006, and once more, with the inclusion of more countries in PISA, this structure might be enhanced, refined and strengthened in the 2006 data.

In similar studies of data from TIMSS 1995 a distinct Eastern-European cluster of countries were present (Angell *et al.*, in press; Grønmo *et al.*, 2004b; Vári, 1997; Zabulionis, 2001) while in the studies of data from PISA 2000 (Kjærnsli & Lie, 2004; Lie & Roe, 2003) the indications for this pattern were somewhat weaker. The dendrogram (and also the total correlation matrix) suggests that Hungary and Poland differ most markedly from their partners in what Zabulionis (2001) labels the ‘post-communist’ group of countries. This was also a characteristic pattern of the science items in PISA 2000 (Kjærnsli & Lie, 2004). Nevertheless, the group of five countries (Czech Republic, Slovak Republic, Russia, Latvia, and Serbia & Montenegro) in the lower end of the dendrogram is much more coherent than for instance the Nordic group (see average correlations and coefficient alphas given in Table 4). 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. Also, in PISA 2006 even more countries from this region will be included, and thus the potential for studying characteristics for this group is promising.

As a result, in the following analyses six main clusters, presented above as clusters I to VI, will be used. However, the results for cluster III, the North-West European cluster, are not always easy to compare with the other clusters since it is a cluster at a higher level and includes many more countries. Therefore, from this cluster two subgroups are included as well: IIIa, the German-speaking countries, and IIIb, the Nordic countries.

Relationship between clusters

Table 4 gives the correlations between the average cluster profiles for the groups of countries mentioned above. The correlations between clusters are taken as measures of the degree to which the profiles are similar or dissimilar. Furthermore, in the diagonal of the table the average correlation between the countries in the group and the coefficient alpha is given.

	EastAsia	English	NorthEur	German	Nordic	SouthAm	EastEur	LessDev
EastAsia	0.77/0.45							
English	0.07	0.85/0.50						
NorthEur	-0.20	-0.06	0.86/0.32					
German	-0.13	-0.20	0.89	0.89/0.65				
Nordic	-0.10	0.00	0.85	0.61	0.59/0.24			
SouthAm	-0.52	-0.05	-0.28	-0.25	-0.35	0.83/0.57		
EastEur	-0.13	-0.64	-0.40	-0.29	-0.28	-0.06	0.81/0.45	
LessDev	-0.27	-0.08	-0.66	-0.65	-0.52	0.16	0.49	0.78/0.57

Table 4: Correlation coefficients between clusters of countries. Coefficients statistically significant at the 0.05 level are boldfaced. In the shaded diagonal are the coefficient alphas/average correlations within the clusters.

As could be expected most correlation coefficients are negative, resulting from the fact that these are groups externally isolated from each other in the cluster analysis. In general, this table of correlations is only a different way of expressing some of the information visualised by the dendrogram in Figure 3, which is also based on correlations as the measure of similarity.

Inspecting Table 4 from a Nordic perspective tells us that the overall Nordic profile is very similar to the German-speaking countries, as expected given the fact that these groups of countries were merged in the cluster analysis. In addition it is evident from this table that the South American countries, East European countries, and the group of less developed countries all have profiles that differ from the Nordic countries. Moreover, it is worth noting that the English cluster is not correlated with the Nordic group of countries.

Beyond a Nordic perspective it is noteworthy that the English profile can be regarded as almost the opposite of the Eastern European profile, and the same type of relationship is found between the East Asian and the South American profile. Possible explanations for these relationships will be returned to shortly. Furthermore, the positive correlation between the East European countries and the less developed countries is coherent with the fact that at larger distances these groups are merged in the cluster analysis.

Stability and validity

A first simple check of the stability of the clusters can be done by an intuitive inspection of the correlations between the countries' residuals. It is noteworthy that the correlation coefficients in Table 3 clearly indicate that countries within the same clusters have similar profiles: Nearly all the significant⁸ positive correlations (boldfaced in Table 3) are between countries from the same cluster. Furthermore, all the significant negative correlations (boldfaced and italicised in Table 3) are between countries belonging to different clusters.

⁸ Even if this article is not written in the 'spirit of' statistical inference, the significant correlations are boldfaced because they represent the *highest* correlation coefficients in the table. Therefore, these have been highlighted in order to visualise pairs of countries with very similar profiles across items.

However, as previously argued and described, a more systematic approach for studying the stability of the solution is required. The first step in the procedure is to analyse the same data combining other proximity measures (Block distance or Manhattan distance and ordinary and squared Euclidian distance) with other clustering methods (single linkage and complete linkage) in order to reveal if the clusters mapped in Figure 3 could possibly be artefacts of the specific method used. Furthermore, the analysis has been repeated excluding some countries one at a time. Also, in order to study the possible floor and ceiling effects, the matrix of logistically transformed residuals has been analysed. Without going into detail none of the alternative analyses came up with totally different clusters. Overall, the clustering method finally used is preferred because it presents a clearer cluster structure which is easier to interpret. In particular, the distances establishing the external isolation of the four labelled clusters are larger when using correlation as a proximity measure in combination with the average linkage clustering method. The most profound features of the alternative methods were:

- The ‘English’ cluster and the ‘German’ subcluster are particularly stable. Also, the ‘EastAsia’, the ‘SouthAm’ and the ‘LessDev’ clusters were always kept together (internal cohesion). However, the external isolation varied.
- The large cluster of countries in North-West Europe stayed large, but other European countries were sometimes included as well. Notably, Spain and Italy were regularly found in this cluster.
- In some analyses a clearer Nordic subcluster (except for Finland) emerged within this larger cluster.
- In several analyses the ‘EastEur’ cluster came up clearer than in the reported analysis, including also Poland and Hungary. However, it was sometimes part of a larger structure which also included some less developed countries.

In addition, coefficient alpha and average correlation within the clusters are reported in the diagonal of Table 4. Usually, coefficient alpha is used to evaluate the internal consistency reliability for constructs or test scores. Here, however, this index is not used as a measure of reliability. Together with the average correlation coefficients between countries within the cluster, the coefficient alpha is used in this analysis as an indicator of the internal cohesion of the clusters. All the four major clusters have relatively high alphas and average correlations supporting the idea that these clusters are internally coherent. In addition two subclusters of the North-West European cluster of countries are included. Primarily this gives further support to the use of a German language cluster with the highest average correlation between the item residuals of all the clusters. In addition, this shows that even though the Nordic cluster is not that well established as an internally coherent group of countries, there is considerable covariation among these countries, as also indicated in the ‘Nordic river’ in Figure 1. And finally, this establishes the five East European countries as a cluster with high internal cohesion. However, this cluster is not very well isolated

from the group of countries labelled as ‘less developed countries’, a fact that also can be seen from inspecting the correlations in Table 3

There are other strategies that can be used to identify the clusters. In particular, a common strategy for identifying clusters is to find the ‘best’ vertical line intersecting the material at the distance with the most distinct cluster structure. There is no single line that can reproduce all the six clusters suggested above. A vertical line at about 14 on the scale at the top of Figure 3 would result in all the suggested clusters except the North European cluster. Placing the cluster at a somewhat larger distance so that this cluster is included (at about 18 on the scale) would as a result merge clusters V and VI.

As a last element of the study of robustness, this cluster structure has been compared to other studies using a similar method. In short, the correspondence between these analyses is quite remarkable. Many of the same clusters reappear in all these analyses, a phenomenon that will be referred to when discussing possible implications. The only cluster that is not adequately supported is the group of Nordic countries. Nevertheless, as previously argued, this cluster will be included in order to describe both the unity and the differences between the Nordic countries.

Evidence for the stability of the presented solution and the validity of the interpretation that there are seven possible clusters of countries can be summarised as follows: The extended procedure to check the stability of the proposed solution, including (a) an inspection of the correlation coefficients within and between the groups and the coefficient alphas for the clusters, (b) performing alternative cluster analyses using different proximity measures and various clustering methods, (c) an analysis of the logistically transformed residuals, (d) replications of the reported analysis by excluding some countries one at a time, (e) application of different strategies for identifying the main clusters, and finally, (f) a comparison with other studies with a similar method on similar data, all points to that the explored pattern is relatively robust and it is unlikely that the clusters reported are artefacts of the specific method chosen.

Exploring the item residuals in the clusters

It is not evident what is required for, and what counts as, an explanation of these findings. The clusters are based on profiles across 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 expected on many of the same items. The most direct approach to explain these clusters would therefore be to study the substantial nature of those interactions having the most profound influence on the cluster structure. That approach is chosen here. However, a more fundamental type of explanation would refer to the possible antecedents of these patterns. The labels chosen for the groups more than indicate possible geographical/linguistic or in a wider sense cultural antecedents. In the other studies using a similar method the main suggestions for explanatory factors have been of a wider cultural type (Grønmo *et al.*, 2004b; Zabulionis, 2001). The position taken here is that one should be careful not to jump to conclusions about such fundamental explanations. One small first step in order to understand the

clusters would be to look more closely at the patterns across items in order to identify differential weaknesses and strengths related directly to the items. In this paper priority will be given to this small step before returning in the discussions to some suggestions for possible mechanisms for the empirical patterns in the p-value residuals across countries. In the analysis below the Nordic perspective will again be emphasised.

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 different clusters were in general too few. 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 the previously presented broad item descriptors, indicating some overall characteristics shared by many items, has been analysed. As stated when presenting the Nordic river, the profile for a cluster across the items is characterised by both the means for all items and the deviations from this mean within the group. While the means indicate the relative *strengths and weaknesses* for the cluster of countries as compared to all other countries, the deviations within the group identify items characterising the *unity* within the group. In order to study both these characteristic features of the clusters, both the cluster means for all items, and the standard deviation within the cluster for each item, has been correlated with the broad item descriptors previously presented.

The degree of unity in the profiles for the clusters

	Descriptives for standard deviations within cluster			Correlation between standard deviation in cluster and item descriptors				
	Min	Max	Mean	Competency	Context	Format	Textdist	p-value
EastAsia	2	13	6	-0.09	0.04	0.07	-0.16	-0.25
English	1	8	4	-0.52	0.20	-0.03	-0.53	-0.22
NorthEur	2	8	5	0.08	0.07	-0.09	-0.13	-0.13
German	2	7	4	0.08	-0.25	-0.10	0.10	-0.37
Nordic	1	10	5	0.00	0.10	0.07	-0.18	0.08
SouthAm	1	12	4	0.14	-0.19	0.00	0.08	0.09
EastEur	1	9	5	-0.14	-0.18	-0.02	-0.10	-0.12
LessDev	1	17	6	-0.12	-0.01	0.22	-0.10	0.44

Table 5: Descriptives and correlations for the standard deviations in item residuals within clusters.

Table 5 summarises the characteristics for the degree of unity within the clusters. The issue about to what degree there is a unity within the clusters is represented here by the standard deviation of the item residuals *within each of the clusters*. In other words, this is a measure of how far the residuals for countries within a cluster are from the average residual in the group. To the left of the table, the

degree of unity across items is described by the minimum, maximum and mean standard deviation across the 34 items. Thinking in terms of the type of visualisation given in Figure 1 (the 'Nordic river'), this is a description of how wide the river is, and how this width of the river varies. The right-hand side of Table 5 shows how the degree of unity is related to the broad item descriptors previously defined.

It is evident that the degree of unity varies across items. The left-hand side of the table tells us that if we had drawn similar rivers to that shown in Figure 1 for the clusters of English- and German-speaking countries they would have been slightly narrower than the Nordic river. If we had drawn the East Asian river and the one for the less developed countries, they would have been slightly wider. Accordingly, the overall unity within the English- and German-speaking clusters, respectively, is slightly higher than in the other clusters. However, the differences are small. The correlations with the item descriptors in the right-hand part of Table 5 show that the degree of unity of the item residuals within a group does not vary systematically as a product of these broad characteristics of the items, with a few exceptions.

The English profile is quite distinct in that the variation in the residuals within this cluster is related to the competency being tested and the degree to which the item requires that the students make use of the stimulus material. The negative signs indicate that (a) the profiles of residuals for the English-speaking countries are relatively more similar (standard deviation low) for items testing their understanding of scientific processes than for items testing their understanding of scientific concepts, and (b) the profiles for the English-speaking countries are relatively more similar for items which require the stimulus material to be actively used in the solution process than for items which could be answered correctly without direct use of the stimulus text⁹. Furthermore, there is an overall tendency for easier items (high p-values) to have a relatively smaller variation in residuals within several of the clusters and especially so for the group of German-speaking countries. However, the exception to this generalisation is that in the group of less developed countries the residuals are more similar for difficult items.

⁹ As mentioned previously these two item descriptors are correlated. In order to check the effect of this, each of the two boldfaced correlations for the cluster of English-speaking countries has been recalculated controlling for the other item descriptor. The correlations are still moderate to high ($r \approx -0.4$)

The relative strengths and weaknesses for the clusters

Table 6 describes the magnitudes of the item residuals and how these are related to the item descriptors. What is evident from the descriptives in the left-hand side is that the average residual in the East Asian cluster and in the cluster of less developed countries varies a lot more across items. From a psychometrical perspective this corresponds to the column on the far right of Table 2 telling us that the standard error of international measurement is larger for these countries. From a science education perspective, where the p-value residuals are considered as important descriptions of differences across countries, this tells us that for some reason the performance of students varies more across items for the East Asian and the less developed countries.

	Descriptives for average residuals within cluster			Correlation between average residual in cluster and item descriptors				
	Min	Max	SD	Competency	Context	Format	Textdist	p-value
EastAsia	-20	23	8	-0.27	0.18	-0.42	-0.15	-0.21
English	-8	9	5	0.44	-0.23	-0.11	0.16	-0.05
NorthEur	-8	8	4	0.19	-0.14	-0.04	0.36	0.08
German	-12	10	5	-0.06	-0.15	-0.08	0.19	0.03
Nordic	-6	9	4	0.23	-0.04	0.07	0.45	0.08
SouthAm	-10	14	6	0.12	-0.13	0.22	0.14	0.13
EastEur	-8	18	5	-0.34	0.29	0.31	-0.34	-0.01
LessDev	-17	28	9	0.04	0.19	0.20	-0.20	-0.13

Table 6: Descriptives and correlations for the item residuals within clusters.

What stands out in Table 6 from a Nordic perspective is that the textual aspect of the PISA items is very important. This means that Nordic countries perform relatively better for items where careful analysis of the text in the stimulus material is vital to reach a solution or, as in an alternative interpretation, the Nordic countries perform relatively more poorly on items where the solution is more independent of the textual material itself. Furthermore, a relatively strong relationship is found with the 'Process' variable, which means that the Nordic countries perform relatively better on items testing the understanding and mastery of scientific processes. 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 tells 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 against 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. The degree to which these characteristics are common for all Nordic countries will be returned to.

A compact characterisation of the other clusters is that the English-speaking 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-speaking 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).

Overall, the text distance and the competency involved in the solution of the item seem to be the item characteristics that most successfully separate countries. Central to competencies II and III are skills related to argumentation (e.g. identifying evidence and identifying questions that can be answered by scientific investigation). Such skills could be expected to be related to the indicator of closeness to the text. Argumentation has, for instance, very much to do with the ability to extract information from different sources. And indeed, these two indicators are correlated ($r \approx 0.5$). As a result, when correlating the Nordic residuals with the variable ‘Competency’ controlling for the closeness to the text (‘Textdist’), this correlation disappears totally. However, when the same is done for the English-speaking cluster the correlation with the ‘Competency’ variable stays more or less unchanged. This suggests that the relationship between these two important item characteristics is not straightforward.

A closer inspection of the Nordic unity and diversity

From Table 3 it could be seen that Denmark is the country which has the most prominent overall Nordic profile, correlations with the other Nordic countries being in the range 0.2–0.5, while Sweden is at the other extreme having weak overall correlations with the Nordic neighbours. The latter is quite surprising given the fact that Sweden has been more centrally placed in the Nordic cluster in similar analysis of other datasets. For instance Kjærnsli & Lie (2004) in a similar analysis of PISA 2000 science items found that the Swedish item-by-country residuals were highly correlated with those in Norway and Iceland. On the other hand, they also found that Sweden was the Nordic country with the overall weakest correlation with the average Nordic profile. One way to study the degree to which individual countries are similar to a group of countries, in this case the group of Nordic countries, is to study the degree to which the residuals in individual countries are correlated with the mean Nordic profile of residuals as presented in Table 7.

Denmark	0.74	Australia	0.20	Russia	-0.20
Switzerland	0.70	Korea	0.01	Uruguay	-0.21
Norway	0.69	Canada	0.01	Greece	-0.22
Iceland	0.66	Italy	-0.02	Hungary	-0.23
Finland	0.58	Czech rep.	-0.04	Brazil	-0.28
Belgium	0.58	Spain	-0.06	Tunisia	-0.29
Liechtenstein	0.53	Japan	-0.07	Slovak Rep.	-0.32
Germany	0.50	Latvia	-0.08	Poland	-0.34
Sweden	0.44	Macao	-0.11	Turkey	-0.35
Austria	0.43	Hong Kong	-0.11	Serbia	-0.37
Luxembourg	0.43	Portugal	-0.12	Thailand	-0.41
New Zealand	0.34	UK	-0.14	Mexico	-0.46
Netherlands	0.28	USA	-0.14	Indonesia	-0.51
France	0.23	Ireland	-0.15		

Table 7: Correlation between the mean Nordic *p*-value residuals and the residuals in all the participating countries. Statistically significant correlations are boldfaced.

Table 7 confirms what has already been stated: that Denmark is the country most closely linked to the average Nordic profile across science items, and Sweden is the Nordic country that has least in common with this average Nordic profile. Moreover, Sweden is actually in this respect ‘less Nordic’ than many countries outside the Nordic region. In particular, the tight link between the Nordic countries and the German-speaking countries is once more emphasised in the figures given in Table 7. Switzerland is remarkably close to the Nordic profile. This was also the case in analyses of TIMSS 1995 data (Lie *et al.*, 1997) and PISA 2000 data (Kjærnsli & Lie, 2004), although not so strongly as here.

The list of the item residuals in Table 2 and the visualisation of the Nordic residuals given in Figure 1 tell us that the degree of unity across the Nordic countries varied across the items. This has been confirmed in the subsequent analyses indicating that the Nordic countries are not more similar to each other than they are to some other countries in the north-western part of Europe. Furthermore, the description of the clusters given in Table 5 has shown that even if the other clusters are more distinct in terms of the cluster analysis, the unity across the countries also within these clusters varies across items. This suggests that in developing descriptions for the clusters of countries it is just as important to describe in addition the differences across the countries within the groups. Here this will be done only for the Nordic group of countries.

	Descriptives for average residuals in country			Correlation between average residual in country and item descriptors				
	Min	Max	SD	Competency	Context	Format	Textdist	p-value
Denmark	-16	19	7	0.42	0.02	0.23	0.38	0.32
Finland	-20	13	8	-0.01	-0.25	-0.24	0.23	-0.12
Iceland	-12	13	7	0.00	-0.12	0.15	0.24	0.00
Norway	-8	13	5	0.21	0.23	0.24	0.11	0.04
Sweden	-13	7	4	0.18	0.17	-0.11	0.45	0.05

Table 8: Descriptives and correlations for the item residuals in the Nordic countries.

Table 8 gives a description of the p-value residuals in the Nordic countries together with correlations between the residuals and the broad item descriptors. As a consequence of the way in which these residuals are calculated their average is 0 in all countries. However, it is evident that the variation across the items is less in Norway and Sweden than in the other Nordic countries, as can be seen from the left-hand side of Table 8.

The correlations with the item descriptors indicate some differences in the profiles across items. The average Nordic profile was characterised by a relative success on items requiring that the textual material provided was interpreted. Table 8 confirms this and gives us some more details about this finding. This characteristic of the profile is particularly strong for Denmark and Sweden and weaker in Norway.

Some interesting contrasts are also indicated. Finland is characterised by performing relatively better on items addressing issues related to life and health, while the Norwegian students perform relatively better on items related to aspects of physical phenomena. When it comes to format, the Finnish students perform better on items requiring the students to construct their own answers, while the Danish and Norwegian students in particular perform relatively better on items asking the students to select an appropriate answer. This tendency should be noted, even if it is moderate or weak. One often hears, both in Danish and Norwegian contexts, that our students are not used to the multiple choice format, while students from many other countries are familiar with this format, and as such this introduces a bias into tests such as those in TIMSS and PISA (see for instance J. V. Olsen, 2002). The results presented here supply abundant evidence that Nordic students are not negatively biased by selected response format. Lastly, it may be noted from Table 8 that the Danish students perform relatively better on easier items.

Discussion and implications

The results presented clearly suggest that there are distinct clusters of countries, and some characteristics of the profiles across items for these groups of countries have been presented by studying how the average p-value residuals in the clusters are related to some broad item descriptors. A particular emphasis has been given to a Nordic perspective. In the following some of these results will be discussed and possible implications for the design of, and the analysis and interpretation of results from, large-scale comparative assessment will be suggested.

A Nordic profile of science achievement?

The results presented are not conclusive regarding the Nordic aspect of the research questions. It is evident that there are many similarities between the Nordic countries. For many items the residuals are very close to each other (Figure 1), and to some extent the magnitude of the item residuals for the Nordic cluster had marked correlations with some of the broad item descriptors (Table 6). It was found that the Nordic profile 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 where the correct response was highly dependent on reading and interpreting the textual material. Even if this correlation describes a particular feature of the link between the average Nordic profile and characteristics of the items, Table 8 identified that in the Nordic frame of reference this link was strongest for Denmark and Sweden, and relatively weak for Norway. The other item descriptors were not linked to the Nordic profile of residuals. Furthermore, the average correlation between the Nordic countries' residuals was moderate to low. Sweden was particularly weakly linked to the other Nordic countries, which was particularly emphasised in Table 7 where the correlations between the average Nordic profile of item residuals and the individual country profiles were given. Since the Swedish residuals were included in the mean Nordic profile, the individual profile for Sweden is automatically correlated with the average Nordic profile. Nevertheless, the Swedish profile is only moderately correlated with the mean Nordic profile. In fact, the mean Nordic profile was more strongly correlated with some non-Nordic countries' profiles, particularly the profiles of some of the German-speaking countries.

In general, 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 3). 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, etc. 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. The initial interpretation of this is that the common profile for this larger group of countries is due to similarities with this German profile. In the Nordic context this means that the commonalities in profiles seen for the Nordic countries in science through several studies (Angell *et al.*, in press; Grønmo *et al.*, 2004b; Kjærnsli & Lie, 2004; Lie *et al.*, 1997) to some extent may possibly be explained by a common reference to the school science in Germany and other German-speaking countries. 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 general educational theory. Kjærnsli & Lie (2004) found the same relationship between the Nordic and German-speaking countries and they suggested that this may be due to German influence on how science as a subject has been established and taught in school, without specifying their argument in any more detail. 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, as will be returned to later.

The finding that the Nordic countries' profiles of item-by-country interactions are linked to the German-speaking countries' profiles is consistent with similar analyses of the science items in PISA 2000 (Kjærnsli & Lie, 2004) and of the science items in TIMSS 1995 (Angell *et al.*, in press; Grønmo *et al.*,

2004b), and shows that this tight link between these two regions' profiles of science achievement is a well-established empirical fact.

Scientific literacy and reading

The main characteristic of the Nordic profile is that students in our region tend to do relatively better on items involving careful reading than on items not directly dependent on reading of the text. The positive interpretation of this is that the Nordic students perform relatively well on a competency generally valued as important in a post-industrial society: the ability to interpret and reflect on textual material. The negative interpretation is that Nordic students do not have a strong knowledge base in science, and the relative success in items requiring reading is related primarily to the fact that many of these items do not require that the student possesses any prior knowledge. In the analysis performed here these two possible interpretations cannot be distinguished.

PISA also has a component testing reading literacy. The concept of reading literacy as defined in PISA goes beyond the technical aspects of reading as such. It focuses upon reading in different modes, or reading for different purposes: to retrieve information from a text, to interpret the meaning of a text, and to reflect on the form and content of the text (OECD-PISA, 1999, 2003). Scientific literacy has been found to be very highly correlated with reading. In PISA 2000 the latent correlation¹⁰ between these two domains was found to be nearly 0.9 (Adams & Wu, 2002). It is therefore interesting to note that all Nordic countries performed relatively better in reading than science, the exception being Finland which had the highest score for any country in both reading and science. It could therefore be expected that a relative strength for the Nordic countries is related to items requiring reading competency of this kind.

Since this textual characteristic of the items in general was the item descriptor that could most successfully account for differences in the achievement profiles of the clusters, it is necessary to sharpen and refine this aspect when more items are available for analysis. Norris & Phillips (2003) have described scientific literacy in a *fundamental sense* as being able to read/write science texts and in a *derived sense* as being knowledgeable and competent in science, and the relationship between the two. The results related to the textual aspect of solving items imply that scientific literacy in its fundamental sense is indeed a component or dimension that requires attention in interpreting achievement scores in scientific literacy reported from the PISA study. Fang (2005) has for instance by using analytic tools derived from a systemic functional linguistic perspective (eg. Halliday & Martin, 1993) studied some representative examples of material from textbooks in school science. The examples clearly demonstrate that these two types of scientific literacy are not only interrelated, but also inseparable.

In the framework for PISA (OECD-PISA, 1999, 2003) linguistic perspectives are not very explicitly linked to the overall trait of scientific literacy,

¹⁰ Theoretically the possible magnitude of a correlation coefficient cannot exceed the reliability with which the variables are measured. Latent correlation coefficients are adjusted so that this is taken account of.

or in other words, scientific literacy is mainly presented in a derived sense. However, 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). For many of these units the stimulus material is an extended piece of text, and the texts are no doubt texts that have the same characteristics as those analysed by Fang (2005): many of the texts have a high informational density; processes and phenomena observed in nature or laboratory are abstracted by the use of nouns (nominalisation); and they include specialised technical language. By this operationalisation of scientific literacy the fundamental sense of the concept deserves more attention in the parts of the framework discussing what it means to be scientifically literate.

We have seen that when using some broad descriptors characterizing the items in PISA to account for the cluster profiles, the rough indicator of how vital reading of the text is for the solution is the item descriptor that most successfully could account for the profiles in several of the clusters. Given the available theoretical discussions within science education research on how learning science in many respects is learning to talk, write and read science, and that being scientifically literate in many ways is to know and understand the language of science (eg. Bisanz & Bisanz, 2004; Fang, 2005; Lemke, 1990; Norris & Phillips, 2003; Roth & Lawless, 2002; Wallace *et al.*, 2004; Wellington & Osborne, 2001; Yore *et al.*, 2003), the claim that the fundamental sense of scientific literacy deserves closer attention in the future frameworks of PISA is further strengthened. Furthermore, the link between this emerging field of science education and the operational definition of scientific literacy in PISA 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 to items across these domains could give valuable insights into how these domains relate to each other.

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 subject, independent of study, independent of year of administration, and largely independent of the specific clustering method used. Furthermore, the larger metacluster of North-West European countries has been present in 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 this part of the world. Also, a

Nordic cluster was more clearly present in the analyses of TIMSS 1995 items. The only cluster which is not seen as clearly in the other studies is the cluster of South American countries. However, the reason for this is that in most other studies there has been only one or two countries from this region. All in all, the consistency across the reported analyses gives further reassurance to the conclusion that the clusters of countries presented above are indeed a collection of countries or school systems with common cultural elements which to a varying degree are relevant for the different clusters.

It is reasonable to suggest that further investigation of this phenomenon is warranted. Central to such an investigation would be theoretical contributions with reviews and further developments of the possible mechanisms that might link possible antecedents to the patterns revealed. In doing this, one should find ways to include items from the questionnaire describing the school systems as explanatory variables for the profiles. Also, it should be possible to develop a more distinct science educational perspective 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.

A psychometrical perspective: Residuals and fair tests

The work presented in this article is part of an overarching framework or rationale for studying the cognitive data collected by large-scale international comparative assessment studies, with a specific link to the PISA scientific literacy items. 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 test considerable efforts are made to produce items with minimal item-by-country interactions. Items with large interactions are consciously removed after the field trial. The cognitive traits being measured in PISA have been developed from an operational assumption that such traits are universal and transcend cultural particulars. This is not to say that specific contexts woven into the tests as such, or more specifically into the textual material, will not interfere with cultures within or across countries. Rather, it is to say that when the items are developed attention is given to the cultural and curricular diversity in the participating countries so that systematic bias is avoided as far as possible.

No item-by-country interactions could be considered as an ideal property of a test in an international comparative assessment study since such interaction could threaten the aim of the test, which is to compare countries by measuring the same trait in all countries. First of all, large interactions are equivalent to saying that the standard error of international measurement is large. Furthermore, if the interactions are systematically skewed across countries they might introduce bias in the measurements. If the item-by-country interaction for a specific item is large in many countries, this could be taken as an indication that the item measures different concepts in different countries. Seen from a didactical or subject-centred perspective the procedure of excluding items with such interactions means that highly interesting information about the differences between countries is consciously *not* collected.

Wolfe (1999) has studied profiles of residuals across content categories in mathematics in the Second International Mathematics Survey (SIMS). He concludes that when the profiles of achievement are too discrepant, the overall comparison is either “fundamentally unfair or essentially random” (Wolfe, 1999, p. 225). Furthermore, he concludes that 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. Unlike SIMS and the sequels TIMSS 1995 and 2003, PISA does not 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’. However, Wolfe’s (1999) argument related to the error component is just as important for PISA as in any other international comparative assessment. 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. Thus 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 regional comparisons, given that the profiles are comparable across the countries in a region. Examples of such comparisons are the regional analyses of TIMSS 1995 data in Vari (1997) viewed from an East European perspective. Similarly, PISA 2000 data have been viewed from a Nordic perspective in a special issue in *Scandinavian Journal of Educational Research* (Lie & Linnakylä, 2004) and in the book *Northern Lights on PISA* (Lie et al., 2003). Following Wolfe’s (1999) advice we could imagine that regional designs, including a total rescaling of the data, would increase the information provided by each item to the scale. On the other hand, such regional designs would remove the contrast with which national data can be compared, and from this perspective potentially interesting information would be lost. However, with the current development in PISA where more and more countries are being included, the argument of regional designs for analyses is highly relevant since this would no doubt introduce even larger analytical problems. Tables 2 and 6 indicate that the magnitudes of the residuals are smallest for the countries that can be labelled as modern western societies and substantially larger for a number of countries outside of this group, for instance the East Asian countries, and even more so in the group of less developed countries.

However, it is important to note that the magnitude of the standard error of international measurement that Wolfe (1999) perceived to be a problem of international comparative assessment was larger in the data from SIMS that he based his arguments on. It is likely that the decrease in this measurement error 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). A more speculative explanation for the decrease of the residuals from a test implemented in the eighties (SIMS) to a test implemented two decades later (PISA) could be that this can be taken as evidence for what some have claimed to be a consequence of the globalisation phenomena of which international assessments

are a part; a standardisation of education worldwide (Goldstein, 2004a; Kellaghan & Greaney, 2001; von Kopp, 2004).

At the centre of this critique is the question of how useful it is to rank countries along one dimension as is usually done in all large-scale international comparative assessment studies of educational achievement:

Finally, any such survey should be viewed primarily not as a vehicle for ranking countries, even along many dimensions, but rather as a way of exploring country differences in terms of cultures, curricula and school organization. To do this requires a different approach to the design of questionnaires and test items with a view to exposing diversity rather than attempting to exclude the 'untypical' (Goldstein, 2004b, p 329).

I would suggest that the data produced by studies like PISA may be used to explore country differences. One suggestion to increase the potential for studies of unity and diversity across countries would be to retain items in the test with clear item-by-country interactions. These items could then be left out when computing the overall scale, and instead be used only in analysis of the international diversity. This would, however, not be a very efficient test design. A more feasible approach would be to utilise the data from the field trials from this perspective. The likelihood is high for having a rather large collection of items with relatively strong item-by-country interactions in the field trials, and thus, analyses like the one presented in this paper will have data that are better suited for studying diversities. I have previously pointed to the fact that the type of analysis presented above will be more feasible with the data from the 2006 cycle in PISA since the number of items will be three times higher than it was in 2003. This point is even stronger for the field trial in 2005 which has an even higher number of items, about twice as many as that in the final cognitive test in the main study in 2006. However, the databases from the field trials are weaker in many other respects, e.g. the sampling procedures are less rigid than in the main studies. But still the data from the field trials are well documented and of a quality that is satisfactory for such analyses.

Some more remarks are needed related to the concept of fairness. 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, and since they link countries in a systematic way, they are clearly not random fluctuations, and therefore they could introduce bias. Item response format is an obvious example of an item characteristic which is not intended to be a part of the trait being measured. If the item format introduces systematic differences in item scores across countries, this could be regarded as a bias. The analyses presented (Table 6) indicate that there might be a possible bias related to format. In an alternative test with only multiple choice items, the most likely prediction is that the large performance gap between the South-East Asian countries and the countries from East Europe (OECD-PISA, 2004) would be reduced. And in a test with only open-ended format the gap would increase. On the other hand, selected response items are easier than constructed response items. It could be that the reported correlations between format and item

residuals for these two clusters are due to a ceiling effect for the selected response items. However, the correlations with format are approximately the same for the logistically transformed data. Also, when controlling for p-value, which is a measure of the difficulty of items, the correlations are more or less unchanged. This suggests that there is a need for additional studies targeting the issue of how different formats interact in different cultural settings. Greenfield (1997) reports for instance that Maya Indians were very confused by the multiple choice format. Instead of perceiving the list as a set of alternative solutions whereby one was the correct or appropriate one, they perceived that the list provided information relevant for the solution of the task, and used strategies for solving the problem that involving utilising all the elements in the list to construct a response. Hambleton (2002) adds to this that this format is very unfamiliar in an African context, and even more relevant for the specific finding of East Asia he reported that in a Chinese context they had to make a minor adaptation in the response format. Instead of filling in the bubbles or circles next to the appropriate answer, the students were instructed to tick their preferred response. In PISA the format is a third one, involving circling a letter next to the preferred response, or circling 'Yes' or 'No' for a selection of statements. One suggestion for studying such effects from a cultural perspective would be to include 'the same' item in different formats in the field trials (R. V. Olsen et al., 2001). A negative side-effect would be that this would occupy a substantial amount of the available testing time, and thus fewer items could be trialled.

The other variables that are also differentially correlated between the clusters of countries (Table 6) are directly related to the definition of the trait being measured, and therefore these correlations could not by themselves be regarded as indicators of a bias. On the other hand, if PISA is *perceived* to function as 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 with an eye 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.

Some possible fundamental explanations of diversities

The countries within most of the clusters obviously have many things in common, and the clusters might be referred to in wider sense as representing different cultures in some way. Also, the striking consistencies across domains, across year of testing, and across assessment designs may be taken as evidence that the observed response profiles are to some degree independent of the domain or subject tested. In the end, however, one has to substantiate how features of a culture might influence students' responses to items testing their scientific literacy. It is not easy to see how such factors can be connected to the empirical findings presented above, but some possible mechanisms can be suggested. In

general such mechanisms will be referred to as cultural antecedents, highlighting the fact that they are thought of as causes of the effects documented in the results presented in this paper. Possible mechanisms will be described more specifically below, but a general statement is that such antecedents are introduced into the response patterns by different agents. Firstly, some of them may have a direct effect and secondly, they may also be mediated and enhanced through curriculum documents, textbooks and assessment systems.

At the most fundamental level, belonging to a culture involves sharing a special way of observing, judging, valuing and participating in the world. Consequently, thinking, values, attitudes and emotions are affected by that culture. This is often referred to as having a certain *world-view*. In general this is a less than precise concept referring to the set of presuppositions or assumptions which predispose you to feel, think, and act in predictable patterns. Such dispositions might be thought of as a culturally dependent, subconscious, fundamental organisation of the mind (Cobern, 1991). Kearney (1984) refers to world-view as

...culturally organized macro-thought: those dynamically inter-related basic assumptions of a people that determine much of their behaviour and decision making, as well as organizing much of their body of symbolic creations ... and ethno-philosophy in general. (p. 1)

In this way, the concept of world-view is related to cognition in general; a world-view inclines one to a particular way of thinking, or as formulated by Kearney (1984) a worldview

...consists of basic assumptions and images that provide a more or less coherent, though not necessarily accurate, way of thinking about the world. (p. 41)

Different worldviews are most often associated with civilisations, religions and eras (Cobern, 1996; Quigley, 1979), e.g. one speaks of a Western worldview, an Eastern or Chinese worldview, a medieval worldview, or a *scientific worldview*. Different worldviews are likely to be more or less coherent with a scientific one (Aikenhead, 1996; Cobern, 1996). In conclusion, students' responses on items are probably somehow affected by fundamental assumptions about how the world actually is (the ontological issue) and how knowledge about the world may be obtained and communicated (the epistemological issue). If this is the case, this would in the end produce item residuals that are clustered as in the PISA 2003 science data.

A more specific and concrete aspect of culture and worldview is the tool by which it is communicated: language. Given that the science items in PISA no doubt to a high degree include competency in reading (as previously discussed), we should not be very surprised that some of the clusters consist of countries with similar languages. In addition to being an important element in preserving and mediating worldview in a culture, language also has a potentially more direct effect on students' responses to test items. Taking the position that direct translation is not completely possible, in other words that all aspects of meaning and companion meaning of a text cannot be kept unchanged in a translation, it is not very likely that the difficulty of an item will be the same in all languages. It is, however, difficult to find specific examples from the science items in PISA

where this obviously has happened. In the process of constructing items, well-known problems from the literature on test adaptation (Hambleton, 2002; Hambleton & de Jong, 2003) have been emphasised (Halleuxd, 2003) and in general this type of potential bias has been taken very seriously in the item development (Adams & Wu, 2002; Grisay, 2003; McQueen & Mendelovits, 2003). Indications that items in the field trial have worked differently in some countries have been reported back to the national centres, followed by recommendations that the items are checked for translational ‘errors’. Very often possible sources for the malfunction of the item have been identified, and the item could be successfully modified. However, the systematic features of the residuals presented here go beyond such ‘errors’. It is highly unlikely that the independently processed translations in countries with similar or the same languages have resulted in identical ‘errors’. If so, they could hardly be called errors, but rather situations where in fact ‘correct’ translations were not available in those languages. One example of how this could produce systematic effects is found in the word ‘scientific’ that appears in several places in PISA items. Translating this word into Norwegian, or any Germanic language, may be problematic. In Norwegian one would have to use either the word ‘vitenskapelig’ or ‘naturvitenskapelig’ depending on the context. Both these terms have a more formal flavour referring to science as a field of academically based research (the German *Wissenschaft*), something done by professional scientists. Such connotations or companion meaning *may* affect the item in a systematic manner. I have to stress that this was only meant as an example to illustrate the general issue. I have no evidence that this example, or any other example, has had such a systematic effect.

Another commonality for most of the clusters is that, to a large degree, they are neighbouring countries, or in other words, the clusters have a geographical character. This can also be used to understand why the p-value residual matrix seems to be well represented by a cluster structure. There are several possible mechanisms for how such neighbouring countries might develop common profiles across science items. Firstly, since many of the items are related to phenomena or issues related to the life of the students, this can create differential item functioning due to differential exposure to the phenomena or differential familiarity with the issues addressed in the items. Some examples of such phenomena or issues are that: climate and weather vary with geography; different sources of energy are used in different parts of the world; environmental problems such as the greenhouse effect, although it is a global issue, may be perceived and experienced as more relevant in some parts of the world. Differential familiarity with such phenomena is not only related to the direct experiences of them, but is probably also strengthened through curricula that to some degree will emphasise aspects in science that are important in the local, national or larger regional context. Furthermore, it is likely that geographical neighbours have a relatively stronger influence on each other in many ways. This might lead to the exchange of a general policy for schools, including, for instance, ideas about how science should be taught, or documents describing the content in science courses. It might also have a direct impact such as in the

exchange of textbooks and other instructional material. An extreme example of the latter is in Iceland where textbooks in science, for instance, are translated from the other Nordic languages.

These were some examples of possible causal links between wider cultural antecedents and the country or region specific achievement profiles. The clusters of countries reported here are very stable and replicates the clusters reported from other studies and it is therefore likely that these patterns are related to such antecedents somehow. In this paper such general and more fundamental characteristics were only brought into the discussion to the extent that they could be more tightly linked to some of the specific findings of this study. It is to be hoped that the description given of how the profiles in the clusters are linked to item descriptors will stimulate the debate and future efforts to find ways of connecting these patterns to fundamental explanations. A further hope is that the various possible antecedents described in this concluding discussion can be used as a starting point for the design of future studies with the aim of developing a more systematic description and understanding of the unity and diversity in students' knowledge and thinking in science across the world.

References

- Adams, R., & Wu, M. (Eds.). (2002). *PISA 2000 Technical Report*. Paris: OECD Publications.
- Aikenhead, G. (1996). Border Crossing into the Subculture of Science. *Studies in Science Education*, 27, 1-52.
- 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.
- Baxter, M. J. (1994). *Exploratory Multivariate Analysis in Archeology*. Edinburgh: Edinburgh University Press.
- Bertin, J. (1981). *Graphics and Graphic Information-Processing* (W. J. Berg & P. Scott, Trans.). Berlin/New York: Walter de Gruyter.
- 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), Vancouver.
- Björnsson, J. K., Halldórsson, A. M., & Ólafsson, R. F. (2004). *Stærðfræði við lok grunnskóla. Stutt samantekt helstu niðurstaðna úr PISA 2003 rannsókninni*. Reykjavík: Námsmatsstofnun.
- Cobern, W. W. (1991). *World view theory and science education research*. Manhattan: National Association for Research in Science Teaching.
- Cobern, W. W. (1996). Worldview Theory and Conceptual Change in Science Education. *Science Education*, 80(5), 579-610.

- 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. (1993). *Cluster analysis* (3rd ed.). London: Edward Arnold.
- 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.
- Goldstein, H. (2004a). Education for All: the globalization of learning targets. *Comparative Education*, 40(1), 7-14.
- Goldstein, H. (2004b). International comparisons of student attainment: some issues arising from the PISA study. *Assessment in Education*, 11(3), 319-330.
- Greenfield, P. M. (1997). You Can't Take It With You: Why Ability Assessments Don't Cross Cultures. *American Psychologist*, 52(10), 1115-1124.
- Grisay, A. (2003). Translation procedures in OECD/PISA 2000 international assessment. *Language Testing*, 20(2), 225-240.
- 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.
- Halleuxd, B. (2003). *Anticipating potential translation problems when writing items*. Unpublished manuscript.
- Halliday, M. A. K., & Martin, J. R. (1993). *Writing science: Literacy and discursive power*. Pittsburgh: University of Pittsburg Press.
- Hambleton, R. K. (2002). Adapting Achievement Tests into Multiple Languages for International Assessments. In A. C. Porter & A. Gamoran (Eds.), *Methodological Advances in Cross-National Surveys of Educational Achievement* (pp. 58-79). Washington, DC: National Academy Press.
- Hambleton, R. K., & de Jong, J. H. A. L. (2003). Advances in translating and adapting educational and psychological tests. *Language Testing*, 20(2), 127-134.
- Kearney, M. (1984). *World view*. Novato: Chandler & Sharp Publishers, Inc.

- Keeves, J. P., & Masters, G. N. (1999). Introduction. In J. P. Keeves & G. N. Masters (Eds.), *Advances in Measurement in Educational Research and Assessment* (pp. 1-19). Oxford: Pergamon.
- Kellaghan, T., & Greaney, V. (2001). The globalisation of Assessment in the 20th Century. *Assessment in Education*, 8(1), 87-102.
- Kjærnsli, M. (2003). Achievement in Scientific Literacy in PISA: Conceptual understanding and process skills, *4th Biannual Conference of European Science Education Research Association (ESERA)*. Noordwijkerhout, The Netherlands.
- 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.
- KUF. (1996). *Læreplanverket for den 10-årige grunnskolen*. Oslo: Nasjonalt læremiddelsenter.
- Kupari, P., Välijärvi, J., Linnakylä, P., Reinikainen, P., Brunell, V., Leino, K., Sulkunen, S., Törnroos, 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*. Oslo: Institutt for lærerutdanning og skoleutvikling, Universitetet i Oslo.
- 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. (2004). Nordic PISA 2000 in a Sociocultural Perspective. *Scandinavian Journal of Educational Research*, 48(3), 227-230.
- Lie, S., Linnakylä, P., & Roe, A. (Eds.). (2003). *Northern Lights on PISA: Unity and diversity in the Nordic countries in PISA 2000*. Oslo: 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., 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*. Boston: TIMSS & PIRLS International Study Center, Lynchs School of Education, Boston College.
- McQueen, J., & Mendelovits, J. (2003). PISA reading: cultural equivalence in a cross-cultural study. *Language Testing*, 20(2), 208-224.
- 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. (2001). *Knowledge and Skills for Life. First results from PISA 2000*. 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, J. V. (2002). Mange fejlkilder i Pisa-undersøgelsen. *Folkeskolen*(4).
- Olsen, R. V. (2004). The Search for Descriptions of Students' Thinking and Knowledge: exploring nominal cognitive variables by correspondence and homogeneity analysis. *Scandinavian Journal of Educational Research*, 48(3), 325-341.
- Olsen, R. V. (2005). An exploration of cluster structure in scientific literacy in PISA: Evidence for a Nordic Dimension? *Nordina*, 1(1), 81-94.
- Olsen, R. V., Lie, S., & Turmo, A. (2001). Learning about students' knowledge and thinking in science through large-scale quantitative studies. *European Journal of Psychology of Education*, 16(3), 403-420.
- Porter, A. C., & Gamoran, A. (Eds.). (2002). *Methodological Advances in Cross-National Surveys of Educational Achievement*. Washington, DC: National Academy Press.
- Quigley, C. (1979). *The Evolution of Civilizations* (Reprint edition, originally published 1961 ed.). Indianapolis: Liberty Fund.
- 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.
- SPSS. (2003). *SPSS® Base 12.0 User's Guide*. Chicago: SPSS Inc.
- Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Boston: Allyn and Bacon.
- Tukey, J. W. (1977). *Exploratory data analysis*. Reading: Addison-Wesley.
- Vári, P. (Ed.). (1997). *Are We Similar in Math and Science? A Study of Grade 8 in Nine Central and Eastern European Countries*. Amsterdam: International Association for the Evaluation of Educational Achievement.
- von Kopp, B. (2004). On the Question of Cultural Context as a Factor in International Academic Achievement. *European Education*, 35(4), 70-98.
- 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), Vancouver.
- 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.
- Yore, L. D., Bisanz, G. L., & Hand, B. M. (2003). Examining the literacy component of science literacy: 25 years of language arts and science research. *International Journal of Science Education*, 25(6), 689-725.
- Zabulionis, A. (2001). Similarity of Mathematics and Science Achievement of Various Nations. *Education Policy Analysis Archives*, 9(33).

Appendix 1: Units released from PISA 2000

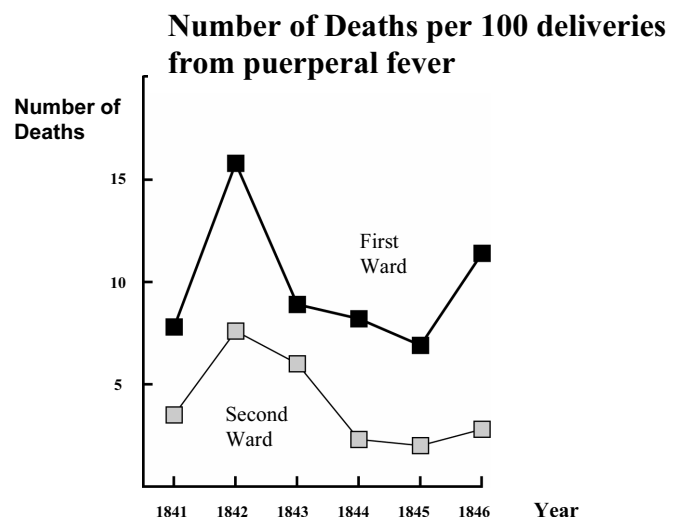
Appendix 1: Units released from PISA 2000

This appendix includes two science units from the 2000 assessment. These units are publicly released from the OECD. Both units are referred to in various places of the dissertation. The framework classifications according to the PISA 2000 framework and the marking rubrics are shown in grey boxes

SEMMELEWEIS' DIARY TEXT 1

'July 1846. Next week I will take up a position as "Herr Doktor" at the First Ward of the maternity clinic of the Vienna General Hospital. I was frightened when I heard about the percentage of patients who die in this clinic. This month not less than 36 of the 208 mothers died there, all from puerperal fever. Giving birth to a child is as dangerous as first-degree pneumonia.'

These lines from the diary of Ignaz Semmelweis (1818-1865) illustrate the devastating effects of puerperal fever, a contagious disease that killed many women after childbirth. Semmelweis collected data about the number of deaths from puerperal fever in both the First and the Second Wards (see diagram).



Diagram

Physicians, among them Semmelweis, were completely in the dark about the cause of puerperal fever. Semmelweis' diary again:

'December 1846. Why do so many women die from this fever after giving birth without any problems? For centuries science has told us that it is an invisible epidemic that kills mothers. Causes may be changes in the air or some extraterrestrial influence or a movement of the earth itself, an earthquake.'

Nowadays not many people would consider extraterrestrial influence or an earthquake as possible causes of fever. We now know it has to do with hygienic conditions. But in the time Semmelweis lived, many people, even scientists, did! However, Semmelweis knew that it was unlikely that fever could be caused by extraterrestrial influence or an earthquake. He pointed at the data he collected (see diagram) and used this to try to persuade his colleagues.

Question 1: SEMMELWEIS' DIARY

Suppose you were Semmelweis. Give a reason (based on the data Semmelweis collected) why puerperal fever is unlikely to be caused by earthquakes.

.....

.....

.....

SEMMELWEIS' DIARY SCORING Q1

QUESTION INTENT: Process: Drawing/evaluating conclusions
 Theme: Human biology
 Area: Science in life and health

Full credit

- Code 21: Refers to the difference between the number of deaths (per 100 deliveries) in both wards
- Due to the fact that the first ward had a high rate of women dying compared to women in the second ward, obviously shows that it had nothing to do with earthquakes
 - Not as many people died in ward 2 so an earthquake couldn't have occurred without causing the same number of deaths in each ward
 - Because the second ward isn't as high, maybe it had something to do with ward 1
 - It is unlikely that earthquakes cause the fever since death rates are so different for the two wards

Partial credit

- Code 11: Refers to the fact that earthquakes don't occur frequently
- It would be unlikely to be caused by earthquakes because earthquakes wouldn't happen all the time
- Code 12: Refers to the fact that earthquakes also influence people outside the wards
- If there were an earthquake, women from outside the hospital would have got puerperal fever as well
 - If an earthquake were the reason, the whole world would get puerperal fever each time an earthquake occurs (not only the wards 1 and 2)
- Code 13: Refers to the thought that when earthquakes occur, men don't get puerperal fever
- If a man were in the hospital and an earthquake came, he didn't get puerperal fever, so earthquakes cannot be the cause
 - Because girls get it and not men
 - Puerperal fever is unlikely to be caused by earthquakes as many women die after giving birth without any problems. Science has told us that it is an invisible epidemic that kills mothers
 - The death is caused by bacteria and the earthquakes cannot influence them
 - Because there aren't any earthquakes by the wards and they still got it [Note: The assumption that there were no earthquakes at that time, isn't correct.]

Continues on next page

Continued from previous page

No credit

- Code 01: States (only) that earthquakes cannot cause the fever
- An earthquake cannot influence a person or make him sick
 - A little shaking cannot be dangerous
- Code 02: States (only) that the fever must have another cause (right or wrong)
- Earthquakes do not let out poison gases. They are caused by the plates of the Earth folding and faulting into each other
 - Because they have nothing to do with each other and it is just superstition
 - An earthquake doesn't have any influence on the pregnancy. The reason was that the doctors were not specialised enough
- Code 03: Answers that are combinations of Codes 01 and 02.
- Code 04: Other incorrect answers
- I think it was a big earthquake that shook a lot
 - In 1843 the deaths decreased at ward 1 and less so at ward 2

SEMMELWEIS' DIARY TEXT 2

Part of the research in the hospital was dissection. The body of a deceased person was cut open to find a cause of death. Semmelweis recorded that the students working on the First ward usually took part in dissections on women who died the previous day, before they examined women who had just given birth. They did not pay much attention to cleaning themselves after the dissections. Some were even proud of the fact that you could tell by their smell that they had been working in the mortuary, as this showed how industrious they were!

One of Semmelweis' friends died after having cut himself during such a dissection. Dissection of his body showed he had the same symptoms as mothers who died from puerperal fever. This gave Semmelweis a new idea.

Question 2: SEMMELWEIS' DIARY

Semmelweis' new idea had to do with the high percentage of women dying in the maternity wards and the students' behaviour.

What was this idea?

- A Having students clean themselves after dissections should lead to a decrease of puerperal fever.
- B Students should not take part in dissections because they may cut themselves.
- C Students smell because they do not clean themselves after a dissection.
- D Students want to show that they are industrious, which makes them careless when they examine the women.

SEMMELWEIS' DIARY SCORING Q2

QUESTION INTENT: Process: Recognising questions
 Theme: Human biology
 Area: Science in life and health

Full credit

Code 1: Having students clean themselves after dissections should lead to a decrease of puerperal fever.

No credit

Code 0: Other

Code 9: Missing

Question 3: **SEMMELWEIS' DIARY**

Semmelweis succeeded in his attempts to reduce the number of deaths due to puerperal fever. But puerperal fever even today remains a disease that is difficult to eliminate.

Fevers that are difficult to cure are still a problem in hospitals. Many routine measures serve to control this problem. Among those measures are washing sheets at high temperatures.

Explain why high temperature (while washing sheets) helps to reduce the risk that patients will contract a fever.

.....
.....

SEMMELWEIS' DIARY SCORING Q3

QUESTION INTENT: Process: Demonstrating knowledge and understanding
 Theme: Human biology
 Area: Science in life and health

Full credit

Code 11: Refers to *killing of bacteria*

- Because with the heat many bacteria will die
- Bacteria will not stand the high temperature
- Bacteria will be burnt by the high temperature
- Bacteria will be cooked [Note: Although "burnt" and "cooked" are not scientifically correct, each of the last two answers as a whole can be regarded as correct.]

Code 12: Refers to *killing* of microorganisms, germs or viruses

- Because high heat kills small organisms which cause disease
- It's too hot for germs to live

Code 13: Refers to the *removal* (not killing) of bacteria

- The bacteria will be gone
- The number of bacteria will decrease
- You wash the bacteria away at high temperatures

Continues on next page

Continued from previous page

Code 14: Refers to the *removal* (not killing) of microorganisms, germs or viruses

- Because you won't have the germ on your body

Code 15: Refers to sterilisation of the sheets

- The sheets will be sterilised

No credit

Code 01: Refers to killing of disease

- Because the hot water temperature kills any disease on the sheets
- The high temperature kills most of the fever on the sheets, leaving less chance of contamination

Code 02: Other incorrect answers

- So they don't get sick from the cold
- Well when you wash something it washes away the germs

Code 99: Missing

Question 4: **SEMMELWEIS' DIARY**

Many diseases may be cured by using antibiotics. However, the success of some antibiotics against puerperal fever has diminished in recent years.

What is the reason for this?

- A Once produced, antibiotics gradually lose their activity.
- B Bacteria become resistant to antibiotics.
- C These antibiotics only help against puerperal fever, but not against other diseases.
- D The need for these antibiotics has been reduced because public health conditions have improved considerably in recent years.

SEMMELWEIS' DIARY SCORING Q4

QUESTION INTENT: Process: Demonstrating knowledge and understanding
 Theme: Biodiversity
 Area: Science in life and health

Full credit

Code 1: Bacteria become resistant to antibiotics.

No credit

Code 0: Other

Code 9: Missing

OZONE TEXT

Read the following section of an article about the ozone layer.

The atmosphere is an ocean of air and a precious natural resource for sustaining life on the Earth. Unfortunately, human activities based on national/personal interests are causing harm to this common resource, notably by depleting the fragile ozone layer, which acts as a protective shield for life on the Earth.

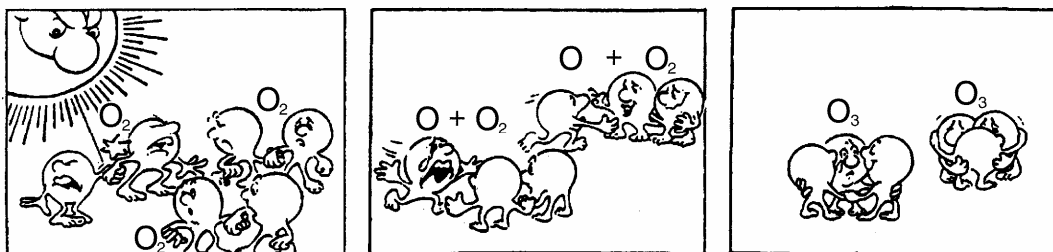
- 5 Ozone molecules consist of three oxygen atoms, as opposed to oxygen molecules which consist of two oxygen atoms. Ozone molecules are exceedingly rare: fewer than ten in every million molecules of air. However, for nearly a billion years, their presence in the atmosphere has played a vital role in safeguarding life on Earth. Depending on where it is located, ozone can either protect or harm life on Earth. The ozone in the troposphere (up to 10 kilometres
- 10 above the Earth's surface) is "bad" ozone which can damage lung tissues and plants. But about 90 percent of ozone found in the stratosphere (between 10 and 40 kilometres above the Earth's surface) is "good" ozone which plays a beneficial role by absorbing dangerous ultraviolet (UV-B) radiation from the Sun.

- 15 Without this beneficial ozone layer, humans would be more susceptible to certain diseases due to the increased incidence of ultra-violet rays from the Sun. In the last decades the amount of ozone has decreased. In 1974 it was hypothesised that chlorofluorocarbons (CFCs) could be a cause for this. Until 1987, scientific assessment of the cause-effect relationship was not convincing enough to implicate CFCs. However, in September 1987, diplomats from around the world met in Montreal (Canada) and agreed to set sharp limits to the use of CFCs.

Question 1: OZONE

S253Q01- 01 11 12 13 21 22 23 31 99

In the text above nothing is mentioned about the way ozone is formed in the atmosphere. In fact each day some ozone is formed and some other ozone disappears. The way ozone is formed is illustrated in the following comic strip.



Suppose you have an uncle who tries to understand the meaning of this strip. However, he did not get any science education at school and he doesn't understand what the author of the strip is explaining. He knows that there are no little fellows in the atmosphere but he wonders what those little fellows in the strip stand for, what those strange notations O, O₂ and O₃ mean and which processes the strip represents. He asks you to explain the strip. Assume that your uncle knows:

- that O is the symbol for oxygen;
- what atoms and molecules are.

Write an explanation of the comic strip for your uncle.

In your explanation, use the words atoms and molecules in the way they are used in lines 5 and 6.

.....

.....

.....

.....

.....

OZONE SCORING Q1

QUESTION INTENT: Process: Communicating
 Theme: Chemical and physical changes
 Area: Science in Earth and environment

Full credit

- Code 31: Gives an answer in which the following three aspects are mentioned:
- First aspect: an oxygen molecule or some oxygen molecules (each consisting of two oxygen atoms) are split into oxygen atoms (picture 1).
 - Second aspect: the splitting (of oxygen molecules) takes place under the influence of sunlight (picture 1).
 - Third aspect: the oxygen atoms combine with other oxygen molecules to form ozone molecules (pictures 2 and 3).

Continues on next page

Continued from previous page

REMARKS ON EACH OF THE THREE ASPECTS

First aspect:

- The splitting should be described using the correct words (see lines 5 and 6) for O (atom or atoms) and O₂ (molecule or molecules).
- If O and/or O₂ have been described only as “particles” or “small parts” no credit should be given for this aspect.

Second aspect:

- The Sun's influence should be related to the splitting of O₂ (an oxygen molecule or oxygen molecules).
- If the Sun's influence is related to the forming of an ozone molecule from an oxygen atom and an oxygen molecule (pictures 2 and 3) no credit should be given for this second aspect.
- Note: Aspects 1 and 2 may typically be given in the one sentence.

Third aspect:

- This aspect should be given credit (one point) if the answer contains any description of an O combining with an O₂.
If the formation of O₃ is described as combining of (three, separate) O atoms this third aspect should not be given credit.
- If O₃ is not described as a molecule or molecules but for example as “a group of atoms” this can be tolerated for the third aspect.

Examples of Code 31:

- When the sun shines on the O₂ molecule the two atoms separate. The two O atoms look for other O₂ molecules to join with. When the O₁ and O₂ join they form O₃ which is ozone.
- The strip illustrates the formation of ozone. If an oxygen molecule is affected by the sun, it breaks into two separate atoms. These separate atoms, O, float around looking for a molecule to link up to; they line up to existing O₂ molecules and form an O₃ molecule, as three atoms are now joined together; O₃ forms Ozone.
- The little guys are O, or oxygen atoms. When two are joined they make O₂ or oxygen molecules. The Sun causes this to decompose into Oxygen again. The O₂ atoms then bond with an O₂ molecule creating O₃ which is ozone. [Note: The answer can be regarded as correct. There is only a slip of the pen (“O₂ atoms” after having mentioned “oxygen atoms” previously).]

Partial credit

Code 21: First and second aspects only correct

- The sun decomposes the oxygen molecules into single atoms. The atoms fuse into groups. The atoms form groups of 3 atoms together.

Code 22: First and third aspects only correct

- Each of the little fellows stand for one atom of oxygen. O is one oxygen atom, O₂ is an oxygen molecule and O₃ is a group of atoms all joined together. The processes shown are one pair of oxygen atoms (O₂) getting split and then each joining with 2 other pairs forming two groups of 3 (O₃).
- The little fellows are oxygen atoms. O₂ means one oxygen molecule (like a pair of little fellows holding hands) and O₃ means three oxygen atoms. The two oxygen atoms of one pair break apart and one joins each of the other pairs and out of the three pairs, two sets of three oxygen molecules (O₃) are formed.

Continues on next page

Continued from previous page

Code 23:	Second and third aspects only correct <ul style="list-style-type: none">• The oxygen is broken up by the sun's radiation. It splits in half. The two sides go and join other oxygen "particles" forming ozone.• Most of the time in pure oxygen (O_2) environments oxygen comes in pairs of 2 so there are 3 pairs of 2. 1 pair is getting too hot and they fly apart going into another pair making O_3 instead of O_2. <i>[Note: Although "one pair is getting too hot" is not a very good description for the sun's influence, credit for the second aspect should be given; the third aspect can also be regarded as correct.]</i>
Code 11:	First aspect only correct <ul style="list-style-type: none">• Oxygen molecules are breaking down. They form O atoms. And sometimes there are ozone molecules. The ozone layer remains the same because new molecules are formed and others die.
Code 12:	Second aspect only correct <ul style="list-style-type: none">• O represents an oxygen molecule, O_2 = oxygen, O_3 = ozone. Sometimes both oxygen molecules, joining each other, are separated by the sun. The single molecules join another pair and form ozone (O_3).
Code 13:	Third aspect only correct <ul style="list-style-type: none">• <u>The 'O' (oxygen) molecules are forced to bond with O_2 (2 x oxygen molecules) to form O_3 (3 x oxygen molecules)</u>, by the heat of the Sun. <i>[Note: The underlined part of the answer shows the third aspect. No credit can be given for the second aspect, because the Sun is not involved in the formation of ozone from $O + O_2$ but only in breaking down bonds in O_2.]</i>
No credit	
Code 01:	None of the three aspects correct <ul style="list-style-type: none">• The sun (ultraviolet rays) burns the ozone layer and at the same time is destroying it as well. Those little men are the ozone layers and they run away from the sun because it is so hot. <i>[Note: No point can be awarded, not even for mentioning something about the Sun's influence.]</i>• The sun is burning the ozone in the first boxes. In the second boxes they are running away with tears in their eyes and in the third box they are cuddling each other with tears in their eyes.• Well uncle Herb it's simple. 'O' is one oxygen particle, the numbers next to 'O' increases the amounts of particles in the group.
Code 99:	Missing

Question 2: **OZONE**

S253Q02

Ozone is also formed during thunderstorms. It causes the typical smell after such a storm. In lines 9–13 the author of the text distinguishes between "bad ozone" and "good ozone".

In terms of the article, is the ozone that is formed during thunderstorms "bad ozone" or "good ozone"?

Choose the answer and the explanation that is supported by the text.

	Bad ozone or good ozone?	Explanation
A	Bad	It is formed during bad weather.
B	Bad	It is formed in the troposphere.
C	Good	It is formed in the stratosphere.
D	Good	It smells good.

OZONE SCORING Q2

QUESTION INTENT: Process: Drawing/evaluating conclusions
 Theme: Atmospheric change
 Area: Science in Earth and environment

Full credit

Code 1: Bad. It is formed in the troposphere.

No credit

Code 0: Other

Code 9: Missing

QUESTION 3: OZONE

S253Q05- 0 1 9

Lines 14 and 15 state: "Without this beneficial ozone layer, humans would be more susceptible to certain diseases due to the increased incidence of ultra-violet rays from the Sun."

Name one of these specific diseases.

.....

OZONE SCORING Q5

QUESTION INTENT: Process: Demonstrating knowledge and understanding
 Theme: Physiological change
 Area: Science in life and health

Full credit

Code 1: Refers to skin cancer or cataracts

- Skin cancer.
- Melonoma *[Note: This answer can be regarded as correct, despite the fact it has a spelling mistake.]*

No credit

Code 0: Refers to other specific type of cancer

- Lung cancer

OR

Refers only to cancer

- Cancer

OR

Other incorrect answers

Code 9: Missing

Question 4: **OZONE**

S270Q03

At the end of the text, an international meeting in Montreal is mentioned. At that meeting lots of questions in relation to the possible depletion of the ozone layer were discussed. Two of those questions are given in the table below.

Which of the questions below can be answered by scientific research?

Circle Yes or No for each.

Question:	Answerable by scientific research?
Should the scientific uncertainties about the influence of CFCs on the ozone layer be a reason for governments to take no action?	Yes / No
What would the concentration of CFCs be in the atmosphere in the year 2002 if the release of CFCs into the atmosphere takes place at the same rate as it does now?	Yes / No

OZONE SCORING Q3

QUESTION INTENT: Process: Recognising questions
 Theme: Atmospheric change
 Area: Science in Earth and environment

Full credit

Code 1: No and Yes, in that order

No credit

Code 0: Other

