PISA 2006 Science testen og danske elevers naturfaglige formåen

Hvad siger PISA science om danske elevers naturfaglige kompetencer – og hvad siger den ikke?

Rapport nummer tre fra Validering Af PISA projektet

af
Lars Brian Krogh og Jens Dolin

18. maj 2011

1. Indledning

Indeværende rapport er den tredje og afsluttende fra forskningsprojektet Validering Af PISA Science (VAP). PISA er en forkortelse for OECD's *Programme for International Student Assessment.* PISA havde sin første testrunde i 2000, og VAP-projektet analyserer science-delen af den tredje runde, som blev gennemført i 2006. Det overordnede formål med VAP-projektet er at placere PISA 2006 science testen i en dansk kontekst, dvs. i forhold til de danske mål for naturfagsundervisningen.

Bevæggrunden for at starte projektet er den enorme opmærksomhed som PISA-resultaterne har fået ved deres offentliggørelse, og den efterfølgende indflydelse hele PISA-apparatet har på det danske uddannelsessystem. Det er derfor naturligt at spørge om PISA-testsystemet berettiger til en så fremtrædende rolle i den uddannelsespolitiske debat, og om der er videnskabelig belæg for de dragne konsekvenser af undersøgelsesresultaterne. Centralt heri ligger spørgsmålet om PISA-testens validitet – måler testen det den intenderer? Lige så vigtigt er spørgsmålet om testens relevans - måler testen det som anses for centralt i det danske uddannelsessystem?

Sådanne spørgsmål rammer ind i en række uddannelsesforhold med meget høj grad af kompleksitet. Der er tale om at vurdere uddannelsesmæssige mål, undervisningsmæssig praksis relateret til læringsteoretiske overvejelser, testteoretiske grundantagelser og testmetodiske forhold – for bare at nævne de mest centrale områder!
For at kunne navigere i dette felt har det været nødvendigt at opstille en række sammenhængende forskningsspørgsmål, som er gennemgået nærmere i afsnit 5, og som har kunnet opdele projektet i en række hver for sig operationaliserbare dele med en logisk progression. De enkelte dele har meget forskellig forskningsdesign, og der har undervejs været behov for metodiske tilpasninger for at sikre den forskningsmæssige lødighed af projektet. Denne rapport lægger derfor relativ stor vægt på beskrivelsen af de anvendte forskningsmetoder, idet udvikling af disse har udgjort en stor del af projektet, og et grundigt kendskab til metoderne er nødvendig for at kunne vurdere de opnåede resultater.

Rapporten indledes med en gennemgang af de danske PISA-præstationer i science og de uddannelsesmæssige tiltag de har forårsaget. Disse perspektiveres med reference til den internationale debat om PISA. På baggrund heraf opridses VAP-projektets teoretiske ståsted og projektets grundlæggende forskningsspørgsmål. Resultaterne fra de to første rapporter refereres som afsæt for den egentlige feltvalidering, hvis metodiske grundlag og vigtigste resultater udgør resten af denne rapport.

Rapporten blev i en tidligere version afleveret til Skolestyrelsen december 2009. Indeværende udgave er færdigskrevet januar 2011, dvs. efter PISA2009-resultaterne er offentliggjort, men uden at de er medtaget i rapporten.

2. English summary

This is the third and last report from the VAP-project (Validation of PISA Science).
The overall aim of the project is to validate the PISA 2006 science test in a Danish context. The first step of the VAP-project was to compare the PISA Science Framework (OECD 2006) with the general understanding of scientific literacy within the science education community and with the general aims of the Danish science education in the compulsory school (the "Common Goals", "Fælles Mål"). The first VAP report gives the results of this comparison (Dolin, Busch and Krogh 2006).
The next step examined the PISA test format in relation to the assessment culture in the Danish compulsory school. Interviews with science teachers and a nationwide survey give a basis for assessing the importance of the differences between the PISA test system and the everyday assessment and test practice in science education in Denmark. The second report gives the results of this comparison (Dolin and Krogh 2008).

These two studies are the foundation of the real validation study: Are the PISA science test results an adequate expression of the Danish students' competencies in science? 120 students were selected among those who went through the PISA 2006 test. Some weeks later these students were subjected to a specially designed VAP-test. Special educated assistants (student teachers) assessed the students (one at a time) about two selected items, which they had been exposed to in the PISA test. The conversations involved relevant artefacts, took 30 minutes, and followed a rather tight scheme for reliability reasons. All the talks were videotaped and scored afterwards. The results from this oral, socio culturally oriented assessment are compared with the paper and pencil PISA test. In excess of assessing the students' knowledge and competencies related to the PISA item content, the conversations also included the subject specific knowledge demanded by the Danish curriculum, but still within the item domain. The students were finally put together two and two to do a practical lab work (a carrying into practice one of the PISA items). The performance was videotaped and scored.

Before we present a summary of this third report, we will bring the results from the first two reports. In Dolin and Krogh 2010 we have given a more expanded version of the comparison between the Danish course of study and the PISA Science Framework together with an analysis of the impact from PISA on the Danish educational system.

2.1 The results of the comparative analysis (VAP Report 1)

The PISA Science Framework and the Danish course of study (Common Goals) are two very different systems of describing the desired outcome from science education. They emphasize different aspects and they use different terminologies - which complicates comparison. But the report concludes:

- *The literacy definition* in the PISA2006 Science Framework is quite different from the literacy definition in 2003, bringing the 2006 definition in accordance with the international used terminology. The items are, apart from the attitudinal questions, completely alike in the 2003 and the 2006 tests.
- *The contexts* for assessment items in the Framework are in good alignment with the recommended contexts in Common Goals.
- The PISA *concept of competency* is not in accordance with the way competency is used in the Danish educational system.

- The correspondence within *the affective domain* is not very good. Both weights interest in science. But Common Goals do not mention support for scientific inquiry (which is a part of the PISA attitudinal area) but gives instead high priority to responsibility towards resources and environments (which is a part of the PISA attitudinal area in the Framework, but not in the items).
- *The knowledge of science* categories has different weight in the PISA science test and in the Danish compulsory school. PISA gives more importance to Biology and Geosciences, while the Danish school gives higher weight to Physics and Chemistry.
- The alignment in *content* varies from subject to subject. For Biology the correspondence is nearly complete, in the Geosciences the correspondence is very high, but in Physics/Chemistry it is very difficult to compare, due to different description formats.
- The PISA *knowledge about science* has a relatively weak position in the Danish curriculum. Aspects of knowledge about science is part of the described working methods and ways of thinking in the Common Goals for the different subjects, but they do not form a systematic, self-contained system.
- *The practical dimension*, which is a major part of Common Goals, is not included in the PISA Framework, although the theoretical aspects are.
- *In general*, the Danish course of study is kept in broader, more general terms than the PISA Framework. Phrases like 'emphasis on understanding of connections', 'the essential points', 'involve the perspectives for' etc. are quite widespread in the Common Goals. This promotes teaching with emphasis on processes, overall views, problem orientation etc. where the teachers adapt the Common Goal curriculum to the concrete class and students. The result is a more heterogeneous science teaching, without concern for standards.

2.2 The assessment culture in the Danish lower secondary school (VAP Report 2)

The aim of this study was twofold:

1. to investigate whether assessment practice in the science subjects in the Danish grade 8 and 9 (year 14-16) is in accordance with the test format used in the PISA 2006 Science test
2. to characterise the present assessment practice in the science subjects with a view to qualify a development of this practice.

Based on two focus group interviews with science teachers and a nationwide, representative survey involving 1159 teachers teaching science in year 8 and 9 in 2005/2006, the short answer to the first question is YES. The assessment practice in science in the Danish lower secondary school is to a large degree in accordance with the PISA test format. Individual, written tests are the most used test format in the science subjects and 75 % of the teachers indicate to test the PISA competencies regularly. Teachers typically include PISA relevant contexts in their assessments and the teachers seem to weight the balance between open/closed answer formats in accordance to the PISA weighting. The Danish science teachers also give a moderate emphasis to the students' use of correct technical language – though a heavier weight on this aspect would be more in accordance with the relatively rigid demand on this in the PISA score handbook.
In the light of these results, it is thought-provoking that the teachers only express moderate confidence in having prepared the students adequately for the PISA test. An explanation could the lack of evidence in our study for any 'teaching to the test' or rehearsal of PISA-like test items (which the concealment of most of the PISA items makes impossible). The teachers do not teach in order to enhance a PISA performance. They teach for more varied goals and they simply use a

much broader palette of assessment tools than the PISA formats – things this study clearly shows. PISA has a relatively little impact on the science teaching in Denmark - or had in 2005/2006, a situation that seems to change thanks to the increasing attention politicians put to PISA ranking.

We have found the typical science teacher to hold a broad repertoire of active assessment forms, with a combination of individual written tests and not-binding talks with the whole class as the dominant form.
Most of the teachers express a student oriented aim with their assessments. They have a marked focus on learning and learning potential in their assessments with as well formative as summative approaches, and they evaluate both during and after the teaching sequences. They often make the criteria for the assessment open for the students, and the students always receive feedback on the tests and assessments.
The study did not uncover any patterns in assessment and evaluation practice due to gender or science subject. Two third of the teachers answer 'no' to whether the school has 'a common attitude toward assessment' which indicates that an assessment culture, if any, is more a national phenomenon than a local.

2.3 Does PISA paint a valid picture of Danish students' Scientific Literacy – Danish students' performance in a more socio-cultural test-format? (VAP Report 3)

The previously presented VAP-investigations demonstrated a considerable overlap in the intentions of the scientific framework of PISA and the goal descriptions of science teaching in Denmark (the so called Common Goals) (VAP Report 1). Moreover, the Danish pupils are somewhat used to PISA-like tests (VAP Report 2). Thus at the level of intention, PISA appears to be *relevant* for Denmark – and at the level of pupils the test format appears to be *familiar.*

With this third part of the VAP project we wish to address the concern, that PISA's results may be shaped by the chosen test-paradigm and the test-formats applied. Using a more socio-cultural test-format we retest students in original PISA 2006 Science items to see whether PISA produces a complete, un-contested, and (in contemporary senses of the word) valid representation of Danish students' competencies within these domains. We consider research based and reassuring answers to this question crucial if PISA should maintain its present position as "prime mover" for the development of science teaching in lower secondary school.

At our starting point, we did have some concerns about the interplay of the restricted post-positivistic test format of PISA and the science competences Danish pupils acquire in the much broader socio-cultural learning paradigm of their schooling. A central principle of modern evaluation theory is precisely that assessments should be designed in accordance with the concepts of learning that have guided the learning process:

Theories of learning have implications for assessment design…. Constructivist models of learning, which see learning as a process of personal knowledge construction and meaning making, describe a complex and diverse process and therefore require assessment to be diverse, examining in more depth the quality of students' learning and understanding. While for example standardized MC or short answer tests are efficient at sampling the acquisition of specific knowledge gained from teachers, more intense, even interactive assessment (e.g. essays, performance assessments, small-

group tasks and projects) is needed to assess the processes of learning and understanding, and to encourage a deeper level of learning.
(Gipps, 1999 p.375).

This quotation clarifies that the choice of assessment format is not purely a technical matter, but in fact decisive for the extent to which the assessment can capture the principal learning processes and values of the teaching and the school system.

Due to a concern that (aspects of) the Danish pupils' science competences could easily remain invisible or underexposed as a consequence of the PISA format we composed a rather complex assessment design to answer the following research questions:

> Q3a: How much will students' performance on PISA-measures change if they are (re)tested within a socio-culturally oriented test format using dialogue, mediating artifacts, and practical enacting.
> Q3b: What picture emerges of the pupils' strengths and weaknesses in 'the expanded window of attention', which the alternative test format constitutes? Does PISA provide a valid picture of Danish 15-year olds' abilities in the tested domains?

Originally, our emphasis was to provide an enhanced and more detailed picture of Danish students' capabilities – while assessment theoretical aspects were of secondary importance and primarily included to set the framework for our investigation. Yet, during the analytical work it became more and more clear that the research-based answers to Q3a and Q3b also contribute with fundamental knowledge about the strengths and weaknesses of the involved assessment paradigms – and point, in particular, to a number of limitations in PISA. Indeed, the overall final picture disputes important aspects of the validity of the PISA test for the Danish context. Moreover, the implemented form of PISA may be less relevant than it was possible to see originally where we, in the first report, compared the intentions that inform the scientific framework of PISA with the Danish school curriculum (the Common Goals). The test methodological problematizations of PISA are gathered at the end of this section.

The basis for our conclusions

120 pupils, who completed the PISA 2006 test, were re-examined in three items: a biological (the PISA 2006 item no S478, 'antibiotics'), a geographical/physical (the PISA 2006 item no S465, 'different climates') and an experimental item (the PISA 2006 item no S477, which imitates a laboratory work, we allowed students to enact in practice). The VAP-assessment also examined possible effects of the re-examination together with the impact of superficial changes in the design of the individual tasks. In accordance with the socio-cultural orientation of VAP, the re-examination was carried out by means of interviews/conversation and a practical task (performed in pairs) – both allowed the pupils to make use of relevant artefacts etc. The examination inquired into specific PISA items as well as broader questions related to Common Goals within the domain. The latter was to ensure VAP validity in relation to the requirements of the aims of the "Folkeskole" and Danish science teaching.
All VAP sessions were recorded on video and the pupils' written responses from the re-examination were collected. To answer research question 3a video-recordings were analyzed using PISA's scoring criteria. This allowed us to quantify our rich data, and to compare the raw and Rasch VAP scores with students' original PISA scores.

To answer research question 3b VAP has conducted qualitative and in-depth subject matter didactic analyses of the pupils' language usage, ability to explain and argue etc. seen from a socio-cultural perspective.
Furthermore, students' video-taped performances were analyzed and scored holistically according to standards of the" Folkeskole".
For all analyses inter-rater reliability have been ensured.

The performance of Danish pupils within the altered test format

Based on comparative analyses of the pupils' performances in the VAP-developed assessment and their performances in the PISA test, we conclude that:

> **When compared directly and following the scoring criteria of PISA, the pupils' performance increase by about 25% when they are allowed to exercise their knowledge in a socio-culturally oriented test format.**

Here, the point of the VAP estimation is that in the interview the pupils "only" need to show that they are capable of fulfilling PISA's criteria for a correct answer by, for instance, independently referring to relevant information or by choosing the correct MC-possibility, given they are presented during the course. *Thus, this score does not necessarily indicate that the pupil can command a coherent model, explanation or argumentation, but simply that the pupil knows enough to provide the correct answer in PISA.* In VAP's socio-cultural re-examination students' performance was significantly better on six of ten PISA as estimated from their 'raw' scores. All in all, for the questions we found an average improvement of 26%. One explanation to this improvement may lie in the various possibilities available to the pupils to actualize their knowledge in this richer test format and more authentic test situation.

The strength of the PISA project is its comparative measurements of pupil performance across countries. The cornerstone of this relative ranking is an empirical Rasch-scale where the OECD average corresponds to a set value of 500. In order to estimate the impact of the altered test format on this scale and according to the standards set by PISA, we have compared the pupils' VAP performances with their PISA performances by means of Rasch modeling. This analysis also serves as a control of the results from our previous analysis based on raw scores, as we made use of a triangulation of the method of calculation.

The unidimensionality of our set of PISA-items in-VAP-format has been checked, and though it was not perfect it was acceptable for at least 8 of the items. VAP item difficulties were found to overlap considerably with the original PISA item difficulties, which allowed us to compare Rasch-scores within the two setups. In concordance with the previous less sophisticated analysis we found:

> **In a socio-culturally oriented test-format pupils improve their performance by 25% on a Rasch-scale. In absolute numbers the improvement is estimated 125 points.**

Such an increase corresponds to a change in the Danish ranking in the PISA 2006-science from 496 to 621 – which is notably higher than Finland's test winning 563 Rasch points.
Naturally, this does not suggest that a more socio-culturally oriented PISA test format would lead to a corresponding increase in Denmark's ranking on the international comparative scale, since the other countries are also expected to benefit from the altered test paradigm. Yet, as the science

teaching in Denmark is normally considered to be more dialogic compared to most of the participating countries and as many countries have a test system that is closer to that of PISA than Denmark, a consequence of the PISA test format is probably that Denmark's performance appears relatively poorer.

That the test format has a certain impact on the result is not particularly remarkable - this is probably intuitively expected by many. What is remarkable is that VAP is capable of calculating this impact quantitatively, and that the impact proves to be of such significance! The pupils' performance in PISA does thus not reflect their full capacity in the subject matter, but is largely influenced by their ability to engage with the particular test-format and unfold their knowledge within its limitations.

By means of the developed research design it is thus possible to give a precise answer to the presented research question Q3a. Yet, the data of the VAP assessment is significantly richer than the corresponding PISA data. Therefore we are able to analyse the capability of Danish pupils in the subject areas tested in VAP, i.e. climate differences, antibiotics and scientific work methods and way of thinking, in relation to the demands of The Aims of the "Folkeskole". This is part of our 'expanded window of attention', and through a number of analyses of the collected data we are able to answer research question Q3b: What are the pupils actually capable of? Contrary to our expectations that a socio-cultural test format would better unfold the Danish pupils' competences in science, a rather discouraging picture emerges from the VAP analyses:

> **We find a significant gap between the Danish pupils' actual capability within a range of scientific topics and the conceptual and procedural requirements stated in the aims of the Danish compulsory school.**

One aspect to mention, among many others, is that the tested pupils had difficulties expressing themselves about relevant biological concepts (bacteria, virus, immune defence, resistance etc.), and their overall understanding of the tested areas in biology is estimated to correspond to 20-35% of full understanding as it is defined in the aims of the Danish compulsory school (the "Folkeskole"). The understanding of geographical concepts such as climate, precipitation and seasons were equally poor. In this field, the students' procedural knowledge within central discipline areas (e.g. the circulation of water and formation of precipitation) are estimated to correspond to about 20-40% of full understanding as described in the Common Goals.
The poorest result was the pupils' knowledge about (aspects of) scientific inquiry and scientific explanations. Less than 5% of the pupils were able to express themselves using relevant technical terms about the nature of scientific experiments and investigations.

These discouraging results emerged from analysis by means of a science didactic system of concepts, including 'words of science' (Wellington & Osborne 2001), 'entities and explanatory stories' (Ogborn et al 1996), 'argumentation' (Osborne 2005), and 'the use of artefacts' (Säljö 2003, Schoultz et al 2001b). Among the results of the socio-culturally informed analysis we find reason to underline that:

> **Generally, pupils lack knowledge of scientific vocabulary apart from the simplest, as well as capacity to use more abstract scientific terms from the curriculum in dialogue. Their appropriation of subject matter entities tend to be inadequate to a degree where it**

impedes their explanation of phenomena and cause-effect relationships within tested domains.

The pupils are able to employ simple technical terms and everyday language in their conversation, but they have little knowledge of more abstract concepts and processes. Often, the concepts and models of explanations on which the pupils must build their accounts are not constructed to a degree that enables them to independently generate satisfactory explanations. It is easier for them to select/deselect among already articulated suggestions. In this context, it was characteristic that:

Pupils' capacities to argue in a scientific manner are modest.

The lacking ability to generate tenable accounts affects the ability to form scientific arguments and technical language and argumentation based on scientific evidence (support and warrant) are not employed naturally by the pupils.
The socio-cultural test format of VAP involved artefacts in the test situation. It appeared that:

Pupils who independently use artefacts to scaffold their explanations perform the better.

Yet, it was also clear that:

The majority of the pupils are not confident with the employed artefacts.

The analysis does not unequivocally uncover what is cause and effect in the connection between using artefacts and proficiency in science. Yet, a reasonable assumption is that only the technically strong pupils have sufficient knowledge of what and how the artefact represent to enable them to use the artefact as tools for knowledge construction and explanation.

The validity of PISA – based on the empirical data of the VAP-project

The PISA test is described in detail and methodically stringently assessed from its test theoretical standpoint. Yet, when diverging from this post-positivistic paradigm to apply the VAP project's more socio-cultural approach, one must query the validity of the PISA's results regarding a number of points – firstly for Denmark but presumably also in a more general context. The VAP analysis of pupils' abilities in an alternative test and scoring format has for instance clarified how test results are generally formed by a wide range of aspects that has little to do with the pupils' proficiency within the given subject matter. Thus, the VAP project documents that by operating within a socio-cultural assessment paradigm and with a test format that is closer to the everyday teaching, it is possible to affect "the test result" by 25% - even measured on the same questions and using the same criteria. From a post-positivistic approach one would probably say that the pupils' improved performance is due to the added assistance provided by interviewers and artefacts, but in a socio-cultural context one would say that the communicative situation has been altered which leads pupils to activate and construct their knowledge differently. Students' capacities are not fixed, but dynamical.
With the demonstration that the pupils' assessed performance is this sensitive to the choice of test format, VAP sets the alarm bells ringing that conclusions based on this type of test require utmost caution:

The choice of assessment paradigm and test format greatly influences the test result. The PISA test results are thus relative and cannot be considered indicative of pupils' capacities in any absolute sense. This knowledge should be reflected in the conclusions and consequences of PISA.

We have argued that the validity of any assessment is enhanced when the chosen assessment corresponds with the paradigm and values that form the foundation for the pupils learning processes. In these respects, PISA does not appear to be the most obvious choice if the goal is a comprehensive understanding of the knowledge Danish pupils acquire during their schooling:

The PISA results are inadequate measures of students' performance in relation to the standards of the Danish school, the "Folkeskole".

Most notably, we have found noteworthy differences in the pupils' knowledge as it was reflected in the 'expanded window' of the VAP test and the restricted window with the specific items in the PISA test. The VAP test found that only 20% had an understanding of the effect of antibiotics on bacteria and viruses that was equivalent to that of the aims of the "Folkeskole", while 45% of the sample pupils provided the correct answer in the PISA 2006. In the VAP test, the pupils' overall understanding of central biological concepts associated with the use of antibiotics (again assessed in accordance with the descriptions in the aims of the "Folkeskole") was estimated to be about 35% of full understanding, while the PISA item that tested the same concepts were solved by 75% of the pupils. In the assignment on climate differences only 42% of the boys and 7% of the girls taking the VAP test were able to account for fundamental conditions associated with climate variation. Nevertheless, 90% of the pupils were able to select the correct one among PISA's possible answers to the corresponding PISA-question.

Put differently, we see that the pupils consistently score higher in the PISA test than when they are assessed according to the aims of the "Folkeskole". PISA thus poorly reflects the fulfilment of objectives set in the aims of the "Folkeskole". It should be mentioned that PISA has never given the impression that it could or would capture the fulfilment of objectives in the national curricula. Nevertheless, the results of the survey have often been presented in the public as if the survey does have predictive value regarding this issue.

We have been wondering about the reason for this considerable discrepancy as we found a reasonable concordance between the intentions of the Danish objectives for sciences and in PISAs Scientific Literacy Framework earlier in our first VAP report. We believe to be able to see a considerable move in opposite directions when the intentions are implemented - into teaching in school and into the PISA tests, respectively. Scientific Literacy is not simply Scientific Literacy, and according to Bybee (Bybee, 1997) it can be understood at different taxonomic levels. It is our impression that the Danish objectives regarding teaching are operationalized as *Conceptual Scientific Literacy,* while the PISA tests measure on a taxonomically lower *Functional Scientific Literacy* (as well as logical-rational thinking). Partly for that reason, PISA does not test the conceptual understanding and procedural knowledge that the pupils are expected to acquire, according to the aim of the "Folkeskole". Yet, a more thorough examination of the items is necessary in order to substantiate this hypothesis.

The issue truly became visible in the 'expanded window of attention' in the VAP test, which also disclosed how PISA often misleadingly reward pupils for a correct choice in the MC-option in

situations where their argumentation and underlying explanations are clearly incorrect. As previously mentioned, the analysis of a random sample of pupil responses in the assignment about climate showed that 80% of the pupils were able to select the correct explanation, while 60% of the pupils were no where near at producing a satisfactory explanation. On the contrary, the contributions within this group would normally be categorised as misconceptions/unauthorised everyday life ideas. Thereby PISA down plays the pupils' actual problems of understanding and explaining scientific phenomena and issues. Moreover, PISA does not capture the fundamental problems of argumentation, using artefacts and reflecting on the methods of the field etc. which VAP uncovered among the Danish pupils. From that perspective it seems reasonable to conclude that:

> **The test format of PISA does not capture essential problems regarding the pupils' mastering of scientific language, their capacity to explain and argue on a scientific basis, use artefacts and reflect on the methods of the field etc.**

At a concrete level this unfortunate situation can be assigned to inadequacies of the PISA- items as well as the response formats and scoring criteria. At a more general level the situation can also be seen as an expression that:

> **Due to the paradigmatic standpoint and the choice of test format, PISA is largely blind to aspects of the socio-cultural paradigm that, from an ideal perspective, support the learning processes in Danish science teaching.**

Furthermore, VAP has pointed to another aspect of the use of the test, which also touches upon core values of the Danish compulsory school: The school must provide equal opportunities for the children. Thus, an ideal test would be neutral in the sense that it does not favour one group of pupils over another. Yet, the direct comparison of the pupils' capacity within the framework of PISA and VAP, respectively, suggests that:

> **The choice of test format significantly favours/discriminates certain groups of pupils. A socio-culturally oriented test format is a benefit for the low achieving pupils. Thereby, the PISA test format becomes a relative choice of benefit to the higher achieving pupils.**

Finally, there is reason to mention the main result of VAP's inquiry into how superficial changes in task design (i.e. the quantity of text, text sequencing, use of figures etc) influence test performance. Here we found that:

> **Superficial changes in the task/item-design, including linguistic phrasing, significantly influence the test-performance.**

In obscure ways, the pupils' performance seems to be affected by minor changes in the superficial design of the PISA assignments. Again this demonstrates that PISA performance is not merely a measure of scientific understanding – it confounds task-decoding and arbitrary task/item characteristics.

To sum up, it is fair to say that the VAP investigation has indeed unfolded the notion that answers are affected by the way you ask. By asking the questions in a different and elaborating communicative format than PISA, we have retrieved very different answers; answers that have

exposed significant weaknesses in the pupils' capacity, which PISA has proven blind to – and what is more, we would argue that these answers are even more adequate, valid, and interesting for the Danish context. Consequently, this triggers a fundamental query about the capability of the PISA test to validly capture Danish pupils' abilities – with its present paradigmatic values, choice of test format, scoring procedures and concrete operationalization of assignments. The investigations have made it clear that PISA is not a neutral test tool of universal scientific literacy. On the contrary, it is inflicted by (educational) policy intentions and a post-positivistic test paradigm, which in many ways conflicts with dominating Danish and international educational paradigms and values. It is not designed to measure central aspects of the demands that are formally put on Danish pupils.

2.4 Perspectives

Our curiosity whether the validity of the PISA test justifies its importance in education policy was the driving force behind the VAP project. This curiosity has indeed been gratified! The PISA test does not accurately reflect the knowledge and capacities within science of Danish 15-year olds. Compared to the developed assignments, the overall PISA test concept cannot measure the outlined scientific literacy objectives and certainly not the Danish aims of the "Folkeskole". Particularly for the Danish situation we have shown how crucial competences – and the lack of these – are not caught in the PISA test.
In part, this strongly questions the international benchmarking and ranking which is one of the stated targets and primary result of the PISA project. Yet, it also calls for greater reservation concerning the role PISA has come to play in education policy. PISA provides us with very general results and can point to some relevant correlations, but as we have seen the results are fragile and provide little insight into the reasons for the results. The span from statistical measures of detached student performance to inferences about what goes on in the classroom is simply too wide, and PISA gives us no possibility to correlate the pupils' PISA performance with the teaching that underlie the test results.

Statistical processing of aggregated data typically covers significant qualitative aspects and the outcome may thereby be misleading results. One cannot apply research results at a level that differs from the level at which the research draws its conclusions. The PISA test addresses a very general level of conclusion and offers knowledge about the educational system in general, but we cannot use it in relation to what happens or should happen in the individual classroom. When PISA is used to inform the plan for the actual teaching, we risk promoting international unification at the expense of a number of valuable Danish educational values.
This is not to say that there is no need for further development of science teaching in Denmark. On the contrary, our research shows that we have plenty of work to do. PISA data may serve as a valuable starting point, but it should be combined with classroom oriented, didactical research that the teachers can apply in their daily teaching.
Yet, an important point here is that the PISA results themselves do not provide a valid basis for sweeping reforms that affect the level of teaching. Consequently, we hope that the results of the VAP project can contribute to a more balanced interpretation of the PISA test results.

3. PISA og det danske uddannelsessystem – specielt hvad angår naturfag

VAP-projektet er et forsøg på at afdække i hvilken udstrækning PISA siger noget relevant og dækkende om danske elevers formåen indenfor naturfag. I den forstand er VAP et forsøg på at kvalificere brugen af PISA. I dette afsnit vil se nærmere på, hvorledes PISA hidtil er blevet brugt i Danmark, og på de konsekvenser PISA på forskellig vis har haft.

Danmark har nu deltaget i de fire første runder af PISA, som alle er analyseret og rapporteret (Andersen m.fl. 2001; Mejding 2004; Egelund 2007, 2010). Forud for den første PISA-test i 2000 blev den danske folkeskole opfattet som værende noget nær "verdens bedste": inkluderende mere end sorterende, med mindst samme vægt på brede dannelsesmål som på snævre faglige mål - og kendetegnet ved at det er lysten, dialogen og fællesskabet, der driver værket. Som bl.a. PISA-undersøgelserne senere skulle afdække, var det en skole som usædvanligt mange elever befandt sig godt i, og som afleverede elever med en i international sammenhæng utypisk stor lyst til at lære mere (Andersen et al 2001, s. 9). Det var samtidig en skole som havde bred opbakning såvel politisk som befolkningsmæssigt: "… der er generelt offentlig tilfredshed med kvaliteten af Folkeskolen" (Egelund 2005, s. 207, egen oversættelse). I den mellemliggende periode er der imidlertid sket radikale forandringer i folkeskolens virke og værdigrundlag, såvel som i opfattelsen af og forventningerne til folkeskolen. En del af disse forandringer henføres med større eller mindre rimelighed til PISA – en diskussion som uddybes nedenfor.

Et blik på den internationale modtagelse og brug af PISA mere end antyder, at PISA har haft størst opmærksomhed og flest implikationer i de tilfælde, hvor et land har klaret sig dårligt i forhold til "sammenlignelige" lande, eller har underpræsteret i forhold til den nationale selvforståelse for uddannelsesmæssig formåen. Det gælder fx Tyskland i kølvandet på PISA Science 2000 og Norge efter PISA Science 2003. Politisk handling synes først og fremmest knyttet til *dårlige* testresultater, for "Test er farlige – for ministre", som den engelske matematik-didaktikprofessor og testudvikler Margaret Brown har udtalt (Månedsmagasinet Undervisere, (04-01-2007)). De danske resultater har været blandede, idet de 15årige danske elever konsistent har ligget over gennemsnittet i matematik, hele tiden har ligget i midterfeltet, hvad angår læsefærdigheder og de første gange har ligget under OECD-gennemsnittet i den naturfaglige test. Udtrykt ved hjælp af den anvendte Rasch-skala, hvor tallet 500 modsvarer det internationale gennemsnit, har Danmark på naturfagsområdet præsteret 481, 475, 496 og 499 points i de fire test-runder (2000, 2003, 2006 og 2009). De sidste scorer placerer Danmark i det relativt snævre midterfelt for PISA Science testningen. Ud fra logikken om "farlige tests" ville man forvente størst uddannelsespolitisk konsekvens af PISA på det danske naturfagsområde. Udover naturfagsområdet forekommer de mest overraskende og ubehagelige resultater at vedrøre centrale aspekter af den hidtidige selvforståelse: PISAs dokumentation af, at den danske folkeskole *ikke* formår at inkludere og danne alle, men efterlader nydanske og socialt dårligere stillede elever i Danmark blandt de dårligste i hele OECD-området. Også på dette ømfindtlige område vil man naturligt forvente konsekvenser i kølvandet på PISA. Dårlige test-resultater er imidlertid ikke kun "farlige" ved deres krav om politisk handlekraft, men kan i nogle sammenhænge også tjene som belejlig legitimering af initiativer og omkalfatringer, selv på områder hvor PISA ikke er leveringsdygtig i belæg. På forhånd synes det derfor relevant at skelne mellem direkte ("*PISA-substantierede"*), indirekte ("PISA-initierede") og *PISA-legitimerede* virkninger. En sådan skelnen forudsætter dog, at man ser virkningerne i deres danske kontekst og historik.

Inden vi går over til en sådan redegørelse, vil vi introducere yderligere et par optikker, som kan tilføre diskussionen perspektiv. For det første har OECD ikke lagt skjul på, at PISA er tænkt som et uddannelsespolitisk beslutningsredskab: *"Its policy orientation, with design and reporting methods determined by the need of governments to draw policy lessons"* (http://www.pisa.oecd.org/dataoecd/51/27/37474503.pdf). OECD fremhæver med vekslende vægt to hensigtsmæssige måder at bruge PISA på: én som udnytter det internationale og komparative – hhv. én som udnytter, at PISAs regelmæssige monitorering gør det muligt at gennemføre og effekt-evaluere nationale tiltag indenfor de berørte dele af uddannelsessystemet. Den internationale sammenligning af testresultater gør det muligt at udskille og lære af lande med Best-Practice, ligesom testresultaterne kan sammenkøres med OECD's øvrige databaser, såsom de årlige rapporter "Education at a glance", og give et indtryk af "Value-for-money". Begge anvendelser forudsætter, at PISAs evalueringskriterier gøres til værdier for praksis og skolesystemet. Dette leder os naturligt over i den anden perspektiverende optik, nemlig det som P. Dahler-Larsen et al (Dahler-Larsen & Krogstrup, 2001) har kaldt *evalueringers konstitutive virkninger.* Hermed mener forfatterne, at evalueringer medvirker til at *"konstituere en praksis som ikke er tilsigtet, og som derfor heller ikke er beskrevet i det officielle formål med evalueringen."* (p.232). Blandt de eksempler på konstitutive virkninger, som forfatterne omtaler, er netop, at *Evalueringskriterierne bliver værdier i sig selv.* Derudover betoner Dahler-Larsen et al, at *Evalueringer sætter deres egne tidsmæssige rammer* (bl.a. fordi initiativer helst skal give synlige effekter på tidspunkter, hvor der finder evaluering sted) – og at *Evalueringer former aktører*, samtidig med at de definerer, hvem der tæller som aktører. Hvor OECD's redegørelser giver et indtryk af den intenderede PISA-brug, giver Dahler-Larsens diskussion således et blik for nogle af de utilsigtede virkninger.

3.1 Anvendelse og konsekvenser af PISA i den danske kontekst.

Undervisningskonsulent Jørgen Balling Rasmussen fra Undervisningsministeriet er som dansk PISA-ansvarlig embedsmand i perioden én vigtig kilde til at etablere sammenhæng mellem PISA og de uddannelsespolitiske beslutninger. Ved en nordisk konference i 2006 gav han en fremstilling af "Policy consequences of PISA – Danish experiences" frem til konferencetidspunktet. Da PISA 2006 ikke afdækkede væsentlige nye problemer anses nedenstående at sammenfatte de mest markante virkninger:

1. *Øget vægt på grundlæggende faglighed* (i traditionel forstand). Udmøntet i ændring af Folkeskolelovens formåls-paragraf i 2006 og bl.a. i udvidede timetal til de "grundlæggende fag" dansk og historie)

2. *Styrkelse af evalueringskulturen i folkeskolen/indførelsen af et omfattende nationalt test-system*

3. *Indførelse af et benchmarkingsystem med bindende trin- og slutmål* ("Fælles Mål")

4. *Ændringer i læreruddannelsen* (med færre og større linjefag, bl.a. indførelse af fysik/kemi og natur/teknik, som store og centrale linjefagsvalg i 2006)

I det følgende vil vi kort uddybe, kontekstualisere og perspektivere de forskellige virkninger af PISA.

Den øgede vægt på grundlæggende faglighed

Undervisningsminister Ulla Tørnæs udsendte straks ved offentliggørelsen af PISA2000 en pressemeddelelse (4.12.2001), hvor det dårlige resultat i naturfag blev imødegået med et fokus på basale færdigheder: *"Det er uacceptabelt, at danske elever er dårligere end gennemsnittet i OECD, når det drejer sig om naturfag. Her skal der altså hankes gevaldigt op i elevernes færdigheder".* Dette var allerede skrevet ind i regeringsgrundlaget for den nytilkomne borgerlige regering. Indtrykket er her, at PISA blev brugt som påskud for at gennemføre et politisk program *direkte imod* undersøgelsens substans, idet regeringens vægtning af traditionel faglighed harmonerede meget dårligt med PISAs grundlæggende begreb om fx Scientific Literacy. En konstitutiv effekt i retning af, at selve evalueringskriterierne er blevet til værdier på det politiske niveau, kan man i hvert fald ikke tale om. Snarere er de konstitutive virkninger til stede som en større opmærksomhed på målbare aspekter af uddannelse. Det følgende år omtalte pressen ministerens ærinde som *"en ideologisk ændring af folkeskolen, der skal kurere alt fra skolebørns dårlige resultater i internationale undersøgelser til social ulighed"* (JP, 4.8.2002). Med henvisning til PISA-resultaterne gjorde undervisningsministeren nu eksplicit op med folkeskolens (angivelige) hidtidige vægtning af bløde værdier som demokrati og rummelighed, til fordel for et fokuseret arbejde med elevernes boglige færdigheder og kundskaber. Tilkendegivelsen blev i første omgang ledsaget af et ønske om større timetal til de "grundlæggende" fag historie og dansk og offentliggørelse af landsresultater af test. Dette blev realiseret ved ændringen af Folkeskoleloven i 2006 (LOV nr 572 af 09/06/2006), hvor den faglige opprioritering bl.a. også førte til nyformulering af folkeskolens formålsparagraf. Her blev den alsidige personlige udvikling (LBK nr 730 af 21/07/2000) nedtonet til fordel for en rettethed mod videre uddannelse.

Naturfagenes relativt ringe præstation gav anledning til betydelig bekymring hos Dansk Industri, som direkte krævede en reform af folkeskolens naturfagsundervisning. De hævdede, at "*de naturvidenskabelige fag er blevet forsømt i lang tid*" (JP, 21.08.2002) og forbandt de dårlige danske PISA-resultater med lærernes manglende linjefagsbaggrund og lave timetal for naturfagene. Som vi skal se nedenfor fik dette konsekvens for læreruddannelsen.

Indførelse af et nationalt test-system

PISA 2000 var med til at konstituere værdien af *at* evaluere, og de konkrete, utilfredsstillende danske resultater fik efterfølgende regeringen til i 2003 at takke JA til OECD's tilbud om en PISA-opfølgende evaluering/review af folkeskolen. I rapporten, som kom i midten af 2004, begrundes review-panelets sammensætning med ønsket om at få adgang til Best-(PISA)-Practice: *"Because the Danish Government was anxious to learn from the experience of 'best practice', experts from three countries, which had performed well in the PISA assessments yet were similar in critical respects to Denmark, were chosen to make up the review team."* (OECD 2004a, s. 14). Landene var UK, Canada og Finland. Evalueringspanelets engelske formand P. Mortimore er kendt for sit målrettede arbejde med "School Effectiveness" og med evaluering som middel – og ret forudsigeligt handlede mange af de 35 rekommandationer om øget evaluering på skole-, lærer- og elevniveau. Bl.a. anbefaledes det at udvikle national monitorering af elevudbytte, kriteriebundne tests og et sæt af benchmarks for forskellige aldersgrupper i folkeskolens vigtige fag. På en opfølgende konference understregede P. Mortimore, at det danske uddannelsessystem hidtil havde gjort det godt: "*Verden er bare blevet mere kompleks, så I er nødt til at foretage visse ændringer uden at ødelægge det gode, I har. Uden at smide barnet ud med badevandet"* (citat Folkeskolen, 18.6.2004). Barnet var det efterhånden vidt berømte glade danske skolebarn, der kan lide at lære – og hvis usædvanlighed indenfor OECD-området PISA netop havde dokumenteret.

Her kan man altså tale om, at PISA initierer en proces, som evalueringspanelet så tilføjer retning. Den resulterende systemiske evaluering blev dog først lanceret med offentliggørelsen af PISA2003 resultaterne, der trak overskrifter a la *Uacceptabelt – Brug for national mobilisering til at genskabe kvaliteten i folkeskolen* (Kristeligt Dagblad, 7.12.2004) og *PISA-rapporten: Ny nedtur for folkeskolen* (Politiken, 7.12.2004). Selvom de danske PISA-forskere forsøgte at komme igennem med et budskab om, at der kun var egentlig tilbagegang på naturfagsområdet meddelte undervisningsministeren, at hun som konsekvens af de dårlige resultater ville indføre obligatoriske tests af børnenes færdigheder (6.12.2004)[1]. En markant PISA-initieret konsekvens er således indførelsen af obligatoriske nationale tests i udvalgte fag:

	Klassetrin								
Fag	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Dansk/læsning		X		X		X		X	
Matematik			X			X			
Engelsk							X		
Geografi								X	
Biologi								X	
Fysik/kemi								X	
Dansk som 2. sprog					X		X		

Det er nærliggende at læse PISA-konstitutive effekter ind i tabellen, først og fremmest via dens udhævning af bestemte fag som relevante "aktører" og bestemte tidspunkter, som testningsmæssigt interessante. Med engelsk som eneste undtagelse er alle testfagene PISA-relevante, og testningstyngden er særlig stor i "PISA-vinduet" omkring 8. klassetrin. Testene er tænkt at være computer-baserede og adaptive, således at de gradvist tilpasser sig den enkelte elevs niveau. Efter planen skulle testsystemet fungere fra afslutningen af skoleåret 2005-2006, men implementeringen har været præget af massiv kritik af naive, factsorienterede første-opgaver, servernedbrud og afbrydelse af internetbaseret eksamen for tusinder af elever, mangelfuld adaptivitet, ubrugelig tilbagemelding til lærerne, for nu blot at nævne nogle af indvendingerne (Jørgensen, 2010a; Jørgensen, 2010b; Fuglsang & Saietz, 2010). Mere dybtgående teknisk analyse af (dele af) det miserable forløb kan findes i Skolestyrelsens egen evalueringsrapport (Devo Team Consulting, 2007). I skoleåret 2010/11 er systemet i store træk oppe og køre.

Indførelse af et benchmarkingsystem med bindende trin- og slutmål

I 2001 indførtes "Vejledende Klare Mål" som et første forsøg på benchmarking af fagligheden i folkeskolen. Dette system har været igennem adskillige revisioner, først en mindre justering ved VK-regeringens tiltrædelse, dernæst en gennemgribende omformulering til trin-målsbeskrivelsen i *Fælles Mål* (2003). Endelig er Fælles Mål blevet gennemskrevet endnu engang for at genopstå som *Fælles Mål 2009*. I modsætning til tilsvarende norske standarder er *Fælles Mål 2009* (som sine forgængere) lavet uafhængigt af gymnasiets målbeskrivelser. Fælles Mål opererer med undervisningsmål (som læreren så må transformere til elevernes læringsmål), mens gymnasiet anvender læringsmål. Alligevel synes der at være et betragteligt overlap mellem folkeskolens slutmål og gymnasiemålene for flere af naturfagene. Benchmarking-systemets trinmål flytter på godt og ondt evalueringsbølgen nedad i systemet; godt fordi det åbner op for formativ evaluering,

[1] Dette selvom man i den danske PISA2003 rapport kunne læse, at *"Resultaterne fra en analyse på alle OECD-lande tyder dog heller ikke på, at standardiserede tests spiller en selvstændig rolle for elevernes færdigheder i matematik, når der tages højde for elevernes hjemmebaggrund" (Mejding & et al, 2004, p.220)*

ondt fordi et udstrakt fokus på *summativ* evaluering (jf. VAP-rapport 2) virker tilbage på undervisningen og nemt trækker folkeskolen i retning af sortering og udskillelse og en mere factsorienteret og unuanceret undervisning. At såkaldte high-stake tests (dvs. tests som har afgørende konsekvenser for eleverne og/eller skolen) har en negativ effekt på undervisningen er forskningsmæssigt veldokumenteret (Nordenbo, Allerup et al. 2009). Nu kan hverken PISA eller de nationale tests siges at være en high-stake test, men da de har så stor uddannelsespolitisk indflydelse vil de i mange henseender have samme effekt på undervisningen. En nylig evaluering af beslægtede nationale tests i Norge peger således på, at de er styrende for undervisningen og har begrænset pædagogisk og læringsmæssig værdi ift. andre evalueringstyper (Skov 2010).

Ændringer i læreruddannelsen

Som tidligere omtalt var der flere forskellige forklaringer i spil, da de utilfredsstillende danske naturfagsresultater i PISA 2000 skulle forklares. Da regeringen i stedet valgte at opprioritere timetallene i "de grundlæggende fag" dansk og historie, var den afskåret fra at gøre noget ved naturfagenes "lave timetal". Derfor var man henvist til et PISA-initieret forsøg på at rette op på lærernes manglende eller relativt svage linjefagsbaggrund i naturfag. Med 2006-reformen af læreruddannelsen blev antallet af linjefag reduceret til 2-3, og natur/teknik samt fysik/kemi fik højere status ved at blive placeret blandt de største (1.2 årsværk) linjefag i øverste valglag. Med denne begunstigede position har man ønsket og forventet at skaffe flere lærere med linjefagsbaggrund i disse fag – men de senere tal for tilvalg og holdoprettelse på seminarierne tyder snarere på, at den nye struktur polariserer, så kun overbeviste naturfagstilhængere i dag kommer i nærheden af linjefag indenfor naturvidenskab. De konstitutive effekter på læreruddannelsens struktur blev ikke forsøgt overført til de nye linjefags indholdsbeskrivelser, i form af de nye Centrale Kundskabs- og Færdighedsområder. I hvert fald finder man ikke her en udtalt vægt på evaluering, endsige nogen eksplicit henvisning til Scientific Literacy eller PISA-lignende kompetencemål og kontekster.

3.2 Konsekvenserne – i det store perspektiv

Det er påfaldende, at det er svært at pege på *direkte* konsekvenser af PISA, altså forandringer hvor PISA har leveret *substansen.* Tættest på har man måske været i diskussionen af små vs. store skoler. Resultaterne i PISA2000 udpegede store skoler, som de læringsmæssigt bedste. Resultatet er blevet flittigt citeret og brugt, fx bekræfter professor N. Egelund i 2007, at store skoler (med 700-800 elever) er fagligt og økonomisk bedre end små (JP Aarhus, 13.3.2007). Evidensen forekommer imidlertid mindre entydig, idet det stort set samtidig i den danske PISA2006-rapport konstateres (Egelund 2007, p. 200): *"PISA 2000 viste, at elever i større skoler har bedre naturvidenskabsfærdigheder, også når der er korrigeret for en række baggrundsfaktorer, men som det fremgår af tabel 6.1, er denne tendens ikke statistisk sikker for PISA 2006."* Lokale lukninger af små skoler med henvisning til PISA-resultaterne, må siges at være en (måske kontroversiel) effekt, hvor PISA-undersøgelsen leverer et vigtigt input. PISAs manglende evne til substantielt at kvalificere de uddannelsespolitiske beslutninger er en naturlig konsekvens af, at undersøgelsen udelukkende fastlægger et "slut"-produkt (elevformåen, år 15) og nogle randbetingelser (fx socio-økonomiske, skolestørrelser m.m.), men er blind overfor de pædagogiske og sociale processer, som reelt har skabt slutresultatet. Man kan derfor med rette diskutere *om* PISA reelt har givet politikerne et bedre grundlag at træffe beslutninger på. Det er derimod temmelig indiskutabelt, at PISA-undersøgelserne i sig selv *ikke* har givet lærere og/eller skoler et bedre grundlag for at bedrive den

daglige undervisning. Et forhold som til dels skyldes PISAs opbygning og formål og til dels kan tilskrives manglende udnyttelse af PISA-dataene og mangelfuld supplerende forskning.
PISA *har* leveret øjenåbnere, vigtigst måske indsigten i, at vi i Danmark har et problem med social arv, og *næppe* har verdens bedste skole: *"PISA udgør en lille brik i det store billede. Mens vi før i tiden gik og troede, at Danmark havde verdens bedste skole, så har samarbejdet i internationale netværk som PISA og andre internationale undersøgelser…åbnet vores øjne for virkeligheden"* (Jan Meiding, Teknologirådets Høring/Diverse Oplægsholdere. 2005). *Indirekte* har denne og andre PISA-afdækninger haft konsekvenser, først og fremmest, hvor de har initieret analyse-, strategi- og reviewarbejde, samt evt. følgeforskning (Andersen, Busch, Horst, & Troelsen, 2003; Andersen et al., 2006; Arbejdsgruppen til forberedelse af en national strategi for Natur, 2008; Mortimore, David-Evans, Laukkanen, & Valijarvi, 2004; Egelund & Andersen, 2006). Det PISA-initierede OECD-review af folkeskolen har således klart leveret nogle af de mest markante aftryk på den danske grundskole, i form af benchmarking og test-systemer. Endelig er PISA blevet brugt til at *legitimere* en opprioritering af en traditionel (fag)faglighed, som egentlig er i modstrid med PISAs Scientific Literacy grundlag.
Der foreligger ikke undersøgelser, som gør det muligt at se, hvorledes PISA har konstitueret *praksis*. Praksis er i det hele taget forbløffende u-undersøgt i Danmark. Men: PISA har i allerhøjeste grad været med til at konstituere feltet, først og fremmest hvad undervisningen i den danske folkeskole skal måles på. Et par citater fra Teknologirådets PISA-høring (Teknologirådet/Diverse Oplægsholdere, 2005) kan belyse nogle aspekter af denne problematik:

- "Det er et problem, hvis fokus på "standardmennesket" i den offentlige debat om skolen helt overskygger betydningen af det individuelle." (H.H. Knoop)

- "utroligt, hvad moderne mennesker søger at lægge ned i tal – fx komplekse relationer mellem mennesker, som de foregår i en undervisningssituation, der jo altid er situationsbestemt og menneskeafhængig. Det bør give stof til eftertanke, at vi ved så lidt om disse sammenhænge.. " (S. Hildebrandt)

- "PISA måler ikke alle de intelligenser, der er brug for i fremtidens samfund" (S. Hildebrandt)

På hver sin vis udtrykker disse citater PISAs evne til at konstituere og reducere undervisningens felt. PISA beskæftiger sig kun med det der kan måles og i-tal-sættes via individuelle paper-and-pencil-tests. Dermed indsnævres feltet af relevante kompetencer betragteligt, og det forbliver et postulat, at PISA virkelig indfanger unges forudsætninger for at begå sig i fremtidens samfund. Endelig forsvinder den enkelte elev i både rapportering og den offentlige debat – til fordel for gennemsnitstyper og nationale scorer. For en skole med traditionel vægt på "den enkelte elevs alsidige personlige udvikling" (§ 1, tidl. formålsparagraf, Lov om Folkeskoleloven 1993) udtrykker dette et markant skred! Dertil skal så lægges pædagogisk-konstitutive effekter af, at test-systemet er nærværende i dagligdagen i en række fag (se fx (EPPI - Evidence for Policy and Practice Information Centre & Assessment and Learning Research SynthesisGroup (ALRSG), 2002; Nordenbo, Allerup, Andersen, Dolin, & et al, 2009). PISA har allerede udvirket store ændringer i folkeskolens værdigrundlag – og potentielt kan konsekvenserne blive endnu større.
Afslutningsvist kan man spørge, hvorvidt PISA er blevet brugt i overensstemmelse med de intentioner og anbefalinger, som OECD har lagt frem. Indtrykkene er her modstridende: På den ene side udtrykker man i OECD-reviewet af folkeskolen ønske om at lære af Best Practice i andre lande, men reelt er det kun den engelske evalueringspraksis, som står tydeligt i reviewet og de

efterfølgende beslutninger - Best Practice fra panel-landene Canada og Finland forbliver usynlige. Undervisningsministrene har gentagne gange slået på, at (PISA)resultaterne ikke står mål med, at vi har "verdens dyreste folkeskole". Value-for-money-argumentet er således blevet fremført, men det er svært at pege på fx direkte overenskomstmæssige konsekvenser heraf. PISA 2006 viste forbedrede naturfagsresultater – uden at nogen egentlig kunne pege på en fuldgyldig forklaring. Samtidig var det svært at se nogen blivende effekt af iværksatte initiativer til fremme af læsefærdigheder og til bekæmpelse af social arv. I sine kommentar fik undervisningsministeren antydet, at PISAs løbende monitorering tænkes brugt til at vurdere effekten af politiske initiativer: "glædeligt, at vi nu kan se, at satsningen på at øge fagligheden begynder at bære frugt, men der er rum til forbedringer" (Kristeligt Dagblad, 5.12.2007). Samtidig afviste han, at der denne gang ville være konsekvenser i form af reformer og nye lovforslag fra regeringens side: "Nu skal dét, vi har gennemført, have lov til at virke i praksis." (Berlingske Tidende, 5.12.2007).

4. International diskussion af PISAs testtekniske forhold

PISA er i disse år den mest anvendte og omtalte internationale test, men også den mest omdiskuterede. Diskussionerne drejer sig mest om de uddannelsespolitiske konsekvenser af PISA, men også PISAs testtekniske opbygning er under debat, i erkendelse af at evalueringers tekniske forhold ikke er neutrale praksisser, men en del af det værdigrundlag, som evalueringer bygger på.

Førende internationale tidsskrifter udgiver særnumre om PISA (fx International Journal of Science and Mathematics Education 2010/8), internationale konferencer om uddannelsesforskning har specielle sessioner om PISA (fx ESERA, EARLI), Nordisk Ministerråd har finansieret en række rapporter som sammenligner de nordiske landes præstationer i PISA (Mejding and Roe (eds.) 2006, Matti (ed.) 2009) og der er udgivet en række mere kritiske artikler og bøger om PISA (fx Goldstein 2004, Hopmann et al 2007). Diskussionerne om PISAs mere testtekniske opbygning og udførelse kan samles i tre grupper: Uhensigtsmæssigheder grundet i kritiske valg, diskussion af hvorvidt PISA måler det den påstår at udtale sig om (dvs. validitetsproblemer) og kritik af den tekniske gennemførelse af testen (dvs. pålidelighedsproblemer). I en undersøgelse af PISAs anvendelighed i en dansk kontekst er validitetsproblemet centralt og omdrejningspunktet for denne rapport, men vi vil kort belyse de to andre diskussionspunkter (for en grundigere gennemgang, se Dolin 2008)

Enhver evaluering har et formål og for at opfylde dette skal der træffes en række teoretiske, praktiske og metodiske valg. I PISA-testen er det fx besluttet at prioritere sammenligning af elevresultater frem for undervisningsformer, og for at det skal være økonomisk muligt, vurderes elevkunnen i en to-timers skriftlig test. Hvad der vurderes, beskrives i et grundlagsdokument, the Literacy Framework (OECD 2006). Disse grundlæggende valg er kritiske, forstået således at de har betydning for både pålidelighed og gyldighed og sætter grænserne for testens udsigelseskraft. Nogle af de vigtigste valg, som PISA er blevet kritiseret for, er:

- Det er meget ambitiøst at måle i hvilket omfang 15-årige er 'fit for life'. En så kompleks størrelse er praktisk taget umulig at operationalisere, og der synes ganske langt fra det valgte proxymål (hvor godt man klarer sig i en to timer individuel, skriftlig test) til det der ønskes et mål for (deltagelse i livet). PISA prøver end ikke at etablere en forskningsbaseret relation mellem testperformence og livsduelighed.
- Manglende sammenlignelighed med andre internationale tests (som fx TIMSS) gør det umuligt at vurdere PISA-resultaterne sammen med andre resultater.
- Årgangs sample i stedet for klasse sample umuliggør korrelationer mellem klassevariable (typisk tilrettelæggelsesformer) og præstationer, hvilket reducerer den umiddelbart pædagogiske brug af PISA.
- Den valgte statistiske model måler kun variation langs en lineær skala, på trods af at virkeligheden er væsentlig mere kompleks. Alle PISAs opgaver skal skalere korrekt på tværs af kulturer, men ved kun at medtage opgaver, der gør det, risikerer man at udelukke opgaver, som illustrerer relevante forskelle mellem lande.

Nogle af disse kritiske valg har konsekvenser for testens pålidelighed og gyldighed, og da det er ganske svært at optimere begge kvalitetskrav samtidigt, vil pålideligheden oftest blive prioriteret højest - når lande rangordnes er det vigtigere at det sker korrekt og efter gennemskuelige regler, end at man måler præcist det man ønsker. Det internationale PISA-konsortium har da også udviklet nogle standarder for den tekniske gennemførelse af testen (oversættelser, elevudvælgelse, opgavescoring etc.), som er velbeskrevne, især i den tekniske rapport som udkom i forbindelse med det første test-gennemløb (Adams and Wu 2001). Her sammenfattes pålidelighedsmålene for de

forskellige processer i testforløbet til en overordnet pålidelighed på 92 %, som dog dækker over meget store udsving mellem landene og mellem de forskellige elementer i testen. Men selv om der sandsynligvis er opstillet så gode procedurer, som muligt, så skal de implementeres i mange forskellige lande med forskellige traditioner, og kompleksiteten er samtidig så stor, at det ikke er muligt at undgå fejl, som der ikke er taget højde for, og som mindsker pålideligheden (se fx Wuttke 2007, Allerup 2007).
Sådanne pålidelighedsproblemer er selvsagt vigtige når politikere anvender deres lands placering i rangordningstabellerne som argument for uddannelsesmæssige ændringer. Men vi anser det dog vigtigere for vurdering af relevansen af PISA, om PISA egentlig leverer data om de overordnede forhold, som den stiller op som de erklærede mål, dvs. gyldighedsmæssige problemstillinger. Fx kan man spørge hvorvidt et lands PISA-resultater siger noget om skolesystemets evne til at forberede eleverne til deres fremtidige virke i samfundet. Sådanne brede, uddannelsespolitiske spørgsmål er blevet behandlet af mange uddannelsesforskere (fx Dolin 2005, Sjøberg 2007), og svarene afhænger i vid udstrækning af ens uddannelsespolitiske ståsted. Mere skolenær og forskningsmæssigt muligt er en undersøgelse af testens overensstemmelse med de nationale mål for undervisningen. Det er en sådan validering indeværende projekt foretager for naturfagene i Danmark.

5. Samlet oversigt over VAP-projektet

VAP-projektet (Validering Af PISA science) forestås af lektor, ph.d. Jens Dolin, Institut for Naturfagenes Didaktik, Københavns Universitet, og adjunkt, ph.d., Lars Brian Krogh, Center for Scienceuddannelse, Århus Universitet. Lektor, ph.d., Henrik Busch, Det Naturvidenskabelige Fakultet, Københavns Universitet, var med fra starten og indtil udarbejdelsen af konceptet for den anden undersøgelse, men har på grund af arbejdsbyrden som prodekan ved Det Naturvidenskabelige Fakultet ved Københavns Universitet ikke kunnet medvirke ved projektets videre forløb.

Grundideen i VAP-projektet er at sammenligne nogle udvalgte danske elevers præstationer i PISA-testen med deres kunnen målt i en mere skolenær testsituation. For at kunne perspektivere denne sammenligning er kravene i det danske skolesystem sammenholdt med PISA-kravene og elevernes erfaringer med PISA-lignende testformater. Disse forudsætninger er tilvejebragt og formidlet i VAP-projektets to foregående rapporter.

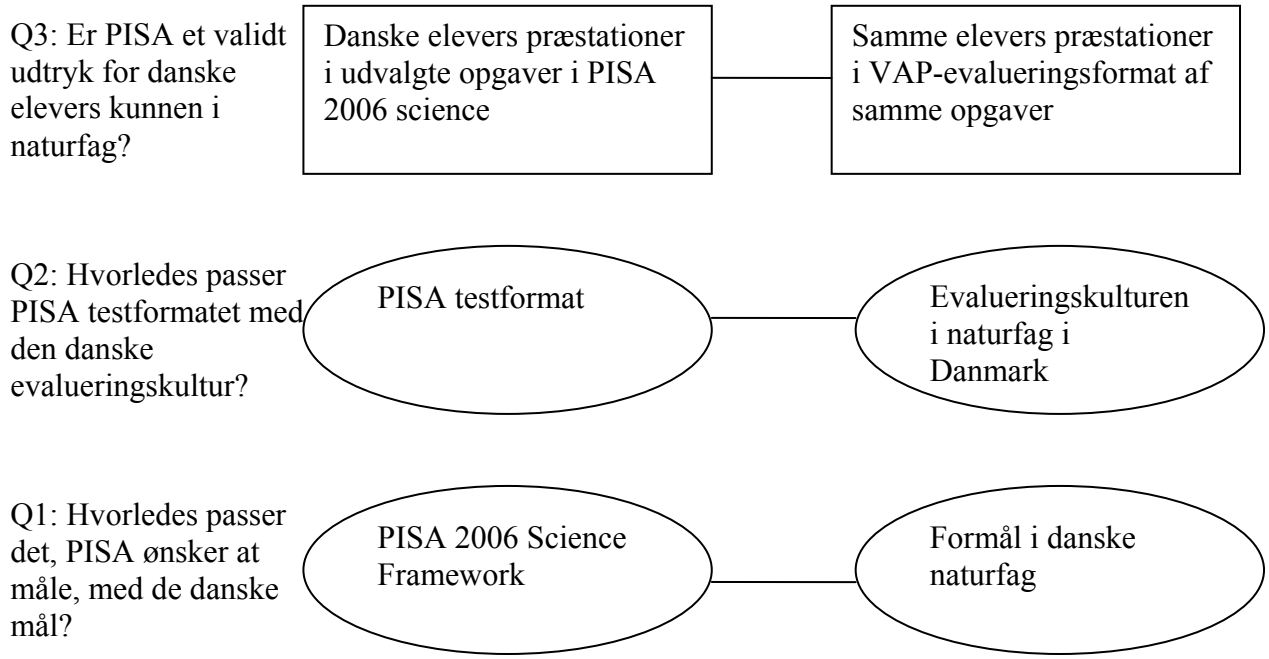
Hovedtrækkene i forskningsdesignet blev udviklet i februar 2005 og arbejdet gik i gang i april 2005. Tidsmæssigt er projektet knyttet til PISAs tredje testcyklus, hvor eleverne i Danmark blev testet i uge 10 og de følgende uger i 2006. En række baggrundsspørgsmål og forberedelser skulle afklares op til PISA-testen, og i ugerne efter testen skulle udvalgte elever udsættes for en specieludviklet VAP-evaluering. Resultaterne heraf skulle så sammenholdes med de samme elevers PISA-testresultater, når disse blev offentliggjort december 2007.

Projektet er delvist finansieret af Undervisningsministeriet (J.nr. 2005-2234-6) med en bevilling på 500.000 kr., og Undervisningsministeriets Styregruppe for PISA fungerede som styregruppe for projektet indtil gruppens nedlæggelse i 2006. Herudover har Danmarks Lærerforening bevilliget 45.000 kr. til undersøgelsen af naturfagslærernes evalueringskultur (VAP-rapport 2). Resten, og dermed langt størstedelen af de samlede udgifter, er dækket af de involveredes forskningstid.

Projektet er tilmeldt Datatilsynet (J.nr. 2007-54-0369), hvor anmeldelsen ikke gav anledning til bemærkninger.

5.1 Forskningsspørgsmål og overordnet forskningsdesign

VAP-projektet er formuleret som en række delprojekter, der tilsammen muliggør en placering af PISA science testen i en dansk kontekst. De tre grundlæggende forskningsspørgsmål hænger sammen således at de to første udgør et grundlag for at kunne svare fyldestgørende på det tredje, se figuren:



Q1. Hvad vil PISA måle – og i hvilken udstrækning er det foreneligt med den intenderede danske naturfagsundervisning?

På ét plan er her tale om en undersøgelse af PISAs formelle relevans for Danmark. Konkret og metodisk handler det om at sammenligne PISAs teoretiske grundlag og testkoncept, det såkaldte Framework (OECD 2006), med de intenderede danske mål, de såkaldte *Fælles Mål,* for undervisningen i fagene Fysik/kemi, Biologi og Geografi. Sammenligningen omfatter såvel kompetencemål som indholdskategorier og afdækker både hvilke af de danske prioriteringer PISA faktisk indfanger – og hvilke PISA udelader.

Analysen er samtidig en overordnet vurdering af, i hvilken grad danske undervisnings-/læreplaner dækker PISAs testområde, et aspekt som burde være afgørende for tolkningen af PISA-resultaterne, i det omfang de lægges til grund for en vurdering af det danske uddannelsessystem. I PISA-konceptet har man bevidst fravalgt at bekymre sig om nationale curricula. I stedet fokuserer man på en bred naturfaglig kompetence, *scientific literacy*, som (uden at dette gøres eksplicit) antages at være af universel natur. Analysen giver således også et fingerpeg om rimeligheden af en sådan antagelse for den danske kontekst. Resultaterne af denne analyse fremlægges i VAP-rapport 1 (Dolin, Busch, Krogh 2006) og opsummeres nedenfor.

Q2. Er PISAs målemetode rimelig og dækkende i forhold til den evalueringspraksis danske elever møder i naturfagsundervisningen?

En nærtliggende hypotese kunne her være, at danske elever underpræsterer i PISA, fordi de ikke er fortrolige med en sådan type testning i naturfag. I Danmark har folkeskolens afsluttende prøve i naturfagene traditionelt været mundtlig og praktisk/performance-orienteret.

Evalueringspraksis i naturfagsundervisningen i Danmark er her undersøgt via fokusgruppeinterviews og en elektronisk survey-undersøgelse omfattende 1159 naturfagslærere. Indholdsmæssigt afdækker undersøgelsen bl.a. anvendelsen af PISA-lignende evalueringsformater (individuel, paper-and-pencil, MC-opgaver), kompetencetyper, kontekstinddragelse og vægtningen af korrekt fagsprog. Resultaterne fremgår af VAP-rapport 2 (Dolin og Krogh 2008) og opsummeres nedenfor.

Q3. Er PISA-resultaterne et validt udtryk for, hvad danske elever faktisk kan indenfor det testede område?

Overordnet set er dette gyldighedsspørgsmålet, som er afgørende for enhver test, og et af VAP-projektets centrale spørgsmål. Derfor må en gyldig måling af elevers reelle formåen foregå på en måde som giver dem mulighed for at vise så stor en del af deres viden som muligt. Dette kan fx ske ved at lade testsituationen afspejle den undervisningsmæssige praksis, som eleverne har været udsat for. Det er fx almindeligt anerkendt at danske elever er vant til en ganske dialogisk og praktisk-eksperimentel naturfagsundervisning, ofte i grupper, i modsætning til PISA-testens individuelle papir-og-blyant-test. Et testformat som afspejler undervisningen må derfor også involvere samtale og give udfoldelsesmulighed for praktisk orienterede kompetencer. At samtale og papir-og-blyant-test giver anledning til forskellige rekonstruktioner af elevers viden er tidligere demonstreret (se fx Schoultz, Saljo, & Wyndhamn, 2001a). VAP-delundersøgelsen knyttet til Q3 er derfor designet som en feltvalidering, hvor en gruppe PISA-testede elever udsættes for en specieludviklet evalueringsproces, der tager hensyn til såvel PISAs egne scientific literacy-mål som de danske mål og dansk skolepraksis. Indeværende rapport omhandler udviklingen og gennemførelsen af denne evaluering, de opnåede resultater af evalueringen og sammenholdelse af disse resultater med PISA-resultaterne.

5.2 Opsummering af rapport 1 og 2

VAP-rapport 1 og 2 beskæftiger sig med ovenstående forskningsspørgsmål 1 og 2, dvs. hvorvidt PISA-testens målkategorier er i overensstemmelse med de danske mål for naturfagsundervisningen, og hvorvidt PISA-testformatet svarer til den danske evalueringskultur i naturfagene.
Det hurtige svar på begge spørgsmål er, at det er ikke helt ved siden af. Der er rimelig stort overlap mellem PISA-projektets mål og indhold og de danske Fælles Mål, og testformatet virker ikke fremmed for de lærere, der underviser danske elever i naturfag.
Men der er en række ikke ubetydelige nuancer, især hvad angår de bagvedliggende værdier og intentioner.

De kontekster, som PISA-testen stiller sine opgaver indenfor, passer fint med de faglige områder, som Fælles Mål omhandler. Biologiindholdet og geografiindholdet er praktisk talt det samme. Fysik-kemi, som har en relativt højere vægt i det danske uddannelsessystem end det har i PISA-testen, er svært at sammenligne, da de danske målbeskrivelser lægger vægt på elevnær behandling af hverdagsfænomener, mens PISA mere vægter konkret faglig viden. Den danske uddannelseskultur, hvor elevers snak om de faglige problemstillinger og egne meninger vægtes højt, giver således et dårligt udgangspunkt for en test, hvor der primært gives point for brug af det korrekte fagudtryk. Således scores PISA-testens åbne spørgsmål i vid udstrækning ved at se om eleven har anvendt nogle på forhånd vedtagne nøgleord, hvorved omskrivninger og egne vendinger ikke tillægges værdi. Viden *om* naturvidenskab vægtes højt i PISA-testen, men det er et område, som er svagt dækket i de danske læseplaner. Det indgår nok på forskellig vis i fagenes arbejdsmetoder og tankegange, men adresseres ikke eksplicit som en selvstændig dimension med et veludviklet begrebsapparat, hvilket stiller danske elever svagt i rigtig mange af PISA-opgaverne.
Den største forskel er uden tvivl fraværet af den praktiske dimension i PISA-testen. De danske læreplaner lægger stor vægt på elevernes praktiske arbejde i form af laboratorie- og feltarbejde, og dette kan ikke indgå i PISA-testens papir- og blyantsformat.
Generelt er de danske læreplaner formuleret i brede vendinger som ”særlig vægt på forståelse af sammenhænge”, ”væsentlige træk ved”, ”inddrage perspektiver for” etc., der lægger op til helhedsorienterede, problembaserede undervisningssituationer. Herigennem fremmes en

undervisning som vægter elevernes evne til at gennemføre processer og se sammenhænge og helheder, måske til en vis grad på bekostning af præcis faglig formulering. Dette hænger sandsynligvis også sammen med at danske lærere i overensstemmelse med den danske dannelsestradition i høj grad tilpasser de Fælles Mål til den enkelte klasse og de konkrete elever – frem for andre uddannelsessystemers opbygning af fælles standarder. Man kan se tendenser til at denne danske tilgang til undervisning ændres, måske bl.a. på grund af PISA.

VAP-rapport 2 viser ganske god overensstemmelse mellem PISAs testformat og den danske evalueringskultur i naturfagene. 1159 naturfagslærere i folkeskolens 8. og 9. klasser tegner således via deres spørgeskemasvar et billede af en evalueringspraksis, som i vid udstrækning anvender PISA-lignende evalueringsformater, fx er individuel og skriftlig testning den hyppigst anvendte organisering og evalueringsformat. På det overordnede niveau angiver i hvert fald 75 % af lærerne regelmæssigt at teste kompetencer svarende til PISA. Efter lærernes oplysninger inddrager den typiske lærer PISA-relevante kontekster i sine evalueringsopgaver - og der synes at være en god overensstemmelse mellem lærernes og PISAs vægtning af opgaver med åbne og lukkede svar-formater. Endelig anser lærerne det for rimelig vigtigt, at eleverne anvender korrekt fagsprog i opgaverne. Muligvis ville en endnu større vægtning af dette aspekt være i overensstemmelse med de krav til fagsprogsbrug, som PISA lægger til grund for sin scoring.
Alligevel er det tankevækkende at lærerne i undersøgelsen kun udtrykker moderat tiltro til at have "klædt eleverne på" til en individuel, skriftlig test som PISA. Hvis man som lærer kun har kendskab til de offentliggjorte, overfladiske træk ved PISA, kan det være svært at tro på, at man har forberedt sine elever på bedste vis.

Opsummerende kan det siges, at PISA-testens faglige områder og testformer ikke ligger voldsomt langt fra de danske læreplaner og evalueringsformer, men at der er ganske stor forskel på de intentioner og normer, der ligger bag PISAs testsystem, og de formuleringer og værdier, der bærer den danske naturfagsundervisning. Danske elever burde derfor på et formelt plan være i stand til at klare PISA-testen på lige fod med andre unge i verden. Men der er indikationer på at den konkrete undervisning i naturfagene forgår på en måde, som reelt fremmer andre, bredere kompetencer (samarbejde, dialog, praktisk arbejde, helhedsforståelse, problemløsning, nysgerrighed, selvsikkerhed etc.) til en vis grad på bekostning af konkret (parat)viden og sproglig skarphed.
Det er derfor interessant at undersøge danske elevers naturfaglige kompetencer på en måde som er i overensstemmelse med skolens krav og arbejdsformer og sammenligne resultaterne heraf med deres præstationer i PISA-testen, for at kunne vurdere i hvilket omfang PISA-testen indfanger det danske skolesystems krav til elevernes kompetencer. Det er netop formålet med denne tredje del af VAP-projektet.

6. VAP-evalueringens teorigrundlag og metoder

6.1 Paradigmatiske blik på evaluering

I de seneste tiår er der sket et skred i opfattelsen af, hvad man kan og bør bruge evaluering til. Skredet kan selvfølgelig aflæses i den måde hvorpå evalueringer gennemføres og anvendes, men vedrører nok så vigtigt selve grundantagelserne for evaluering. Flere forfattere (Gipps, 1999; Buhagiar, 2007) taler således direkte om paradigmeskift, fra et traditionelt, psykometrisk funderet, empirisk-rationelt post-positivistisk evalueringsparadigme til et alternativt, konstruktivistisk, interpreterende eller socio-kulturelt paradigme (Guba & Lincoln, 1994). Grundtræk i denne udvikling sammenfattes i et lidt længere citat af Buhagiar (Buhagiar, 2007, p.42-43):

"This shift from the psychometric model of assessment to the educational model draws on the postmodern condition that requires a suspension of belief in the absolute status of 'scientific' knowledge (see Gipps, 1993). This means that no matter how much we try to calibrate the 'measuring' instrument, we can still never know what is inside a student's head. Assessment can instead only tell us what the student can do in particular circumstances.

Keeping in mind that domains and constructs are multidimensional and complex, that assessment is not an exact science, and that the interaction of student, task and context is sufficiently complex, we still cannot know what the student can do in other circumstances. This renders the generalizability of assessment to other tasks and contexts limited, if not dubious."

The embedded denial within the new paradigm of the existence of a 'true score', however, reconceptualizes rather than bans the use of tests and examinations in assessment. Constructivist theories demand, in fact, that tests show what students know and can do, as well as facilitate good learning—what Glaser (1990) calls 'placing tests in the service of learning'. Within this emerging framework, tests should consequently be 'ambitious instruments aimed at detecting what mental representations students hold of important ideas and what facility students have in bringing these understandings to bear in solving their problems' (Shepard, 1991, p. 9)."

Citatet kommer omkring de væsentligste begrundelser for, at der er sket et skred i grundsynet på evaluering. For det første har post-moderniteten og den moderne videnskabsfilosofi ført til nye epistemologiske grundantagelser, dvs. et nyt syn på, hvilken natur og status viden har og kan have, samt på hvilke måder man kan erhverve sig viden. Intet nok så godt forsknings- eller evalueringsdesign anses længere at kunne give sand/rigtig/endegyldig viden om noget fænomen. Viden og videnskabelse erkendes samtidig at være af social og kontekstuel/kulturel natur. Parallelt hermed er der sket en udvikling i synet på, hvordan individer lærer. I en traditionel forståelse er undervisning og læring et spørgsmål om at overføre objektificeret viden til eleven. Dette traditionelle "lærings"-syn er nu erstattet af et billede af læring, som elevens personlige konstruktion af mening. Konsekvensen for evaluering af elever er fundamental: langt henad vejen vil standardiserede opgaver i begrænsede responsformater (fx multiple choice, korte åbne svar) være tilstrækkeligt til i denne forståelsesramme at checke om den autoriserede viden er modtaget. Omvendt skal der meget mere åbne, interaktive og autentiske evalueringsformer til for at indfange elevers mangfoldige, socialt og kontekstuelt forankrede videns-konstruktioner. De erkendelsesmæssige og læringsteoretiske nybrud slår kraftfuldt igennem i kravene til praktiske evalueringsformater.

En tredje type af argumenter for fremkomsten af et nyt evalueringsparadigme kommer fra selve evalueringsforskningen. Her har talrige studier påvist, hvorledes testning og eksaminer på utilsigtet og uhensigtsmæssig vis virker ind på både elever, læringsmiljø og lærere (EPPI - Evidence for Policy and Practice Information Centre & Assessment and Learning Research SynthesisGroup (ALRSG), 2002; Broadfoot & Black, 2004; Buhagiar, 2007). Rekommandationen i forlængelse heraf er at udvikle evalueringsformer som understøtter læreprocessen, frem for blot at fastholde elever og lærere på at afregne det lærte. Slogan-agtigt handler det om at gå fra *proving learning* til *improving learning* (Bell, 2007).

Gipps opsummerer udviklingen således (Gipps 1999, p. 384):

"Developments in cognition and learning are telling us to assess more broadly, in context,and in depth. This requires methods of assessment that do not lend themselves readily to traditional reliability, highlighting the tension between types and purposes of assessment".

Den omtalte spænding kan tydeliggøres ved at modstille karakteristika for evaluering indenfor hhv. traditional/post-positivistisk og socio-kulturel evaluering (se fx Gipps, 1999; Bell, 2007):

POST-POSITIVISTISK EVALUERINGS- PARADIGME:	SOCIO-KULTURELT EVALUERINGS-PARADIGME:
• Non-Interaktiv • Non-Kollaborativ • Statisk • Produkt-orientering • Begrænset brug af værktøjer (symbolske, fysiske) • Løsrevet fra situeret & autentisk praksis	• Interaktiv • Kollaborativ • Dynamisk • Proces-orienteret • Udstrakt brug af værktøjer (symbolske, fysiske) • Indlejret i situeret & autentisk kulturel praksis

Det post-positivistiske evalueringsparadigme er karakteriseret ved, at ville udmåle objektivt, pålideligt, sammenligneligt og med størst mulig generaliserbarhed. Dette tilstræbes bl.a. gennem standardisering af testningsprocedurer og ved at dyrke reproducerbarhed som primært kvalitetsmål. Lidt groft sagt er det vigtigere at målingerne fremstår pålidelige ("reliabilitetsspørgsmålet"), end at de indfanger målkategorien på helt dækkende vis ("validitetsspørgsmålet"). Viden/Læring anskues objektivistisk, dvs. som et produkt, som den enkelte tænkes at have erhvervet sig i større eller mindre grad. Målingens sigte er derfor væsentligst at fastslå i hvilken grad. For ikke at ødelægge den objektive måling, bør måleproceduren minimere interaktion med målingens genstand, dvs. eleven. Samarbejde mellem elever anses på samme måde at forstyrre udmåling af, hvor meget viden den enkelte elev har modtaget. Ydermere har den viden som er værd at måle en karakter, som gør den gyldig på tværs af kontekster og gør det meningsfuldt at kortlægge den uafhængigt af bestemte værktøjer og praksisser (inklusive den type af situationer, hvor viden er tilegnet/genereret). Det giver således i dette paradigme mening at måle elevernes viden i en "kunstig" test-situation i fx en gymnastiksal, frem for i en undervisningsnær situation i naturfagslokalet.

Herover for står det socio-kulturelle paradigme, som "*Stresses how knowledge is conditioned and constrained by the technology, information ressources, representation systems, and social*

situations with which people interact" (Mislevy 2003). Viden er her ikke kun et flytbart objekt, som den enkelte besidder i større eller mindre grad, men findes indlejret i de situationer og processer, som man indgå i sammen med andre. Interaktion og samarbejde er her ikke barrierer for en objektiv måling, men selve læringens medium – og en naturlig forudsætning for evaluering af læreprocesser. I dette dynamiske læringsbillede er det også utilstrækkeligt blot at udmåle en statisk øjebliksbillede af den nuværende viden; lige så vigtigt er det at få et indblik i elevernes beredthed til at tage næste læringsskridt, hvilket indebærer, at evalueringen også skal søge at indfange elevernes nære udviklingszoner. I dette evalueringsperspektiv er sigtet ikke at tilvejebringe objektiv og reproducerbar viden (hvilket anses for omsonst), men at skaffe valid information, som er meningsfuld og potentielt brugbar for de direkte aktører med kendskab til situationen.

Der er således tale om to meget forskellige evalueringsparadigmer, som hver især måler noget meget forskelligt - og det er disse forskelle vi forsøger at indkredse og at kvantificere og kvalificere. Vi begiver os derfor i den første del af rapporten (kulminerende i afsnit 7.3) ud i den metodisk meget vanskelige opgave at "lukke" en sociokulturelt baseret vurdering og sammenligne den med en post-positivistisk. I den anden del (afsnit 7.4) tager vi de sociokulturelle briller på og vurderer danske 15-åriges naturfaglige kunnen med dette særlige blik.
For at præcisere de to tilgange vil vi først klargøre de bagvedliggende paradigmer yderligere.

6.2 PISA og VAP i et evalueringsparadigmatisk lys

PISA konsortiet har så vidt vides aldrig deklareret sit evalueringsparadigmatiske ståsted, endsige ekspliciteret det videns-/læringssyn, som ligger til grund for PISA's Scientific Framework. Med Pellegrino et al's ord er dette uhensigtsmæssigt:

"*A model of cognition and learning, or a description of how people represent knowledge and develop competence in a subject domain, is a cornerstone of the assessment development enterprise. Unfortunately, the model of learning is not made explicit in most assessment development efforts, is not empirically derived, and/or is impoverished relative to what it could be."* (Pellegrino, Chudowsky, & Glaser, 2001, p.176)

Vi anser det imidlertid for temmelig uproblematisk at karakterisere PISA's test-setup i forhold til de dimensioner, som ovenfor tjente til at tydeliggøre spændingen mellem det post-positivistiske hhv. det socio-kulturelle evalueringsparadigme. I PISA-testningen interageres der kun med eleven via den låste opgave-tekst. Eleven laver sin egen individuelle besvarelse, og der gøres intet forsøg på at indfange elevens læringspotentiale eller på nogen måde at anbringe vedkommende i en stilladseret læringssituation. PISA hævder at kortlægge naturvidenskabelige processer (fx forklare fænomener ud fra naturvidenskab, anvende naturvidenskabelig evidens), men gør sig meget begrænsede anstrengelser for reelt at indfange disse processer. I stedet antages, at elever som vælger det rigtige produkt/svar, er nået så vidt via anticiperede processer. PISA-testningen hviler på et grundlag af tekst (m. tilhørende billedflade), samt papir og blyant. Der er således et minimum af generelle symbolske og fysiske værktøjer, samt antydninger af fagspecifikke repræsentationer. Egentlige fagspecifikke værktøjer forekommer ikke. Endelig foregår PISA-testen løsrevet fra de aktiviteter, der præger elevernes hverdag, skoleliv og naturvidenskabelige undervisning. PISA leverer ikke reel kontekst, men "opgaver con-text", for nu at bruge W.M. Roth's kritik af paper-and-pencil-test. PISA synes derfor at placere sig helt konsistent i skemaets venstre side, svarende til en post-positivistisk orientering. Med P-angivelser har vi i tabellen nedenfor søgt at indikere, hvorledes vi efter bedste overbevisning ville placere PISA ift. de forskellige dimensioner.

VAP-projektets bestræbelse er at vurdere i hvilket omfang PISA-resultaterne afspejler hvad eleverne rent faktisk kan i naturfag, hvis man måler deres kunnen i andre og mere realistiske sammenhænge end PISA gør. Udgangspunkt er derfor skolekonteksten, men i modsætning til PISA-projektet ses viden ikke som en dekontekstualiseret evne til opgaveløsning, men mere som en kompetence indlejret i en konkret situation:

From the viewpoint of situated cognition, competent action is not grounded in individual accumulations of knowledge but is, instead, generated in the web of social relations and human artefacts that define the context of our action.
(St. Julien 1997)

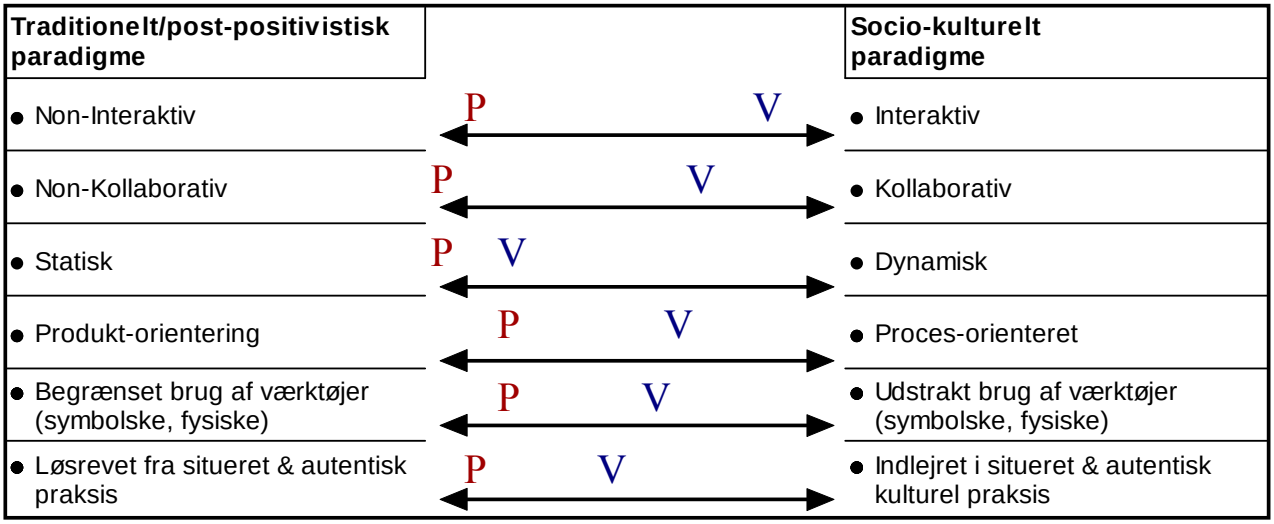
Den opbyggede viden afhænger således af læringssituationen, og idealet er derfor at testsituationen ligge så tæt herpå som muligt. Kun ved at tilnærme testsituationen til skolehverdagen vil man ud fra et socio-kulturelt perspektiv få resultater som med rimelighed afspejler elevernes faglige formåen frem for deres evne til slet og ret at overvinde testformatet. Det er aldrig muligt fuldt ud at genskabe samme situation, men forskning påpeger at især mulighed for italesættelse, dialog og brug af artefakter er afgørende elementer i en sociokulturel læringskontekst. Hos Säljö (Säljö 2003) fastslås det således:

Beherskelse af sproglige eller intellektuelle redskaber spiller således en central rolle i et sociokulturelt perspektiv på læring og udvikling
(p. 104)

VAP-testformatet har derfor italesættelse og dialog (med interviewer, med anden elev) som sit omdrejningspunkt. Et sådant sociokulturelt udgangspunkt for VAP-evalueringen giver desuden mulighed for at indfange aspekter af elevernes viden (og mangel på samme!) som PISA-testens format ikke kan rumme. Det er muligt i en dialogisk situation at spørge ind til hvorfor elever svarer som de gør, fx om det er rent gætteri eller udtryk for en egentlig viden, om svaret er rigtigt selv om det er baseret på faglige misforståelser osv. Det er ligeledes muligt at undersøge om elevviden er baseret på skoleundervisningen eller på andre erfaringer, det er muligt at se om de har viden ud over den efterspurgte etc. Der åbnes gennem en dialog- og artefaktbaseret samtaleform op for, hvad vi vil kalde "et udvidet opmærksomhedsvindue", som kan give en væsentlig mere nuanceret indsigt i elevernes viden end en skriftlig test. Det er i denne sammenhæng også vigtigt, at VAP med konsekvent brug af video-optagelse også *fastholder* størstedelen af informationsrigdommen i det udvidede opmærksomhedsvindue – til gavn for den senere analyse.

VAP-testningen udvider også *udfoldelsesvinduet*, idet der er indføjet en komponent med praktisk-eksperimentelt arbejde i elevpar. Her udfoldes både den dialogiske interaktion (mellem elev og interviewer, mellem elever), et kollaborativt islæt og en udvidet brug af artefakter. Både interview og praktisk-eksperimentelt arbejde giver rum for en vis processuel udvikling af forståelse hos eleverne, mens dynamikken i betragtelig grad begrænses af den semi-strukturerede interview-manual, som var nødvendig for at sikre pålidelighed i vurderingen. I udgangspunktet vil en veludført evaluering af denne type kunne opnå en rimelig høj validitet, mens udfordringen inden for dette paradigme til gengæld er at sikre pålideligheden.

Test-formatet og VAP's bestræbelse på at optimere pålideligheden vil blive beskrevet i større detalje nedenfor; beskrivelsen her skulle primært tjene til at rimeliggøre den positionering af vores VAP-evaluering, som V'erne i nedenstående tabel indikerer.



Som det fremgår af tabellen, anser vi ikke, at vi med VAP har begået en aldeles socio-kulturel evaluering. Mere mådeholdent betegner vi i denne rapport VAP som socio-kulturelt *orienteret,* for at pointere at VAP-formatet har indføjet socio-kulturelle elementer i større eller mindre grad.

VAP's midterposition mellem de to paradigmatiske yderpoler er ikke en tilfældighed, men et bevidst forsøg på at finde en position, hvorfra det stadig er muligt at sammenligne elevers PISA præstation med deres præstation indenfor det udvidede socio-kulturelt orienterede udfoldelses- og opmærksomhedsvindue. Dette indebærer en sammenligning af resultater fra to meget forskellige paradigmer, og da PISAs format er givet på forhånd, har VAP-evalueringen måttet indrettes, således at den *også* giver mulighed for at uddrage simplificerede kvantitative scorer, som kan sammenlignes med PISA. Dette kig på tværs af evalueringsparadigmer stiller krav til såvel test-gennemførelsen som analyse, hvilket vil blive beskrevet mere detailleret nedenfor.

6.3 Gyldighed i VAP og PISA

Validitet eller gyldighed i evaluerings- og forskningssammenhæng er et ganske sammensat begreb, med en betragtelig udviklingshistorie. Grundlæggende handler det om, at testen faktisk indfanger den type viden eller kompetence, den foregiver at udtale sig om – og er i stand til at gøre dette på dækkende vis. Harlen formulerer det således:

"Validity means how well what is being assessed corresponds with the behaviour or learning outcome that it is intended should be assessed" (Harlen, 2007, p.18)

Denne intuitivt simple definition har det imidlertid med at sløre, hvor mange forskellige aspekter der kan indgå i en validitetsbetragtning. Er fx det konstrukt, som man vil måle operationaliseret på dækkende vis? Er det fx rimeligt at reducere "livslang læring" og "scientific literacy" til nogle få del-kompetencer (jf. PISA's definition af Scientific literacy)? Eller på næste trin: er opgaverne udformet, så de rent faktisk indfanger disse kompetencer?; sker evalueringen på en måde, som tillader udfoldelse og udmåling af kompetencerne?; Er der anvendt metoder, så konklusionerne er konsistente med de data, som er indsamlet? Alle disse spørgsmål adresserer på forskellig vis dét, som nogen tidligere har kaldt *indre* validitet. Andre aspekter af validitet handler om, hvorvidt andre

velanskrevne metoder ville føre til samme resultat, om testningen loyalt tester træk som undervisningen og kulturen har stillet til rådighed for eleverne o.s.v. I erkendelse af, at de mange former med hver deres betegnelse skaber mere forvirring end godt er, har de toneangivende publikationer *Educational Measurement* og *Standards for Educational and Psychological Testing* siden 1985 anbefalet, at man kun anvender begrebet *construct validity*. De forskellige aspekter af validity henregnes så til forskellige *evidenstyper.* De mange forskellige forståelser og mulighederne for paradigmatisk perspektiv-forskydning gør det imidlertid stadig nødvendigt at overveje i hvilken forstand VAP sætter validitetspørgsmålet til undersøgelse.

PISA forholder sig primært til *construct validity* i den forstand, som er den psykometriske Rasch-analyses. dvs. som et spørgsmål om, at de enkelte items skalerer på een og samme dimension, om "item-difficulties" matcher elevsamplets formåen, samt i bedste fald også om empiriske item-difficulties modsvarer opgavestillernes forventninger til opgavernes relative sværhed, samt endelig om opgavernes relative sværhed er konsistent på tværs af lande. Med kvantitative indikatorer godtgør PISA derved, at de måler een bestemt ting, som har med sværhed *i en eller anden forudsigelig og konsistent forstand* at gøre. Dette teknisk-matematiske forsøg på at objektivisere diskussionen om validitet er typisk for en traditionel/post-positivistisk tilgang.

Det er tankevækkende, at de to mest autoritative publikationer i feltet (*Standards for Educational and Psychological Testing* ("Testing Standards") hhv. "Validity Chapter" i *Educational Measurement*[1]) parallelt hermed og siden 1985 har defineret validering på en måde, som langt overskrider PISA-forståelsen. Således hedder det i Testing Standards (American Educational Research Association, A. P. A. &. N. C. o. M. i. E. 1985, p.9), at validering adresserer:

> *"the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores"* (Testing Standards, 1985, p. 9).

Her er "appropriateness" en relativt nøgtern kategori, som til en vis grad nærmer sig psykometriens ønske om objektive mål for validitet. De øvrige kategorier er imidlertid en bevidst afspejling af, at man i forskningsverdenen i øvrigt har erkendt, at forskning (og evaluering) *ikke* lader sig gennemføre objektivt, men indeholder elementer af fortolkning. *Meningsfuldhed* og *Brugbarhed* er således ikke kvantitative kategorier, men derimod noget der skabes af subjekter og vurderes i lyset af en anvendelseskontekst. Det sidste træk forstærkes af Messick's trendsættende diskussion af forskellige evidenstyper, hvor man især skal bemærke hans indførelse af evidens knyttet til konsekvenserne af en evaluering ("*the consequential basis"*):

> *"We can [a] look at the content of the test in relation to the domain [about which inferences are drawn]…, [b] probe the ways in which individuals respond to items or tasks…, [c] examine relationships among responses to the tasks, items, or parts of the test, that is the internal structure of test responses…, [d] survey relationships of the test scores with other measures and background variables, that is the tests external structure…, [e] investigate differences in these test processes and structures over time, across groups, and settings, and in response to experimental interventions, and [f] trace the social consequences of*

[1] Testing Standards publiceres af American Educational Research Association, The American Psychological Association og National Council on Measurement in Education, mens Educational Measurement udgives af sidstnævnte samt National Council on Education. Typisk udkommer disse med 10-15 års interval med normsættende indvirkning på feltet.

interpreting and using the test scores in particular ways, scrutinizing not only the intended outcomes but also unintended side effects" (p. 16). (Messick, 1989):

Ifølge Messick omfatter en test-validering således også et blik på mulige sideeeffekter af i øvrigt planmæssigt afviklet testning. "Teaching for the test" og "test-anxiety" er eksempler på sådanne sideeffekter, som trækker fra i (konsekvens-)vurderingen af validiteten for en test der er tænkt at teste og tjene læringen i folkeskolen. Kapitel 3 i denne rapport kan ses som et forsøg på at belyse denne evidenstype, idet utilsigtede konsekvenser af intenderet PISA-testning her diskuteres tillige med konsekvenser foranlediget af politisk (mis)brug af testresultaterne. De sidste falder udenfor validitetsopgørelsen, også i Messick's forstand.

Af ovenstående skulle det gerne fremgå, at validitetsbegrebet har udviklet sig gennem de seneste tiår – parallelt med og langt henad vejen i overensstemmelse med periodens evalueringsparadigmatiske skred. Da PISA og VAP kan henføres til forskellige evalueringsparadigmatiske positioner er det ikke så overraskende, at VAP anser andre sider af validitetsspørgsmålet for relevante end dem PISA rutinemæssigt afregner på.

Vores VAP-optik deler i udgangspunktet den intuitive forståelse af validitet, som lanceredes med henvisning til Harlen ovenfor. Sammen med den bredere (evidens *og* konsekvens) og blødere (mere interpreterende/argumenterende/kontekstualiserende) tilgang til validitetsspørgsmål indikerer det den position, hvorfra VAP sætter gyldigheden af PISA i en dansk kontekst til debat.

En lang række spørgsmål validitetsspørgsmål kan rejses fra denne position. I og med at VAP ikke har haft uendelige ressourcer vil der naturligt være nogle, som vi finder væsentlige og principielt interessante, men alligevel *må undlade at adressere i VAP*. Væsentligst i denne kategori er, at vi principielt og overordnet set gerne ville have belæg for, at PISA er valid i Harlen's "face-value" forstand, nemlig at PISA-testen faktisk indfanger, hvad den intenderer at teste:

"… knowledge and skills that are essential for full participation in society." (OECD 2004b)

" …not merely in terms of mastery of the school curriculum, but in terms of important knowledge and skills needed in adult life." (OECD 1999)

Ingen kender imidlertid for indeværende sammenhængen mellem elevers naturfaglige kompetencer, som de demonstreres i skolen, og deres evne til senere i livet at kunne klare sig i situationer, som involverer naturvidenskabelige problemstillinger. Der er ganske langt fra en isoleret papir-og-blyant test til senere livssituationer, og ved at fastholde denne intention med PISA, giver opdragsgiverne i OECD deres testudviklere gevaldige validitetsproblemer, som de i alt væsentligt har undladt at forholde sig til (se i øvrigt Dolin 2005). Mere jordnært & stadig principielt ville vi i VAP-optikken gerne have belæg for, at PISA's reduktion af Scientific Literacy til et sæt af 4 statements og siden 3 kompetencer er eksternt gyldig, hvilket imidlertid vanskeliggøres af, at PISA "bryder ny grund": der foreligger således ikke evidens fra sammenlignelige"measures" i allerede etablerede tests (Messick's evidence-type [d] ovenfor). Endelig ville vi principielt gerne se en validitetsundersøgelse der godtgjorde, at PISA's opgaver faktisk operationaliserer kompetencemålene, herunder fx hvilken forståelse af *Explaining phenomena scientifically* de indlejrer. Er der fx tale om naturvidenskabelige kompetencer baseret på indholdsviden eller tale om evne til logisk-rational tænkning? Desværre har PISA-konsortiet bidraget meget lidt til at belyse disse kritiske

validitetsaspekter, og desværre falder de også stort set udenfor rammerne af VAP. Reelt giver VAP kun stof til spekulation omkring det sidste punkt.

I hvilken forstand kan VAP så siges at sætte PISA's validitet på dagsordenen?

Overordnet og formuleret i evidenstyper er det formentlig rigtigt at sige, at mens PISA vel primært beskæftiger sig med Messick's evidenstyper [c] og [e], leverer VAP evidensbidrag af typerne [a], [b] og [e].

VAP's tredje del søger først og fremmest at tilvejebringe empiri/evidens for, hvor meget testformatet/testparadigmet betyder for test-"resultatet". Det afgørende er her: *sker evalueringen på en måde, som tillader udfoldelse og udmåling af kompetencerne?* Hvor meget ændres "resultatet" (scorer m.m.), hvis *ikke* dette er opfyldt? Dette kan måske bedst ses som et bidrag til Messick's punkt [e] (*"investigate differences in these test processes and structures over …settings*). I det omfang man anerkender, at den daglige undervisning i folkeskolen langt henad vejen er drevet af et socio-kulturelt læringssyn leverer VAP hermed reelt evidens af den type, som tidligere ville være benævnt *instructional validity* ((McClung, 1979).

VAP går også med sit dialogiske og udvidede opmærksomhedsvindue længere i retning af at *" probe the ways in which individuals respond to items or tasks"* (Messick [b]), fx ved høre elevernes kommentarer og forklaringsmodeller bagom valget af MC-optioner – og ved pilotmæssigt at afdække, hvor meget ændringer i opgavernes "overflade" (som intet ændrer ved deres faglige indhold) påvirker "resultaterne".

Endelig stiller VAP sig ikke kun tilfreds med at få elevernes svar indenfor de meget snævre item-domæner, men spørger til elevernes viden indenfor det større, relaterede område af pensum. På denne måde giver VAP et bidrag til at belyse *content of the test in relation to the domain* (Messick's [a]). Selvom PISA ikke selv påberåber sig en pensumdækning, vil de fleste brugere af PISA nok forvente, at PISA's resultater udtrykker folkeskolens formåen indenfor de nært relaterede pensum-domæner, hvor sådanne findes.

Kun hvad angår det første vil VAP levere *kvantitativ* evidens. De øvrige VAP-bidrag er kvalitative og vil kunne indgå som belæg i et mere udbygget validitets-argument, af den type, som moderne validitetsteori også plæderer for (fx (Kane, 2006).

6.4 VAP-evalueringens to forskningsspørgsmål

VAP-projektets tredje forskningsspørgsmål kan på baggrund af de foregående evalueringsteoretiske overvejelser således udfoldes til følgende to forskningsspørgsmål:

> Q3a: Hvor meget vil PISA-resultatet ændres, såfremt elevernes formåen hvad angår originale PISA-opgaver efterprøves indenfor et mere sociokulturelt evalueringsparadigme med dialog, adgang til medierende artefakter og mulighed for konkret praksis?
> Q3b: Hvilket billede tegner der sig af elevernes styrker og svagheder i "det udvidede opmærksomhedsvindue", som det ændrede testformat konstituerer? Leverer PISA et dækkende billede af danske 15-åriges naturfaglige formåen?

Begrebet 'dækkende' skal her ses i lyset af rapportens diskussion af validitet.
I det følgende vil vi gennemgå de metodiske valg i forbindelse med opgaveudvælgelse, test-setup, opbygning af en faglig standard (baseline) og testgennemførelse.

6.5 Testformat

Udvælgelse af opgaver
Vi fik gennem det danske PISA konsortium adgang til den engelske udgave af opgavehæfterne til PISA2006 testen. Science-opgaverne er fordelt på 7 clusters med hver 5-6 opgaver, der hver består af typisk 3 delspørgmål (varierende fra 1 til 4). De 7 clusters er delt ud på 13 opgavehæfter, således at hver cluster indgår i 4 opgavehæfter. For at have en passende faglig bredde skulle vi gerne genteste tre opgaver, og en af dem skulle kunne omformes til eksperimentelt arbejde. Ved at vælge opgaver fra eet og samme cluster blev det muligt at finde elever, som havde mødt samtlige opgaver i den oprindelige PISA-test. Herved kunne vi minimere antallet af elever, som var nødvendige for en bæredygtig gentestning.

Vi gennemgik alle opgaverne med hensyn til deres egnethed til at inddrage eksperimenter i den behandlede problemstilling, deres egnethed til at initiere en faglig samtale, hvilke fag/fagområder, de behandler og hvilke kontekster og kompetencer de inddrog. Ud fra kriterier om faglig bredde og faglig relevans i forhold til de danske læreplaner, udvalgte vi i cluster 5 opgaverne S465 ("*Different climates"*) og S478 ("*Antibiotics"*) til den faglige samtale og S447 (*"Sunscreens"*) til øvelsen, fordi den illuderede et forsøg, som nemt og direkte kunne gennemføres af elever i praksis. Vi vil fremover referere til disse opgaver som *Klimaforskelle, Antibiotika* og *Solcreme.*

Beskrivelse af test-setup
Gentestningen af den enkelte elev (dvs. hele VAP-evalueringen) skulle af praktiske grunde kunne holdes inden for én sammenhængende to timers periode. Den blev foretaget af assistenter, som blev uddannet til opgaven (se senere). Evalueringen af eleverne faldt i første omgang i tre dele, men en fjerde del blev hurtigt tilføjet. De to første komponenter knytter sig direkte til VAP-spørgsmålene Q3a og Q3b, mens de andre er af mere testteknisk karakter. Testdel 3 var oprindelig mest en eksplorativ tilføjelse, men da undersøgelsen faktisk belyser et aspekt af PISA-testens validitet - og da resultaterne tilmed har vist sig at være overordentlig interessante – har vi valgt at tage den med i vores redegørelse.

Test 1: *Sociokulturelt funderet samtale om PISA-opgaver*
Individuelle samtaler med eleverne om de to udvalgte opgaver (*Klimaforskelle* og *Antibiotika*). For begge opgaver gælder:

1. Vi analyserede det faglige indhold af opgaverne i sammenhæng med PISAs scoringskriterier for de enkelte spørgsmål. Med dette som udgangspunkt kunne vi sikre os, at evaluerings-samtalen undervejs kom forbi de af PISA honorerede aspekter og pointer.
2. Når eleverne fik vist opgaven, startede vi med at undersøge hvorledes de tolkede opgaven. Hvad mente de at den gik ud på, hvad opfattede de som svært.
3. Med udgangspunkt heri og i opgaveformuleringerne indledte vi en faglig samtale med eleverne. Diverse medierende artefakter blev inddraget i takt med at samtalen skred frem. Formålet var at komme rundt om alle de faglige aspekter, som opgaven skal teste ifølge PISAs egen opfattelse. Vi spurgte dels på samme måde som PISA for at kunne foretage en direkte sammenligning

mellem elevperformance i VAP og i PISA og dels til den bagvedliggende forståelse af opgavens faglige krav. *Denne del muliggør valideringen af om PISA science måler den viden, som eleverne rent faktisk har (i en normal skolesituation, jfr. ovenfor).*
Endelig blev eleverne spurgt, om de havde anden relevant viden inden for det nære opgaveområde. Da de enkelte PISA-spørgsmål oftest er ganske afgrænsede, gav det mening at afdække om eleverne "brænder inde med" supplerende og relevant faglig viden.

4. Til slut blev der spurgt ind til de faglige aspekter, som man normalt dækker i Folkeskole-undervisningen i det pågældende emne, og som ikke er blevet berørt i pkt. 3.
Denne del skal afdække PISAs dækning i forhold til den danske Folkeskoles krav.

Hele Test 1 varede ca. 30 minutter. Det blev optaget på video med henblik på senere scoring. Oprindeligt troede vi, at testassistenterne kunne score løbende i et testark, så videooptagelserne kun skulle tjene som pålidelighedstjek og basis for uddybende forskning. Tilbagemeldinger fra pilot-interviews viste imidlertid, at den løbende scoring gik ud over samtalens kvalitet.

Test 2: *Praktisk opgave*
Eleverne arbejdede to og to på at løse den praktiske opgave (*Solcreme*). De arbejdede sammen, i dialog med en assistent, som skulle sikre grundlag for individuel vurdering af eleverne.
De fik udleveret en lille kasse med relevante materialer, en kort deklaration af indholdet, samt en opgaveformulering der handlede om at designe og gennemføre en undersøgelse af, hvilken solcreme der bedst beskytter mod sollys.
Opgaven *Solcreme* er offentliggjort i den danske PISA-rapport. Vi kan derfor vise den danske udgave af opgaven i Bilag 1.

Hele test 2 varede ca. 30 minutter og eleverne blev optaget på video og scoret efterfølgende (se nedenfor).

Test 3: *Opgavedesignets betydning for testresultater*
Udgangspunktet for denne test var den kritik, som har været fremført af sproget og tekstmængden i PISA-opgaverne. Marcus Puchhammer (Puchhammer, 2007) citerer den tyske PISA2000 rapport for at udregne, at 76 % af variansen i matematikresultaterne kan forklares ved faktorerne socioøkonomisk status, køn, generel kognitiv evne, matematisk selvopfattelse og læsekompetence – og at over halvdelen af denne varians kunne tilskrives læsekompetence. Puchhammer udregner desuden tekstlængde og ordalmindelighed for såvel den tyske som den engelske udgave af de offentliggjorte matematikopgaver, for at dokumentere hvorledes oversættelse øger sværhedsgraden. Her viser det sig, at de tyske opgaver i gennemsnit har 670 karakterer i sammenligning med de engelskes 583 karakterer, og at de tyske ord har en væsentlig højere vanskelighedsgrad end de engelske originalord. Danske undersøgelser tyder på at samme effekt gør sig gældende i den danske oversættelse (Henningsen 2005).
PISA-konsortiet har ganske vist bestræbt sig på at mindske tekstmængden i 2006-scienceopgaverne:

"… to more clearly distinguish scientific literacy from reading literacy, the set of PISA 2006 science test items required, on average, less reading than did the sets of science items used in the two earlier PISA surveys." (Thompson & De Bortoli, 2008, p.19)

Men vores gennemlæsning af alle 2006-scienceopgaverne afslørede ganske mange uklarheder og lange tekster, så vi bearbejdede (andre) 3 opgaver (6 items) i cluster 5, altså det cluster, som de VAP-testede elever havde regnet i PISA-testen. Bearbejdningen betød i nogle tilfælde, at informationen i stedet blev præsenteret via figurer m.m., i andre tilfælde at sekvenseringen af informationen blev ændret og i atter andre at overflødig tekstinformation blev fjernet. Mængden af faktuel information, det faglige indhold og selve PISA-spørgsmålet forblev uændret (se i øvrigt uddybning nedenfor). De "overflade-modificerede" opgaver indgik i VAP-testen i et vanligt PISA-format, dvs. som en sekvens med individuel, skriftlig besvarelse. Tiden blev tilpasset den tid, som eleverne fik til opgaverne i PISA-testen. Ca. 15 min.
Denne del skal vise i hvilket omfang opgavedesign og -layout er en hindring for nogle elevers demonstration af faglige viden.

Test 4: *Betydning af gen-testning*
De elever, som deltog i VAP, var med i den oprindelige PISA2006-test nogle få uger forinden. I den forstand havde de set og arbejdet med opgaverne tidligere, hvilket umiddelbart burde give dem en fordel og føre til et forbedret resultat. Mod en sådan forestilling taler, at eleverne i den oprindelige testsituation er blevet "tæppebombet" med opgaver i et tempo, så opgaverne aldrig har nået at bevæge sig fra elevernes korttidshukommelse til langtidshukommelsen (kilde: privat kommunikation med prof. P. Allerup, DPU). Endvidere har de ikke haft mulighed for at bringe opgavetekst m.m. med sig fra testlokalet til efterfølgende snak med kammerater, naturfagslærere m.m. Dermed har de i realiteten været afskåret fra at modtage "præstationsfremmende" feedback. Hvis en sådan udlægning holder, udtrykker det at PISA er ren "udmåling" af elevers faglige formåen i den pågældende situation – ikke en situation eleverne lærer noget af.
Med VAPs intention om at studere, hvorledes ændringer i selve testformatet påvirker/forbedrer elevernes resultat, er det selvsagt kritisk, at kunne udskille betydningen af testformatet fra den eventuelle betydning af gen-testningen. For at imødekomme dette behov har vi i VAP-setuppet indlagt 2 forskellige kontrolmekanismer:

- 2 spørgsmål om *elevernes oplevelse* af gen-testning og fordelene herved. Spørgsmålene var af multiple choice-karakter, og rent praktisk blev de stillet som afslutning på den eksperimentelle del af gentestningen.
- *En direkte PISA-autentisk gen-testning*: Yderligere et antal oprindelige opgaver (typisk 4 pr elev) blev genbesvaret i et helt PISA-autentisk format, dvs. paper-and-pencil og med en responstid pr. spørgsmål, som i den oprindelige test. Dette giver os mulighed for direkte at talsætte, hvor stor en eventuel gen-testningsgevinst i VAP måtte være.

Skematisk så det samlede testningsforløb for to elever således ud:

	Tid	Organisering	Dokumentation
Introduktion	5 min	2 elever sammen med 1 ass.	
Test 1: Individuel samtale af klimaforskelle og antibiotika	30 min	2 x (1 elev sammen med 1 ass.)	Hver opgave-session filmet m. fast kamera-opstilling
Test 2: Fælles solcremeøvelse	30 min	2 elever med 1 ass.	Manuel videooptagelse foretaget af assistent
Test 3: Re-designede opgaver	15 min	2 elever sammen med 1 ass.	Indsamlede skriftlige svarark

Test 4: Individuel gentestning	20 min	2 elever sammen med 1 ass.	Indsamlede skriftlige svarark

Opbygning af en faglig baseline og udarbejdelse af samtaleskemaer

Til brug for Test 1 og Test 2 etablerede vi en benchmarking af de tre opgavers faglige områder. Vi skrev til fagkonsulenterne for de tre fagområder og til de faglige foreninger, som henviste til nogle erfarne lærere, som vi kontaktede.
På baggrund af de indkomne svar (to-tre for hver opgave) udarbejdede vi den benchmarking, som eleverne skulle vurderes efter, og den spørgerække, som skulle følges.

Da de to opgaver (*Klimaforskelle* og *Antibiotika*) ikke er offentliggjorte, kan vi ikke vise hele denne proces og den endelige spørgeguide, men vi havde for hvert emne en række faglige sammenhænge og faglige begreber, som vi skulle sikre os at eleverne blev spurgt ind til. PISA-opgaven strukturerede tidsfølgen, og ved hvert delspørgsmål var angivet de relevante begreber og sammenhænge, som fagfolkene havde sammensat. Assistenterne skulle sikre sig at elevernes svar på PISA-spørgsmålene blev klare, og at eleverne kendskab til og inddragelse af de tilgrundliggende begreber og sammenhænge blev afdækket. Ud over PISA-spørgsmålene blev der desuden formuleret spørgsmål i de dele af emnet, som PISA-spørgsmålet ikke dækkede, men som fageksperterne havde sagt indgik i kravene i Fælles Mål.
Når vi taler om opfyldelse af Fælles Mål skal man være opmærksom på at Fælles Mål udgør et for Danmark helt særegen slags styringsdokument. Alle andre skolesystemer, vi kender til, har curriculumdokumenter som beskriver hvad eleverne skal kunne efter endt undervisning, typisk i form af idealmål inden for forskellige områder. Fælles Mål angiver derimod *undervisningsmål*, dvs. *mål for, hvad undervisningen skal lede frem mod, at eleverne skal kunne*. Men hvad siger denne sætning andet end at eleverne skal kunne noget (læringsmålene), og undervisningen skal gøre det muligt for dem at lære dette, så undervisningsmål og læringsmål må være meget tæt på hinanden. Ellers må undervisningsmål opfattes som standarder for undervisningen - og det skal så være undervisningen, der evalueres ved fx afgangsprøverne. Men Fælles Mål skriver fx om brugen af Fælles Mål: ”Lærerne … evaluerer elevernes læreprocesser og resultater i forhold hertil.” (http://pub.uvm.dk/2006/faellesmaal/faelles_maal.pdf s.4)
Vi har i VAP-evalueringen undersøgt i hvilket omfang der er overensstemmelse mellem PISAs måling af elevernes kunnen og graden af opfyldelse af de mål som folkeskolens fagfolk opstiller som relevante i termer af Fælles Mål.

I Bilag 2 er vist samtaleskemaet og assistentinstruktionerne for *Solcreme*. Som man kan se, er der tale om en ganske detaljeret styring af samtalerne for at kunne sikre en høj pålidelighed i den efterfølgende scoring. Samtidig er der taget højde for at såvel alle PISA-testens som Fælles Mål - kravene blev berørt.

6.6 Dataindsamling

Logistik

I slutningen af 2005 begyndte vi at forberede vores gentestning.
Vi fik hjælp af Thomas Yung, SFI, til en række testtekniske forhold, fx afklaring af at et samlet elevsample på ca. 120 elever ville være nok til vores formål. Vi fik også adgang til et stratificeret

sample på 168 elever på 33 sjællandske skoler, som alle havde besvaret de udvalgte PISA2006-opgaver.
Vi udvalgte 6 lærerstuderende (af 11 ansøgninger) som havde et naturfag som linjefag, og uddannede dem i løbet af februar-marts 2006 til at kunne gennemføre VAP-evalueringen.
Finn Horn, DPU, blev hyret til at kontakte skolerne, som skulle samle de ønskede elever til gentestning 2-8 uger efter PISA-testen og samtidig sørge for to nærtliggende lokaler, hvoraf et kunne bruges til den eksperimentelle øvelse. Finn udarbejdede også besøgsplan for assistenterne.
Vi indkøbte videoudstyr og relevante artefakter (glober, atlas, (UV)lysfølsomt papir, solcreme med forskellig solfaktor, petriskåle, pipetter etc.) og samlede seks kufferter med videoudstyr og samtalemateriale.

Uddannelse af assistenter og pålidelighedstjek

Hele projektet stod og faldt med assistenternes evne til at foretage en faglig solid samtale med eleverne efter fælles retningslinjer. Vi gjorde derfor meget ud af assistenternes faglige fundering og deres samtaleforståelse, og at de fulgte samtaleskemaet på samme måde. Det er en ganske svær balance at skulle gennemføre en 'naturlig' samtale om et fagligt emne og samtidig sikre at bestemte elementer gennemføres, således at der efterfølgende er mulighed for at score samtalerne efter den samme manual.
Vi gennemgik først den opbyggede benchmarking for de tre faglige områder og gav nogle generelle retningslinjer for god samtale, der tog udgangspunkt i Olga Dysthes (2000) begreber autentiske spørgsmål, optag og høj værdisætning. Så lod vi assistenterne gennemføre samtaler med hinanden som elever samtidig med at de øvede videooptagelse. Herefter gennemførte en af assistenterne et fuldt testforløb med en indforskrevet 15-årig (tak Stine!). Det blev optaget på video og efterfølgende diskuteret i hele gruppen.
På baggrund af erfaringer herfra måtte vi udarbejde en revideret samtaleguide for at sikre en højere styring af samtalen med henblik på at øge pålideligheden. Vi droppede også den løbende scoring af eleverne for at gøre samtalen mere dialogisk og mindre vurderende. Endelig måtte vi sikre et større fagligt overskud hos assistenterne, så de fik udleveret supplerende fagligt materiale, og vi gennemgik nogle centrale faglige problemstillinger med dem.

Assistenterne testede parvis elever på ca. 12 skoler per par. I den periode, hvor assistenterne testede på skolerne, mødtes alle hver anden-tredje uge for at diskutere og harmonisere testforløbet. Dette foregik ved at holdene på skift viste optagelser for hinanden af samme testsituation, som vi så sammenlignede og analyserede for at optimere og harmonisere samtaleformen. Ved det første møde efter teststart gav det fx anledning til følgende ændringer (uddrag af brev fra projektledelsen til assistenterne):

> "Det var en nyttig gennemgang af **øvelsesopgaven**. Og også lovende for projektet. Jeg husker følgende som de meget overordnede forhold vi (dvs. I) skulle være opmærksomme på:
> 1. Få afklaret hvorvidt eleverne forstår problemet (altså at de skal teste hvor godt forskellige solcremer virker) – og hvis de ikke gør det, så forklar dem det.
> 2. Få afklaret hvorvidt eleverne forstår forsøget (altså ideen i forsøget og forsøgsgang/plan) – og hvis ikke så gennemgå det med dem.
> 3. Sørg for at der et grundlag for at vurdere hvorvidt de har forståelse for anvendelse af referencer og hvorvidt de kender ordet (hvad der ikke er så vigtigt i sig selv).
> 4. Sørg for at der et grundlag for at vurdere hvorvidt de har forståelse for anvendelse af variabelkontrol (her: at de varierer via forskellige cremer, men holder tykkelse og eksponeringstid konstant).
> 5. Spørg til hvad pletternes tykkelse/areal betyder. Fx: 'hov, når I trykker klatterne flade, så bliver de jo dobbelt så store – betyder det ikke noget?').

6. Inddrag elevernes erfaringer i samtalen: 'bruger I selv solcreme? Hvilken faktor? Forskellige faktorer forskellige steder på kroppen?' etc.
7. Sørg for at der er basis for at vurdere begge elever. Især hvis der er én der dominerer, er det vigtigt at inddrage den anden: 'er du enig?' 'hvad mener du?' etc.
8. Prøv at fylde så lidt selv i samtalen, men giv plads til eleverne.

Vi er nødt til at **udvide evalueringskonceptet** med en almindelig opgavetest for at kunne korrigere for elevernes erindring om de opgaver, vi bruger. Det forgår efter vedhæftede procedure. I vil i starten af næste uge få sendende opgavesættene. Udfyldelsen af dette sæt (bestående af tre opgaver, som kan være forskellige for forskellige elever) skal tage 20 minutter.
Dvs. en samlet evalueringsrunde består nu af:

Fælles introduktion	1 minut
Fælles solcremeøvelse	30 minutter
Individuel samtale	30 minutter
Individuel bevarelse af strippede opgaver	15 minutter
Individuel besvarelse af opgavesæt	20 minutter"

Et par af de første VAP-testninger levede ikke helt op til de udstukne normer og de blev udeladt af databehandlingen. Der var enkelte assistenter, som havde svært ved at opfylde normerne, så vi måtte gennem hele forløbet arbejde med at optimere pålideligheden og samtaleformen.

Samling og registrering af data

Med jævne mellemrum blev de optagne videobånd og opgavebesvarelser samlet sammen og registreret. Videooptagelserne blev til sidst overført på en harddisk, som blev kopieret i tre eksemplarer med henblik på senere scoring.
Der blev opbygget en excelfil med alle gentestede elevers navn, skole og cpr.nr. til indføring af alle testresultater.
Samtlige data, dels videooptagelserne, dels elevernes besvarelser og dels de scorede elevpræstationer, opbevares på de involverede forskningsinstitutioner.

6.7 DATAPRODUKTION

Dataindsamlingen i form af videooptagelser og elevbesvarelser foregik i løbet af foråret 2006, og herefter kunne det store arbejde med databehandlingen og yderligere datakonstruktion begynde.

Klimaforskelle

Scoringen af klimaopgaven blev foretaget af stud.scient. Ellen Berg Jensen, som skrev speciale i geografididaktik ved Københavns Universitet (Jensen 2007), i samarbejde med en af projektets forskere. Videooptagelserne af samtalerne om klimaopgaven udgjorde empirien for opgaven. Scoringen tog udgangspunkt i det udviklede samtaleskema. Ud over faktuelle elevoplysninger blev hver elev scoret på 19 spørgsmål/områder. Hvert spørgsmål/område indeholdt en forklaring, en scoringskode og en vejledning i brug af koden. Som eksempler kan nævnes:

Anvendelse af [geospecifik][1]figur: Ved eleverne, hvad figurerne kan bruges til og kan de aflæse den? Her vurderes det, om en [geospecifik]figur er noget, de kender til og kan bruge. Har de en viden, uden at intervieweren hjælper dem? Score: 1: ja 2: noget 3: nej 9: mangler info.

[1] Navnet på denne figur kan ifølge Skolestyrelsen ikke offentliggøres.

(1: Gives når eleven ved, at figuren viser [klimavariabel1][1] og [klimavariabel2], samt at kurven er [klimavariabel1] og søjlerne [klimavariabel2] og de kan aflæse den. 1 gives også hvis de bytter om på hvad søjlerne og kurven viser, men de selv kan se, at det er galt og får det rettet.
2: Gives når eleven har kendskab til figuren, men bliver ved med at bytte rundt på, hvad kurven og søjlerne viser eller aflæser både [klimavariabel1] og [klimavariabel2] på [klimavariabel1]skalaen.
3: Gives når eleven ikke ved, hvad figuren viser og ikke selv ræsonnerer sig frem, men får det forklaret af interviewerен.)
Forståelse af [klimatype1]- og [klimatype2]klima[2]: Det er ikke altid, at eleven selv forstår disse termer, selvom vedkommende nævner dem. Derfor undersøges det, om eleven viser forståelse af termerne [klimatype1]- og [klimatype2]sklima. Scoring: 1: ja 2: noget 3. nej 9: mangler info
(1: Eleven forstår og forklarer forskellen på de to typer af klima. Dvs. i tilfælde hvor eleven ikke bliver spurgt om kendskab til termerne, men har svaret og forklaret korrekt på ovenstående spørgsmål gives fuld kredit.
2: Forklaringen er ufuldstændig og der hersker tvivl om eleven forstår begreberne til fulde.)

Forståelse af klima-forskelle: Her testes om eleverne opfylder kravene i Fælles mål. Elevernes viden vurderes ud fra forskellige parametre hentet i Fælles Mål og fagfolks udtalelser. Blandt andet ud fra om de kender og kan angive, samt vise beliggenheden af forskellige klimazoner, temperaturværdier for disse, plantebælter, hvad højden over havet betyder for klimaet etc. Den samlede viden vurderes ud fra en skala fra 1 til 5, hvor 5 er bedst.)

Som det ses, anvendes mange forskellige scoringsformater. Typisk har VAP-formatet flere svarkategorier i de spørgsmål, som er sammenfaldende med PISA-testen, fx i form af en mellemkategori ”nogen”, hvilket muliggør en mere nuanceret vurdering end PISA-testen gør. Fx indgik en vurdering af elevernes selvstændighed som en scoringsparameter. Desuden scores de spørgsmål, som vurderer elevernes opfyldelse af mere komplekse faglige mål, som hovedregel efter en femtrinskala.

Ellens Bergs scoring blev reliabilitetstjekket af en af projektets forskere. Syv tilfældigt udvalgte elever blev scoret af Ellen og forskeren uafhængigt af hinanden. Der viste sig ret fin overensstemmelse i de fleste spørgsmål. På de ’faktuelle’ spørgsmål var der fuld overensstemmelse. På femtrinskalavurderingerne var der typisk ét skalatrins forskel (i samme retning, hvor forskeren konsekvent lå lavere) ved 3-4 af de 7 scorede, hvilket afspejlede forskellige opfattelser af hvilke krav man kan/bør stille til eleverne.

For disse syv første scoringer var der følgende afvigelser mellem scoreren og forskeren:
I de tre spørgsmål, hvor der var tre scoringstrin, var der uoverensstemmelse på et trin i fire af de 21 scoringer (svarende til Cohen’s Kappa 0.71).
I de 10 spørgsmål, hvor der var fire scoringstrin, var der uoverensstemmelse på et trin i 25 af de 70 scoringer (svarende til Cohen’s Kappa 0.53).
I de 6 spørgsmål, hvor der var fem scoringstrin, var der uoverensstemmelse på et trin i 15 af de 42 scoringer (svarende til Cohen’s Kappa 0.56).

Ellen og forskeren så videoen af den elev, hvor der var den største uoverensstemmelse, sammen, og blev enige om hvilke score der var den rigtige. Dette ændrede Ellens vurdering på enkelte områder, men havde som et vigtigere udbytte, at der skete en harmonisering af forståelsen af de få kategorier, hvor der var uoverensstemmelse.

Den efterfølgende stikprøvekontrol viste næsten 100 % overensstemmelse mellem Ellens og forskerens vurderinger.

[1] Benævnelserne på de to klimavariable må ikke offentliggøres.
[2] Benævnelserne på de to klimatyper må ikke offentliggøres.

Antibiotika

Scoringen af biologiopgaven blev foretaget af lærerstuderende Elizabeth Juhler i samspil med en forsker. Elizabeth var også assistent og skrev sin bacheloropgave under inddragelse af VAP-resultaterne. Scoringsforløbet var magen til forløbet ved geografiopgaven. Der blev scoret 12 områder og en forsker sikrede løbende pålideligheden efter samme metode som ved opgaven om klimaforskelle.

Solcreme

Denne opgave er scoringsmæssigt mere kompleks og metodisk mere udfordrende end de foregående. To forhold er af afgørende betydning for dette: for det første opgavens *praktiske* karakter, og dernæst, at eleverne *arbejder i par*. Det praktiske islæt betyder principielt, at praktisk kompetence/tavs viden i princippet bør indgå i bedømmelsesgrundlaget. Ved at vælge videooptagelse som empirisk grundlag har vi fastholdt muligheden for at studere ikke-verbale aspekter, hvorved vi i princippet vil kunne håndtere denne type kompleksitet. Pararbejdet frembyder imidlertid anderledes metodisk kompleksitet og større principielle problemer for den direkte PISA-sammenligning: PISA opererer med *individuel accountability*, altså at man på pålidelig vis kan vurdere *den enkeltes* præstation – hvorimod VAP-formatets pararbejde med størst pålidelighed tillader, at man udtaler sig om *elevparrets samlede formåen*. I visse situationer kan man af videoen følge, hvorledes eleverne på lige fod og synkront lægger de relevante brikker på plads, hvorfor man med stor overbevisning kan bryde parbedømmelsen ned til at begge bør have fuld score. I andre situationer vil der være en faglig asymmetri, som gør at den ene elev bidrager med mest, mens den anden hænger på, men efterlader én med indtrykket af, at den selvstændige præstation ville være blevet ringere. Hvilken score skal vedkommende så have i den direkte PISA-sammenligning? Endelig vil der være situationer, hvor fx pardynamikken i det nyetablerede elevsamarbejde foranlediger den ene elev til at være mere stille end den anden. Her vil der være direkte faglige udsagn og handlinger til at bedømme den udfarende, men kun nikken, berigtigende småord og vagere indikatorer som grundlag for bedømmelse af den anden. Endelig er der samspillet med forskningsassistenten og de medieringsmuligheder, som ligger heri og udnyttes forskelligt (og som deles med interviewformatet fra de foregående opgaver).

For bedst muligt at beskrive og imødegå dette iboende problem har vi udviklet et kodningssystem, som inddrager *selvstændighed*, *mediering* (peer, forskningsassistent), *usikkerhedsangivelse* i tildelingen af en individuel score, samt muligheden for at helt at undlade at tildele en score – i situationer, hvor eleven faktisk gennemfører pararbejdet, men på en måde, så usikkerheden i bedømmelsen bliver for stor. Udviklingen af et pålideligt scoringssystem i denne sammenhæng har været et mindre forskningsarbejde i sig selv, men hovedtrinnene har været:

- Fase 1: Indkredsning af problemfelter i forhold til PISA-sammenlignende scoring. Til dette formål er et antal videoer (10 stk) blevet gennemset med henblik på PISA-sammenlignende scoring (jf. oprindelige scoringsark).
- Fase 2: Udarbejdelse af foreløbigt kodesystem
- Fase 3: Afprøvning og tilpasning af kodesystemet i dialog mellem flere kodere og med løbende inter-rater reliabilitetscheck

I Bilag 4 er der detailleret redegjort for kodningsprincipperne for de enkelte delspørgsmål.

Da først et optimalt kodningssystem var etableret, blev scoringen foretaget af en forskningsassistent i dialog med én af projektets forskere. Scoringens pålidelighed blev sikret ved, at i alt 15 videoer blev scoret af de to aktører uafhængigt af hinanden. Overensstemmelsen mellem de to sæt af scorer blev opgjort via inter-rater-korrelation og Cohen's Kappa. Talværdien for begge blev (tilfældigt) 0,73. Typisk anser metodelitteraturen værdier af Cohen's Kappa i lejet > 0,70 for acceptable og substantielle. Det er i den grad betryggende, at scoringen i så metodisk udfordrende en situation alligevel tilfredsstiller gængse kriterier.

Til sidst blev de foretagne scoringer omregnet til indeksværdier for hvert af de fire PISA-delspørgsmål, så sammenligning med de tilsvarende PISA-testværdier blev mulig.
Her er S447Q03 speciel ved *ikke* at kunne transformeres til praktisk aktivitet. Eleverne har derfor besvaret denne del af opgaven *i det oprindelige PISA-format* i en pause undervejs i det praktiske forløb. Reelt bliver dette item således en gentestning af eleverne.
Hvorledes de øvrige *Solcreme*-spørgsmål er indekseret er beskrevet i Bilag 5.

7. Dataanalyser og resultater

Det samlede datamateriale giver som tidligere omtalt mulighed for en række forskellige analyser.

Test 1 og 2 udgør den egentlige validering. Til brug for den direkte PISA-sammenligning registreres 19 aspekter af elevers viden og videnudfoldelse inden for området klimaforhold, 12 aspekter inden for området antibiotika og 34 aspekter knyttet til arbejdet med den eksperimentelle opgave. Disse modsvarer enkeltvist eller i kombination (i form af "indeks-scorer") de 4 item-scorer i PISA-opgaven S465 (*Klimaforskelle*), de 4 item-scorer i PISA-opgaven S478 (*Antibiotika*), samt de 4 item-scorer i PISA-opgaven S447 (*Solcreme*). Sammenligningen af hvorledes elever klarer disse 12 opgaver i VAP-gentestningen og PISA-testen udgør svaret på forskningsspørgsmål *Q3a: Hvor meget vil PISA-resultatet ændres, såfremt elevernes formåen hvad angår originale PISA-opgaver efterprøves indenfor et mere sociokulturelt evalueringsparadigme med dialog, adgang til medierende artefakter og mulighed for konkret praksis?*

Resultaterne fremlægges og analyseres på to forskellige måder, nemlig dels i form af en sammenligning af de rå opgave-scorer i VAP-testen og PISA-testen, og dels i form af en sammenligning af de tilsvarende Rasch-scorer, som muliggør en perspektivering i forhold til hele PISA-projektet.

Derudover er der lavet en mere omfattende sociokulturel analyse af elevernes præstation ud fra et antal af videoerne. Sammen med alle de registrerede aspekter giver det mulighed for at svare på forskningsspørgsmål Q3b: *Hvilket billede tegner der sig af elevernes styrker og svagheder i "det udvidede opmærksomhedsvindue", som det ændrede testformat konstituerer? Leverer PISA den fulde sandhed (endsige den væsentlige del af sandheden)?*
Disse mere kvalitative resultater fremlægges til sidst.

Test 3 og 4 afdækker i hvilket omfang elever hæmmes af opgavernes design, fx tekstmængde, og i hvilket omfang gentestningen påvirkes af at eleverne tidligere har regnet opgaverne under PISA-testen. Resultaterne heraf fremlægges først.

7.1 Drager eleverne fordel af at VAP er en gen-testning?

Nogle af de første elever, der blev VAP-testet, gav udtryk for at kunne genkende opgaverne. Vi fandt det derfor nødvendigt at tjekke, om eleverne kunne huske opgaverne i et sådant omfang, at det påvirkede deres performance i VAP-testen. Vi spurgte eleverne om deres grad af genkaldelse, og gennemførte derefter vores test 4, hvor vi testede dem i et antal opgaver helt identiske til de, de havde besvaret i PISA-testen.

Elevernes oplevelse af gentestning og evt. fordele heraf

Vi har spurgt generelt til elevernes genkaldelse af opgaverne og deres oplevelse af at kunne bruge de forudgående PISA-svar i forbindelse med VAP-testen. Fordelingen af elevsvar (N=113) fremgår af nedenstående tabel:

	Kan du huske opgaven fra PISA-testen?	*Føler du, at du har kunnet bruge dine svar fra PISA-testen til at løse denne*

		opgave?
NEJ, Overhovedet ikke	22	53
I nogen grad	57	39
JA, klart	34	21

Tabel 1. Elevers genkaldelse af PISA-opgaverne ved VAP-gentestningen.

Langt størstedelen af eleverne (91 af 113, dvs. ca. 80 %) angiver at have *nogen* eller endda *klar* erindring om gentestens opgaver. Spørgsmålet er så, om det er en erindring om overfladetræk ved opgaverne eller om der er tale om en mere indholdsorienteret genkaldelse. Af elevernes svar på spørgsmålet *Føler du, at du har kunnet bruge dine svar fra PISA-testen til at løse denne opgave?* fremgår det, at 53 % af eleverne mener at have haft klar eller nogen fordel af at have set opgaverne tidligere. Hvis man således skal tage elevernes udsagn for pålydende, bør VAP korrigere for en sådan gen-testningsfordel – inden effekterne af testformatet i øvrigt udlægges. Heldigvis giver *den direkte PISA-identiske gen-testning* mulighed for at vurdere *om* elevernes vurdering holder og i bekræftende fald, hvor stor en sådan korrektion skal være.

Den faktiske fordel ved at have set opgaverne tidligere – resultater af den PISA-identiske gen-testning.

Opgaver fra den oprindelige PISA-test blev fotokopieret, så eleverne kunne arbejde med dem præcist som i PISA2006-testgangen. Opgaverne var udvalgt så diverst som muligt indenfor de booklets, som VAP havde valgt at fokusere på, hvorfor der både var spørgsmål af lukket og åben art. De åbne spørgsmål blev scoret i henhold til PISAs egne scoringskriterier. De resulterende "rå scorer" blev herefter sammenlignet med elevernes oprindelige PISA2006-scorer.
Resultatet blev, at 70 % af elevernes svar på enkelt-spørgsmål var uforandrede, 17 % forbedrede og 13 % forringede. Der er altså tale om en næsten symmetrisk fordeling, hvilket indikerer, at den faktiske fordel er minimal. *Når man kun medtager datasæt, hvor eleverne med vished har afgivet svar i både PISA og VAP er der tale om en forbedring på sølle 1 % i den rå elevscore!*
Som en sidste kontrol blev elevernes to test-resultater underkastet en *Paired T-test* i SAS-programpakken, og resultatet blev, at man med 90 % sandsynlighed må sige, at der ingen forskel er. *På trods af elevernes oplevelse af genkendelighed af opgaverne og anvendelighed af dette kendskab, er det altså ikke muligt at påvise, at det har nogen reel indflydelse på deres test-resultat.*
Med denne argumentation vil vi herefter glemme alt om gentestningseffekter – og i stedet relatere ændringer i elevernes test-resultat til VAPs brug af andre test-formater.

7.2 Opgavedesignets betydning for elevpræstationer

På eet niveau kan man sige, at VAP-undersøgelsen afdækker, hvor meget PISA-resultaterne afhænger af det valgte test-paradigme og det konkrete testformat. Som et sidespor i samme retning har vi også ønsket at få et indtryk af, hvor følsomt test-resultatet er overfor konkrete *design-træk* ved selve opgaverne. Dette sidespor har vi valgt at forfølge på eksplorativ vis ved at ændre en række designtræk ved tre opgaver fra booklet 5. Selve opgaveindholdet og de afgørende spørgsmål var derimod uændrede, ligesom testformatet i denne del af VAP-prøven var helt igennem PISA-autentisk (individuel, paper-and-pencil, tidsknap o.s.v.). De indarbejdede ændringer omfatter:

- Opgaven *Skeer* (S256Q01) er en multiple choice opgave om varmeledning. I originalen er den udelukkende båret af 2 liniers tekst, efterfulgt af de forskellige valgmuligheder. Faktisk er teksten *for minimal,* idet visse forhold for betydning for besvarelsen ikke ekspliciteres.

I VAP-bearbejdningen har vi tilføjet en figur, som en alternativ grafisk repræsentation af problemstillingen. Figuren anskueliggør og præciserer samtidig ordløst de aspekter, som den oprindelige opgave undlod at præcisere.

- Opgaven *Gode Vibrationer* (S131Q02/S131Q04) er en åben-svar opgave, som i originalen indeholder 15 liniers tekst. Teksten er placeret som en optakt i to afsnit, hvorefter følger de to spørgsmål med åbent respons format. Selv teksten indeholder både en række facts om hørelse og en beskrivelse af et forsøg med insekter. Spørgsmålene henviser eksplicit til bestemte linier i hvert sit tekstafsnit, men afslører ikke at det er tilstrækkeligt at holde sig indenfor det pågældende afsnit, når svaret skal syntetiseres. I vores bearbejdede version har vi reduceret *tekstmængden* og *etableret en endnu tydeligere relation mellem den enkelte opgave og den tilhørende tekst.* Konkret minimerede vi tekstmængden til godt 8 linier, uden *i øvrigt at ændre på tekstens faktuelle indhold*. Dette gjorde vi, fordi især de tidlige PISA-opgaver blev kritiseret for at være for tekstrige og for at teste læsefærdighed i højere grad end scientific literacy. Derudover brød vi teksten op, så det blev tydeligt for eleverne, hvilket tekstsegment der skulle bruges til at besvare hvilket spørgsmål. Til gengæld droppede vi den direkte liniehenvisning i hvert delspørgsmål. Alt andet lige ville man forvente, at dette gør det nemmere for eleverne at nå frem til selve opgaverne og at overskue grundlaget for den enkelte opgave.

- Opgaven *Plastikalderen* består af tre delopgaver (benævnt hhv. S413Q06, S413Q04 og S413Q05). Fælles for opgaverne er, at eleverne ud fra tre forskellige data-tabeller over plastmaterialer fysisk-kemiske egenskaber, skal forholde sig til stoffernes opførsel og anvendelighed i forskellige situationer. Alle delopgaver er af lukket respons-type, men kun den ene er en simpel multiple-choice opgave.
    a. Den første delopgave har vi udvidet med en figur, som indeholder originalens tabel-data OG skitserer den efterprøvningsproces, som eleverne skal udtale sig *om udfaldet* af.
    b. I den anden delopgave har vi arbejdet med at forsimple den teknisk-naturvidenskabelige sprogbrug i tabellen. Sammen med den større nærhed til hverdagssprog er lix nedbragt.
    c. Spørgsmålet er et MC-spørgsmål, som vi har ladet stå uændret. Men: vi har sikret os, at eleven fastholdes lidt længere og lidt mere refleksivt ved opgaven, idet vi har *bedt eleverne fortælle, hvordan de nåede frem til det valgte svar.*

	S256Q01 (Skeer)	S131Q02T (Vibrationer)	S131Q04T (Vibrationer)	S413Q06 (Plastik)	S413Q04T (Plastik)	S413Q05 (Plastik)
Antal m. ringere præstation i VAP	5	8	16	4	10	4
Antal m. uforandret præstation i VAP	89	71	70	52	64	73
Antal m. bedre præstation i VAP	4	19	12	42	24	20
Konklusion: Hvilken version har eleverne	Ingen forskel	VAP signifikant nemmest	Ingen forskel	VAP signifikant nemmest	VAP signifikant nemmest	VAP signifikant nemmest

klaret bedst? (Paired T-test signifikans-niveau)	p=0.74	p=0.049 (*)	p=0.45	p=0.0001 (***)	p=0.016 (*)	p=0.014 (*)

Tabel 2. Elevpræstationer i de re-designede opgaver i forhold til PISA-formuleringerne. (* angiver signifikans på p<0,05 niveau)

Af Tabel 2 fremgår det særdeles bemærkelsesværdige*, at ændringer på opgavernes "overflade"-niveau faktisk i 4 af de 6 tilfælde giver signifikante ændringer af test-resultatet.* Denne lille eksplorative undersøgelse dokumenterer således, at underordnede og principielt ligegyldige design-træk ved opgaverne kan få stor betydning. I sin mest vidtgående form problematiserer dette i høj grad, hvad en sådan test måler: er det indsigt i et naturvidenskabeligt indhold eller simpelthen evne til at tolke og overkomme en mere eller mindre tilfældig opgave-indpakning?

I alle tilfælde, hvor der er sket en forandring, er det interessant, at eleverne klarer sig bedst med *VAP-formuleringerne.* Helt overraskende er dette ikke, al den stund vi bevidst har søgt at foretage *hensigtsmæssige* design-ændringer. I den forstand er det måske mest tankevækkende, at opgaven 'Skeer' *ikke* blev synligt forbedret ved at få tilføjet ekstra semiotiske ressourcer. Det er dog en mulighed, at den uforandrede præstation skyldes "mætning", idet 89 % af eleverne allerede havde svaret rigtigt i denne opgave i PISA-testens oprindelige version. Evt. forbedringer har tilsyneladende ikke været vidtgående nok til at fungere for den marginale gruppe af fejlsvarere i den oprindelige test. I de øvrige opgaver har fra 30 % til 70 % af eleverne forlods svaret rigtigt, hvilket giver synligt "plads til forbedring/-værring".

Hvilke design-forbedringer har så gjort det nemmere for eleverne at besvare opgaverne? Mønstrene er bestemt ikke nemt gennemskuelige. Ifølge vores forhåndstænkning skulle *begge* delspørgsmål i opgaven om *Vibrationer* være blevet tættere sammenknyttet med opgaveteksten, som i begge tilfælde både var omformuleret og reduceret. Da præstationen kun bliver bedre i det første spørgsmål, synes tekstmængden ikke at være det kritiske aspekt i denne opgave. Den absolut mest markante VAP-forbedring opnås i plastikspørgsmålet S413Q06 (Plastik). Her ligger den væsentligste forandring efter vores udgangstænkning i, at data og forsøgssituation angives med en figur. *Hvis* det er figuren, som udgør den reelle faciliterende omstændighed, rejser det spørgsmålet, hvorfor man så ikke ser en tilsvarende effekt i opgaven med *Skeer*, hvor den afgørende ændring efter hensigten også var knyttet til brugen af en figur? Er det subtile detaljer ved figurerne som afgør deres succes, eller er der snarere tale om figurer er særligt nyttige i forbindelse med bestemte opgavetyper/-krav og/eller i forhold til bestemte elevkategorier? Fx kunne man opstille hypotesen, at restgruppen af fejlsvarere i opgaven *Skeer* er for fagligt svag til at tilegne sig informationer fra en teknisk orienteret figur. I så fald er det tendensen til mætning og en differentiel eleveffekt, som til sammen begrunder, at figuren ser ud til at gavne i denne opgave. Svarfordelingerne for de to andre Plastik-spørgsmål antyder, at den teknisk orienterede sprogbrug faktisk øger opgavekompleksiteten for eleverne, samtidig med at et simpelt krav om, at eleverne faktisk forklarer, hvorfor de svarer, som de gør, tilsyneladende virker "præstationsfremmende". I analysen her, har vi kun medregnet elevernes afkrydsning, og altså ikke forsøgt at checke kvaliteten af deres nedskrevne forklaringer. Vi har heller ikke haft mulighed for at fastlægge elevernes faktiske tidsforbrug på opgaven + forklaring. Det får således stå åbent, om forbedringen skyldes længere fastholdelse omkring denne opgave – eller om forklaringskravet sikrer en mere intensiv bearbejdning af opgavens problemstilling.

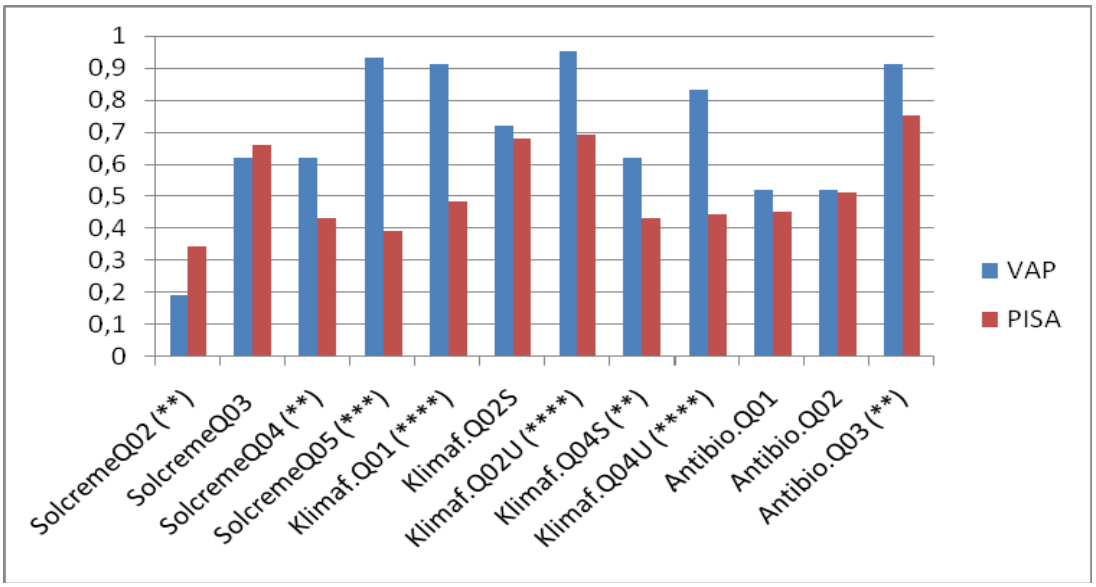
Dette lille eksplorative undersøgelsesintermezzo har i høj grad synliggjort, at djævelen meget vel ligger i test-detaljen. Det efterlader flere væsentlige spørgsmål end svar – men antyder i høj grad, at testresultater er skrøbelige og nemt vil kunne overfortolkes.

7.3 Sammenligning af elevsvar i VAP-testen og PISA-testen

Den direkte sammenligning mellem elevernes præstationer i PISAs testformat og i VAP-formatet kan foretages på mere eller mindre avanceret vis. Indledningsvist vil vi foretage en simpel sammenligning af elevernes rå item-scorer indenfor de to formater. På grundlag af dette vil man på gennemskuelig vis kunne diskutere både generelle tendenser og item-specifikke forhold. Imidlertid hviler en del af PISAs politiske succes på evnen til at reducere kvalitetsmålingen af Scientific Literacy til eet tal, den nationale Rasch-score. Indenfor denne forståelse giver det rigtig god mening at vurdere, hvor mange Rasch-point den samlede præstation flytter sig som følge af det ændrede testformat. Til sammenligning er usikkerheden på en typisk lande-Rasch-score ca. +/-3 - og 15 rasch-point udgør forskellen på, om et land ligger signifikant under hhv. over OECD-gennemsnittet i PISA Science 2006. Vi vil derfor også foretage en Rasch-analyse på VAP-dataene med henblik på at fastslå, hvor meget den samlede Rasch-score ændres.

Sammenligningen af rå item-scorer

Figur 1 viser for samtlige VAP-testede PISA-spørgsmål, hvorledes eleverne har klaret sig i de to forskellige test-setup.



Figur 1. Sammenligning mellem elevernes gennemsnitlige score i VAP-testen og i PISA for forskellige opgaver.

Et par ord til deklaration af figuren: de blå søjler angiver VAP-resultater, mens røde søjler repræsenterer elevernes oprindelige præstation i PISA2006. Søjlehøjderne angiver, hvor stor en andel af eleverne, som har fået fuld score (1 svarer til 100 %). Som tidligere omtalt er adskillige af VAP-scorerne i udgangspunktet beregnet som indeksværdier, dvs. ved at lægge flere delscorer sammen. For at kunne sammenligne elevpræstationerne i VAP og PISA er slutscoren for samtlige

spørgsmål blevet re-normaliseret, så *Fuld Score* modsvarer et 1-tal. Derfor kan man også sige, at en søjlehøjde udtrykker den gennemsnitlige elevscore i spørgsmålet.

En opmærksom læser vil hurtigt bemærke, at spørgsmålene *Klimaf.*Q02 og *Klimaf.*Q04 hver afbildes i en 'S' og en 'U'-variant. Reelt er det dog kun VAP-søjlen, som er forskellig i de to varianter. Varianterne er taget med her, fordi de meget godt illustrerer, hvorledes valget af scoringsprincipper *også* influerer resultatet. VAPs 'S' og 'U' søjler repræsenterer to forskellige måder at score en elevs faglige præstation i VAP-samtalen på, begge forekommer rimelige – og dog kan man argumentere for, at de henholdsvis må anses for hårdere og mildere end PISAs egen. Dén VAP-værdi, som på mest fair vis kan sammenlignes med PISA, ligger rimeligvis et sted mellem de to ekstremer 'S' og 'U'. I 'U'-varianten får eleven *i løbet af samtalen* mulighed for at forholde sig til de 4 konkrete svaroptioner som indgik i de oprindelige PISA-opgaver. Hvis eleven vælger korrekt svar tildeles vedkommende fuld score. I forhold til PISA får eleven dermed en begunstigelse, i form af den ind-tuning og stilladsering som samtale-optakten leverer. Nogen vil derfor kunne hævde, at resultatet bliver misvisende godt og ikke er et fair grundlag for sammenligning med PISA. 'S'-varianten har på samme vis samtalen som forudsætning og potentiel mediering, men strammer kravene til en fuld score betragteligt. For at opnå en sådan skal eleverne i løbet af samtalen selv formulere et synspunkt svarende til den korrekte PISA-option *og tilmed levere relevante bidrag til forklaring af fænomenet.* I løbet af samtalen skal eleven altså selv demonstrere en vis *forståelse* – eet krav som ikke indgår i den ordinære håndtering af multiple-choice-opgaverne i PISA-formatet. Hvis dette forståelseselement ikke foreligger, opfattes svaret i 'S'-sammenhæng kun som Delvist Rigtig, hvorfor scoren halveres. Denne procedure udelukker, at eleven får fuld score i situationer, hvor der blot gættes helt tilfældigt. Dette er en afgørende pointe her, idet de testede opgaver – med én enkelt undtagelse - alle er af multiple-choice-typen. Når man ser de gennemsnitlige scorer i de røde PISA-søjler, skal man altså medtænke, at en naturfagselev i mange af opgaverne faktisk vil kunne opnå de første 0.25 ved blot at krydse helt tilfældige svarmuligheder. Fx er der i *Solcreme*Q02 25 % 's sandsynlighed for et korrekt svar ad tilfældighedens vej – og tilsyneladende er der kun ca. 33 % korrekte svar i samplet.
Denne gratis-gevinst af 'tilfældigt-rigtige' forekommer ikke rigtigt i VAP's sociokulturelle test-setup – og da slet ikke med den illustrede 'S'-kodningsprocedure.
Derfor er der grund til at mene, at 'S-scorerne underspiller elevernes formåen i en sammenligning med PISA.

VAP og PISA-resultaterne for hvert spørgsmål sammenlignes via parvise T-tests, og hvor der er signifikant forskel er dette indikeret med et antal *' i en parentes efter koden på spørgsmålet nederst på figuren. Jo flere stjerner desto mere sandsynligt er det, at de er forskellige. Mindste signifikante forskel er på vanlig vis p=0.05, hvilket modsvarer en enkelt *-notation[1].

Vurdering af resultaterne på delopgaveniveau.

Ser vi for et øjeblik væk fra første spørgsmål, SolcremeQ02, synes tendensen klar: *overalt hvor der er signifikant forskel klarer de danske elever sig bedst i VAPs sociokulturelt orienterede test-format!* Dette formats udvidede respons- og opmærksomhedsfelt giver tilsyneladende eleverne bedre muligheder for at artikulere og få godskrevet svar, der efter PISAs egne scoringskriterier bør give point. Som ovenstående diskussion af 'S' og 'U'-scorer antyder, så er der imidlertid ingen

[1] Hvert ekstra nul som indskydes i signifikansniveauet afspejles i en ekstra *.

garanti for, at fuld og PISA-sammenlignelig score dækker over en fyldestgørende forståelse hos eleven. Dette forhold vender vi tilbage til senere i vores 'PISA-overskridende' kvalitative analyse.

SolcremeQ02-spørgsmålet falder uden for denne tendens, ved som det eneste spørgsmål at have en signifikant højere PISA-score end VAP-score. Dette undergraver dog ikke den generelle konklusion, idet VAP her klart er kommet af sted med at stille større krav end PISA. I den oprindelige opgave skal eleverne blot vælge ét af 4 MC-udsagn om, hvorfor mineralolie og zinkoxid anvendes i et forsøg (referencer). I PISA har selv elever, som ikke kender nogen af delene, 25 % sandsynlighed for at få point! I VAP derimod har vi forudsat, at man *både* skal kende *ordet* 'reference/referencestof' *og funktionen* af et referencestof for at kunne afgive et meningsfuldt svar og fortjene fuld score. Det er derfor ikke overraskende, at VAP-scoren er betydelig lavere end den oprindelige PISA-score.

*Solcreme*Q03 *om* naturvidenskab handler om at identificere spørgsmål, som kan undersøges naturvidenskabeligt – og da dette *om-science*-spørgsmål ikke lod sig omsætte til praktisk undersøgelse blev det gennemført præcist som i PISA: samme spørgsmål og med individuel paper-and-pencil-afkrydsning blandt de originale svarmuligheder. I praksis er der altså tale om yderligere et check på, om gentestning foranlediger bedre resultater. Og atter viser det sig, at man ikke kan se afgørende forskel på første og anden testningsrunde. Billedet af effekten af gentestning er altså internt konsistent her.

I de øvrige spørgsmål indenfor den praktisk/eksperimentelle opgave er scoren klart forbedret.

I geografiopgavens spørgsmål *Klimaf.*Q02 og *Klimaf.*Q04 ser man, at VAP-scoren er signifikant højest i 3 af de 4 scoringsvarianter. Som omtalt er forskellen i *Klimaf.*Q02S underdrevet, og hvis fx man omkoder, så korrekt & selvstændigt svar *uden* forklaring giver *fuld* score, som i PISA, vil også denne søjle vise en signifikant forskel til fordel for VAP.

Præstationen i biologi-spørgsmålene *Antibio.*Q01 og *Antibio.*Q02 er ikke signifikant ændret. Også her er det muligt at pege på specifikke træk ved opgaverne, som kan tjene som forklaring på at VAP-testformatet her ikke gavner eleverne; fx er den første opgave langt henad vejen et spørgsmål om faktuel viden: virker antibiotika på bakterier eller virus? Stillet overfor et sådant kontant spørgsmål – uden krav om forklaring, virkemåde m.m. - er samtale ikke indlysende faciliterende.

Sammenligning af elevernes samlede præstation i VAP-testen og i PISA 2006 Science.

A. Sammenligning på basis af samtlige opgaver og simple gennemsnit:

Samlet set lader de rå scorer ingen tvivl om, at eleverne klarer sig bedre efter PISAs kriterier, når evalueringen sker i et VAP-format. Og forbedringen er ikke blot marginal: eleverne går fra en gennemsnitlig score på 54 % rigtige, til VAP-værdier på 68 % hhv. 71 % rigtige – alt efter om 'S' eller 'U' varianter lægges til grund. Taget over samtlige 12 indikatorer (dvs. både u og s-varianter, samt de spørgsmål, hvor der ikke måles forbedret præstation) er middelforbedringen 0.17 (på en skala, hvor fuldstændig korrekt score er 1.00). Hvis vi dropper de mest positive u-varianter fås en middelforbedring 0.14. Udgangspunktet for samplet i PISA-testens rå scorer er 0.54 – så en forbedring på 0.14 udgør 26 %!

En ændring af testformatet fra PISAs post-positivistiske tilgang til det mere sociokulturelt orienterede VAP-format giver således en præstationsfremgang på godt 25 %! Med samme elever,

faglige spørgsmål og evalueringskriterier bliver det endog meget tydeligt, at valget af testformat i vid udstrækning "skaber" test-resultatet.

B. Sammenligning på basis af Rasch-analyse:
I den foregående analyse har vi anvendt rå scorer, og forudsat at alle opgaver måler én og samme underliggende kompetence-dimension (en slags én-dimensional 'Scientific Literacy'). Derved har vi meningsfuldt kunnet summe delscorer op, uden at det får karakter af at lægge æbler og pærer sammen. Ydermere har vi forudsat, at alle opgaver *i samme grad* indfanger den kritiske kompetence, og derfor bør veje lige tungt i et samlet præstationsmål (som fx den gennemsnitlige score). Disse grundlæggende antagelser virker måske nok intuitivt rimelige, men PISA sætter nye psykometriske standarder ved sin brug af såkaldt Rasch-statistik til empirisk at undersøge og omgå evt. problemer med disse grundantagelser. I PISA-regi bruges Rasch-analyse i pilotfasen til at udvælge opgaver og items med fornuftige "skala-egenskaber", herunder at de måler én og samme dimension. Derved frasorteres i realiteten 50 % af potentielle opgaver. Derudover bruges analysen til at fastlægge en empirisk sværhedsgrad for hver del-opgave, og dermed en relativ vægt for opgavens bidrag til en samlet score. Den relative sværhedsgrad af en enkelt opgave skal helst være uafhængig af, om det er en 'dygtig' eller en 'svag' elev, der besvarer opgaven, og en god test indeholder et spænd af relative sværhedsgrader. Ydermere er det et krav til en god international og komparativ test, at de relative sværhedsgrader er konsistente henover de deltagende lande (i parentes bemærket, er det i øvrigt en Rasch-analyse-praksis, der gør at den internationale gennemsnitspræstation i PISA altid har scoren 500 og en standardafvigelse på 100). Rasch-interesserede henvises i øvrigt til (Bond & Fox, 2001).

Vi har ikke selv det store kendskab til Rasch-procedurerne, og nærer muligvis af samme årsag en vis forkærlighed for mere umiddelbart gennemskuelige statistiske analyse-gange. Imidlertid har vi ønsket at udmåle præstationsforbedringen på PISAs "egen skala" – for derved samtidig at imødegå indvendinger gående på, at vore konklusioner baserer sig på inadækvate og usammenlignelige metoder. Derfor har vi entreret med Professor Peter Allerup fra DPU, der om nogen er den danske kapacitet indenfor området empirisk Rasch-modellering.

Vi har valgt at anbringe dele af den tekniske dokumentation for Rasch-analysen i et "Rasch-bilag" (bilag 6.1-6.4), tillige med Allerup's medfølgende kommentarer. Ved at give indsigt i Allerups uredigerede kommentarer giver vi samtidig udenforstående mulighed for at vurdere, hvorvidt vi misforstår eller skævvrider konklusionerne. For yderligere at øge troværdigheden vil vi forsøge i størst mulig udstrækning at bruge direkte citater af P. Allerup.

Om den anvendte analyseprocedure hedder det:

> *"Selve fittet af Raschmodellen er gennemført ved hjælp af RUMM 2020 (se PDF filen, [Bilag 3 nedenfor[1]]), det samme program som anvendes ved analyserne af de nationale tests. Resultatet af test of fit er vedlagt og viser ikke alvorlige fejl! (det er udmærket, at man kan finde svagere tilpasning, hvis man anvender stærkere metoder, men nu har vi valgt at gennemføre 'standard' på området= de programmer, der anvendes fx i de nationale tests)"* (se bilag 6.1)

[1] Refererer til bilag i kommunikationen. Her modsvares dette af bilag…

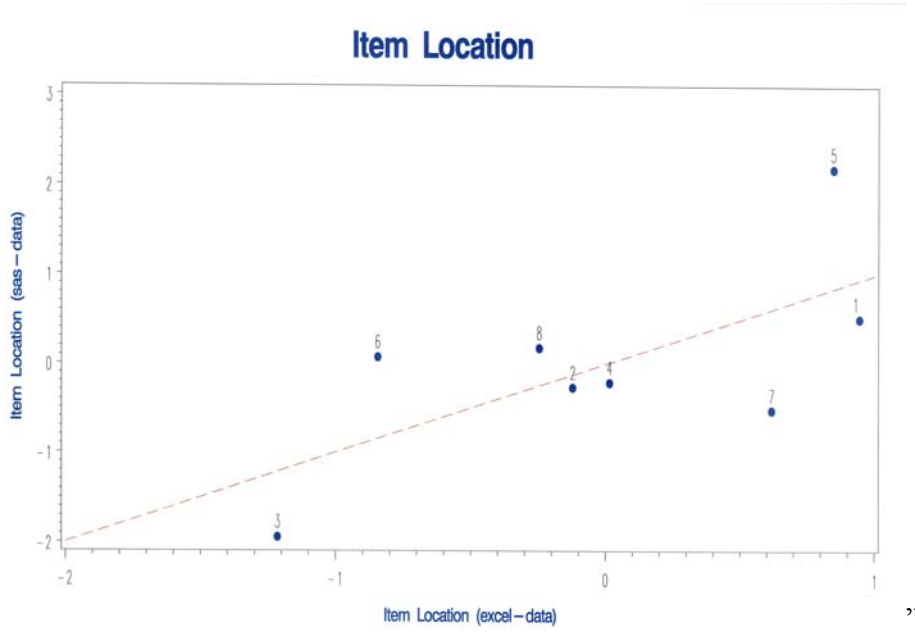
VAP-opgavernes evne til at skalere på en enkelt dimension er ikke perfekt, men brugbar, jf. ledsagekommentarerne:

> *"…den statistiske Rasch Model peger på flere problemer mht. til at indplacere de 8 VAP opgaver på én dimension. Fx er det ikke helt klart, at den relative sværhedsgrad af enkelte opgaver er uafhængig af, om det er en 'dygtig' eller 'svag' elev, der besvarer opgaven… Problemerne er samlet vurderet til at være af et sådant omfang, at man kan gå videre med alle 8 opgaver, som 'godkendt' af Rasch analysen."*(se bilag 6.2)

> " *Der er helt sikkert en vis grad af lokal afhængighed – ikke overraskende – på grund af den mediering, som finder sted under det alternative regime, men det er ikke så markant, at RUMM –analysen falder på gulvet af den grund!"* (se bilag 6.1).

Om muligheden af at lave en direkte sammenligning mellem VAP -scorer og de oprindelige PISA-scorer hedder det:

> *"For at sammenligningen derefter kan gennemføres på skala-niveau kræves, at PISA opgaverne og VAP opgaverne for, i den mindste nogle få opgavers vedkommende, lapper over. Dvs. har samme relative sværhedsgrad. Dette er undersøgt og fundet acceptabelt, lidt overraskende, for de fleste af opgaverne. Tegningen herunder viser de 8 items relative sværhedsgrader, vurderet i VAP skala og i PISA skala. Ens relative sværhedsgrader opnås, hvis punkterne ligger på den stiplede linje.* (se bilag 6.2, "SAS"-data refererer til VAP, mens "Excel" refererer til PISA)



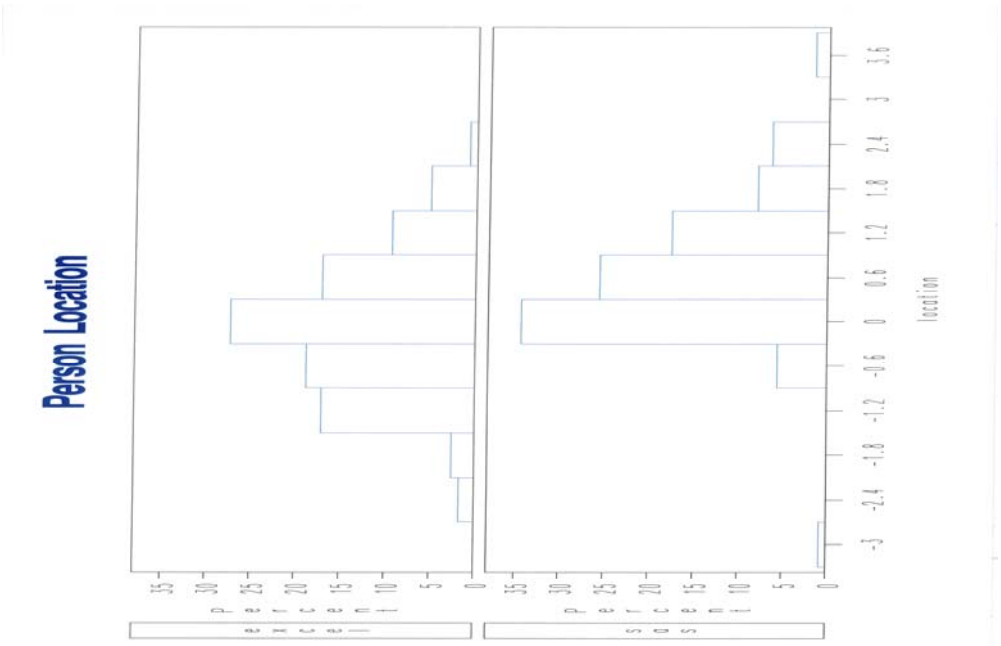
"

Det er således rimeliggjort, at man kan sammenligne elevers scorer i de to test-setups. Om resultatet af denne sammenligning heddet det:

> *"Som det fremgår, ligger majoriteten, en signifikant stor del, af eleverne over den indlagte stiplede identitetslinje, hvilket betyder, at eleverne vurderet med VAP skalaen er 'dygtigere' end elevpræstationerne beregnet ved hjælp af de originale PISA opgaver"*. (se bilag 6,2).

De generelt højere scorer i VAP-versionen kommer bl.a. til udtryk i:

> " *to histogrammer, der viser systematisk forskydning af nye scores sammenlignet med de 'gamle'"* (bilag 6.1).



Tilbage er blot at estimere, hvor stor forbedringen er i et VAP-format. Dette er gjort i bilag 6.3 og 6.4. Opsummerende hedder det således:

> *"den procentvise forskel mellem de to gruppe ligger på ca 25% rigtige*
>
> *- at denne 'rå' forskel udmålt på en PISA målestok (med de sædvanlige internationale 500 i midten og en standardafvigelse på 100) beløber sig til ca 125 point"*
>
> (bilag 6.3)

Den samlede Rasch-forbedring på ca. 25 % modsvarer aldeles den vurdering, som blev lavet på grundlag af samtlige opgaver og den simplere deskriptive statistik. I en vis forstand udgør dette en seriøs metodisk triangulering – og så er det da betryggende, at resultaterne falder så ens ud.

Målestoksforbedringen på 125 Rasch-point skal ses i lyset af, at med undtagelse af 2 lande ligger samtlige nationale scorer i PISA Science 2006 i intervallet fra 473 points (Grækenland) til højeste score 563 (Finland). Testformatet har altså en betydning som overgår enhver national variation!

Konklusionen på forskningsspørgsmål 3a må være, at

- i en direkte sammenligning efter PISAs scoringskriterier klarer eleverne sig i gennemsnit 25 % bedre, når de får lov til at udfolde sig i et VAP-lignende sociokulturelt orienteret testformat. Undersøgelsen dokumenterer således, at billedet af elevernes performance i PISA ikke udelukkende udtrykker deres formåen i forhold til opgavernes faglige indhold, men i særdeles signifikant grad også skabes af deres evne til at gennemtrænge et bestemt test-formats nåleøje!
- Der er dog stor variation henover opgaverne. På kun delvist gennemskuelig vis synes specifikke aspekter af opgaverne, evt. i interaktion med bestemte medierende omstændigheder, at påvirke sammenligningen.

Hvilke elevgrupper får størst fordel af et ændret testformat?

VAP-samplet har en størrelse, som i princippet gør det muligt at undersøge om ændringer i testformatet i særlig grad tilgodeser elever af bestemt *køn, fagligt niveau* og/eller *etnicitet.* Med udgangspunkt i elevernes rå scorer vil vi nedenfor sammenligne forskellige elevgruppers præstationsforbedring ved overgangen til VAP's testformat.

Køn og betydningen af test-format

Tabel 3 viser præstationsforbedringen for de to køn på hvert af VAP-testens spørgsmål (negative tal i kolonne 2 og 3 betyder altså, at eleverne har klaret sig *dårligere*). Kolonnen yderst til højre angiver, hvor der foreligger signifikante forskelle i drenge og pigers forbedring.

	Forbedring (drenge, N=58))	Forbedring (piger, N=60)	Signifikante kønsforskelle
*Solcreme*Q02	-0.1369 (A)	-0.15 (A)	
*Solcreme*Q03	-0.09 (A)	0.02 (A)	
*Solcreme*Q04	0.26 (A)	0.06 (A)	
*Solcreme*Q05	0.53 (A)	0.55 (A)	
*Kllimaf.*Q01	0.43 (A)	0.43 (A)	
*Kllimaf.*Q02s	0.14 (A)	-0.07(B)	0.03*
*Kllimaf.*Q02u	0.30 (A)	0.23 (A)	
*Kllimaf.*Q04s	0.38 (A)	0.03 (B)	0.002**
*Kllimaf.*Q04u	0.56 (A)	0.23 (B)	0.009**
*Antibio.*Q01	0.02 (A)	0.13 (A)	
*Antibio.*Q02	-0.10 (B)	0.13 (A)	0.02*
*Antibio.*Q03	+0.19 (A)	0.13 (A)	

Tabel 3. Præstationsforbedringer efter køn på hver af de gentestede opgaver.

Tre ud af fire signifikante forskelle ligger i geografi-opgaven – og i alle tre tilfælde nyder drengene i særlig grad godt af samtale-testformen. Eneste sted hvor pigerne går mest frem er indenfor ét af de biologisk orienterede spørgsmål (*'Antibio.*Q02'). I den oprindelige PISA-test klarer pigerne i samplet sig marginalt bedre end drengene i geografi-opgaven (*'Klimaf.Q02'*), mens det forholder sig omvendt i den biologi-orienterede opgave. Set under eet kunne resultaterne således indikere, at den mundtlige test i særlig grad begunstiger de svageste faglige elever indenfor et domæne. Denne hypotese forfølger vi nærmere i næste afsnit.

Fagligt niveau og betydningen af test-format

Vi har valgt at inddele eleverne efter fagligt niveau, alt efter hvor godt de besvarede de relevante items i den oprindelige PISA-test. Helt konkret dækker kategorierne Høj, Middel, Lav over scorer i følgende intervaller:

Høj:	rå PISA-score >8
Middel:	4< rå PISA-score ≤8
Lav:	rå PISA-score ≤4

På basis af denne inddeling kan man undersøge de tre elevgruppers relative fremgang i mødet med VAPs testformat. Tabel 4 viser resultaterne af ANOVA-analyse[1], hvor elevgrupperne parvist er sammenlignet. Som man kan se, er der anført bogstaver (A, B og evt. C) i parentes i alle rækker, hvor der er fundet signifikante forskelle. Forskellige bogstaver viser, at elevgrupperne er signifikant forskellige, mens samme bogstav udtrykker, at man ikke signifikant kan se forskel. A betegner elevgruppen med signifikant størst fremgang, B den med næststørst fremgang o.s.v.. I fx tilfældet *Klimaf.*q02s har gruppen af Lave elever signifikant større fremgang end de to andre elevgrupper, som i øvrigt på uskelnelig vis har haft samme fremgang (i situationen faktisk en tilbagegang, da tallet er negativt).

	Elever med lavt niveau (N=49)	Elever med middel niveau (N=44)	Elever med højt niveau (N=23)	Signifikante forskelle knyttet til fagligt niveau
*Solcreme*Q02	0.01	-0.24	0.20	(0.11)
*Solcreme*Q03	-0.18	0.06	0.06	(0.26)
*Solcreme*Q04	+0.22	0.23	0.06	
*Solcreme*Q05	0.91 (A)	0.59 (B)	0.07 (C)	0.0001***
*Kllimaf.*Q01	0.73 (A)	0.41 (B)	-0.01 (C)	0.0001***
*Kllimaf.*Q02s	0.23 (A)	-0.07 (B)	-0.10 (B)	0.005**
*Kllimaf.*Q02u	0.50 (A)	0.19(B)	0.05 (B)	0.001**
*Kllimaf.*Q04s	0.15	0.23	0.14	
*Kllimaf.*Q04u	0.31	0.50	0.25	(0.24)
*Antibio.*Q01	0.12	0.07	0.00	
*Antibio.*Q02	0.27 (A)	-0.08 (B)	-0.32 (B)	0.0001***
*Antibio.*Q03	0.18	0.17	0.09	

Tabel 4. Relativ fremgang i VAP-testen i forhold til PISA-testen for elever med forskelligt fagligt niveau (målt via PISA-testen).

[1] Beregninger foretaget via SAS statistisk software, med proceduren GLM og den særlige'Tukey'-option, som muliggør systematisk parvis sammenligning af gruppescorerne.

Det er bemærkelsesværdigt, at *alle signifikante træk peger på, at det er de svageste elever, som får størst fordel af den ændrede prøveform.*

Der anes at være *tendenser* til en modsat konklusion, i særdeleshed i den praktisk-eksperimentelle opgaves spørgsmål *Solcreme*q03. At netop denne opgave skiller sig ud, er imidlertid interessant og en pointe her, da vi i VAP-forberedelsen ikke rigtig formåede at transformere den til et sociokulturelt setup – og følgelig valgte at afvikle den på vanlig individuel-paper-and-pencil-vis-på-basis-af-det-originale-PISA-spørgsmål. *I netop dette spørgsmål er test-formatet ikke ændret, hvilket tenderer til at være til ugunst for de svagere elever. Altså præcist samme konklusion som ovenfor!*

Det forekommer umiddelbart logisk, at de svageste elever har størst behov for det sociokulturelle setups medierende ressourcer og får størst gavn af dem. I studier af hvem der har størst gavn af gruppearbejde, har man da også fundet noget tilsvarende (Webb, 1997). For yderligere at konsolidere konklusionen har vi imidlertid undersøgt, i hvilken udstrækning *de fagligt dygtigste elevers relativt mindre resultatforbedring kunne skyldes "mætning".* Såfremt de allerede i PISA-testen har opnået fuldt point-tal i et spørgsmål, vil de jo have svært ved at forbedre sig, uanset hvilket nyt testformat de udsættes for efterfølgende. Umiddelbart ser man, at der er spørgsmål, hvor de gode elevers præstation direkte bliver ringere. Dette kan åbenlyst ikke opfattes som et mætningsfænomen. Mere systematisk har vi for den høje gruppe udregnet, hvor mange procent rigtige den manglede i hvert spørgsmål i PISA-testen – samt hvor stor en del af denne mulige forbedring ("Gain" (Hake, 1998)), de faktisk opnår i VAP-testen. Disse tal fremgår af Tabel 5. "Mætning" er en sandsynlig ingrediens, når Gain-værdien nærmer sig 100 % - og i særdeleshed såfremt den systematisk befinder sig i dette leje.

	Gain(high)
*Solcreme*Q02	-49 %
*Solcreme*Q03	-2,7 %
*Solcreme*Q04	9,2 %
*Solcreme*Q05	47 %
*Klimaf.*Q01	-13 %
*Klimaf.*Q02s	-148 %
*Klimaf.*Q02u	100 %
*Klimaf.*Q04s	+56 %
*Klimaf.*Q04u	100 %
*Antibio.*Q01	16 %
*Antibio.*Q02	-187 %
*Antibio.*Q03	100 %

Tabel 5. Andel af mulig forbedring (mætningsgrad) for elever med højt fagligt niveau (jfr. tab 3)

Der er i hvert fald ikke tale om systematiske tegn på mætning, om end der væsentligst synes at forekomme mætning i 3 af de 12 spørgsmål. Men de svage elevers signifikante fremgang kan ikke entydigt forklares med mætning. Den praktiske opgave nærmer sig også mætning i det sidste spørgsmål, hvor de høje elever høster ca. halvdelen (47 %) af en ganske lille mulig scoreforbedring på ca. 10 %. To af de 'mætningstruede' spørgsmål vedrører scoringsform 'U' i geografiopgaven, og man kan da godt fundere over om der gives dybe forklaringer på dette. Det tangerer imidlertid ren spekulation, hvorfor vi ikke kaster os ud i det.

Afslutningsvist har vi sammenlignet de tre elevgruppers *samlede forbedring* henover samtlige spørgsmål (igen via SAS, med proceduren GLM og brug af Tukey-optionen). Resultatet fremgår af tabel 6.

PISA-niveau for elevgruppen	Forbedring for gruppen (gennemsnit, på en skala, hvor fuldstændig korrekte svar =1,00)	Tukey-gruppe
Lavt niveau	0,36	A
Middel niveau	0.23	A, B
Højt niveau	0.03	B

Tabel 6. Forskellige elevgruppers forbedring i det sociokulturelt orienterede VAP-format

Tabellens resultater skal tages med et vist forbehold, idet vi har måttet frafalde at score størstedelen af eleverne på i hvert fald eet af *Solcreme*-opgavens underspørgsmål. Det betyder, at vi kun har et komplet indeks for *den samlede forbedring* for 18 af samplets elever her. Tabellen understøtter imidlertid meget godt de foregående analyser, idet den indikerer, at gruppen af Lave elever har signifikant større fremgang end gruppen af Høje elever. Middelgruppen synes også i denne sammenhæng at indtage en mellemstatus. Formentlig ville et lidt større sample bevirke, at middelgruppen på signifikant vis udskilles fra de øvrige. Det er bemærkelsesværdigt, at den Høje gruppe i modsætning til de øvrige grupper ikke forbedrer sin præstation i det sociokulturelle format.

Overordnet dokumenterer VAP-undersøgelsen således, at valget af test-form på signifikant måde tilgodeser/diskriminerer bestemte elevgrupper. VAPs sociokulturelle setup er således klart mere gunstigt/inkluderende for de svagere elever.

Etnicitet og betydningen af test-format

Denne del af analysen er lidt uskarp i kanten, idet vi ikke har aldeles entydige data for elevernes etniske oprindelse. I mangel af bedre er eleverne klassificeret som "nydanske", såfremt de har navne, som *klart* er udenlandske af natur og/eller såfremt de indlysende ser 'fremmedartede' ud på vore videooptagelser (typisk hudfarve, ansigtstræk m.m.). Disse kriterier har placeret 20 af eleverne i samplet i den "nydanske" kategori.

Der er ikke gjort noget forsøg på at estimere, hvor godt eleverne taler dansk, om de er adopterede ind i en dansk familiekontekst, om de er 1. eller 2. generationsindvandrere osv. Med dette noget udifferentierede blik og det relativt beskedne sample er der en betragtelig sandsynlighed for, at vi kommer til at udviske/overse væsentlige og eksisterende træk – men vi har valgt at gøre forsøget alligevel.

	PISA			Forbedring via VAP-format		
	Danske (N=97)	Nydanske (N=20)	Signifikans af forskel	Danske	nydanske	Signifikans af forskel
*Solcreme*Q02	0,36	0,11	0,04 *	-0,16	-0,06	
*Solcreme*Q03	0,73	0,28	0.0001***	-0,08	0,25	0,10
*Solcreme*Q04	0,53	0,50		0,18	0,12	

*Solcreme*Q05	0,37	0,32		0,55	0,45	
*Klimaf.*Q01	0,47	0,47		0,42	0,50	
*Klimaf.*Q02s	0,69	0,53		0,02	0,13	
*Klimaf.*Q02u				0,23	0,47	0,07
*Klimaf.*Q04s	0,42	0,60		0,25	-0,13	0.02*
*Klimaf.*Q04u				0,43	0,14	0,10
*Antibio.*Q01	0,47	0,33		0,06	0,17	
*Antibio.*Q02	0,57	0,21	0,0035**	-0,04	0,31	0,01*
*Antibio.*Q03	0,77	0,65		0,17	0,11	

Tabel 7. Forskellige etniske elevgruppers præstationer i PISA-testen og deres forbedringer i VAP-testen, fordelt på opgaver. Signifikansniveauet anført hvor der er tale om signifikante, eller næsten signifikante, forskelle.

Af tabel 7 fremgår at nydanske elever i samplet klarer en række af PISA-opgaverne markant dårligere end traditionelt danske elever. En nærmere analyse godtgør således, at ca. 70 % af de "nydanske" elever tilhører den svageste faglige kategori ("Lav") vurderet ud fra deres PISA-præstation.

Der viser sig ikke noget entydigt mønster, når man studerer forbedringerne via VAP-formatet, fx trækker de to eneste signifikante forbedringer hver sin vej. Det er egentlig lidt overraskende, når man betænker den store andel af 'lave elever' i gruppen af "nydanske". Tidligere har vi jo netop set, at fagligt svagere elever har relativt størst fordel af VAP-formatet. En mulig forklaring på, hvorfor den svagere gruppe af 'nydanske' ikke i samme grad høster VAPs sociokulturelle fordele, kan være at det sociokulturelle vindue på samme tid øger kravene til selv at kunne formulere ting og gøre sig forståelig. Altså krav som også implicerer en sprogbeherskelse, som adskillige af de Nydanske elever har relativt sværere ved at honorere. *Dette kan vise sig at være en væsentlig nuancering af den foregående konklusion om, at VAP test-formatet begunstiger svagere elevgrupper*.

7.4 PISA-Overskridende analyser

VAP-testningen giver både mulighed for et bredere og et anderledes rigt billede af elevernes formåen. Billedet er bredere i den forstand, at vi bevidst ikke kun har søgt at afdække elevernes formåen indenfor de snævre PISA-items, men tillige har spurgt ind til væsentlige videnselementer fra det bredere område af Fælles Mål. En analyse af dette materiale giver derfor et bedre indtryk af, hvorledes eleverne lever op til nationale læringsmål indenfor de pensumområder, som opgaverne berører. Denne ekstra information er selvsagt ikke relevant, når PISA i udgangspunktet ikke forsøger at udtale sig om elevernes mestring af et nationale mål og pensa, men udelukkende forholder sig til en universel forståelse af scientific literacy. Dvs. for internationale, komparative formål kan PISA godt stå alene, især hvis PISA tester nogle for de enkelte lande relevante kompetencer. Det var dette den første VAP-rapport undersøgte, og hvor den nok fandt stort overlap mellem PISA-frameworket og Fælles Mål, men også vigtige principielle forskelle. Danske elevers performance i PISA er derfor ikke nok til at vurdere, hvorvidt eleverne opfylder de nationale krav, hvilket det især er vigtigt at tage hensyn til, i det omfang PISA-resultaterne lægges til grund for uddannelsespolitiske tiltag.

Det anderledes og rigere VAP-billede af elevernes formåen fremkommer først og fremmest, fordi vi har valgt et andet og mere sociokulturelt orienteret evalueringsformat. Typiske træk ved en sociokulturel evaluering er fx, at den er interaktiv, kollaborativ, måler læringspotentiale lige så

meget som udgangsniveau, procesorienteret, anvender symbolske og fysiske værktøjer, samt oftest foregår i en autentisk praksis (se fx oversigtsartiklen af Gipps (Gipps, 1999)). På varierende vis tilgodeses de fleste af disse elementer hen gennem VAP-evalueringen af den enkelte elev. Dialog, samarbejde, autentiske (undersøgende) processer og muligheden for at inddrage artefakter er således elementer i VAP-forløbet. PISAs evalueringskoncept er noget nær det stik modsatte, idet en dækkende beskrivelse her fremkommer ved at negere alle karakteristiske træk ved den sociokulturelle evaluering. Dette tydeliggør, at der vitterligt er tale om to meget forskellige blik på evaluering.

VAP-undersøgelsens større validitet kan måske simplest henføres til vores forudgående snak om et "udvidet opmærksomhedsvindue". Eleverne udfører her en langt bredere vifte af fagrelaterede handlinger – diskursive, dialogiske, meningsskabende, undersøgende m.m., dvs. der er ganske enkelt indblik i mere. Eleverne har mulighed for at inddrage artefakter i deres meningsdannelse. De kan svare i åbent format, frem for at forholde sig til præfabrikerede MC-svarmuligheder. Og sidst, men ikke mindst, giver den interaktive form mulighed for at intervieweren kan spørge uddybende/opfølgende og afklarende, således at der opnås et skarpere blik i analysen af handlingerne. Såvel styrker som svagheder i det diskursive, dialogiske, meningsskabende, undersøgende m.m. må forventes at træde tydeligere frem i VAP-vinduet. Det er her VAP-vinduet har sit fokus og sin force.

I den foregående direkte PISA-sammenligning måtte vi holde det righoldige, udvidede billede op mod PISA-scoringens mere snævre scoringskriterier, der så at sige fungerede som et monokromt filter. Mange strukturer og nuancer forsvinder i filtreringen. I nærværende PISA-overskridende analyse slipper adskilligt mere igennem.
Først analyseres elevernes formåen i lyset af Fælles Mål (FM). Videooptagelserne af VAP-gentestningen af de tre PISA-opgaver – *Antibiotika*, *Klimaforskelle* og *Solcreme* – analyseres med udgangspunkt i de baselines, som blev udformet for de pågældende faglige områder (se afsnit 6.1). Disse analyser baserer sig på alle de gentestede elever.
Dernæst analyseres et tilfældigt udvalgt sample af elever med hensyn til argumentationsevne og brug af fagsprog og artefakter. Der kastes herigennem et mere sociokulturelt blik på naturfagsundervisningen i den danske folkeskole.
Tilsammen tilbyder disse analyser et naturfagsdidaktisk blik på elevers kompetencer inden for de undersøgte fagområder, som kan anvendes som afsæt for egentlig didaktisk udvikling af undervisningen.

Elevernes formåen i lyset af Fælles Mål

Nedslag i området Biologi
Den biologi-relaterede opgave i PISA, *Antibiotika*, tester, om eleverne kender antibiotika og mulige konsekvenser ved brug af det. Ikke noget med, hvad der sker. Endelig er tredje spørgsmål en evidens-logisk slutning på basis af specifikke laboratoriemålinger.

I Fælles Mål (FM, 2006-versionen, Biologi, 8.kl.) skal undervisningen gøre det muligt for eleverne at "kende til, hvordan kroppen forsvarer sig mod bakterier og vira", samt "kende og beskrive forskellige organismer og deres systematiske tilhørsforhold samt anvende begreber om livsytringer, herunder fødeoptagelse, respiration, vækst, formering og bevægelse i forbindelse med forskellige typer organismer." Efter konsultation med ressourcepersoner i folkeskolens biologiundervisning og

lærebøger for årgangen blev det ekspliciteret, at eleverne i henhold til intentionerne i FM burde kunne

- diskutere forskellen på bakterier og vira ud fra forskellige klassifikationskriterier
- fortælle om infektioner og smitteveje
- fortælle om immunforsvar
- redegøre for antimikrobiel behandling og udvikling af resistens.

Det er umiddelbart[1] synligt, at Fælles Mål er bredere og først og fremmest kvalitativt sigter adskilligt højere end PISA-opgaven tester. Kendskabet til klassifikationskriterier, samt forståelsen af mekanismer og forklaringer i fx udviklingen af [biologisk begreb knyttet til brug af antibiotika][2] er taksonomisk langt mere fordringsfulde mål end PISA-opgavens.

Analysen af de video-tapede interviews godtgør, at mange elever har svært ved at udtrykke sig om de relevante biologiske begreber (bakterier, virus, immunforsvar, resistens etc.). En stor del af eleverne kendte således ikke ordet *antibiotika,* som er "titlen" på den originale PISA-opgave. De fleste kendte til *penicillin* og blev straks mere fortrøstningsfulde, når de hørte, at dette er et eksempel på et antibiotikum. I almindelighed var der stor usikkerhed omkring begrebsindhold og faglige distinktioner – et problem som naturligt nok gjorde det mere eller mindre umuligt for langt størstedelen af eleverne at foretage en faglig skelnen mellem bakterier og vira.

For at få et indtryk af elevernes viden i forhold til Fælles Mål blev elevernes *forståelse* af de faglige begreber helhedsbedømt, som man ville gøre det til sædvanlig mundtlig eksamen. To uafhængige iagttagere/"forskere" (specialestuderende + forskningsvejleder) vurderede på basis af videooptagelserne graden af forståelse på en skala fra 1-5, hvor 5 var "fuld forståelse" på det relevante lærebogsniveau. I gennemsnit blev elevernes forståelse af forskellen mellem bakterier og virus vurderet til 1.8 på denne skala. Den tilsvarende helhedsorienterede vurdering af elevernes forståelse af [biologisk begreb knyttet til brug af antibiotika]begrebet blev 2.4. *Hvis* skalaerne regnes lineære svarer det til hhv. 20 % og 35 % af *fuld forståelse* (i forhold til Fælles Mål). Til sammenligning var rigtighedsgraden blandt eleverne ca. 45 % i PISA-spørgsmålet, hvor antibiotikas indvirkning på bakterier og vira blev testet (ca. 51 % i VAP). PISA-opgaven, som testede kendskab til *at* fænomenet [biologisk begreb knyttet til brug af antibiotika] forekommer, gav tilsvarende rigtighedsgrader på 75 % (PISA) og 90 % (VAP). *Danske elever fremstår her betydeligt dårligere i forhold til Fælles Mål, end de gør efter PISA-kriterier (i begge testformater)! Resultatet indikerer, at PISA-resultaterne kun i ringe grad udtrykker målopfyldelse i forhold til kravene i Fælles Mål.*

Hvorfor nu dette misforhold? I første VAP-rapport konkluderede vi, at målene i de danske læreplaner et godt stykke hen ad vejen modsvarede intentionerne i PISA Science Framework'et. Sammenligningsgrundlaget var dengang målkategorier, kompetencebeskrivelser og indholdsmæssige vægtninger og perspektiver. De to sæt af målsætningsbeskrivelser forekom at udtrykke stærkt overlappende forståelser af Scientific Literacy. Hvis man derimod sammenligner de krav, som er implementeret i PISA-opgaverne, med de fortolkninger af danske målbeskrivelser, som danske lærebøger og VAPs referencegruppe repræsenterer, så er der tale om to meget forskellige *niveauer* af en scientific literacy. R. Bybee, som har været en central aktør i både debatten omkring Scientific Literacy og udviklingen af PISA Science Framework 2006, har påpeget, hvorledes forskellige opfattelser af Scientific Literacy ikke kun varierer i deres indhold,

[1] Det umiddelbare er noget forsvundet efter Skolestyrelsens krav om sløring af PISA-opgaveindholdet.
[2] Begrebet slettet efter krav fra Skolestyrelsen.

men også i niveauet af de faglige kompetencer, som indgår (Bybee, 1997). Han opererer med forskellige niveauer af scientific literacy der karakteriseres således:

Nominel scientific literacy: Eleven kan genkende og placere en vending, et spørgsmål eller et emne som naturvidenskabeligt, men har ingen/ringe forståelse for det naturvidenskabelige indhold. Vil afsløre hverdagsforståelser og misforståelser, når termerne bruges.

Functional scientific literacy: Eleven har et fragmenteret og overvejende faktuelt kendskab til naturvidenskab. Eleven kan måske definere visse fagtermer, men har begrænset forståelse af eller erfaring med dem. Eleven kan bruge naturvidenskabelige vendinger, men ofte kun i bestemte situationer, som fx testsammenhænge, hvor passende begreber kan genkaldes. Fagsproget er ingen hjemmebane for eleven.

Conceptual scientific literacy: Eleven har forstået hvorledes enkeltbegreber relaterer sig til hele fagområdet og har forstået fagområdets grundlæggende principper og lovmæssigheder.
Procedural scientific literacy: Eleven kan bruge naturvidenskabernes processer og ideer, såsom udspørgen, undersøgelser, diskussioner.[1]

Multidimensional scientific literacy: Eleven kan inddrage perspektiver på naturvidenskab ud over fagsprog, begrebssæt og procedurer, såsom den historiske udvikling af de videnskabelige begreber, naturvidenskabens natur og naturvidenskabens betydning i personens eget liv og i samfundet, ligesom faget kan relateres til andre fag og discipliner.

I lyset af disse kategorier kan man forstå kravene i PISA-opgaven som udtryk for *funktionel* scientific literacy, fx tester PISA-opgaven udelukkende om eleverne har fat i de funktionelle pointer: hvad kan man behandle med antibiotika? Hvad er det, der går galt, hvis man overdriver brugen af antibiotika? Evalueringen af det mundtlige interview efter Fælles Mål-målestokken ovenfor er derimod klart en test på *konceptuel og proceduremæssig* scientific literacy niveau.

Hvorfor bliver denne niveauforskel først synlig nu, frem for i den indledende VAP-rapports analyse af de formelle målbeskrivelser? Det videnskabelige Framework er bygget op omkring følgende fagligt orienterede kompetencemål: *"possesses scientific knowledge and uses that knowledge to identify questions, acquire new knowledge, explain scientific phenomena and draw evidence-based conclusions about science-related issues"*. Man ser altså, at Fælles Mål-relevante mål som *evne til at forklare* og *argumentere evidensbaseret* indgår i de formelle PISAs intentioner. Derfor er det nærliggende at antage, at forskydningen først sker i næste led, altså i operationalisering af PISA-målene til konkrete evalueringsopgaver. Det rejser således spørgsmålet, om den aktuelle PISA-opgave (og PISA-opgaverne i almindelighed) på valid måde operationaliserer PISA Frameworket. Hvad angår den aktuelle PISA-opgave om antibiotika, gælder det, at den kan besvares med langt mindre end én naturvidenskabelig forklaring: Hverdagskendskab (fx fra Tv-avis eller tilsvarende) der etablerer et associativt link mellem *mikroorganisme-penicillin-resistens* er alt tilstrækkeligt – eleven behøver ikke vide noget om mikroorganismer, antibiotika eller deres interaktion og de mekanismer, som frembringer [biologisk begreb knyttet til brug af antibiotika]en. Dvs. opgaven *kan* besvares uden den forståelse, som en naturvidenskabelig forklaring forudsætter. Og som opgavestillerne måske har forestillet sig, at opgaven tester for.

[1] Bybee har slået den begrebsmæssige og procedurale literacy sammen til én.

Hvis analysen af denne opgave udtrykker en mere generel tendens, så er der al mulig grund til bekymring – over validiteten af PISA-opgaverne i forhold til Frameworket, over danske elevers formåen i forhold til Fælles Mål, evt. over det rimelige i at lade danske målbeskrivelser m.m. spille ud på et relativt højt scientific literacy niveau (”Konceptuel og proceduremæssig”), over brugen af PISA-resultater til at legitimere tiltag i forhold til undervisning rettet mod nuværende Fælles Mål mm.

Nedslag i området naturgeografi-fysik
PISA-opgaven *Klimaforskelle* tester kendskab til tre forskellige aspekter af klima, herunder evnen til at anvende en af de bærende repræsentationer i faget og evne til at vælge en korrekt forklaring på variation i klimaet over tid og på forskellige steder.

Opgaven ligger i hvert fald indenfor rammerne af Fælles Mål, idet de på daværende tidspunkt indeholdt følgende mål:

- kende hovedtræk af solsystemets opbygning og forbinde dette med dagslængde, årstider, klimaforskelle, tidevand *(Natur/teknik efter 6. Klassetrin)*
- beskrive jordens inddeling i klimazoner og plantebælter
- anvende enkle begreber i beskrivelsen af vejr og klima
- anvende kortet som et væsentligt arbejdsredskab til at søge viden om og svar på geografiske spørgsmål
- kende verdensdele, lande, byer m.m. på kort og globus, herunder væsentlige danske lokaliteter og deres placering *(geografi, efter 8. kl.)*
- kende jordens og månens bevægelser og nogle af de virkninger, der kan iagttages på jorden som årstider, tidevand og formørkelser *(Fysik/kemi efter 8.kl)*

I lighed med den forudgående opgave vurderede to uafhængige personer ved hjælp af VAP-dataene elevernes forståelse af de begreber, processer og mekanismer, som indgår i en grundlæggende naturvidenskabelig forklaring af fænomenerne. Alt sammen på et niveau som typisk lærebogsmateriale for elevgruppen præsenterer det. Desuden skrev som tidligere nævnt stud.scient. Ellen Berg Jensen geografididaktisk speciale: *15-åriges viden om klimaforskelle* (Jensen 2007) baseret på analyse af VAP-dataene. En del af resultaterne herfra inddrages i det følgende.

Forståelsen af [geospecifik]figur vurderes ud fra Fælles Mål til et gennemsnitligt forståelsesniveau på 2.9, målt på en skala fra 1-5 (dvs. ca. 50 %). Lejlighedsvist sker der aflæsningsfejl, men størstedelen af eleverne er i stand til grundlæggende at beskrive klimaet ud fra denne type repræsentation. Forståelse i Fælles Mål-forstand indebærer yderligere, at eleverne kan skelne mellem vejr og klima, herunder at de forstår, at klima er en gennemsnits-konstruktion af vejrmønstre over en årrække. Denne mere begrebsrettede forståelse står det ikke så godt til med. Dette skal måske ses i lyset af, at en stor del af eleverne først og fremmest erindrer at have set [geospecifik]figurerne i deres matematikbøger! Hvis man først og fremmest kender figurerne fra en matematikkontekst, er det meget sandsynligt, at deres geografifaglige forståelsesbaggrund er udeklareret og/eller utilegnet. Vurderingen af et forståelsesniveau på ca. 50 % harmonerer fint med den oprindelige PISA-testscore, mens VAP-testformatet giver indtryk af, at langt flere rent faktisk lever op til PISAs relativt beskedne krav (se Figur 1 med sammenligning af rå item-scorer).

Elevernes procesforståelse indenfor Fælles Mål-området *vandkredsløb og dannelsen af nedbør* er tilsvarende vurderet. Størstedelen af eleverne kommer her frem til, at vandet fra havet fordamper og

kommer op og bliver til skyer, hvorefter det falder ned igen. Hvordan vanddampen bliver til skyer, forklares oftest rimeligt godt og hverdagseksemplet med ånden, når det er koldt, og vand, der koger, bliver flittigt brugt. Men når der bliver spurgt ind til, hvordan det så bliver til regn, kommer de mere kreative forklaringer i brug, fx er den mest udbredte, at skyen bliver for tung og så regner det. Kun 7 elever har en fyldestgørende Fælles Mål-forklaring her.

Interviewene af eleverne afslørede ganske mange 'morsomme' misforståelser. De vil senere blive behandlet mere seriøst som manglende elementer til at danne fagsprog. Her vises nogle eksempler, der om ikke andet har en vis underholdningsværdi.

I: ja, hvordan er det nu regn det dannes?
E1: ja (virrer med hovedet)… hvis du spurgte min kemilærer, så ville han sige det var to grundstoffer der var blevet dannet.. eller samlet [aflæser synlig skepsis hos I]… men det er jo ikke det vi skal svare på her…

I: hvad plejer der at være på himlen, når det regner..
E2: ja, det ved jeg jo godt, det er skyer… men jeg ved jo ikke, hvordan skyerne holder regnen oppe.. for det er jo sådan noget der bare ryger igennem
I: men hvad er skyerne lavet af?
E2: ja, det ved jeg så heller ikke,,,
I: det er vanddamp
E2: okay, så det er simpelthen bare vanddamp… for jeg vidste jo godt, det er jo klart, at vandet det fordamper på et tidspunkt… og så stiger det jo bare til vejrs… men jeg ved så ikke lige
I: det er simpelthen bare vanddamp..

I: hvordan opstår regn?
E3: fra vand og søer..
(herefter lang tavshed)
I: … Solen skinner ned på os.. [kraftig scaffolding, markerer kritiske aspekter]
E3: og så fordamper det .. op og bliver til skyer.. og så ...skal det jo også falde ned på et tidspunkt … og så bliver det til regn.
I: hvorfor falder det ned igen, ved du hvad der sker?
E3: fordi at det bliver varmere.. eller sådan et eller andet.. det kan vel kun være deroppe et antal dage?... nej, jeg ved det ikke… jeg ved kun, hvordan det gør (snurrer rundt med blyanten i en cirkelbevægelse)… sådan en cirkel

Procesforståelse af vandets kredsløb og dannelse af nedbør samt af havets betydning for kystklima blev totalt set vurderet til et gennemsnitsniveau på 2,65 (ca. 40 %). Da PISA ikke direkte tester dette område, foreligger der her ikke noget sammenligningsgrundlag.

Analysen viser tillige, at 42 % af drengene og kun 7 % af pigerne (!) er i stand til at skelne mellem døgn- hhv. årstidsvariation afledt af Jordens bevægelse. Ca. 1/3 er selv i stand til at bringe Jordaksen i spil, som element i en forklaringen (men som vi skal se, ikke nødvendigvis på korrekt vis). Videoanalysen viser samtidig, at eleverne trods store problemer med de underliggende mekanismer med rimelig sikkerhed når frem til at vælge det rigtige svar, når de bliver præsenteret for de oprindelige PISA-svarmuligheder (ca. 90 %).

På ny indikerer denne analyse, at målopfyldelsen i forhold til Fælles Mål er temmelig ringe, og at PISA-opgaven undlader at teste den forståelse og de proces-aspekter, som efter gængs opfattelse indgår i Fælles Mål.

Nedslag i arbejdsmåder og tankegange

I tilknytning til det praktisk-eksperimentelle arbejde med *solcreme*-opgaven blev eleverne spurgt, hvad de forbandt med et naturvidenskabeligt eksperiment eller naturvidenskabelig undersøgelse. Spørgsmålet ligger på kanten af Fælles Mål, idet 8. klasses trinmål (i fysik-kemi) primært handler om at *udføre* forskellige elementer af undersøgelse (*formulere, planlægge og gennemføre undersøgelser*) – men ikke nødvendigvis have en klar fornemmelse for karakteristika for sådanne undersøgelser. Alligevel kan det næppe være tilfredsstillende, at analysen viser:

- 4 elevsvar, som er indholdsmæssigt relevante og formuleres i fagrelevante termer:
    - eksempler: *"undersøger naturen"* og *"laver forsøg ud fra en hypotese"*
- 39 elevsvar som har en vis relevans, men er temmelig common-sense af natur og som ikke inddrager fagsprog:
    - eksempler:
        - *Sample r1: "laver nogle forsøg for at se hvad der sker"*
        - *Sample r2: "noget man kan teste fra naturen"*
        - *Sample r3: "Fysik og kemi"*
        - *Sample r4: "udendørs ting, noget med natur"*
- 33 direkte misforståede eller tautologiske elevsvar:
    - eksempler:
        - *Sample r1: "noget man laver i naturen - ikke i laboratorier"*
        - *Sample r2: "hvad der er bedst for naturen"*
        - *Sample r3: "gøre det professionelt"*
- 31 elever angav, at det havde de ingen ide om.

*I betragtning af, at Arbejdsmåder og tankegange er et centralt kundskabs- og færdighedsområde synes eleverne forbløffende uvidende om naturen af naturvidenskabelige eksperimenter og undersøgelser***.** På dette alderstrin burde naturfaglig undersøgelse være mere end ureflekteret gøren, men meget lidt tyder på, at det er tilfældet.

Det praktisk-eksperimentelle islæt var ikke voldsomt vidtgående, men gav alligevel mulighed for at vurdere aspekter af elevernes evne til at *formulere, planlægge,gennemføre og evaluere undersøgelser*. Disse træk udgør kernen i 8. klasses trinmål for arbejdet med Arbejdsmåder og tankegange i faget Fysik/kemi. Konkret blev elevernes formåen indenfor denne dimension vurderet i relation til 5 delaspekter (se nedenstående tabel). Det er vigtigt at pointere, at Solcreme-opgaven og VAP-setuppet ikke muliggør en fuld kortlægning af elevernes eksperimentelle kompetencer. Der er tale om at VAP leverer "det mulige" og relevante indblik, snarere end det fuldstændige.

Delaspekt af naturfaglige arbejdsmåder og tankegange	Antal elever, som demonstrerer at kunne dette aspekt (total: 125)
K1: Kan lave en håndværksmæssigt hensigtsmæssig realisering af den indledende plan	97
K2: tænker i variabelkontrol	5
K3: tænker i fejlkilder og minimering af sådanne	7
K4: kan revurdere forsøget/fremsætte forbedringsforslag	16
K5: har en forståelse af forsøget som model	18

Tabellen viser tydeligt, at *når først planen er lavet*, går det meget godt med at realisere den. Omvendt demonstrerer tabellen med al ønskelig tydelighed, at evnen til at formulere, planlægge og revurdere forsøg er svag. Det peger mest indlysende på, at elevforsøg gennemføres efter

”kogebøger”, med vægt på at følge beskrevne procedurer og uden eksplicit opmærksomhed på fagets arbejdsmetoder.

Undervejs i VAP-testningen af denne opgave spurgte vi også mere bredt ind til det underliggende område af Fælles Mål. Fælles Mål indeholder bl.a. krav om kendskab til *Energiformer* (Fysik-kemi) og *aktuelle miljøproblemer og deres betydning for menneskets sundhed og den omgivende natur* (Biologi). Først i 9. kl. trinmål i fysik skulle eleverne kunne beskrive virkningen af ioniserende stråling. Den bredere VAP-afdækning gik på viden om, hvad Uv-stråling er, samt kendskab til skadevirkninger på mennesker og miljø. Det sidste bedømt på et funktionelt niveau, uden at kræve indsigt i mekanismerne for skadevirkning. Ikke alle elevpar blev spurgt om samtlige aspekter, og undertiden gjorde elev-elev-dialogen det umuligt at tildele individuel score. Idet vi kun angiver de klassificerbare situationer fås følgende resultater:

	intet kendskab, blanke	udtrykker hverdagssproglig viden (a la nyhedsmedier)	Udtrykker viden, som involverer relevante fagtermer
Uv-strålingens natur (bølger, energi)	14	4	1
Uv-stråling som skadelig (mennesker, dyr)	24	72	4
Uv-stråling og det globale miljø (ozonlag)	31	59	2

Det er svært at se dette som en succes for den formaliserede naturfagsundervisning!

Opsummering på evalueringen i forhold til Fælles Mål
Analysen demonstrerer et stort gab mellem danske elevers faktiske formåen og de konceptuelle og proceduremæssige krav som Fælles Mål udtrykker. Hvis man vil fastholde den gældende tolkning af Scientific Literacy som naturfagenes rette mål, må man sikre sig, at eleverne får et kvalitativt anderledes udbytte af naturfagsundervisningen. Undervisningen kan med fordel styrke elevernes indsigt i forklaringer, argumenter og arbejdsmåder og nedtone facts og fastlagte procedurer. Denne pointe vil træde endnu tydeligere frem i næste afsnit.

Analysen indikerer, at PISA Science-opgaverne tester relevante dele af Fælles Mål, men tester dem på et funktionelt niveau, som ligger under de forståelsesstandarder der gælder det nuværende Fælles Mål. *Målopfyldelse i forhold til PISA-kriterier synes således at være en ringe målestok for målopfyldelsen i forhold til Fælles Mål.*
Den foreløbige tolkning er her, at PISAs opgaver ikke på valid måde implementerer de kompetencemål, som PISAs Scientific Framework tilsiger at ville. En mere definitiv afgørelse på dette punkt kræver imidlertid detaillerede studier af et større sæt af opgaver, end vi for nærværende har mulighed for at udføre.

Den sociokulturelle evaluerings særlige blik på fagsprog, forklaringer, argumentation og brug af artefakter.

Indenfor et sociokulturelt læringsperspektiv er læring af naturfag et spørgsmål om at have tilegnet sig naturfaglige tale- og synsmåder, tillige med at have udviklet brugen af naturvidenskabens værktøjer (fysiske såvel som symbolske). Bl.a. er dette tydeligt formuleret i amerikaneren J. Lemke’s (Lemke. 1990) hyppigt citerede udsagn: ”Learning science is learning to talk science”. For en sociokulturelt orienteret *evaluering* er det derfor indlysende at have et vist fokus på, hvorvidt og

hvorledes eleverne faktisk formår at tale om naturfaglige problemstillinger på måder, som overskrider hverdagssproget. Dette fokus er mindre udtalt, men dog tilstedeværende, i de danske læreplaner, fx skal eleverne allerede i henhold til trinmålene for Natur/Teknik efter 6. klasse kunne *benytte fagsprog og anvende abstrakte begreber.*

Mere generelt betyder VAP-evalueringens sociokulturelle opmærksomhedsvindue, at evalueringen giver et skarpere blik og et øget fokus på særlige områder, hvoraf vi her vil komme ind på :

- *Anvendelsen af fagtermer* (art, hyppighed o.s.v., fx inspireret af Wellington m.fl. (Wellington & Osborne. 2001)).
- *Entiteter til at bygge forklaringer op om* (’Entities’ & ’Explanatory stories’, inspiration fra Ogborn m.fl. (Ogborn, Kress, Martins, & McGillicuddy. 1996))
- *Argumentation* (jf fx Osborne m.fl. ((Osborne. 2005))
- *Brug af artefakter* (fx Säljö (Säljö. 2003; Schoultz, Saljo, & Wyndhamn. 2001b))

Yderligere er der i VAP-analysen en øget opmærksomhed på, hvorledes VAP ændrer den kommunikative situation og dermed rekonstruerer kravene til eleverne. Specifikt søges det belyst, hvordan dialogen lemper/skærper elevens muligheder for at forstå, hvad selve opgaven går ud på (”Opgave-tilegnelse”), elevernes konstruktion af et bud på et svar (”Meningsskabelse”), samt kravene til artikulation af et forståeligt svar (”Artikulation”). En umiddelbar hypotese kunne være, at VAPs medierende dialog gør det nemmere at forstå, hvad opgaverne går ud på, kan understøtte konstruktionen af svarmuligheder – men at kravene til selvstændig artikulation samtidig øges.

Anvendelsen af fagtermer

I PISA 2006 Science testen indgår de sproglige aspekter, men implicit og sædvanligvis udenfor den gængse rapportering. Fx forudsætter et kvalificeret svar i visse opgaver (som fx Solcreme-opgavens første *Multiple Choice*-spørgsmål), at man kan afkode trigger-ord i de fortrykte valgmuligheder. I andre opgaver af *Open Response*-typen får eleven i en del tilfælde kun point, såfremt bestemte fagtermer indgår i elevens selvstændige svar (ofte uanset svarets kvaliteter i øvrigt).

Der findes en række empiriske forskningsstudier, der fokuserer på, hvilke enkeltord og ordtyper eleverne i naturfag hyppigst oplever som svære (fx Wellington & Osborne. 2001). På basis af disse studier er der bl.a. formuleret en ”Taxonomy of the words of science”, som fremgår af nedenstående boks (Wellington & Osborne. 2001)

BOX 2.4 A taxonomy of the words of science

Level 1: Naming words
1.1 Familiar objects, new names (synonyms).
1.2 New objects, new names.
1.3 Names of chemical elements.
1.4 Other nomenclature.

Level 2: Process words
2.1 Capable of ostensive definition, i.e. being shown.
2.2 Not capable of ostensive definition.

Level 3: Concept words
3.1 Derived from experience (sensory concepts).
3.2 With dual meanings, i.e. everyday and scientific: for example, 'work'.
3.3 Theoretical constructs (total abstractions, idealizations and postulated entities).

Level 4: Mathematical 'words' and symbols

Vanskeligheden af ordene vokser med niveauet her. Dele af Wellington m.fl.'s studier – og især de dybtgående sprogstudier af Halliday m.fl (Halliday & Martin. 1993) peger dog på, at naturvidenskabens særegne sprogkarakter og for så vidt også dens kompleksitet, i højere grad er forbundet med sætningsstruktur end med enkeltord (Halliday & Martin. 1993). I VAP-analysen har vi holdt os til at undersøge brugen af enkeltord – samt kaste et nærmere blik på to specifikke genrer, som i særlig grad bærer naturvidenskabens bestræbelse og sociale praksis, hhv. *naturvidenskabelige forklaringer* og *naturvidenskabelig argumentation*.

Entiteter til at bygge forklaringer op om
Ogborn et al (Ogborn, Kress, Martins, & McGillicuddy. 1996) har begået et af de mere seriøse fagdidaktiske (i modsætning til filosofiske) forsøg på at indfange essensen af naturvidenskabelige forklaringer. I deres optik er en forklaring en særlig historie, som

- bygger på et sæt af aktører ("protagonists")/hovedrollehavere, som i naturvidenskab oftest *ikke* er personer, men begreber.
- normalt forudsætter, at det er klart, hvad den enkelte aktør *er* og *kan*.
- indebærer en serie af hændelser, hvor aktørerne gør noget af det de kan
- hændelserne har en konsekvens eller et udfald i overensstemmelse med naturen af aktører og hændelsesforløbet. Her indføjes et element af årsag-virkning til den forklarende historie. Underforstået ligger heri også, at ikke hvad-som-helst anerkendes som årsag. Nutidig naturvidenskab beskæftiger sig kun med *den virkende årsag, eller causa efficiens*, som hos Aristoteles kun er én af 4 årsagstyper (se fx Den Store Danske Encyklopædi)

Afgørende træk i den naturvidenskabelige forklaring fremgår af følgende citat:
"Everyday explanations are in terms of familiar entities doing familiar things. Scientific explanations are often in terms of unfamiliar entities doing unfamiliar things, and the student is a stranger in an unknown world. It follows that much explanation in science classrooms is not the explanation of phenomena, but is the explanation of resources the student needs in order to explain phenomena." (p. 13).

Det pointeres altså, at naturvidenskabelige forklaringer først bliver mulige og meningsfulde, når man har konstrueret de nødvendige *entities*[1], dvs. har det nødvendige minimum af klarhed over, hvad den enkelte aktør er og kan. Som forfatterne anfører: "Before we can explain respiration we have to tell about lungs, blood, oxygen, carbon-dioxide and hemoglobin." (p.13). Hvis ikke eleven har grebet de grundlæggende træk af et begreb, har styr på 'forklaringsenhederne' bag begrebet, vil vedkommende ikke kunne forklare noget som helst med brug af det. En vis grundlæggende begrebslæring er en forudsætning for, at eleverne kan håndtere naturvidenskabelige forklaringer og/eller argumentation – hvad enten de selv skal frembringe den eller gennemskue holdbarheden af forklaringer/argumentation de måtte møde.

Argumentation

Videnskabshistorikeren S. Toulmin (Toulmin. 1969) har leveret en meget kendt analyse og model for naturvidenskabelig argumentation, som har tjent som grundlag for en længere række studier af elevers argumentationsevne i naturfag (fx (Osborne. 2005; Erduran, Simon, & Osborne. 2004)). Bl.a. har Osborne m.fl. argumenteret for, at argumentationsevne er den fundamentale komponent i *scientific literacy* og udviklet et omfattende undervisningsmateriale til både lærere og elever med henblik på at udvikle denne komponent. Osborne et al har bl.a. udviklet en taksonomi til at beskrive forskellige grader af sofistikation i elevargumentation. I den mindst differentierede form skelner de mellem tre niveauer af argumentationsformåen:

- simple påstande
- påstande med forskellige grader af belæg/hjemmel/rygdækning
- påstande med forskellige grader af belæg/hjemmel/rygdækning OG mulige gendrivelser

I denne generelle beskrivelse er der ingen specifik reference til naturvidenskab. Genren *naturvidenskabelig argumentation* træder først frem, idet man specificerer hvilke typer af belæg (fx empiriske, laboratoriefrembragte, tankeeksperimenter m.m.) man meningsfuldt kan bringe i spil, at gyldig hjemmel hviler på eksistensen af forklarende mekanismer af universel natur – og at naturvidenskabens særlige valideringsmekanismer (replikation, peer-review etc.) indgår som et vigtigt aspekt i en brugbar rygdækning.

Brug af artefakter

Vygotsky fremhæves ofte som sociokulturel læringsteoris grundlægger - og hans fundamentale betoning af, at menneskelig læring medieres af fysiske og symbolske redskaber står uanfægtet. Mest centralt står sproget som både et symbolsk redskab OG et medium for tænkning - derfor det dominerende fokus på sprog og sproghandlinger i VAP-konceptet. Men: en række nyere studier (se (Säljö. 2005)) har i forlængelse af Vygotsky dokumenteret, at også fysiske artefakter understøtter elevers begrebsdannelse og deres præstation i internationale test-sammenhænge (fx (Schoultz, Saljo, & Wyndhamn. 2001b)). Bl.a. påviser den sidst anførte reference, hvorledes inddragelse af en globus undervejs i et interview med elever totalt ændrer billedet af elevernes formåen: med denne 3D-repræsentation inden for rækkevidde har eleverne generelt ikke svært ved at forstå, hvorledes man kan leve i Australien – med benene mod Jorden og uden at falde af. Inspireret af dette arbejde har vi i VAP fundet det relevant og interessant at undersøge, hvordan konkrete artefakter som kort og globus (geografiorienteret opgave), podeplader (biologiorienteret opgave) og eksperimentelt udstyr (solcreme-opgaven) synligt understøtter elevernes forståelse og svar-konstruktion.

[1] Entity har på engelsk en dobbeltbetydning som både eksistens/væren/væsen og et selvstændigt hele. Et dækkende dansk udtryk kunne være betydningsenhed eller forklaringsenhed. Vi vælger det sidste.

Scaffolding interactions
Stilladseringsbegrebet (engelsk: "Scaffolding") har fyldt ganske meget i den pædagogiske diskurs i Danmark i det seneste tiår (se fx (Tønnes Hansen & Nielsen. 1999)). Det har været mere sparsomt med konkrete bud på, hvilke elementer en hensigtsmæssig stilladsering omfatter, og hvilken struktur den kan have. Langt henad vejen forekommer den klassiske introduktionsartikel af Wood et al (Wood, Bruner, & Ross. 1976) stadig at være et af de mere brugbare afsæt for en analyse af, hvordan stilladseringen foregår. Heri beskrives stilladseringen ud fra et lærer/tutorperspektiv med følgende elementer:

- vække den lærendes interesse for opgaven
- reducere frihedsgrader og valg til et passende niveau
- fastholde det overordnede mål med opgaven
- fremhæve/tydeliggøre kritiske aspekter af opgaven
- modellere en hensigtsmæssig tilgang til opgaven
- hjælp til at håndtere frustration

Man kan diskutere *om* der overhovedet forekommer stilladsering i VAP-testsituationen, i særdeleshed hvis man forbinder dette med en bevidst pædagogisk praksis. Imidlertid forekommer der social mediering, hvor væsentlige funktionelle aspekter kan indfanges af de samme termer: forskningsassistenternes input til eleverne i de individuelle interviews kan ses i dette lys – ligesom elevernes kommentarer og hjælp til hinanden undervejs i pararbejdet omkring *Solcreme*-opgaven. Der er derfor foretaget en VAP-analyse på grundlag af disse kategorier – og med et primært fokus på den første type medierende interaktion, altså vekselvirkningen mellem forskningsassistent og gymnasielever.

VAP-billedet af elevernes evne til at anvende fagtermer
Det empiriske grundlag for at udtale os om elevernes evne til at anvende fagsprog er flersidigt, dels det tidligere omtalte geografispecialestudium (Jensen. 2007) med udgangspunkt i VAP og med støtte af én af denne rapports forfattere, dels en mere dybtgående stikprøve-analyse. Specialestudiet inddrager samtlige test-interviews i tilknytning til opgaven *Klimaforskelle*, mens stikprøven går i noget større detalje med et mindre antal (N=10) af disse. Lejlighedsvist vil resultater fra den direkte PISA-sammenligning blive brugt til at belyse dele af dette afsnits problemstillinger.

Det overordnede billede baserer sig på elevernes evne til at "reproducere viden", fx ved på rimelig vis at benytte sig af fagtermer. Elevernes evne blev *helhedsvurderet* på en 5-punkts Likert-skala (meget dårlig, dårlig, middel, god, meget god), hvor *meget dårlig* gives til en elev, der ikke selv benytter nogle fagtermer og begreber, og som stiller sig uvidende overfor dem. Ca. ¼ af eleverne placeres i denne kategori og tilsammen falder ca. 55 % af eleverne i én af de to dårligste kategorier. Til sammenligning falder kun ca. ¼ af eleverne i én af de to bedste kategorier. Helhedsvurderingen af elevernes anvendelse af fagsprog er således klart negativt tonet.

Fra analyserne af test 1 og 2 kan vi tilføje, at kun 4 (af 125) elever havde kendskab til ordet "[geospecifik][1]", mens kun 3 kendte på rimelig vis til ordet "reference"/"referencestof". Sidstnævnte fagterm er et centralt meningsbærende ord i én af PISA-sættets multiple choice-opgaver (solcreme-opgaven) - uden et dækkende kendskab til termen udvikler besvarelsen af denne opgave sig til noget nær gætteri. Man kan diskutere, hvorvidt netop disse specifikke fagtermer retteligen hører hjemme i alle elevers aktive vokabularium, men det massive "udfald" er alligevel

[1] Strøget efter krav fra Skolestyrelsen.

bekymrende – og *kunne* antyde, at præcise termer og mere fagspecifik sprogbrug i almindelighed er blevet et forsømt område i naturfagsundervisningen. Noget sådant kunne i hvert fald forklares som en konsekvens af hverdagskulturens hastige indmarch i skolens liv og den i øvrigt positive udvikling i retning af større elevcentrering i naturfagsundervisningen.

For at nå til større afklaring omkring problemets omfang og natur har vi for et stikprøve-sample kortlagt *hvilke* fagtermer der drages ind i interviewene, og i hvilken udstrækning eleverne kan siges at kende til dem. Følgende sondring er blevet anvendt:

En fagterm anses for kendt, når eleven enten selv inddrager den på adækvat vis i samtalen, eller er i stand til direkte at forklare betydningen, evt. respondere på en måde som overbeviser om en *basal* forståelse af termens indhold.
En fagterm opfattes som ukendt, når eleven ikke er i stand til på fyldestgørende vis at deklarere eller respondere på termen.

Analyse af stikprøvesamplet viser, at:
*Der er stor variation mht. hvor mange fagtermer der drages ind i interviewene***:** fra 3 til 15 af de nedenfor listede ord bringes i spil i den enkelte interview-samtale. 2 af de 10 stikprøveelever anvender stort set ikke fagtermer i samtalen - og når de undtagelsesvist gør det, er det forkert. Her skal det bemærkes, at en vis (mindre) del af variationen kan henføres til forskningsassistenterne og de semi-strukturerede forskningsinterviews. Ikke alle interviews gav helt samme afsæt for inddragelse af bestemte fagtermer. Sammenligning af hyppigheder på tværs af interviews er derfor lidt mere usikre end generelle træk og pointer indenfor det enkelte interview.

Stikprøven førte til følgende akkumulerede forekomst af fagtermer:
Forekomsten af fagtermer i stikprøve-interviews:
Første tal i parentes angiver total antal forekomster – andet tal angiver, hvor ofte termen var *ukendt* for eleven (jf. definitionen ovenfor)

[geospecifik]/[geospecifik][1]-figur (6/6)
Nedbør (6/4)
Atmosfæren (1/1)
Front (vejr) (1/1)
Lufttryk (1/0)/Højtryk "/(1/0)/Lavtryk (1/0)
Varmeenergi (1/1)
Vanddamp (1/1)(/Fordampning (6/1)/Fortætning (2/1)
Kredsløb (1/1)
Tidszoner (1/0)
Varmefylde (2/2)/varmekapacitet (1/1)
Massefylde (1/1)
Gulf-strømmen (1/0)
Ækvator (4/0)
Sydlig/nordlig halvkugle (2/0)
Længdegrad (3/0)/breddegrad (3/0)
Greenwich ("central tid") (1/0)
Klima (3/2)/Klimazone (3/1) /Klimabælte (3/1)
Kystklima (1/0)/ Fastlandsklima (2/0)
Tempereret/subtropisk/tropisk (klimabælte) (12/3)
Nåleskov/ørken/savanne/regnskov (6/1)
Monsun/regntid (2/2)
Arktis (1/0)/Antarktis (3/2)/Sydpolen (2/0)
Brasilien (1/0)/Sydafrika (1/0)
Himalaya (1/0)

Umiddelbart ser fordelingen af kendte vs. ukendte meget tilforladelig ud: 57 kendte fagterm-forekomster overfor 31 ukendte. Men: der tegner sig et billede af, at eleverne *først og fremmest er stærke, hvad angår simple faglige "betegnere"* ("Naming words", jf. Wellington) knyttet til klimabælter & - zoner. *[geospecifik]/[geospecifik][2]-figur* er tilsyneladende ikke en term, som dukker hyppigt op i undervisningen. *Nedbør* er selvfølgelig i sin hverdagsforståelse velkendt af eleverne, men de færreste opfatter det som en kategorial fællesbetegner for regn, sne og slud. På samme måde kan alle elever indgå i en "hverdagsagtig" samtale om *Klima,* men en meget stor del af eleverne vil ikke have bidt mærke i, at man i naturfag faktisk forbinder termen med gennemsnitlige årsvariationer af en række parametre og udledt på basis af data fra mindst 30 år. Det er et gennemgående træk, at hvor hverdagssprog og videnskabssprog overlapper med forskelligt meningsindhold i et "Naming Word", vil den almindelige elev orientere sig i henhold til den hverdagssproglige forståelse. De kendte "Naming words" er - udover deres lave taksonomiske niveau – karakteristiske ved, at de typisk vedrører 8. klasses pensum, altså henviser til den umiddelbart foregående undervisningshorisont. Det antyder, at der kan være problemer med fagligt niveau og mere langtidsholdbar vidensopbygning.

Det er symptomatisk, at problemerne opstår i relation til mere abstrakte begreber og processer, herunder adskillige af de for PISA-opgaverne kritiske fagtermer & 'entities'**.** Ingen elever i stikprøven leverer således en redegørelse for varmefylde/varmekapacitetsbegrebet, som gør det muligt at forstå og forklare forekomsten af [klimatyper]. En pæn del af eleverne ved godt, at skyer er fremkommet ved fordampning af (flydende) vand, men dette er hverken koblet til

[1] Strøget efter krav fra Skolestyrelsen.
[2] Strøget efter krav fra Skolestyrelsen.

energibetragtninger, forestillinger om tilstandsformers mikroskopiske natur, reversibilitet eller indsigt i hvad der driver en proces som *fortætning*.
Indtrykket er således blandet – JO, mange elever kan godt samtale med brug af et vist fagsprog – men NEJ, de kritiske 'entities' forekommer ikke tilstrækkeligt tilegnede. Dvs. eleverne behersker det nominelle niveau, men ikke det konceptuelle.

VAP-billedet af elevernes evne til at konstruere naturvidenskabelige forklaringer.
Som pointeret ovenfor, kan man ifølge Ogborn m.fl. kun konstruere naturvidenskabelige forklaringer i det omfang, man har konstrueret et rimeligt billede af de størrelser, som qua deres natur indgår som aktører i den kæde af hændelser, som udgør forklaringens narrativ. Man skal, så at sige, have tilegnet sig et basalt kendskab til "forklaringens byggeklodser". Analysen ovenfor er en første indikation af, at mange elever savner det nødvendige, basale kendskab.

Det er en væsentlig pointe, at PISAs MC-opgaver typisk er bygget op omkring et antal formulerede *bud på forklaringer*. Det er så op til eleverne at udskille "den rigtige" forklaring – ved at genkende meningsfulde og relevante protagonister og sikre, at deres iboende handlingspotentialer udfoldes i en kausal-logisk handlingsfølge svarende til svarmuligheden. Ved at studere de forskellige MC-optioner er det muligt at analysere, hvad der er de kritiske forklaringsenheder i de forskellige PISA-opgaver – og hvad der er tilstrækkelige hhv. fyldestgørende konstruktioner af dem. Hvis man gør dette for de forskellige delspørgsmål indenfor opgaven om *Klimaforskelle* tegner der sig således følgende billede af relevante forklaringsenheder, hvor kun det med sort er nødvendigt for at besvare PISA-opgaverne:

Item	Forklarings-enheder	Hvad det er	Hvad det kan
Q01 (klima-beskrivelse)	Klima	• Et begreb • Sammenfatter årsvariationen i bl.a. *temperatur* **OG** *nedbør* på en bestemt lokalitet – som et *gennemsnit* over de seneste 30 år	• Afbildes i en grafisk repræsentation ("[geospecifik]-figuren") • Bruges til at beskrive klimaet, dvs. typiske/gennemsnitlige træk ved det situerede "vejr".
Q02 (årstids-variation)	Jorden	• Kuglelignende objekt i rummet – ikke så langt fra Solen • tillægges en nord-syd-ende og en skillelinie ("Ækvator") • På et bestemt sted er der tilbagevendende årstidsvariation • Årstidsvariationen mere eller mindre modsat i nord-syd-enderne • Når der er mørke på den ene side, er der lyst på den anden	• bevæge sig i forhold til Solen • bevægelsen sikrer, at nord-enden og syd-ende på skift vender ind mod Solen. • opvarmes af lys fra Solen • bevægelsen foregår *om* Solen – og omløbstiden er (ca.) et år. • samtidig med bevægelsen *om* Solen roterer Jorden om sin egen nord-syd-akse. Rotationstiden er eet døgn. • Rotionsaksen er skrå i forhold til Jord-Sol-

			baneplanet • Rotationsaksen ændrer *ikke* retning på noget tidspunkt • kan repræsenteres i form af kort og globus
	Solen	• Stjerne	• udsender uophørligt sollys i retning af Jorden (i alle retninger)
	Sollys	• en slags energi	• kan opfanges af Jorden • kan omdannes til varme og opvarme ting det rammer
Q04 (Kyst- hhv. fastlandsklima)	Australien	• Kontinent (i hvert fald Stort land, d.v.s midten er kystfjern) •	
	Varmekapacitet	• udtryk for et systems "vanskelighed" ved at opvarmes og afkøles, når det tilføres energi • Et begreb, der har en eksplicit definition i varmelæren • formelt den varmetilførsel, der skal til for at få en temperaturstigning på en grad i et system. Forudsætter et billede af underliggende energitransport.	• kan bruges til at sammenligne temperaturstigning i forskellige stoffer, når de tilføres energi i form af varme. Vand er fx "vanskeligere" at opvarme og afkøle end jord. • kvalificeret sammenligning forudsætter, at man har kendskab til stofmængder og *specifikke* varmekapaciteter
	Breddegrad	• Tværgående stribe på Jordoverfladen • Parallelt med Ækvator	• Afbildes på kort og glober med et vist interval • Breddegraden udtrykker indirekte afstanden til Ækvator - og har betydning for solindstrålingen

For at konstruere forklaringer, der udtrykker reel forståelse, bør eleven have styr på disse forklaringsenheder i hele tabellens omfang. Farvekoden i skemaet er her indført for at anskueliggøre, at PISA-opgaverne lader sig besvare med langt mindre end det fulde omfang: kun det med sort markerede er nødvendigt her. I første spørgsmål er det således datagrundlagets karakter og forholdet mellem situeret vejr og gennemsnitlig klimabeskrivelse, som lades ude. I det næste *behøver* man ikke nogen forståelse for Solsystemets mekanik – man skal blot have et billede af, at nord- og sydpolen på skift vender ind mod Solen. I det sidste spørgsmål gør opgavestillerne opmærksom på, at to lokaliteter som skal sammenlignes ligger på samme breddegrad. Underforstået: solindstrålingen må forventes at være den samme, hvorfor eventuelle klimatiske forskelle må tilskrives noget andet – konkret altså tilstedeværelsen af hav i den ene og ikke den

anden situation. Hele denne forudsætning om variabelkontrol som et grundlag for en fyldestgørende forklaring "opdager" de færreste elever – og alligevel kan de få fuldt point. De kan også klare sig igennem med et noget naivt/"anløbent" billede af varmekapacitet.

PISA intenderer at indfange elevernes evne til at forklare naturvidenskabelige fænomener. Med Klimaopgaven som konkret eksempel tydeliggør vores analyse, at det er tvivlsomt, om man lever op til intentionen. For det første er dét at vælge et bud på forklaring klart forskelligt fra – og sædvanligvis nemmere end - selv at konstruere en forklaring. For det andet kan eleverne komme frem til den korrekte svarmulighed med forklaringer, som enten er utilstrækkelige, hverdagsagtige eller direkte forkerte. Dette sidste er blevet ganske tydeligt i nærstudier af elevernes argumentation. Vi giver her nogle interview-sekvenser, som kan belyse og belægge denne påstand. "I" betegner intervieweren og "E" eleven i den enkelte sekvens. Rød font er brugt til at fremhæve kritiske dele af elevudsagn, lilla indeholder deskriptive træk ved situationen, mens blå font angiver analysator-kommentarer. Q2 er et spørgsmål om årstidsvariation, mens Q4 er spørgsmålet om kyst- og fastlandsklima.

Q2 (E2): I betoner i sit oplæg, at Bynord og Bysyd (som her bruges som betegnelse for de to konkrete byer, der indgik i PISA-opgaven) ligger på nogenlunde samme længdegrad, og nogenlunde lige langt fra Ækvator. Den ene på den nordlige, den anden på den sydlige halvkugle.

I: Hvis vi kigger specielt på temperaturen, hvad kan vi så sige?
E: at det er generelt varmere i Bysyd … og de har sommer og vinter lige præcist modsat hinanden.
I: …. Og hvad kan det skyldes?
E: at den ene ligger mod syd og den anden ligger mod nord (I: ja)….
I: viser de 4 mulige PISA-forklaringer.. læser mulighederne højt…
E: så ville jeg tage C! [korrekt, og med stor overbevisning]
I: hvorfor ville du tage den?
E: fordi man kan se på Bynord (peger på grafen), at der er varmest i juni-juli… og det må så skyldes at de er tættere på Solen [E anvender således empirisk belæg i sin argumentation, om end koblingen til afstand forbliver uudsagt]...
I: hvad betyder det, at den nordlige halvkugle hælder mod Solen…. Du må godt bruge globussen til at forklare..
(stor globus med synligt skrå akse placeret lige foran eleven)
E: at Jorden har en.. skrå akse.. og jo tættere på Solen jo varmere er der … fordi Solen udsender varmeenergi

Herefter følger en sekvens, hvor eleven forsøger at rotere globussen OG med hånden angive en position for Solen, som modsvarer sommer på den nordlige halvkugle. Sommer på den nordlige halvkugle repræsenteres ved at Solen placeres højt over Ekliptika. E omtaler ikke døgn-rotationen i en diskussion af lys og skyggesider. Derefter er der tomgang. I ender med at dreje hele globeopstillingen, således at nord-syd-aksen skifter retning. Må selv fastslå, at dette svarer til at gå fra nordlig sommer til nordlig vinter. Herefter konstaterer E, at i vores vinterposition er temperaturen højest i Bysyd:

E: "fordi linien mellem Bysyd og Solen .. den er kortere (peger)" [tydeligt, at den sydlige halvkugle nu er tættere på hånden, der illuderer Sol. Det forkerte afstandsforhold forstærker indtrykket af, at afstand er den afgørende faktor].
E: Den sydlige halvkugle har ligesom en fordel ved hele tiden at være længere fremme mod Solen.

[E vælger altså det korrekte svar & med stor overbevisning– men har faktisk ikke den korrekte underliggende forklaring, og vi nærmer os aldrig et svar på, hvad der ligger omme bag, at Jord-aksen skifter orientering i forhold til Solen. Afstanden til Solen anses fejlagtig at være forklaring på årstidsvariationen – og denne afstand reguleres tilmed ikke af Jordens banebevægelse, men af Jordhældningen. Jorden som entitet i bevægelse om Solen og i samtidig rotation om sin egen skrå akse virker ikke konstrueret. Den sammenhængende mekaniske model af Solsystemet mangler]

Q2 (E3):
I rammesætter på ny opgaven om Bysyd og Bynord:

I: kan du se nogle forskelle

E: Ja, der er koldest I Bysyd I juni, juli, hvor det er varmest I juni, juli, august i London.. og det har nok også… og det har noget at gøre med Jordens hældning… fordi.. de ligger jo på hver sin side af Ækvator [først da henter I sin globus frem, den har hidtil stået under bordet]
E: … så derfor: når det er sommer heroppe (peger nordlig halvkugle] så skinner Solen heroppe.. og når det er vinter, så skinner Solen hernede…
(Eleven markerer med hånden, at Solen flytter sig fra en placering over Ekliptika til en position under Ekliptika).
E:... p.g.a. hældningen….. (E drejer lidt hjælpeløst på globen, udstråler ikke helt at kunne få sammenhæng i, hvordan det med hældningen kan bringes I spil. Vi kommer det aldrig nærmere. Da I efterfølgende læser de 4 PISA-muligheder op udpeger E prompte den korrekte.] [Eleven her har altså selv en viden om, at hældningen er det afgørende. Han mangler imidlertid at forene dette med en konstruktion af Jorden i årlig bevægelse om Solen & med rotation om en uforandret akse. I mangel af en mere sammenhængende konstruktion lader eleven Solen bevæge sig – og levere en slags løsning på problemet. Elevens kropsprog afslører dog samtidig klart, at han ikke selv er fornøjet med sin "forklaring"]

Q2 (E4) Efter indledende strukturering udpeger E det korrekte svar blandt PISAs bud på forklaringer. Heri indgår Jordens hældning. Opgaven er nu at redegøre for, at denne forklaring er rimelig.
(forevist stor fin globus):
I: hvilken betydning har det, at den ligger ned?
E: …at der ikke er en konstant .. Sol det samme sted… at der også bliver nat.. [ups] .. og at det også bliver vinter….(smil)…jeg ved ikke præcist, hvad jeg skal svare på det, altså…

Herefter må I kæmpe en sej (og ikke ligefrem) kamp for at få E til at sige, at lyset falder *mere skråt i Danmark* om vinteren, men da hun spørger, om det har betydning for temperaturen, lyder svaret:E: jaah.. det må det jo ha.. men jeg ved ikke lige præcist hvordan..?

[Strengt taget *kunne* man have strikket en slags sammenhængende og relevant (del)forklaring sammen – uden eksplicit reference til Solsystemet: Jordaksens hældning ift. Solen varierer henover året. Derfor falder sollyset i vekslende grad skråt ind. Jo mere skråt desto mindre energi afsættes, og desto køligere må klimaet forventes. Som man kan se af svaret, mangler der flere trin i denne forklaring. Mest problematisk er det måske, at dag-og-nat-variationen her *også* forbindes med Jordens hældning]

Q4 (E5):
I: hvis vi nu kigger på to byer… Kystby som ligger *helt ud til kysten* i Land (viser på kort) og Kontinentby som ligger *midt inde i* Kontinent (viser på ny) [igen anvendes synonymer for konkret anvendte navne].[I laver scaffolding her, ved at high-lighte kritiske aspekter].
Herefter beder I E om at beskrive klimaet de to steder:
E: i Kystby er der store temperaturforskelle
I: er der store temperaturforskelle?
E: øh.. jeg kom lige til at bytte om på dem.. der er næsten *ingen* [temperaturforskelle. …de har meget nedbør, det skyldes nok at luften over jorden er… varmere end det luft der er over havet.. fordi det bliver kølet ned.. og så mødes de to jo lige der.. ud til kysten.. og så regner det … [altså i stand til at forklare ud fra en konstrueret entitet: varm-luft-møder-kold-luft-udløser regn]

Efter oplæsning af 4 svarmuligheder:
E: der ville jeg jo nok tage B så [den korrekte]…
I: hvorfor?
E: fordi at øh.. fordi at det tager lang tid at opvarme havet.. fordi der hele tiden er strømme… der fører fx ..koldt vand ned fra Nordpolen .. og Sydpolen.. op til det andet vand… men på land ... på landet tager det ikke så lang tid, fordi at der er sand ... og jord og sådan noget ... der suger varmen til sig…
I: ja.. ja.. er det kun pga. strømmene, at vandet er længere tid om at varme op?... eller har det noget med vands varmefylde at gøre? [her får E faktisk det hele foræret på et sølvfad]
E: ….det ved jeg ikke helt… [men griber det ikke, E ender med at frafalde uddybning]

[dialogen er interessant her, fordi eleven faktisk genererer en *alternativ* naturvidenskabelig forklaring – den langsommere opvarmning af havet tilskrives her kolde havstrømme. Ikke den vigtigste forklaring, men i princippet *en mulig* forklaring. E har et billede af at nogle stoffer "suger varmen til sig", hvilket lyder som en "farlig"

sammenblanding af varmeledningsevne, varmefylde og en særlig affinitet for varme!? Svaret afdækker i hvert fald, at varmefyldebegrebet ikke er konstrueret på en måde, der gør det anvendeligt for E]

Q4 (E6):
E kan placere Kystby i Land, men mener, at Kontinentby ligger i "*Amerika, et eller andet sted"* [hvad den ikke gør]

E beskriver klimaet de to steder ud fra [geospecifik]figurerne præcist og relevant.
I: hvordan kan det være, at der er så stor forskel.. de ligger jo lige langt fra Ækvator?
E: jeg tror, at.. det er det samme.. som Danmark.. det ligger tæt ved vandet ... og … Kontinentby det er så ... jeg kan ikke huske, hvad det hedder … fastlandsklima eller sådan noget …..
I: hvad har det af betydning, at det er fastlandsklima?
E: æh… (fortrækker ansigtet gevaldigt).. vandet det er langsommere til at varme op.. i forhold til ... i forhold til Jorden…og…men det holder vist også længere på varmen.. [E leverer således selv en brugbar forklaring, uden efterfølgende at kunne udtrykke dette i termer af varmefylde etc.]

[Her kan vi umiddelbart glæde os over et stk. acceptabel forklaring. Det virkelig interessante er imidlertid, at da E efterfølgende bliver bedt om at udpege den korrekte forklaring blandt PISAs bud på forklaring, vælger E forkert. Og der skal faktisk en temmelig omstændelig stilladsering til, før eleven indser, at svarmulighed B svarer til E's egen foregående forklaring. Nogen gange kan man faktisk vælge det forkerte bud på en forklaring, på trods af at man selv har skruet en brugbar version sammen! Selv om den slags tilfælde ikke er voldsomt hyppige i samplet, er det yderligere et eksempel på, at PISA-resultaterne kan være misvisende]

Q4 (E7):
I: hvad tænker du om de to temperaturkurver i forhold til hinanden? [fastholder, highlighter med henvisning til [geospecifik]figurer]
E: jeg tænker, at den her (peger på Kystby) er stabil i temperatur.. der er der sådan 20 grader … hele året…ca.. men stadigvæk meget regn.
I: ja
E:….. kunne skyldes dampen fra havet.. når det køler af.. eller når det bliver aften, ikke? [en pudsig blanding af noget korrekt og noget uudgrundeligt].
I: men hvad med temperaturerne? (henviser til kurverne og placeringerne på samme halvkugle)..... i Kystby er der ikke den store variation.. det ser ud som de nærmest aldrig har vinter…[medieringen tydeliggør, at der er noget der fortjener forklaring her]
E: …..nææh…..
I (viser nu PISA' bud på forklaringer og læser dem højt).
E vælger C, retter til B.. [den korrekte] og derefter "nej, nu skal jeg lige tænke mig om"… går derefter i udelukkelse af forskellige svarmuligheder. Klart i tvivl om brugbarheden af C (som er et deskriptivt statement, som måske kobler temperatur og nedbør til hinanden – og i hvert fald ikke synligt *begrunder* temperaturforløbet). .. og pludselig:
E: .. så sku det være den der.. fordi at æh… vandet …æh..vand det har jo en massefylde, som.. der skal man bruge meget energi for at varme det op… [bortset fra brug af termen "massefylde" i stedet for varmefylde, ligner det en brugbar konstruktion af det relevante begreb]
I: okay
E:.. hvorimod en ørken.. der skal du ikke bruge lige så meget..
I: nej
E: og derfor… hvis du lægger mere energi i vand, end du gør i sand.. så vil du jo også få mere energi ud af den… og det vil så give damp… [her bliver man på ny noget desorienteret, hvordan indgår dampen i ræsonnementet? Og ydermere: kun så længe solenergien bindes i havet kan man bruge et simpelt varmefylde-argument for kystklima. Hvis man først lader energien frigive på ny er vi jo lige vidt. Derfor er forklaringen ikke aldeles overbevisende].

Eksemplerne illustrerer meget godt, at de centrale forklaringsenheder gennemgående ikke er konstrueret på et niveau, der gør det muligt for eleverne at give fyldestgørende forklaringer. Tendensen er imidlertid, at eleverne trods dette oftest alligevel er i stand til at udpege den korrekte PISA-svarmulighed. Faktisk er de danske elever i PISA-samplet så gode til at vælge svar, at de

fremstår som bedre end verdensgennemsnittet i den originale PISA 2006 Science (+5 Rasch-point), hvad angår delkompetencen *Explaining Phenomena Scientifically!*
I rapporteringen her har vi kun fremlagt nogle få eksempler, men et overslag på tværs af samtlige analyserede klima-besvarelser indikerer, at eleverne mere generelt har store problemer med at generere fyldestgørende forklaringer: *I hele analysesættet er skønsmæssigt 60 % af elevsvarene ikke i nærheden af en korrekt forklaring. Knap 20 % har ansatser til en forklaring, mens kun godt 20 % producerer acceptable forklaringer - sammenholdt med at ca. eleverne i ca. 80 % af tilfældene faktisk kan udpege det korrekte PISA-bud på en forklaring!*

Før vi generaliserer alt for vidtløftigt på dette grundlag, er vi nødt til at kaste et kritisk blik på nærværende analyses begrænsninger: For det første er der kun foretaget detaljeret sociokulturel analyse på en mindre del af samplet, og man kan derfor stille spørgsmålstegn ved repræsentativiteten af det analyserede sample. Men når man undersøger de analyserede elevers scorer i såvel VAP som PISA ligger de faktisk en anelse *over* gennemsnittet i hele samplet. Derfor er der intet der tyder på, at vi med valget af sample overdriver elevernes problemer med at forklare! Det andet væsentlige forbehold er, at kun klimaopgaven i øjeblikket er analyseret. Klimaopgaven tilhører vidensområdet *Earth and Space Systems*, hvor danske elever ifølge den danske PISA 2006-rapport klarer sig relativt dårligt (9 Rasch-point under OECD-gennemsnittet). Her er der således grund til at tro, at en bredere analyse kunne opbløde konklusionerne en anelse. Problemerne indenfor det aktuelle analysesample er imidlertid så markante, at opblødningen højst kan mildne *graden* af problematisering, ikke essensen af problemet, som synes at have to dele:

- *Folkeskolens "forklaringsproblem"*: Eleverne har ganske enkelt ikke konstrueret de faglige begreber på et niveau og/eller en form, som gør det muligt for dem at anvende dem i en faglig samtale om fænomener i den nære og fjernere omverden. Problemet er formentlig størst i geografi. Hvis man (som fx PISA) anser evnen til at forklare sådanne fænomener for at være en central del af Scientific Literacy, så har naturfagsundervisningen vitterligt et problem. Udover den mangelfulde konstruktion af bærende forklaringsenheder kan man også stille spørgsmålstegn ved, om eleverne i fornødent omfang er trænede i at generere forklaringer – og i at forklare sig til nogen!?
- *PISAs "forklaringsproblem"*: I analysesamplet kan ca. 80 % af eleverne udvælge de korrekte svar blandt PISAs bud på forklaringer[1]. Alligevel er max. 40 % i nærheden af at have en forståelse på et niveau, der må ses som en forudsætning for selv at levere en nogenlunde fyldestgørende forklaring. PISAs mål om at evaluere evnen til forklaring er således realiseret på en meget tvivlsom måde!

Formuleringen af disse to forklaringsproblemer kan forekomme barsk, og vi ville gerne have haft mulighed for at sammenholde disse VAP-konklusioner med resultaterne af tilsvarende analyser i andre lande. Så vidt vides findes der imidlertid ikke parallelle undersøgelser i tilknytning til PISA-testningen. Det nærmeste vi kan komme, er et enkelt sammenlignende studium af elevers formåen i to naturfaglige TIMMS-opgaver, i det sædvanlige TIMMS-setup og udsat for en sociokulturel testsamtale (Schoultz, Saljo, & Wyndhamn. 2001a). Her konkluderede forfatterne bl.a.:

> "Even though most of the students choose the correct response alternative in our study, they talk about atoms that 'dissolve', 'die' and 'disappear' when discussing the problem of a decomposing animal, and they use these terms without realizing that they are problematic from the point of view of atomic theory.

[1] mod godt 70 % i VAPs totale sample

> In a similar vein, a student might carry out an appropriate analysis that signals the recycling of matter, yet choose the wrong response alternative" (p.234)
>
> "However, virtually nobody gives an explanation that is grounded in scientific principles of optics or even comes close to using such accounts. The explanations accepted in the TIMSS manual are mostly indicative of everyday thinking and of knowing what a flashlight looks like." (p.233)

Schoultz et al er altså nået frem til samme helt grundlæggende pointer, både hvad angår elevernes evne til at frembringe fagligt funderede forklaringer, og hvad angår kritikken af testsystemets mål for evnen til forklaring. Det tilfører vores undersøgelse en vis ekstern validitet, som er ganske betryggende. Selve resultaterne er til gengæld bekymrende på alle andre måder!

VAP-billedet af elevernes evne til argumentation

Indtrykket af elevernes evne til at argumentere i relation til klimaopgaven er af flere grunde mere diffust end det foregående billede af evnen til naturvidenskabelig forklaring. For det første er der et metodisk problem med at udskille argumentation fra forklaring – i en interview-situation, hvor elevernes opgave netop er at *vælge en forklaring* og *argumentere for dennes rigtighed*, subsidiært at *argumentere imod mulige forklaringer*. Allerede i udgangspunktet er genrerne vævet sammen – hvilket vil blive tydeligere i diskussionen af konkrete eksempler nedenfor. En anden og mere væsentlig grund er, at *man i en stor del af interviewene rent faktisk kan have svært ved at lokalisere udfoldede eksempler på at elever argumenterer naturvidenskabeligt, dvs. med brug af faglige termer, samt naturvidenskabelige belæg og hjemmel*. Efter endt video-analyse sidder man i op mod halvdelen af de analyserede tilfælde tilbage med et helhedsindtryk, der bedst kan sammenfattes med betegnelsen" argumentativt vakuum".

Bedst ser det ud i den indledende opgave, hvor klimaet beskrives ud fra [geospecifik]figurer. Her fremsætter eleverne deskriptive udsagn, som kan forstås som "simple påstande". Det typiske er, at eleverne sjældent selv, spontant nævner eller henviser til specifikke træk ved empirien, men oftest lader de empiriske belæg være *implicitte*, indtil de bliver "gået på klingen". Man skal imidlertid nok ikke overfortolke dette, da eleverne fint og med rette kan have konstrueret det som en præmis for selve interviewsituationen, at intervieweren er bekendt med empirien og i øvrigt lige har udpeget den figur, som er udgangspunkt for samtalen. Derfor kan de have skønnet det overflødigt at præcisere yderligere. Under alle omstændigheder gælder det for langt den overvejende del af eleverne, at de godt kan udpege og etablere belæg for deres udsagn i denne sammenhæng.

Et par af de mere vellykkede eksempler på empirisk grundet argumentation med udgangspunkt i [geospecifik]figurerne:

(E8)

E: Dér kan man se (placerer fingeren på et relevant sted på [geospecifik]figuren), at det regner og sner.. og at de har … meget kold vinter.. når vi normalt har sommer.. og så har de én.. den varmeste periode … der står 0 grader, der hvor vi normalt har vinter (henviser til konkret træk ved [geospecifik]figuren]
I: hvad kan man sige generelt om klimaet
E: at det er koldt… og… nedbørsfattigt

(E9)

E: fordi man.. så også kan se på graferne her.. på de her [geospecifik]figurer (peger), at øh.. temperaturen er højst på den sydlige halvkugle i øh..januar, februar og december, altså i vinteren.. derfor hælder den altså mod Solen .. og derfor bliver der højest temperatur der. .. og modsat med den nordlige. [elev der selvstændigt argumenterer med belæg]

Tilsvarende et par eksempler på mindre vellykkede argumentationer:

(E10)
I prøver så, at tydeliggøre at regnmængden i Kystby er langt mere konstant end i de øvrige afbildede byer. Peger på [geospecifik]er og det fremlagte kort samtidig.
I: "hvordan kan det være?"
E: jamen.. her I Europa, der har vi jo ikke rigtig nogen sådan.. tørke og regntid.. så der ligger regnen rimelig jævnt fordelt [ikke noget godt argument, snarere en tautologi]

(E11)
I: Hvordan er klimaet i Bykold?
E: yes … det er kun koldt…
I: det er kun koldt, ja..
E: det varmeste tidspunkt det er kun nul grader… og det er i januar.. .. det forstår jeg så ikke lige… fordi januar det er jo koldt for os.. er det fordi de regner med at det er januar for vores… vores..? [E mangler efter alt at dømme ordet "halvkugle". Argumentet er lokalt-logisk: januar så er det koldt, ergo: det varmeste tidspunkt kan ikke være januar. Når vi har januar på den nordlige halvkugle, så har de nok juli på den modsatte? Den hjemmel der entydigt sammenknytter klima og kalender er ikke holdbar]

Det sidste eksempel udstiller, at selv i den første opgave, *kan* man rode sig ud i problemer, som involverer hjemmel. Opgaven er imidlertid utypisk, idet den har indrammet den empiri, som eleven med hjemmel kan bruge (de relevante [geospecifik]figurer). Hvis man blot aflæser på rette sted, giver resten af argumentationen for de deskriptive udsagn næsten sig selv[1]. Dermed forbliver det usynligt, om eleverne *af sig selv* ville inddrage tilsvarende naturvidenskabeligt-empiriske belæg. I andre dele af materialet bliver det tydeligt at visse typer af belæg falder eleverne mere naturligt end andre:

(E12)
Elev: Det er altid varmt her (peger på Afrika på kortet)
I: Er der det?
Elev: Næsten da. Altså hvis man ser de der programmer fra Afrika, altså så synes jeg da næsten, at de aldrig har noget tøj på og så må de jo have det varmt.

E(13)
Om en svarmulighed, der påpeger, at der aldrig er høje temperaturer steder langt fra Ækvator
Elev: Det er ikke rigtigt. Fordi oppe i Danmark har vi det jo også varmt om sommeren, der har vi op til 30 graders varme og det synes jeg da, er okay varmt. Folk render rundt i shorts. Den er ikke rigtig

Eksemplerne viser elever, som dels argumenter med et enkelt belæg og dels gendriver en mulig forklaring. I begge situationer er samtalen ikke længere aldeles indrammet af [geospecifik]figurer – og så holder dagligdagserfaringer og common sense belæg deres indtog i elevernes argumentation. *Det er et meget hyppigt træk i materialet, at dagligdagen tager over, så snart situationens indramning giver mulighed for det*. Hverdagsviden er mere aktivt til rådighed – og formentlig mere velkonsolideret end den naturfaglige viden.

En del af argumenterne falder til Jorden – simpelthen fordi det aldrig lykkes, at etablere et sammenhængende link mellem udsagn og belæg. Der er ofte tale om, at det er en slags kausal forklaring, der skal opbygge forbindelsen, altså fungere som hjemmel. Her ser man netop sammenvævningen af forklaring og argumentation – og får demonstreret, hvorledes *manglende evne til at generere holdbare forklaringer influerer evnen til at argumentere naturvidenskabeligt*.

[1] Eneste mulighed for forplumring ligger i, at man bruger [geospecifik]erne til at udtale sig om situeret vejr. Se eksemplet nedenfor.

De tidligere konstaterede problemer med at etablere forklaringer slår altså også igennem i denne analyse.

(E14)
I: hvordan kan det være, at nedbøren er så stabil?
E: det kan være Gulf-strømmen..[det aktuelle udsagn: stabiliteten i nedbør skyldes Golfstrømmen]
I: okay… hvad.. hvilken betydning har det?
E: Det er den der sørger for varm luft.. mener jeg nok…[belæg]
I: ja… og hvad betyder det for nedbøren? [spørger ind til hjemlen]
E: … at der ikke er så meget nedbør, hvis der er varm luft.. højtryk tror jeg det er …. [hjemlen bliver aldrig konstrueret til ende: formentlig er ideen, at varm luft svarer til højtryk (?) svarer til fint vejr svarer til stabilt lav (?) nedbør. De sidste led i argumentationen mangler – og de steder hvor der er markeret spørgsmålstegn holder implikationskæden ikke. I forsøger at gendrive disse – men E anerkender ikke gendrivelsen]

(E15)
I: To byer på samme breddegrad – forskelligt klima, hvordan kan de være?
E: (piller ved globus)… det er fordi de… de ligger på hver sin side af Jorden..[udsagn] hvis du ser på.. hvis Solen skinner herpå, ikke (bruger hånden som sol i.ft..globus ud for Bysyd), så rammer den der.. i Bysyd … men derovre (viser bagsiden af globus).. derovre er den i skygge…[belæg: den ene del i skygge, mens den anden belyses]
I: hvad er der så deromme? Hvad er det Jorden gør i løbet af et døgn? (drejer globus) [I starter sit angreb på hjemlen, der forbinder de to dele. Pointe: Klima udtrykker årsvariation – belæggets lys-skygge-situation refererer til døgnvariation. Derfor har belægget ingen relevans.]
E: der er jo nat lige nu..
I: men er det her (peger nu på grafforløbet) over et døgn eller over et år?
E: det er over et år.. men stadigvæk.. jeg vil umiddelbart tro.. at det betyder et eller andet.. (svagere overbevisning hørligt).. det kan man jo se… [E anerkender måske nok med hovedet at hjemlen ikke holder, men er alligevel ked af at slippe belægget. Efter udtrykket at dømme leder E intenst efter alternative belæg]

Argumentationen udvikler sig ikke altid pænt og forudsigeligt, der kan undertiden ske pludselige forskydninger i argumentationen – uden at det bliver helt klart, hvorfor dette sker:

(E16)
I betoner, at Bysyd og Kontinentby ligger på "præcist samme breddegrad".
I: hvorfor er der denne forskel på de to [[geospecifik]]figurer?
E: Jeg vil mene, at det er fordi Kontinentby ligger i Ørkenklima..[belæg a] tørt klima.. og så.. hernede i Kystland …der kan man godt nok ikke se, at der er meget skov.. og sådan nogen ting [begrunder klima med plantebælter?].. men det vil jeg gå ud fra, at der er…[belæg b] … og så er det derfor [prøver altså at bruge plantebælter som begrundelse for klimatiske forhold. Kausalt er det måske snarere den anden vej rundt – og der er ikke nogen 1-1-korrelation mellem de to. Så hjemlen har tvivlsom gyldighed]
I: mnh.. hvorfor tænker du, at der er meget skov?
E: fordi, at…øh.. når der er meget nedbør.. så skal det jo komme fra et eller andet?.....det ku også være, at det er fordi, at det ligger lige ud til havet [i eet kvantespring fra en aldeles uudgrundelig og noget suspekt argumentation direkte til et mere holdbart belæg. Hjemlen bliver imidlertid aldrig ordentlig udfoldet for det nye belæg]

Evnen til gendrivelse anses af mange (fx Osborne et al, se tidligere omtale) at udtrykke argumentativ formåen på højeste taksonomi-niveau. I det analyserede sample er der relativt få eksempler, hvor elever *eksplicit* gendriver udsagn. Imidlertid bruger mange af eleverne i større eller mindre grad udelukkelsesmetoden, for at nå frem til deres foretrukne svar blandt MC-optionerne. Udelukkelsesmetoden kan i mange tilfælde være en implicit og uudfoldet gendrivelse. Af hensyn til tidsforbruget i VAP-testningen af den enkelte elev har vi ikke haft tid til systematisk at spørge ind til elevernes udelukkelse af svarmuligheder. Evnen til gendrivelse *kan* således godt være bedre, end den fremstår i vore resultater. Typiske træk i niveau og karakteren af de observerede gendrivelser fremgår af nedenstående eksempler:

E(17)
I spørger til svarmulighed B, som E udelukker med at gendrive den påstand, som indeholdes i svarmuligheden.
[Danmark bruges som en falsifikation af det generelle udsagn]:
E: "Danmark ligger jo ret langt fra Ækvator.. og vi har da nogen gange også rimelig høje temperaturer".
Efterfølgende udpeges mulighed C korrekt.

(E18)
I: kan afstanden til havet have nogen betydning?
E: JA, det har det … fordi havet det holder mere på varmen.. end… end hvis du har jorden.
I: ku det så have noget med sagen at gøre?
E: ja, det ku det .. for der er temperaturen konstant [Kystby] … på den anden side… det er den jo ikke i Danmark [fornuftig undren, tilløb til *gendrivelse* som i det foregående eksempel].. men det er jo heller ikke et lige så stort hav….men det er jo også længere….ligemeget! [opgiver at få strikket en argumentation sammen, som han selv kan se sammenhæng i].

Opsummerende kan man sige, at elevernes evne til at argumentere naturvidenskabeligt for flertallets vedkommende er meget ringe. Mange elever argumenterer stort set ikke med inddragelse af naturvidenskabelige belæg, hjemmel og rygdækning. Hvis ikke diskussionen er direkte linket til naturvidenskabelig empiri eller meget stærkt indrammet af naturvidenskabelighed, så er tendensen at hverdagsargumenter tager over. Det falder de færreste elever naturligt at referere spontant og eksplicit til empiriske træk i en samtale. Det slår igennem på meget afgørende vis, at eleverne generelt er dårlige til at levere fyldestgørende forklaringer på naturvidenskabelige fænomener, idet sådanne kausale sammenhænge ofte er afgørende for, om der kan etableres gyldig forbindelse mellem et udsagn og et angiveligt belæg. Der er således mange eksempler på ugyldig hjemmel i det undersøgte materiale.

Hovedindtrykket fra analysen af elevernes evne til forklaring og argumentation er, at eleverne ikke i tilstrækkelig grad er trænede i at levere sådanne i den forstand, som er naturvidenskabens. Overdrevet simpelt kunne man ønske sig, at eleverne i højere grad var blevet konfronteret med HVORFOR-spørgsmål, med tid & rum til at formulere udfoldede svar – og en formativ feedback, som sikrer, at de faglige begreber og sammenhænge bliver aktive dele af forklaringer og belæg.

Elevernes brug af artefakter
I forbindelse med klimaopgaven havde eleverne ikke kun rådighed over PISA-opgavens [geospecifik]figurer, men også mulighed for at udnytte en globus, samt diverse geografiske kort, herunder kort med klima- og plantebælter. Ca. halvdelen af eleverne valgte *selv* at inddrage dele af disse ekstra ressourcer. Resten blev i vekslende grad "inspireret"/opfordret til at forholde sig til disse artefakter undervejs i interviewet. Resultaterne demonstrerer signifikant, at de der selv inddrager artefakterne præsterer højest scorer (Pearson korrelation 0.40 mellem grad af selvstændig artefakt-brug og VAP-score i opgaven). På nuværende tidspunkt har vi ikke søgt at afklare, i hvilken udstrækning disse elever klarer sig godt, fordi de får gavn af artefakterne – eller de kun vælger artefakterne, fordi de har fagligt overskud, og altså *er* gode. Begge varianter kan ses i materialet.

Imidlertid står det klart, at *størstedelen af eleverne ikke er fortrolige med disse artefakttyper, og at mange har svært ved at orientere sig*. Det er i den sammenhæng næppe overraskende, at eleverne ikke er vant til mere specielle kort, fx over tryk og vejrsystemer – men nok overraskende, at en globus kan være ukendt land for så mange elever. Uden et grundlæggende kendskab til, hvordan de

forskellige artefakter repræsenterer viden, vil man selvsagt ikke kunne få gavn af dem til at skabe mening, heller ikke i en PISA-opgavesammenhæng. Der er eksempler i materialet, hvor det ikke er detailkendskabet til vidensrepræsentationer, men et overfladetræk ved artefaktet (aktuelt: globen), der leder eleven på sporet af det rette svar. Helt konkret er globus-ophænget med en synlig skrå jordakse med til at henlede opmærksomheden på Jordens hældning – som dermed aktualiseres som mulig forklaring. I et tidligere eksempel (E2) har vi set, hvorledes hældningen som overfladetræk i kombination med et forkert skalaforhold (Jord i forhold til Solsystem) faktisk kan understøtte dén fejlforestilling, at afstanden til Solen på afgørende vis influerer klimaet.

I *Antibiotika*-opgaven skulle eleverne tolke på *afbildede* agarplader i petriskåle. For også her at give eleverne mulighed for at inddrage artefakter, stillede vi her konkrete versioner af de kritiske komponenter til rådighed. Artefakterne får en anden karakter her, idet de *ikke* på indlysende måde kan bidrage til elevernes viden*skonstruktion* – men i bedste fald vil kunne understøtte *genkaldelse* af viden. Imidlertid viste det sig, at under 1/3 af eleverne overhovedet kendte til artefakterne – og af dem havde kun godt halvdelen reelt udført arbejde og skaffet sig erfaringer med disse. Dermed er det ikke så overraskende, at man ikke kan se nogen synlige fordele af artefaktkendskab i denne sammenhæng.

I *Solcreme*-opgaven var hele udstyret til den praktiske undersøgelse til rådighed som artefakter. Vigtigste enkelt-genstand var efter al sandsynlighed det uv-følsomme fotopapir, som i model-forsøget agerede "test-hud" og med sit farveskift fra blåt til hvidt indikerede, hvor meget Uv-stråling, der var gået igennem filteret. Ingen af eleverne kendte til sådant papir i forvejen, men de fik indledningsvist mulighed for at studere papiret, overbevise sig selv og hinanden om dets funktion, samt evt. afprøve dets egenskaber. En forudgående erfaring a la "JO, den bliver altså mest hvid, hvor den er ramt af mest Uv-stråling" er en klar fordel, når man bliver konfronteret med PISA-kravet om at fortolke det eksperimentelle mønster af hvidheds-grader. Det er i hvert fald på denne måde de hyppigste artefakt-henvisninger forekommer.

Artefakter behøver ikke nødvendigvis være fysiske og håndværksmæssige, der kunne i princippet også være tale om symbolske produkter og processer udsprunget af menneskelig aktivitet. Heuristikker og procedurebeskrivelser, fx for hvorledes man bedriver god naturvidenskabelig undersøgelse, *kunne* i princippet også mediere elevernes arbejde i den eksperimentelle situation. Imidlertid har vi tidligere set, at eleverne nok kan håndtere de løbende håndværkmæssige aspekter af undersøgelsen, men at de ikke har meget begreb skabt om naturvidenskabeligt eksperimentelt arbejde, og at kun en ubetydelig del af dem synes at have opmærksomhed på elementer som variabelkontrol, fejlkilder, forbedringsforslag osv. Der er således ingen tegn i materialet på, at eleverne agerer på basis af etablerede heuristikker på området, dvs. ud fra et billede af standarder og kvalitetsprocedurer i forbindelse med eksperimentelt arbejde.

8. Konklusioner

I de foregående VAP-undersøgelser har vi vist, at intentionerne i PISAs videnskabelige framework og i de danske naturfags målbeskrivelser overlapper betragteligt, samt at danske elever i rimeligt omfang må siges at være vant til PISA-lignende testning. På det intenderede niveau ser PISA altså ud til at være *relevant* for Danmark – og på elevniveau ser test-formatet ud til at være *velkendt.*

Med denne tredje del af VAP-projektet har vi ønsket at undersøge, hvorvidt PISAs testformat og scoringsprocedurer producerer *retvisende* og *dækkende* resultater for Danmark: dvs. er der dækning for dét PISA siger? – og giver PISA et fuldstændigt, dækkende billede af danske elevers formåen indenfor naturfag? Kvalificerede og betryggende svar på begge spørgsmål er nødvendige forudsætninger for at lade PISA indgå med vægt i diskussionen om udviklingen af naturfagene i den danske folkeskole.

I udgangspunktet har vi haft en bekymring for samspillet mellem PISAs snævre post-positivistiske testformat og de naturfagskompetencer, som danske elever tilegner sig i en hverdag med et meget bredere sociokulturelt læringsparadigme. Det er nemlig en central grundsætning i moderne evalueringsteori, at evalueringer bør udformes i overensstemmelse med de forestillinger om læring, som har guidet læreprocessen:

Theories of learning have implications for assessment design…. Constructivist models of learning, which see learning as a process of personal knowledge construction and meaning making, describe a complex and diverse process and therefore require assessment to be diverse, examining in more depth the quality of students' learning and understanding. While for example standardized MC or short answer tests are efficient at sampling the acquisition of specific knowledge gained from teachers, more intense, even interactive assessment (e.g. essays, performance assessments, small-group tasks and projects) is needed to assess the processes of learning and understanding, and to encourage a deeper level of learning.
(Gipps. 1999 p.375).

Citatet gør det tydeligt, at valget af evalueringsformat ikke kan anskues som slet og ret et teknisk anliggende, men at det faktisk er afgørende for, i hvilket omfang evalueringen er i stand til at rumme de læreprocesser og værdier, som er bærende for undervisningen og skolesystemet.

Fornemmelsen af, at (dele af) danske elevers naturfagskompetence nemt kunne forblive usynlige eller underbelyste i PISAs format, fik os til at designe et ret komplekst evalueringsdesign til besvarelse af følgende forskningsspørgsmål:

> Q3a: Hvor meget vil PISA-resultatet ændres, såfremt elevernes formåen hvad angår originale PISA-opgaver efterprøves indenfor et mere sociokulturelt evalueringsparadigme med dialog, adgang til medierende artefakter og mulighed for konkret praksis?
> Q3b: Hvilket billede tegner der sig af elevernes styrker og svagheder i "det udvidede opmærksomhedsvindue", som det ændrede testformat konstituerer? Leverer PISA et dækkende billede af danske 15-åriges naturfaglige formåen?

Som det er fremgået af det foregående har VAP i udgangspunktet haft fokus på billedet af danske elevers formåen – mens det evalueringsteoretiske har været sekundært og på sin vis blot har sat

rammerne. Undervejs i analysearbejdet er det imidlertid blevet mere og mere klart, at de forskningsmæssige svar på Q3a og Q3b samtidig leverer grundlæggende information om styrker og svagheder ved de involverede evalueringsparadigmer – og i særdeleshed peger på en række begrænsninger ved PISA. Det samlede slutbillede anfægter i høj grad, hvor *retvisende eller dækkende* PISA er i en dansk sammenhæng. Tilmed er PISA måske *mindre relevant i sin implementerede* form, end det oprindeligt var muligt at se, da vi i den første rapport sammenholdt intentionerne i PISAs videnskabelige Framework med de danske læreplaner. De testmetodiske problematiseringer af PISA er samlet til sidst i konklusionsafsnittet.

8.1 Grundlaget for vore konklusioner

120 elever, der havde gennemført PISA2006-testen, blev gentestet indenfor et biologisk, et geografi/fysisk og et eksperiment-relateret område. VAP-evalueringen undersøgte endvidere evt. effekter ved gentestning, samt betydningen af overfladiske ændringer i enkeltopgavers design. I overensstemmelse med VAPs sociokulturelle orientering foregik gentestningen via interviews og en praktisk øvelse i elevpar - begge dele med mulighed for at eleverne kunne inddrage artefakter m.m.. Interviewet omfattede dels snævert PISA-relevante spørgsmål, dels spørgsmål indenfor det bredere område af Fælles Mål, som PISA-spørgsmålene laver punktnedslag i. Dette sidste blev gjort for at sikre VAP en vis gyldighed i forhold til Fælles Mål - kravene og den danske naturfagsundervisning. Alle VAP-sessioner blev videotapet, og diverse skriftlige elevbesvarelser fra gentestningsundersøgelsen m.m. blev indsamlet. Video-optagelserne er efterfølgende blevet analyseret kvantitativt ud fra PISAs egne scoringskriterier, men for en stor dels vedkommende også mere helhedsorienteret i lyset af Fælles Mål. Endelig er der foretaget kvalitative, mere dybgående fagdidaktiske analyser af elevernes sprogbrug, evne til forklaring, argumentation m.m. ud fra et sociokulturelt perspektiv. Alle analyser er kvalitetssikret, således at pålideligheden er størst mulig.

8.2 Danske elevers formåen i det ændrede testformat

På baggrund af analyser af elevernes præstationer i den udviklede VAP-evaluering sammenholdt med de samme elevers præstationer i PISA-testen, kan vi konkludere:

> ***I en direkte sammenligning efter PISAs scoringskriterier klarer eleverne sig ca. 25 % bedre, når de får lov til at udfolde sig i et sociokulturelt orienteret testformat.***

Pointen i VAPs opgørelse er her, at eleverne gennem interviewet "blot" skal vise, at de er i stand til at honorere PISAs kriterier for rigtigt svar, fx ved selv at omtale de relevante informationer eller vælge den korrekte MC-mulighed, såfremt disse forevises undervejs i forløbet. *Denne score udtrykker altså ikke nødvendigvis at eleven har en sammenhængende model, forklaring eller argumentation – blot at eleven ved nok til at svare rigtigt i PISA.* I 6 af de 10 sociokulturelt gentestede PISA-spørgsmål var præstationen opgjort i "rå" elevscorer signifikant bedre - og samlet set over alle spørgsmål blev den gennemsnitlige forbedring på 26 %!
En forklaring på denne forbedring skal sandsynligvis søges i de muligheder det rigere testformat og den mere autentiske testsituation giver eleverne for at aktualisere deres viden.

PISA-projektets styrke er dets komparative målinger af elevperformance i forskellige lande. Her rangeres landene med udgangspunkt i en Rasch-model og en skala hvor OECD-gennemsnittet har værdien 500. For at kunne estimere betydningen af det ændrede testformat på denne skala og efter den standard som PISA selv har sat, har vi sammenlignet elevernes VAP-præstationer med deres

PISA-præstationer ved hjælp af Rasch-modellering. Denne analyse fungerer samtidig som et check af vore resultater fra den foregående analyse baseret på rå scorer, idet der er tale om en triangulering af beregningsmetoden.

De i VAP-testen anvendte opgaver skalerer tilfredsstillende til at kunne underkastes en Rasch-analyse, og da de har sammenlignelige relative sværhedsgrader i VAP-testen og i PISA-testen, kan vi sammenligne elevernes præstationer i de to Rasch-scalaer. Sammenligningen giver betryggende overensstemmelse med det foregående:

Ved et sociokulturelt orienteret testformat forbedrer elever deres præstationer på en Rasch-skala med 25 %, svarende til en forøgelse på 125 PISA-point.

En sådan forøgelse svarer til at den danske placering i PISA2006-science ændres fra 496 til 621 - markant over Finlands testvindende 563 Rasch-point.
Dette betyder naturligvis ikke at Danmark ved et mere sociokulturelt orienteret PISA-testformat ville øge sin relative placering i den internationale sammenligningstabel tilsvarende, idet også andre lande må tænkes at have fordel af det ændrede testparadigme. Men da naturfagsundervisningen i Danmark normalt anses for at være mere dialogisk tilrettelagt end i de fleste af de andre deltagende lande, og mange lande har et mere PISA-lignende testsystem end Danmark, betyder PISA-testformatet formentligt, at Danmarks præstation fremstår relativt dårligere.

Det bemærkelsesværdige er her ikke, *at* testformatet har en vis betydning for resultatet, det ville mange nok intuitivt forvente. Det bemærkelsesværdige er, at VAP er i stand til at opgøre betydningen kvantitativt – og at betydningen viser sig at være så voldsomt stor! Elevernes performance i PISA udtrykker således ikke til fulde deres formåen i forhold til opgavernes faglige indhold, men skabes i betragtelig grad også af deres evne til at gennemtrænge og udfolde sig indenfor et bestemte test-formats nåleøje!

Det har således ved hjælp af det udviklede forskningsdesign været muligt at svare præcist på det opstillede forskningsspørgsmål Q3a. Men VAP-evalueringens datamateriale er væsentlig rigere end de tilsvarende PISA-data. Vi har derfor været i stand til at analysere danske elevers kunnen inden for de VAP-testede fagområder, nemlig klimaforskelle, antibiotika og naturvidenskabelige arbejdsmåder og tankegange, i relation til de krav, som opstilles i Fælles Mål. Dette er en del af vores "udvidede opmærksomhedsvindue", og gennem en række analyser af det indsamlede materiale kan vi svare på forskningsspørgsmål Q3b: Hvad kan eleverne, når man kigger lidt mere udførligt efter? Her tegner VAP - stik imod vore forventninger om, at et sociokulturelt testformat vil kunne udfolde danske elevers naturfaglige kompetencer - et endog ganske nedslående billede:

Der er et stort gab mellem danske elevers faktiske formåen i en række naturvidenskabelige fagområder og de begrebsmæssige og proceduremæssige krav, som Fælles Mål udtrykker.

Her kan i flæng nævnes, at de testede elever havde svært ved at udtrykke sig om relevante biologiske begreber (bakterier, virus, immunforsvar, resistens etc.), og deres samlede forståelse inden for de testede biologiområder blev vurderet til at svare til mellem 20 % og 35 % af fuld forståelse, som den opstilles i Fælles Mål!
Forståelsen af geografiske begreber som klima, nedbør, årstider var tilsvarende ringe. Her vurderes centrale procesforståelser (som fx vandets kredsløb og dannelse af nedbør) til at ligge på ca. 20-40 % af fuld forståelse som formuleret i Fælles Mål.

Værst stod det til med elevernes viden om (aspekter af) naturvidenskabelige arbejdsmåder og tankegange. Under 5 % af eleverne var således i stand til at formulere sig om naturen af naturvidenskabelige eksperimenter og undersøgelser i fagrelevante termer!

Disse nedslående resultater blev analyseret frem ved hjælp af et naturfagsdidaktisk begrebsapparat. Blandt resultaterne fra den sociokulturelt funderede analyse er der grund til at fremhæve:

Eleverne mangler kendskab til og evne til anvendelse af især abstrakte fagtermer og de besidder ikke i tilstrækkelig grad centrale fagforklaringsenheder, hvilket gør det vanskeligt for dem at forklare faglige sammenhænge.

Eleverne kan godt samtale med brug af simple fagtermer og inddragelse af hverdagssprog, men de har et ringe kendskab til mere abstrakte begreber og processer. De begreber og forklaringsenheder som skal bære elevernes forklaringer er ofte ikke konstruerede i en grad, så de selv kan generere fyldestgørende forklaringer. De har meget nemmere ved at vælge/fravælge blandt allerede artikulerede bud. I sammenhæng hermed var det karakteristisk, at

Eleverne er i ringe grad i stand til at argumentere naturvidenskabeligt.

Den manglende evne til at generere holdbare forklaringer influerer evnen til at argumentere naturvidenskabeligt, og fagsprog og argumentation med naturvidenskabelige belæg og hjemmel anvendes ikke naturligt af eleverne.
VAPs sociokulturelle testformat inddrog artefakter i testsituationen. Det fremgik, at

Elever der selv inddrager artefakter i deres forklaringer præsterer højest.

Men det var også tydeligt, at

størstedelen af eleverne er ikke fortrolige med de anvendte artefakter.

Analysen afdækker ikke entydigt, hvad der er årsag og virkning i sammenhængen mellem inddragelse af artefakter og faglig formåen. Det er dog en nærliggende hypotese, at kun de fagligt stærkeste elever har tilstrækkelig indsigt i, hvad og hvordan artefakterne repræsenterer, til at de kan bruge dem som redskaber til videnskonstruktion og forklaring.

8.3 Validiteten af PISA – med udgangspunkt i VAP-projektets empiri.

PISA-testen er velbeskrevet og metodisk stringent vurderet ud fra sit testteoretiske ståsted. Men bevæger man sig ud over dette post-positivistiske paradigme og anvender VAP-projektets mere sociokulturelle tilgang, er man nødt til at problematisere validiteten af PISAs resultater på en række punkter – i første omgang for Danmark, men formentlig langt mere generelt. F.eks. har analysen af elevernes formåen i et alternativt test- og scoringsformat tydeliggjort, hvorledes test-resultater generelt formes af en lang række forhold, som har meget lidt med elevernes faglige formåen at gøre. VAP-projektet leverer således dokumentation for, at man ved at arbejde inden for et sociokulturelt evalueringsparadigme og med et mere undervisningsnært testformat faktisk kan forskyde testresultatet 25 %. Vel og mærke målt på de samme opgaver og i henhold til de samme kriterier! Ud fra en postpositivistisk tilgang ville man nok sige, at elevernes bedre præstationer skyldes at de

er blevet hjulpet, men i en sociokulturel sammenhæng vil man sige at den kommunikative situation er ændret og dermed elevens kunnen.
Med påvisningen af at elevers målte performance er så følsomt overfor det valgte test-format, er VAP en blinkende advarselslampe om, at man skal være ekstremt påpasselig, når man konkluderer på tests af denne type:

> ***Valget af evalueringsparadigme og test-format influerer i høj grad test-resultatet. PISAs test-resultat er således relativt og siger meget lidt om elevernes formåen i absolut forstand. En sådan indsigt bør afspejles i de konklusioner og konsekvenser man drager af PISA.***

Størst udsigelseskraft må resultaterne anses at have, når den valgte evaluering er i overensstemmelse med det paradigme og de værdier, som lå til grund for elevernes læreprocesser. Umiddelbart forekommer PISA på disse punkter ikke at være det mest indlysende valg, hvis man ønsker et dækkende indtryk af den viden, som danske elever har tilegnet sig i folkeskolen:

> ***PISA-resultaterne udtrykker kun i ringe grad, hvad eleverne kan i henhold til Fælles Mål.***

Vi har fx fundet bemærkelsesværdige forskelle på elevers viden, som den kommer til udtryk i VAP-testens "udvidede opmærksomhedsvindue" og i PISA-testens snævre vindue med punktvise nedslag. I VAP-testen blev der fundet 20 % med adækvat Fælles Mål-forståelse af antibiotikas indvirkning på bakterier og vira – mens 45 % af eleverne i samplet svarede rigtigt i PISA 2006. Elevernes helhedsorienterede forståelse af centrale biologiske begreber knyttet til brug af antibiotika (igen vurderet efter Fælles Mål beskrivelserne) blev i VAP-testen vurderet til ca. 35 % af fuld forståelse, mens PISA-opgaven som testede de samme begreber blev løst af 75 % af eleverne i PISA-testen. I klimaforskelle-opgaven kunne kun 42 % af drengene og 7 % af pigerne i VAP-testen redegøre for grundlæggende forhold knyttet til klimavariation, men 90 % af eleverne kunne alligevel vælge det rigtige af PISAs svarmuligheder i det tilsvarende PISA-spørgsmål.
Vi ser altså, at eleverne konsekvent fremstår bedre i PISA-testen end de gør vurderet i forhold til Fælles Mål. PISA udtrykker dermed i ringe grad målopfyldelse i forhold til Fælles Mål. Det skal understreges, at PISA aldrig har givet udtryk for at kunne eller ville indfange målopfyldelse i forhold til nationale curricula. Imidlertid er undersøgelsens resultater ofte i offentligheden blevet brugt, som om der var udsigelseskraft på dette punkt.

Vi har gjort os nogle forestillinger om, hvorfor denne betragtelige uoverensstemmelse forekommer, når vi ellers tidligere (jf. første VAP-rapport) har fundet en pæn overensstemmelse mellem intentionerne i de danske mål for naturfag og i PISAs Scientific Literacy Framework. Vi mener at kunne se en betragtelig skridning i modsatte retninger, når intentionerne skal omsættes til hhv. undervisning i skolen og PISA-opgaver. Scientific Literacy er ikke bare Scientific Literacy, men kan ifølge Bybee (Bybee. 1997) forstås på forskellige taxonomiske niveauer. Det er her vores opfattelse, at de danske mål undervisningsmæssigt operationaliseres som *Conceptual Scientific Literacy,* mens PISA opgaverne måler en taxonomisk lavere *Functional Scientific Literacy* (samt logisk-rationel tænkning). Bl.a. derfor tester PISA simpelthen ikke den begrebslige forståelse og den procesbeherskelse, som forudsættes lært ifølge Fælles Mål. En grundigere undersøgelse af opgaverne er nødvendig for at underbygge denne hypotese.

Denne problematik er for alvor blevet synlig i VAP-testens udvidede opmærksomhedsvindue, som også har udstillet, hvorledes PISA ofte misvisende honorerer elever for et korrekt valg af MC-option – i situationer hvor begrundelser og underliggende forklaringer er klart forkerte. Analysen af

et tilfældigt udvalg af elevbesvarelser i opgaven om klima viser som tidligere omtalt, at 80 % af eleverne kan vælge den korrekte forklaring, mens 60 % af eleverne aldrig er i nærheden af at producere en fyldestgørende forklaring. Tværtimod ville denne gruppes bidrag sædvanligvis blive kategoriseret som fejlforståelser/uautoriserede hverdagsforestillinger. PISA underdriver dermed elevernes reelle problemer med at forstå og forklare naturvidenskabelige fænomener og problemstillinger. Ligesom PISA ikke indfanger de grundlæggende problemer med at argumentere, bruge artefakter, reflektere fagets metoder osv., som VAP har afdækket hos de danske elever. I den forstand forekommer det rimeligt at konkludere, at

PISAs testformat indfanger ikke de essentielle problemer der er, hvad angår elevernes evne til at bruge naturvidenskabeligt sprog, forklare og argumentere videnskabeligt, bruge artefakter, reflektere fagets metoder osv.

På det konkrete niveau kan miseren henføres til såvel opgaver, responsformater som scoringskriterier. På et mere overordnet plan kan man også se det som udtryk for, at

PISA er med sit paradigmatiske ståsted og sit valg af test-format stort set blind overfor aspekter af det sociokulturelle paradigme, som ideelt set bærer læreprocesserne i dansk naturfagsundervisning.

VAP har også tydeliggjort et andet aspekt af test-anvendelsen, som også berører centrale værdier i den danske folkeskole. Skolen skal give børnene lige muligheder. En ideel test er derfor også neutral i den forstand, at den ikke begunstiger én elevgruppe i forhold til en anden. Den direkte sammenligning mellem hvad eleverne kan indenfor rammerne af hhv. PISA og VAP peger imidlertid på, at

Valget af testformat tilgodeser/diskriminerer på signifikant vis bestemte elevgrupper. Et sociokulturelt orienteret testformat er til gunst for de fagligt svagere elever. Dermed er PISAs valg af test-format et relativt valg til fordel for de fagligt stærkere elever.

Endelig er der grund til at omtale hovedresultatet af VAPs eksplorative undersøgelse af, hvorledes overfladetræk ved opgavernes design (tekstmængde, tekst-sekvensering, brug af figurer m.m. til at præsentere information) påvirker test-resultatet. Her fandt vi at

Ændringer i opgavernes overfladetræk og sproglige formuleringer giver signifikante ændringer i testresultatet.

På svært gennemskuelig vis synes elevernes præstation at blive påvirket af mindre ændringer i overfladedesignet af PISA-opgaverne. Det gør det endnu vanskeligere at se test-"resultatet" som et signifikant udtryk for elevernes "sande" faglige formåen.

Sammenfattende kan man sige, at VAP-undersøgelsen langt henad vejen har vist sig at udfolde det gamle mundheld "Som man spørger får man svar"! Ved at stille spørgsmålene i et anderledes og uddybende kommunikativt format end PISA, har vi fået meget anderledes svar; svar som udstiller betydningsfulde svagheder ved elevernes formåen, og som PISA har vist sig blind for – og svar som vi mener at kunne argumentere for er nok så retvisende, dækkende og interessante i den danske kontekst. Det fører til en grundlæggende problematisering af PISA-testens evne til - med dets

paradigmatiske værdier, dets valg af testformat, scoringsprocedurer og konkrete opgaveoperationalisering – at indfange danske elevers formåen på valid vis. Undersøgelserne har tydeliggjort, at PISA ikke er et neutralt måleinstrument af en universel scientific literacy. Det er derimod baseret på (uddannelses)politiske valg og et post-positivistisk testparadigme, som på mange områder er i konflikt med herskende danske og udenlandske uddannelsesparadigmer og -værdier, og det er ikke designet til at måle centrale områder af de krav, der formelt stilles til danske elever:

> ***Valget af evalueringsparadigme og test-format influerer i høj grad test-resultatet. PISAs test-resultat er således relativt og siger meget lidt om elevernes formåen i absolut forstand. En sådan indsigt bør afspejles i de konklusioner og konsekvenser man drager af PISA.***

Størst udsigelseskraft må resultaterne anses at have, når den valgte evaluering er i overensstemmelse med det paradigme og de værdier, som lå til grund for elevernes læreprocesser. Umiddelbart forekommer PISA på disse punkter ikke at være det mest indlysende valg, hvis man ønsker et dækkende indtryk af den viden, som danske elever har tilegnet sig i folkeskolen:

> ***PISA-resultaterne udtrykker kun i ringe grad, hvad eleverne kan i henhold til Fælles Mål.***

Vi har fx fundet bemærkelsesværdige forskelle på elevers viden, som den kommer til udtryk i VAP-testens "udvidede opmærksomhedsvindue" og i PISA-testens snævre vindue med punktvise nedslag. I VAP-testen blev der fundet 20 % med adækvat Fælles Mål-forståelse af antibiotikas indvirkning på bakterier og vira – mens 45 % af eleverne i samplet svarede rigtigt i PISA 2006. Elevernes helhedsorienterede forståelse af centrale biologiske begreber knyttet til brug af antibiotika (igen vurderet efter Fælles Mål beskrivelserne) blev i VAP-testen vurderet til ca. 35 % af fuld forståelse, mens PISA-opgaven som testede de samme begreber blev løst af 75 % af eleverne i PISA-testen. I klimaforskelle-opgaven kunne kun 42 % af drengene og 7 % af pigerne i VAP-testen redegøre for grundlæggende forhold knyttet til klimavariation, men 90 % af eleverne kunne alligevel vælge det rigtige af PISAs svarmuligheder i det tilsvarende PISA-spørgsmål.
Vi ser altså, at eleverne konsekvent fremstår bedre i PISA-testen end de gør vurderet i forhold til Fælles Mål. PISA udtrykker dermed i ringe grad målopfyldelse i forhold til Fælles Mål. Det skal understreges, at PISA aldrig har givet udtryk for at kunne eller ville indfange målopfyldelse i forhold til nationale curricula. Imidlertid er undersøgelsens resultater ofte i offentligheden blevet brugt, som om der var udsigelseskraft på dette punkt.

Vi har gjort os nogle forestillinger om, hvorfor denne betragtelige uoverensstemmelse forekommer, når vi ellers tidligere (jf. første VAP-rapport) har fundet en pæn overensstemmelse mellem intentionerne i de danske mål for naturfag og i PISAs Scientific Literacy Framework. Vi mener at kunne se en betragtelig skridning i modsatte retninger, når intentionerne skal omsættes til hhv. undervisning i skolen og PISA-opgaver. Scientific Literacy er ikke bare Scientific Literacy, men kan ifølge Bybee (Bybee. 1997) forstås på forskellige taxonomiske niveauer. Det er her vores opfattelse, at de danske mål undervisningsmæssigt operationaliseres som *Conceptual Scientific Literacy,* mens PISA opgaverne måler en taxonomisk lavere *Functional Scientific Literacy* (samt logisk-rationel tænkning). Bl.a. derfor tester PISA simpelthen ikke den begrebslige forståelse og den procesbeherskelse, som forudsættes lært ifølge Fælles Mål. En grundigere undersøgelse af opgaverne er nødvendig for at underbygge denne hypotese.

Denne problematik er for alvor blevet synlig i VAP-testens udvidede opmærksomhedsvindue, som også har udstillet, hvorledes PISA ofte misvisende honorerer elever for et korrekt valg af MC-option – i situationer hvor begrundelser og underliggende forklaringer er klart forkerte. Analysen af et tilfældigt udvalg af elevbesvarelser i opgaven om klima viser som tidligere omtalt, at 80 % af eleverne kan vælge den korrekte forklaring, mens 60 % af eleverne aldrig er i nærheden af at producere en fyldestgørende forklaring. Tværtimod ville denne gruppes bidrag sædvanligvis blive kategoriseret som fejlforståelser/uautoriserede hverdagsforestillinger. PISA underdriver dermed elevernes reelle problemer med at forstå og forklare naturvidenskabelige fænomener og problemstillinger. Ligesom PISA ikke indfanger de grundlæggende problemer med at argumentere, bruge artefakter, reflektere fagets metoder osv., som VAP har afdækket hos de danske elever. I den forstand forekommer det rimeligt at konkludere, at

> ***PISAs testformat indfanger ikke de essentielle problemer der er, hvad angår elevernes evne til at bruge naturvidenskabeligt sprog, forklare og argumentere videnskabeligt, bruge artefakter, reflektere fagets metoder osv.***

På det konkrete niveau kan miseren henføres til såvel opgaver, responsformater som scoringskriterier. På et mere overordnet plan kan man også se det som udtryk for, at

> ***PISA er med sit paradigmatiske ståsted og sit valg af test-format stort set blind overfor aspekter af det sociokulturelle paradigme, som ideelt set bærer læreprocesserne i dansk naturfagsundervisning.***

VAP har også tydeliggjort et andet aspekt af test-anvendelsen, som også berører centrale værdier i den danske folkeskole. Skolen skal give børnene lige muligheder. En ideel test er derfor også neutral i den forstand, at den ikke begunstiger én elevgruppe i forhold til en anden. Den direkte sammenligning mellem hvad eleverne kan indenfor rammerne af hhv. PISA og VAP peger imidlertid på, at

> ***Valget af testformat tilgodeser/diskriminerer på signifikant vis bestemte elevgrupper. Et sociokulturelt orienteret testformat er til gunst for de fagligt svagere elever. Dermed er PISAs valg af test-format et relativt valg til fordel for de fagligt stærkere elever.***

Endelig er der grund til at omtale hovedresultatet af VAPs eksplorative undersøgelse af, hvorledes overfladetræk ved opgavernes design (tekstmængde, tekst-sekvensering, brug af figurer m.m. til at præsentere information) påvirker test-resultatet. Her fandt vi at

> ***Ændringer i opgavernes overfladetræk og sproglige formuleringer giver signifikante ændringer i testresultatet.***

På svært gennemskuelig vis synes elevernes præstation at blive påvirket af mindre ændringer i overfladedesignet af PISA-opgaverne. Det gør det endnu vanskeligere at se test-"resultatet" som et signifikant udtryk for elevernes "sande" faglige formåen.

Sammenfattende kan man sige, at VAP-undersøgelsen langt henad vejen har vist sig at udfolde det gamle mundheld "Som man spørger får man svar"! Ved at stille spørgsmålene i et anderledes og uddybende kommunikativt format end PISA, har vi fået meget anderledes svar; svar som udstiller

betydningsfulde svagheder ved elevernes formåen, og som PISA har vist sig blind for – og svar som vi mener at kunne argumentere for er nok så retvisende, dækkende og interessante i den danske kontekst. Det fører til en grundlæggende problematisering af PISA-testens evne til - med dets paradigmatiske værdier, dets valg af testformat, scoringsprocedurer og konkrete opgaveoperationalisering – at indfange danske elevers formåen på valid vis. Undersøgelserne har tydeliggjort, at PISA ikke er et neutralt måleinstrument af en universel scientific literacy. Det er derimod baseret på (uddannelses)politiske valg og et post-positivistisk testparadigme, som på mange områder er i konflikt med herskende danske og udenlandske uddannelsesparadigmer og -værdier, og det er ikke designet til at måle centrale områder af de krav, der formelt stilles til danske elever.

9. Perspektiver

Vores drivkraft for at starte VAP-projektet var en nysgerrighed i forhold til om PISA-testens uddannelsespolitiske betydning var berettiget i forhold til testens validitet. Og den nysgerrighed er blevet stillet! VAP-undersøgelsen indikerer, at PISA-testen ikke giver et retvisende billede af danske 15-åriges viden og kunnen inden for det naturvidenskabelige område. Hele PISA-test-konceptet sammenholdt med de udviklede opgaver er ikke i stand til at måle de opstillede scientific literacy mål og slet ikke de danske Fælles Mål, og specielt for danske forhold har vi vist hvorledes væsentlige kompetencer - og mangler på samme - ikke indfanges.

Dette stiller dels et stort spørgsmålstegn ved den internationale benchmarking og rangordning, som er et at PISA-projektets erklærede mål og primære resultat, men det giver anledning til større betænkelighed vedrørende den uddannelsespolitiske anvendelse af PISA. PISA giver nogle meget overordnede resultater og kan påpege nogle relevante sammenhænge, men dels er resultaterne skrøbelige, og dels giver de ikke megen forståelse af årsagerne til resultaterne. Der er simpelthen for langt fra et statistisk udtræk af 15-årige til hvad der sker i klasserummet, og PISA giver ikke mulighed for at korrelere elevernes PISA-performance med den undervisning, som er medvirkende baggrund for testresultaterne. Statistisk behandling af aggregerede data vil ofte skjule kvalitative forhold af væsentlig betydning og dermed kunne lede til misvisende resultater.

Man kan ikke bruge forskningsresultater på et niveau, som er forskelligt fra det niveau, som forskningen drager slutninger på. PISA er en test på et meget højt slutningsniveau, som kan sige noget om uddannelsessystemet i sin helhed, men det kan ikke bruges til at udtale sig om hvad der sker eller skal ske i den enkelte klasse. I det omfang PISA anvendes retningsgivende for udformningen af den konkrete undervisning, risikerer man at fremme en international harmonisering på bekostning af en række værdifulde danske uddannelsesværdier.

Dette betyder ikke, at der ikke er behov for at udvikle naturfagsundervisningen i Danmark. Tværtimod viser vores forskning, at der er nok at tage fat på. Her er der et stort potentiale i PISA-datamaterialet, men især hvis det kombineres med klassenær fagdidaktisk forskning, som kan bruges af lærerne i den daglige undervisning.

Men det er en vigtig pointe at PISA-resultaterne ikke leverer et validt grundlag for vidtgående reformer som påvirker undervisningsniveauet. Det er derfor vores håb at VAP-projektets resultater vil bidrage til en mere nuanceret tolkning og anvendelse af PISA-testens resultater.

10. Litteratur

Adams, R. and M. Wu (2001). *PISA 2000 Technical Report*. Paris: OECD.

American Educational Research Association, A. P. A. &. N. C. o. M. i. E. (1985). *Standards for educational and psychological testing* Washington: AERA, APA & NCME.

Allerup, P. (2007). Identification of Group Differences Using PISA scales - Considering Effects of Inhomogenous Items. In Hopman et al (Eds) (Ed.), *PISA zufolge PISA - PISA according to PISA* (pp. 175-201). Wien: LIT VERLAG GmbH & Co KG.

Andersen, A.M., N. Egelund, T.P. Jensen, M. Krone, L. Lindenskov & J. Mejding (2001). *Forventninger og færdigheder - danske unge i en international sammenligning*. København: Socialforskningsinstituttet.

Andersen, N. O., Busch, H., Horst, S., Andersen, A. M., Dalgaard, I., Dragsted, S. et al. (2006). *Fremtidens Naturfag i Folkeskole - rapport fra udvalget til forberedelse af en handlingsplan for naturfagene i folkeskolen* Undervisningsministeriet, DK.

Andersen, N. O., Busch, H., Horst, S., & Troelsen, R. (2003). *Fremtidens Naturfaglige Uddannelser - Bd.1: Strategiplan 2003-2008 og videre frem* København: Undervisningsministeriet (In Danish).

Arbejdsgruppen til forberedelse af en national strategi for Natur, T. o. S. (2008). *Et fælles løft* Copenhagen: Danish Ministry of Education.

Bell,B. (2007). Classroom Assessment of Science Learning. In S.K.Abell & N.Lederman (Eds.), *Handbook of Research on Science Education.* (pp. 965-1006). Mahwah, N.J.: Lawrence Erlbaum Associates.

Berg, E. (2007).

Bond,T.G., & Fox,C.M. (2001). *Applying the Rasch Model - Fundamental Measurement in the Human Sciences.* Mahwah, N.J.: Lawrence Erlbaum Associates, Inc.

Broadfoot,P., & Black,P. (2004). Redefining assessment? The first ten years of *Assessment in Education*. *Assessment in Education, 11,* 7-27.

Buhagiar,M.A. (2007). Classroom assessment within the alternative assessment paradigm: revisiting the territory. *The Curriculum Journal, 18,* 39-56.

Bybee, R. (1997). Towards an understanding of scientific literacy. In W.Graeber & C.Bolte (Eds.), *Scientific Literacy.* (pp. 37-68). Kiel: IPN.

Dahler-Larsen,P., & Krogstrup,H.K. (2001). Evalueringers konstitutive virkninger. In P.Dahler-Larsen & H.K.Krogstrup (Eds.), *Tendenser i evaluering.* (pp. 232-245). Odense: Odense Universitetsforlag.

Devo Team Consulting. Anonymous. De nationale it-baserede test i folkeskolen - rapport fra reviewpanelet. Skolestyrelsen. (2007).

Dolin, J. (2005). PISA og fremtidens kundskabskrav. In: *PISA-undersøgelsen og det danske uddannelsessystem*. Folketingshøring om PISA-undersøgelsen 12. september 2005. Teknologirådet.

Dolin, J. (2008). PISA - komparativ evaluering i storskalaformat. I Borgnakke, K. *Evalueringens spændingsfelter*. Aarhus: Klim.

Dolin, J., H. Busch og L. B. Krogh (2006). *En sammenlignende analyse af PISA2006 science testens grundlag og de danske målkategorier i naturfagene*. Første delrapport fra VAP-projektet. Odense: IFPR/Syddansk Universitet.

Dolin, J. og L. B. Krogh (2008). *Den naturfaglige evalueringskultur i folkeskolen.* Anden delrapport fra VAP-projektet. INDs skriftserie nr. 17. København: Institut for naturfagenes Didaktik/Københavns Universitet.

Dolin, J. og L. B. Krogh (2010).The Relevance and Consequences of Pisa Science in a Danish Context. *International Journal of Science and Mathematics Education*, 8, 565-592.

Dysthe, O. (2000). *Det flerstemmige klasserum - skrivning og samtale for at lære*. Aarhus (In Danish): Klim.

Egelund, N. (2005). Educational assessment in Danish schools. *Assessment in Education, 12*(2), 203-212.

Egelund, N. & Andersen, T. Y. (2006). PISA og de 16 1/2 årige uddannelsessøgende. Aarhus: Aarhus Universitetsforlag.

Egelund, N. (red.)(2007). *PISA 2006 – Danske unge i en international sammenligning*. København, Danmarks Pædagogiske Universitets Forlag.

Egelund, N. (red.)(2010). *PISA 2009 – Danske unge i en international sammenligning*. København, Danmarks Pædagogiske Universitets Forlag

EPPI - Evidence for Policy and Practice Information Centre, & Assessment and Learning Research SynthesisGroup (ALRSG). Anonymous. *A systematic review of the impact of summative assessment and tests on students' motivation for learning*. EPPI-Centre, Institute of Education, University of London. (2002).

Erduran,S., Simon,S., & Osborne,J. (2004). TAPping into argumentation: Developments in the application of Toulmin's argument pattern for studying science discourse. *Science Education, 88,* 915-933.

Fuglsang,J., & Saietz,D. 2010. Skoleledere: Nationale test er spild af tid og penge. *Politiken,* (2010)

Gipps,C. (1999). Socio-Cultural Aspects of Assessment. *Review of Research in Education, 24,* 355-392.

Goldstein, H. (2004). International comparisons of student attainment: some issues arising from the PISA study. *Assessment in Education* **11**(3).

Guba,E., & Lincoln,Y. (1994). Competing Paradigms in Qualitative Research. In N.Denzin & Y.Lincoln (Eds.), *Handbook of Qualitative Research.* (pp. 105-117). London: Sage Publications.

Hake,R.R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *Am.J.Phys., 66,* 64-74.

Halliday,M.A.K., & Martin,J.R. (1993). *Writing Science - literacy and discursive power.* Pittsburgh: University of Pittsburgh Press.

Harlen, W. (2007). Criteria for Evaluating Systems for Student Assessment. *Studies in Educational Evaluation, 33,* 15-28.

Henningsen, I. (2005). PISA - et kritisk blik. *MONA* (1).

Hopmann, S. T., Brinek, G., & Retzl, M. (2007). *PISA zufolge PISA - PISA according to PISA. Does PISA keep what it promises?* Vienna: LIT Verlag.

Jensen, E.B. (2007). *15-åriges viden om klimaforskelle*. Speciale ved Institut for Naturfagenes Didaktik, Københavns Universitet. www.ind.ku.dk. Lokaliseret 4. januar 2009.

Jørgensen,P. (2010a). Lærerformand: Drop nu de elendige test. *Politiken,* Uddannelse.

Jørgensen,P. (2010b). Nationale tests giver blank skærm igen. *Politiken,* Uddannelse.

Kane, M. T. (2006). Validation. In R.L.Brennan (Ed.), *Educational Measurement 4th Edition* (Westport: American Council on Education and Praeger Publishers.

LBK nr 730 af 21/07/2000 (www.retsinformation.dk)

Lemke,J.L. (1990). *Talking science : language, learning and values.* Norwood, N. J.: Ablex Publ. Corp.

LOV nr 572 af 09/06/2006 (www.retsinformation.dk)

Matti, T. (Ed.) (2009). Northern Lights on PISA 2006. Copenhagen: Nordic Council of Ministers.

McClung, M. S. (1979). Competency Testing Programs: Legal and Educational Issues. *Fordham Law Review, 47,* 651-712.

Mejding,J., & et al. (2004). *PISA 2003 - danske unge i en international sammenligning*. Danmarks Pædagogiske Universitets Forlag.

Mejding, J. & A. Roe (eds.) (2006). Northern Lights on PISA 2009. Copenhagen: Nordic Council of Ministers.

Messick, S. (1989). Validity. In R.L.Lin (Ed.), *Educational Measurement (3rd Edition)* (pp. 13-103). New York: American Council on Education/Macmillan.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the Structure of Educational Assessments. *Measurement: Interdisciplinary Research and Perspectives, 1,* 3-62.

Nordenbo,S.E., Allerup,P., Andersen,H.L., Dolin,J., & et al. (2009). *Pædagogisk brug af test - Et systematisk review*. Danmarks Pædagogiske Universitetsforlag og Dansk Clearinghouse for Uddannelsesforskning.

Mortimore, P., David-Evans, M., Laukkanen, R., & Valijarvi, J. (2004). *OECD-rapport om grundskolen i Danmark - 2004* København: OECD/Uddannelsesstyrelsen DK

OECD (1999). *Measuring Students' Knowledge and Skills* France: Organisation for Economic Co-operation and Development (OECD).

OECD (2004a). *Special Session of the Education Committee: Pilot review of the Quality and Equity of Schooling Outcomes in Denmark. Examiner's report.* 5. marts 2011 lokaliseret på http://pub.uvm.dk/2004/oecd/final.pdf.

OECD (2004b). *First Results from PISA 2003 - Executive Summary* Paris, France: Organisation for Economic Co-operation and Development (OECD).

OECD (2006). *Assessing scientific, reading and mathematical literacy: A framework for PISA 2006* OECD Homepage/Publishing: Organisation for Economic Co-operation and Development.

Ogborn,J., Kress,G., Martins,I., & McGillicuddy,K. (1996). *Explaining science in the classroom.* Milton Keynes: Open University Press.

Osborne,J. (2005). The Role of Argument in Science Education. In K.Boersma & et al (Eds.), *Research and the Quality of Science Education.* (pp. 367-380). The Netherlands: Springer.

Pellegrino,J.W., Chudowsky,N., & Glaser,R. (2001). *The Nature of Assessment and Reasoning from Evidence.* Washington,D.C.: National Academy Press.

Puchhammer,M. (2007). Language-Based Item Analysis - Problems in Intercultural Comparisons. In Hopman et al (Eds) (Ed.), *PISA zufolge PISA - PISA according to PISA.* (pp. 127-137). Wien: LIT Verlag GmbH & Co.

Säljö,R. (2003). *Læring i praksis - et sociokulturelt perspektiv.* København: Hans Reitzels Forlag.

Säljö,R. (2005). *Lärande och kulturella redskap : om lärprocesser och det kollektiva minnet.* Stockholm: Norstedts akademiska förlag.

Schoultz,J., Saljo,R., & Wyndhamn,J. (2001a). Conceptual knowledge in talk and text: What does it take to understand a science question? *Instructional Science, 29,* 213-236.

Schoultz,J., Saljo,R., & Wyndhamn,J. (2001b). Heavenly talk: Discourse, artifacts, and children's understanding of elementary astronomy. *Human Development, 44,* 103-118.

Sjøberg, S. (2007). PISA and "Real Life Challenges": Mission Impossible? In Hopman et al (Eds) (Ed.), *PISA zufolge PISA - PISA according to PISA* (pp. 203-224). Wien: LIT VERLAG GmbH & Co KG.

Skov, P. (2011). Brugen af evaluering i norsk grundskole: en undersøgelse med resultater der sandsynligvis også er relevante i Danmark. *Unge Pædagoger, 71, 57-64.*

Standards for Educational and Psychological Testing 1985

Statsministeriet (2001). *Regeringsgrundlag 2001. Vækst, Velfærd, Fornyelse*. København: Statsministeriet.

Teknologirådet/Diverse Oplægsholdere (2005). *PISA-undersøgelsen og det danske uddannelsessystem*. Teknologirådet.. 2005/12, Teknologirådets Rapporter.

Thompson & De Bortoli, 2008. *Exploring Scientific Literacy: How Australia Measures Up. The PISA 2006 Survey of Students' Scientific, Reading and Mathematical Skills,* Australian Council for Educational Research, Melbourne.

Tønnes Hansen,J., & Nielsen,K. (1999). *Stilladsering - en pædagogisk metafor.* Aarhus (In Danish): Klim.

Toulmin,S. (1969). *The Uses of Argument.* Cambridge, England: Cambridge University Press.

Uno,G.E., & Bybee,R.W. (1994). Understanding the dimensions of biological literacy. *Bioscience, 44,* 553-557.

Webb,N.M. (1997). Assessing Students in Small Collaborative Groups. *Theory Into Practice, 36,* 205-213.

Wellington,J., & Osborne,J. (2001). *Language and Literacy in Science Education.* Buckingham, Philadelphia: Open University Press.

Wood,D., Bruner,J.S., & Ross,G. (1976). The role of tutoring in problem solving. *Journal of child psychology and psychiatry, 17,* 89-100.

Wuttke, J. (2007). Uncertainties and Bias in PISA. In Hopman et al (Eds) (Ed.), *PISA zufolge PISA - PISA according to PISA* (pp. 241-263). Wien: LIT VERLAG GmbH & Co KG.

Bilag 1 *Solcreme*-opgaven i PISA sættet (m. scoringskriterier på engelsk)

SOLCREMER

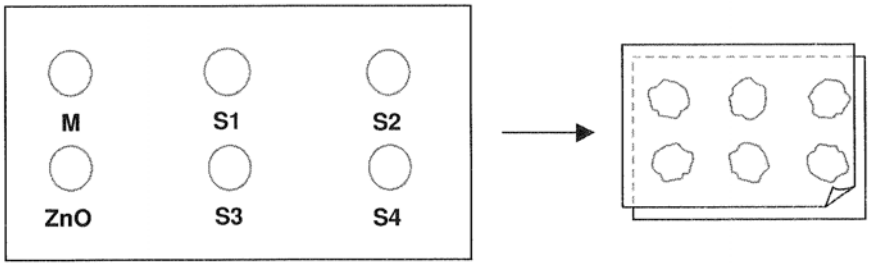
Mia og Dan tænkte over, hvilken solcreme der giver deres hud den bedste beskyttelse. Solcreme har en solbeskyttelsesfaktor *(SPF)*, som viser, hvor godt hvert produkt absorberer solens ultraviolette stråler. En solcreme med høj SPF beskytter huden i længere tid end en solcreme med lav SPF.

Mia udtænkte en måde at sammenligne nogle forskellige solcremer på. Hun og Dan samlede følgende ting:

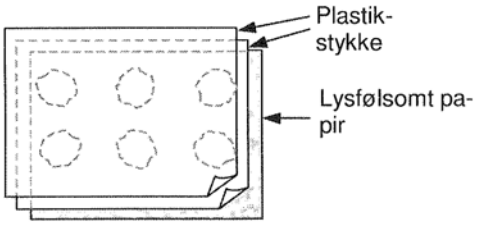
- to stykker klar plastik som ikke absorberer sollys
- et stykke lysfølsomt papir
- mineralolie (M) og en creme indeholdende zinkoxid (ZnO), og
- fire forskellige solcremer, som de kaldte S1, S2, S3, og S4.

Mia og Dan tog mineralolie med, fordi den lader det meste af sollyset komme igennem, og zinkoxid, fordi det næsten blokerer totalt for sollys.

Dan kom en dråbe af hvert stof ind i en cirkel, som var markeret på det ene stykke plastic, og lagde det andet stykke plastic oven på. Han lagde en stor bog ovenpå begge stykker og pressede ned.



Herefter lagde Mia plastikstykkerne ovenpå det lysfølsomme papir. Lysfølsomt papir ændrer farve fra mørk grå til hvid (eller meget lys grå) afhængigt af, hvor længe det udsættes for sollys. Til sidst stillede Dan stykkerne på et solrigt sted.



Spørgsmål 12: SOLCREMER

S447Q02

Hvilket af disse udsagn er en videnskabelig beskrivelse af mineralolien og zinkoxidens rolle, når de bruges i sammenligningen af solcremernes effektivitet?

A Mineralolie og zinkoxid er begge faktorer, som bliver testet.
B Mineralolie er en faktor som bliver testet, og zinkoxid er et referencestof.
C Mineralolie er et referencestof, og zinkoxid er en faktor, der bliver testet.
D Mineralolie og zinkoxid er begge referencestoffer.

Spørgsmål 13: SOLCREMER

S447Q03

Hvilket af disse spørgsmål prøvede Mia og Dan at besvare?

A Hvordan er hver solcremes beskyttelse sammenlignet med de andre?
B Hvordan beskytter solcremer din hud mod ultraviolet stråling?
C Er der nogen solcreme, som giver mindre beskyttelse end mineralolie?
D Er der nogen solcreme, som giver mere beskyttelse end zinkoxid?

Spørgsmål 14: SOLCREMER

S447Q04

Hvorfor blev det andet stykke plastik presset ned?

A For at forhindre dråberne i at tørre ud.
B For at sprede dråberne så meget ud som muligt.
C For at holde dråberne indenfor den markerede cirkel.
D For at dråberne skulle have samme tykkelse.

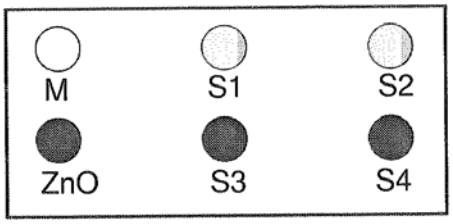
Spørgsmål 15: SOLCREMER

S447Q05 – 0 1 2 9

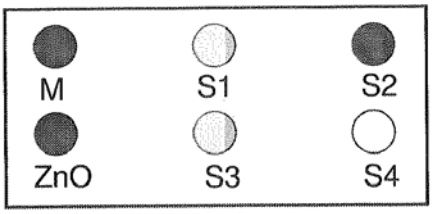
Det lysfølsomme papir er mørkegråt, og det falmer til en lysere grå, når det bliver udsat for sollys, og til hvidt når det bliver udsat for meget sollys.

Hvilket af disse diagrammer viser resultater, man kunne opnå? Forklar, hvorfor du vælger det.

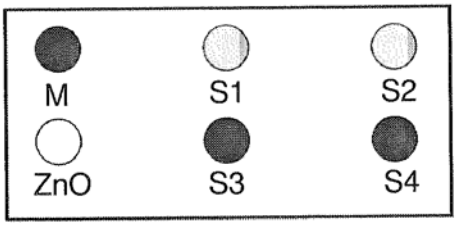
A



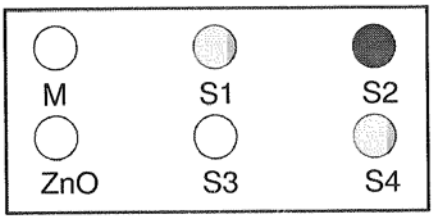
C



B



D



Svar: .....

Forklaring: .....

.....

.....

PISA scoringskriterier for SOLCREMEopgaven
(kun engelsk version tilgængelig)

Question 2: SUNSCREENS ("Spørgsmål 12" ovenfor) *S447Q02*

Full Credit
Code 1: D. Mineral oil and zinc oxide are both reference substances.
No Credit
Code 0: Other responses.
Code 9: Missing.

Question 3: SUNSCREENS ("Spørgsmål 13" ovenfor) *S447Q03*

Full credit
Code 1: A. How does the protection for each sunscreen compare with the others?
No credit
Code 0: Other responses.
Code 9: Missing.

Question 4: SUNSCREENS ("Spørgsmål 14" ovenfor) *S447Q04*

Full Credit
Code 1: D. To make the drops the same thickness.
No Credit
Code 0: Other responses.
Code 9: Missing.

Question 5: SUNSCREENS ("Spørgsmål 15" ovenfor) *S447Q05 – 0 1 2 9*

Full Credit
Code 2: A. With explanation that the ZnO spot has stayed dark grey (because it blocks sunlight) **and** the M spot has gone white (because mineral oil absorbs very little sunlight).
*[It is **not** necessary (though it is sufficient) to include the further explanations that are shown in parentheses.]*
• A. ZnO has blocked the sunlight as it should and M has let it through.
• I chose A because the mineral oil needs to be the lightest shade while the zinc oxide is the darkest.

Partial Credit
Code 1: A. Gives a correct explanation for either the ZnO spot **or** the M spot, but **not** both, **and** does not give an incorrect explanation for the other spot.
• A. Mineral oil provides the lowest resistance against UVL. So with other substances the paper would not be white.
• A. Zinc oxide absorbs practically all rays and the diagram shows this.

No Credit

Code 0: Other responses.

- A because ZnO blocks the light and M absorbs it.
- B. ZnO blocks the sunlight and mineral oil lets it through.

Code 9: Missing.

Bilag 2 Solcreme øvelsen i VAP evalueringen

Solcremer – en eksperimentel opgave

Opgave: I skal lave en undersøgelse, hvor I sammenligner forskellige solcremers evne til at beskytte huden mod solens uv-stråling.

NB: Det er vigtigt, at I laver undersøgelsen "så naturvidenskabeligt" som muligt!

1. **Planlæg undersøgelsen (ca. 5 min):** I skal selv diskutere Jer frem til, *hvordan* I vil lave undersøgelsen. I har de materialer til rådighed, som findes i den lille kasse I har fået udleveret. De vigtigste informationer om materialerne står på det vedlagte ark.
    - Hvis I for alvor kører fast, kan I bede om at få et hint hos observatøren.
2. **Fortæl observatøren, hvad I har tænkt Jer at gøre – og hvorfor.**
3. **Gennemfør Jeres undersøgelse (ca. 10 min)**
4. **Diskutér resultaterne med hinanden og observatøren**

Lidt baggrund:
Solcreme beskytter mod solens uv-stråler. Man kan lave solcremer som beskytter i forskellig grad – og som derfor har forskellig "solbeskyttelsesfaktor" (undertiden blot kaldt "SPF"). Jo større SPF-værdi desto bedre beskyttelse.

Materialerne,
som I kan bruge til Jeres undersøgelse:

- *Et stykke (blåt) papir, som er følsomt overfor sollysets uv-stråling.* Det bliver mere og mere hvidt, jo mere uv-lys det modtager.
  NB: Glas, fx i vinduer, fjerner uv-lyset fra solen. Derfor kan selve eksperimentet ikke gennemføres inden døre. På en lidt overskyet gråvejrsdag skal i regne med, at der går 5-10 min. i sollyset før man kan se, at farven er skiftet i rimelig grad.

- *To stykker klar plast ("transparenter")*. Vi antager, at de ikke vil absorbere solens uv-lys.

- *Mineralsk olie*. Praktisk taget al uv-lys vil kunne passere igennem et tyndt lag af mineralsk olie.

- *En creme med zinkoxid (ZnO)* – som stort set blokerer for uv-stråling

- *De fire solcremer, som I skal undersøge*. For nemheds skyld er de blotmærket som: S_1, S_2, S_3 og S_4.

- Hvad I nu ellers har ved hånden af almindeligt skolegrej, såsom blyanter, bøger m.m…

Bilag 3 Samtaleskema for *Solcreme*-opgaven

10-03-2006

Tid	Samtale-fokus:	Relevante spørgsmål:
0-2	**Igangsætning:** **Handling**: Udlevér opgave-arket og materiale-boxen. **Samtale:** Indgå i en samtale, **så det bliver klart, *om* de har forstået opgaven.** Hjælp med at afklare opgaven – men *uden at fortælle, hvad de skal gøre* (og uden at komme ind på 'reference' m.m.	• **"Er I med på, hvad opgaven går ud på?"**
3-6	Lad dem diskutere, hvad de vil gøre	• (bare hold dem på sporet og mind dem om tiden)
7-10	**Afklaringssamtale:** **1.** *Model-aspektet*: Er de stand til at se, hvordan **fotopapiret i princippet kan bruges som en model for hud** – og at solcremen i princippet "bare skal smøres på papiret/huden"? **2.** ***'Reference' –* kender de ordet og/eller ved de, hvad en sådan gør godt for?** Forstår de betydningen af, at have en (ca) 0%-plet (mineralolie) og 100% (ZnO)-absorptionsplet på papiret? **3. "Et naturvidenskabeligt eksperiment?"** Få dem til at fortælle, hvad de forbinder med "et naturvidenskabeligt eksperiment" – og hvordan dette indgår i deres plan for undersøgelse	1. "Opgaven var jo at sammenligne deres evne til at beskytte **'hud' – men jeg ser ingen hud i Jeres eksperiment?"** o Evt opfølgning: "hvad mener I?", "Fortæl mere….. "(indtil det er klart om modellen foreligger) 2. "**Kender I ordet reference?"** o Evt. opfølgning: "Nu siger I.. Hvad mener I i grunden med det…?", "Uddyb…" o **"hvorfor er det nyttigt at have zinkoxid og mineralolie med i materialeboxen** – når det nu engang er solcremer I skal undersøge?" o Evt. opfølgning: "Hvad mener I?", "Fortæl mere.." o Evt. **vis dem PISA-figur1: "hvorfor kan det være hensigtsmæssigt at sætte mineralolie og zinkoxid på filmen også?"** o Evt. fortæl: "disse stoffer er *referencer og* med for at have noget at sammenligne med, fx en plet svarende til 0% og en plet for 100%" **3. "I skulle planlægge undersøgelsen *så naturvidenskabeligt som muligt.* Hvad forbinder I med det?"** o **"Hvad er "naturvidenskabeligt" ved Jeres planlagte eksperiment?"** o Evt.: Vis dem PISA-figur2 med filmen som sandwich mellem plast og med samtlige stoffer

	4. Har de styr på, at man må have **en ensartet lagtykkelse for at kunne sammenligne – og at dette opnås ved at presse plaststykkerne (ensartet) sammen?**	påsat. Spørg: **”Hvorfor kan det være hensigtsmæssigt, at gøre sådan?”** o NB: Hvis de vil smøre solcreme direkte på filmen: **”Hvorfor mon der var transparenter af plast med i materialeboxen?”** **3.** NB: Hvis de ikke selv kommer på at arbejde med ensartet lagtykkelse**: vis PISA-figur3 og spørg: ”Hvilket formål kan det have at presse plasten ned, fx ved at lægge en bog ovenpå”?** **a.** Evt opfølgning: ”Hvad kunne det ellers være?”, ”uddyb lige……” (indtil det er klart, om de har et kvalificeret bud)

Facilitator-instruks: Gennemførelsesfasen.

Tid	Samtale-fokus:	Relevante spørgsmål og kommentarer:
11-15	At få dem i gang med det **praktiske i en fart.** At **spørge opklarende** undervejs, hvis de laver overraskende ting. Få filmen **UDENFOR efter 5 min.**	o ”I har **kun 5 min til at sætte forsøget** i gang!” o **Hvorfor gør I det**? o Er dét nu **hensigtsmæssigt?** o Er det i **overensstemmelse med Jeres plan** o.s.v.? (Der gives kun en egentlig tilrettevisning, hvis de foretager sig ting, som gør det tvivlsomt om deres undersøgelse vil lykkes)
	Samtale-fokus	Relevante spørgsmål & kommentarer
16-22	Eleverne **besvarer spørgsmålsark individuelt og skriftligt**. Dvs. ikke rigtigt nogen samtale her.	Du må helst ***ikke besvare*** **deres spørgsmål** her**.** Hvis de har spørgsmål må du til gengæld, ***gerne spørge opklarende****:* ”Hvorfor vil du gerne have det at vide?”, ”Hvordan tænker du, når du spørger sådan….?”

Facilitator-instruks: Fortolkningsfasen

	Samtale-fokus:	Relevante spørgsmål:
22-25	1. Forsikr dig, om at **eleverne har en plade, hvor ZnO optræder mørk, mineralolien lys** – og solcremerne nuancemæssigt derimellem. (hvis dette ikke er tilfældet, så vis eleverne PISA-figur 4). Fokus er: **hvad viser pladen** i relation til opgaven? **Har eleverne forståelse for sammenhængen, at jo mere blå filmen forbliver, desto bedre beskytter solcremen mod uv-stråling?**	1. o Åben intro**: "hvordan vil I fortolke filmen?** Hvad siger den om solcremernes evne til at beskytte mod uv-lys? o Specifik opfølgning: **"Hvad siger den om disse to (udpeg dem) solcremer i forhold til hinanden?"** o "Har **solcremerne givet en rimelig farve i forhold til referencepletterne?"** o "Kan man på filmen **afgøre *om* mineralolien slipper "praktisk taget" al uv-stråling igennem?"** o **Spørg opklarende**, hvis eleverne er uklare, inkonsistente eller i åbenlys modstrid med hinanden.
26-	2. Har de forbedringsforslag til udførelsen af eksperimentet?	2. **"Har I nogen forslag til, hvordan undersøgelsen kan blive endnu bedre?"** Uddyb.
-28	3. Er de i stand til at formulere forbehold overfor modellen (fotopapir = hud)?	3. **"Er der nogen svagheder eller indbyggede begrænsninger ved den undersøgelse I har lavet?"**

Afrunding

	Samtale-fokus:	Relevante spørgsmål:
29-30	4. **Har de viden, som gør at de kan se forsøget i faglig sammenhæng eller i et større perspektiv?**	4. o **"Har I arbejdet med sollys eller uv-stråling? Kan I fortælle noget mere…?** (fx at uv-stråling er elektromagnetisk stråling med rimelig høj energi og lille bølgelængde) o **"Ved I noget om, hvad uv-stråling gør ved huden?"** (fx at strålingsenergien overføres til molekyler (bl.a. DNA), som skydes i stykker (til radikaler), med risiko for mutation og hudkræft) o **"Beskyttelse mod Solens uv-stråling er også et globalt/verdensomspændende *miljø*-problem. Kan I fortælle noget om det?"** (fx at Jorden er omgivet af et tyndt lag ozon som også absorberer uv-stråling. Dette lag nedbrydes af menneskeskabte udledninger, fx kvælstofoxider fra biler, CFC-gasser fra bl.a. køleskabe m.m., haloner i brandslukning m.m.)

Bilag 4. Principper for kodning af *Solcreme*-opgaven.

Grundprincipper:

- Individuel scoring anvendes kun i relation til de direkte PISA-relaterede spørgsmål (sp.44, sp.46 og sp.47). I øvrige spørgsmål vurderes blot parrets *forenede anstrengelser.*
- I hvert PISA-spørgsmål er der anført *et antal del-aspekter*, som *scores separat:*
  - Et korrekt og selvstændigt delaspekt bidrager med 2 point til scoren.
  - Et korrekt aspekt, som fremkom efter *lidt hjælp* (fra interviewer eller makker): 1
  - Et ukorrekt, blankt eller *massivt hjulpet* aspekt: 0
- Ofte vil kun den ene elev nå at ytre sig eksplicit omkring et aspekt, mens den anden måske vil nikke eller på anden måde tilkendegive tilslutning/uenighed. Ud fra en holistisk vurdering af kropsprog m.m. tilstræbes det at give hver enkelt elev en score på samtlige delaspekter.
- Hvert delaspekt registreres med en to-cifret score, hvor det andet ciffer angiver sikkerheden i fastlæggelsen af scoren:
  - 2. ciffer = 0: sikker score
  - 2. ciffer = 1: nogenlunde sikker score
  - 2. ciffer = 2: meget usikker vurdering.

  Som tydeliggørelse: scoren 20 står for en korrekt, selvstændig score med en sikker fastlæggelse. Scoren 21 angiver en korrekt score – som er behæftet med lidt usikkerhed, fx fordi partneren *først* var inde med dele af indsigterne..
- Den to-cifrede score 05 bruges, når det er aldeles umuligt at sige noget om det pågældende aspekt.
- For hvert aspekt vurderes mængen af:
  - Faglig Interviewer-mediering (FIM): skala 0-2 (0: ingen, 1: moderat, 2: stor)
  - Faglig Elev-elev-interaktion (FEEI): skala 0-2 (som ovenfor)

Præciseringer efter pilot-scoringer:

1. Forslag til tydeliggørelse af Medieringskategorierne:

FIM:
Score 0: kun den hjælp som ligger i det oprindelige PISA-spørgsmål.
D.v.s. en interviewer, som stiller spørgsmålet: "kender I ordet reference?" mediererer reelt *ikke*.
Score 1: gives når intervieweren giver dem hjælp til at se vejen (enkelttrin/-aspekter)

- Tilbyder extra formuleringer/omformulering
- Stiller relativt *åbne* spørgsmål, som kan lede dem på vej
- Henleder elevernes opmærksomhed på et relevant aspekt

Score 2: gives når intervieweren *baner* vejen

- når vejen struktureres fuldt af intervieweren og når der er hjælp på adskillige af dennes trin
- når intervieweren decideret sænker kompleksiteten og fx leverer et delaspekt af svaret.

Score 0.5 gives, når intervieweren decideret foregriber svaret.
[se filen VideoValidering2Opsummering]

En logbogsoptegnelse fra udviklingen og valideringen af scoringsmanualen:

"Kanon-god påpegning af et muligt scorings-/konsistensproblem – umiddelbart kan det løses indenfor de eksisterende kategoribeskrivelser ved at sp46.1 tildeles scoren 1 og samtidig FIM=1. Alternativt sp46=2 og FIM=0. Overvej lige ud fra videoen om nogen af disse er helt tilfredsstillende. Jeg vil gerne, at elever som faktisk synes at kunne besvare spørgsmålet korrekt får scoren 2 - også selvom assistenten måske stiller spørgsmål o.s.v. Det afgørende er her, om assistenten *indholdsmæssigt* leverer bidrag (jf. score FIM =1 og FIM-score 2 i opsummeringen *VideoValideringOpsummering*). Pointen må i hvert fald være, at *såfremt* assistenten leverer en FIM-indsats på niveau 1 eller 2 vil man nok ikke samtidig kunne give sp.46.1-scoren 2. Så min kombination af scoringer er i hvert fald inkonsistent. Umiddelbart tror jeg, at selve kodningsbeskrivelserne er gode nok!? [se filen *ScoringsdialogmGunver*, hvor der i øvrigt også er præcisering af de eksperimentelle kompetencer]"

Scoringen af de enkelte items og del-indikatorer mhp. direkte PISA-sammenligning

Udvikling i forståelsen af, hvad forsøget *egentlig indebærer* (dvs. ikke kun problemstillingen "at undersøge, hvilken der beskytter bedst", men snarere forståelsen af *hvorledes denne problemstilling kan undersøges eksperimentelt!* *NB:* • Første del-score kan meningsfuldt foretages efter det praktiske arbejde er udført. Variablen 'Taskudv1' er indføjet på det relevante sted. • Tilsvarende er 'Taskudv2' indføjet efter fortolkningen af de tilvirkede film.	KOLLEKTIV, LÆNGDEGÅENDE VURDERING. Forståelsen af den eksperimentelle undersøgelse registreres i tre punkter undervejs for at se, om der sker en udvikling: - *før eleverne går i gang* med det praktiske - *undervejs* i det praktiske - ved den *afsluttende fortolkning* af pladerne Der scores: - er der tegn på, at forståelsen udvikler sig under det praktiske? Ja - nej – uvis - er der tegn på, at forståelsen udvikler sig undervejs i den afsluttende fortolkning? Ja- nej - uvist
Forsøgsdesign: **Modelaspektet (film som hud)** **Skala-etablering/"sammenligningsgrundlag" via referencerne (PISA-sp.44)** **Variabelkontrol (tykkelse via transparenter + bog, PISA-sp.46)**	Droppes i scoringen Droppes i scoringen INDIVIDUELLE SCORER: PISA-Sp.44: Delaspekter, som scores for hver elev og hver for sig, jf "grundprincipper" ovenfor: - Sp44-1:Kendskab til ordet reference (0-2, +usik) - Sp44-2: Forståelse af referencernes funktion (0-2, +usik) Usikkerhed angives på skalaen 0-2 KOLLEKTIVE VURDERINGER: - Sp44-1/-2-FIM (Faglig interviewer mediering) (0-2) - Sp44-1/-2-FEEI (Faglig Elev-Elev-interaktion) Generelt: Kode 05 bruges til at angive, at det er der ikke belæg til at mene noget om. INDIVIDUELLE SCORER: PISA-Sp.46: Delaspekter, som scores individuelt og enkeltvist: - Sp46-1: Lagtykkelse betyder noget for solbeskyttelse (0-2) - Sp46-2: Lagtykkelsen må være ens for at man kan sammenligne SBF (0-2 - Sp46-3: En ensartet lagtykkelse kan i praksis tilvejebringes via pres med en bog (af et transparentindeklemt lag) (0-2) KOLLEKTIVE SCORER: På samme måde som ovenfor angives for hvert aspekt en vurdering af FIM og FEEI. Kode 05 anvendes på samme måde som ovenfor

"Et naturvidenskabeligt eksperiment"?	KOLLEKTIVE SCORER: Elevernes svar vurderes i talsatte kategorier efter arten *og* mere detailleret med stikord/statements **a. Med tal:** 0: Blank 1: skæve/irrelevante bidrag 2: relevante, men common-sensical 3: relevante m. fagbegreber/fagtermer b. **Med stikord (**gerne korte statements, som giver mening i sig selv)
Praktisk eksperimentel kompetence	KOLLEKTIVE SCORER: Der er tegn på, at eleverne *selv:* K1: Kan lave en håndværksmæssigt hensigtsmæssig realisering af den indledende plan K2: tænker i variabelkontrol K3: tænker i fejlkilder og minimering af sådanne K4: kan revurdere forsøget/fremsætte forbedringsforslag K5: har en forståelse af forsøget som model K6: andet relevant (skriv stikord) Flere kompetencer må gerne anføres i samme felt
Fortolkning af de eksponerede plader: **(PISA-sp.47)**	INDIVIDUELLE SCORER: PISA-Sp.47: Delaspekter, som scores individuelt og enkeltvist: - Sp47-1: Sammenhængen mellem beskyttelsesgrad og papirfarve (jo mørkere desto bedre beskyttelse – eller tilsvarende) (0-2. +usik) - Sp47-2: Identifikation af referencepletterne (0-2, +usik) - Sp47-3: En kvalificeret diskussion af solcremernes relative beskyttelse/rangordning ud fra mønstret) (0-2,+usik) KOLLEKTIVE SCORER: På samme måde som ovenfor angives for hvert aspekt en vurdering af FIM og FEEI. Kode 05 anvendes på samme måde som ovenfor
Forbedringsforslag og/eller problematiseringer	droppes
Faglige perspektiveringer	KOLLEKTIVE SCORER: Inden for hvert af indholdsområderne - P1: UV-strålingens natur (bølger, energi) - P2: UV-Stråling som skadelig (mennesker,dyr) - P3: UV-stråling og det globale miljø (ozonlag) Scores elevernes kendskab. Der anvendes følgende tal-scorer: - 0: intet kendskab, blanke - 1: udtrykker hverdagssproglig viden (fx medieviden) - 2: udtrykker viden, som involverer relevante fagtermer
Test-par-dynamik & asymmetri	As1: asymmetri af faglig art: 0: ikke nogen påfaldende asymmetri, begge bidrager 1: E1 klart den der leverer de faglige bidrag 2: E2 klart den der leverer de faglige bidrag As2: asymmetri grundet status og/eller dominerende personlighedstræk: 0: ikke nogen påfaldende asymmetri 1: E1 klart den der sætter sig på ordet, materialerne, initiativet.. 2: E2 klart den der sætter sig på ordet, materialerne, initiativet…

Bilag 5 Etablering af VAP-indeksværdier for solcremeopgaven

Som tidligere anført skal eleverne demonstrere fyldestgørende opfyldelse af 2 delaspekter for at få fuldt pointtal i sp.44-opgaven osv.. Mere generelt findes scoren i et PISA-spørgsmål ved simpelt at addere del-scorerne til en indeksvariabel, jf. oversigten i nedenstående tabel.

PISA-spørgsmål	**VAP-Empiri (delaspekter)**	**Modsvarende VAP-indeksvariabel**
Solcreme (reference/funktion) (S447Q02)	Video + scoring: (Sp44-1, Sp44-2)	VS447Q02=Sp44-1+Sp44-2 (værdiinterval: 0-4)
Solcreme (spørgsmål *om* science) (S447Q03)	Skriftlig besvarelse (individuel, efter praktisk start): (Sp45skr)	VS447Q03=Sp45skr
Solcreme (tykkelseskontrol) (S447Q04)	Video + scoring: (Sp46-1, Sp46-2, Sp46-3)	VS447Q04= Sp46-1 + Sp46-2 + Sp46-3 (værdinterval: 0-6)
Solcreme (resultatfortolkning) (S447Q05)	Video + scoring: (Sp47-1, Sp47-2, Sp47-3)	VS447Q05= Sp47-1 + Sp47-2 + Sp47-3 (værdiinterval: 0-6)

Her bemærker/erindrer man, at S447Q03 er speciel ved *ikke* at kunne transformeres til praktisk aktivitet. Eleverne har derfor besvaret denne del af opgaven i den oprindelige PISA-version i en pause undervejs i det praktiske forløb.

Den metodiske usikkerhed afføder flere ufuldstændige indeksvariable – og tynder ud i sammenligningsgrundlaget
I det sædvanlige PISA-testformat er der et antal blanke, som udtryk for, at usikkerhed får elever til at undlade at svare – eller fordi eleverne ikke når omkring det pågældende spørgsmål.

I VAP-undersøgelsens to andre *interview-bårne* opgaver skyldes forekomsten af blanke primært, at interviewerne ikke tydeligt får fulgt op på uklare elevtilkendegivelser, eller at de undervejs i samtalen kommer til at afskære eleverne fra at levere de kritiske videnselementer, som PISA vil honorere. Hyppigheden af blanke indenfor et enkelt spørgsmål ligger i intervallet 8-11 for disse spørgsmål - ud af et totalt sample på 125.

I den praktiske paropgave inducerer metodologien flere og anderledes begrundede blanke: Hvert PISA-spørgsmål modsvares af en sammensat indeksvariabel med 2-3 delaspekter. Heri ligner opgaven de foregående. Men: usikkerheden på at vurdere det enkelte delaspekt er noget større her – og hvis blot eet af de indgående delaspekter ikke kan bedømmes, så kan indeksvariablen ikke dannes. *Derfor må man forvente – og leve med – at den anvendte metodologi producerer flere blanke (de "overlevende" skulle så til gengæld gerne have betragtelig pålidelighed).* Konkret er antallet af blanke 10 (VS447Q02), 16 (VS447Q05), 17 (VS447Q03) – og 50 (VS447Q04)! Mens de første tre opgaver har en acceptabel hyppighed af ufuldstændige/"blanke" på samme niveau, som de interview-bårne opgaver, så er vi ude af stand til at levere et fuldt indeks for 50 elevers præstation i spørgsmål VS447Q04. Af empirien fremgår det med stor tydelighed, at diskussionen og demonstrationen af, hvorledes man kontrollerer lagtykkelsen af solcreme (i en test for relativ solbeskyttelsesfaktor) må have en særlig subtil karakter, som vanskeliggør en pålidelig scoring af

begge elever. Det er et metodisk interessant problem at indkredse, hvori vanskeligheden består, men for denne sammenhæng er det væsentlige, at i en simpel sammenligning af PISA-item-respons og modsvarende VAP-indeks vil S447Q04 være noget svagere datamæssigt funderet end de øvrige. De mange "udfald" af scorede elever på dette spørgsmål betyder selvsagt også, at man ikke kan etablere en *samlet* (rå) VAP- score for disse elever. For at få tilstrækkeligt mange med i den statistiske analyse er det derfor tjenligt at sammenligne PISA og VAP via deres rå scorer for *hvert enkelt item*, frem for blot at studere forskelle i total-scorer. Denne pointe præger analysen, som fremlægges i næste afsnit.

Bilag 6 Rasch-analyse-bidrag og kommentarer

(privat korrespondance med P. Allerup, kronologisk med de nyeste først)

Bilag 6.1: Kommentarer i tilknytning til eksternt review (punkterne i teksten refererer hertil)

Fra: Peter Allerup [mailto:nimmo@dpu.dk]
Sendt: 5. oktober 2010 00:16
Til: Jens Dolin
Cc: Peter Allerup; André Torre
Emne: SV: Hjælp til PISA-projekt - igen!

Kære Jens,

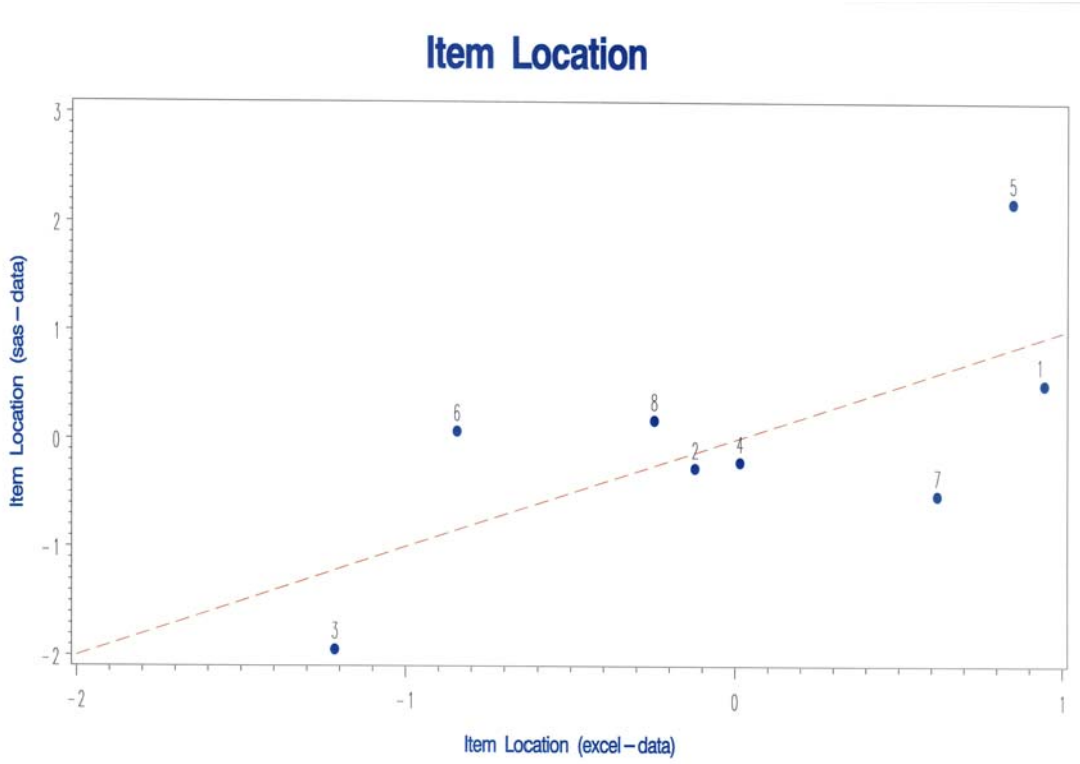
Jeg har fundet de sider frem, som jeg – også dengang – synes, at I skal/skulle benytte som dokumentation for Rasch analyserne. Så punkt 2 side 7 bør 'håndteres' ud fra følgende:

Den ene tegning viser (item locations) at de 'nye' items ,indbyrdes relative sværhedgrad ikke er så meget forskellig fra de gamle – dvs den nye skala fungerer så nogenlunde på samme måde som den gamle. Det er ekstremt vigtigt, ellers giver efterfølgende sammenligninger mellem person scores absolut ingen mening (anden tegning med to histogrammer, der viser systematisk forskydning af nye scores sammenlignet med de 'gamle')!

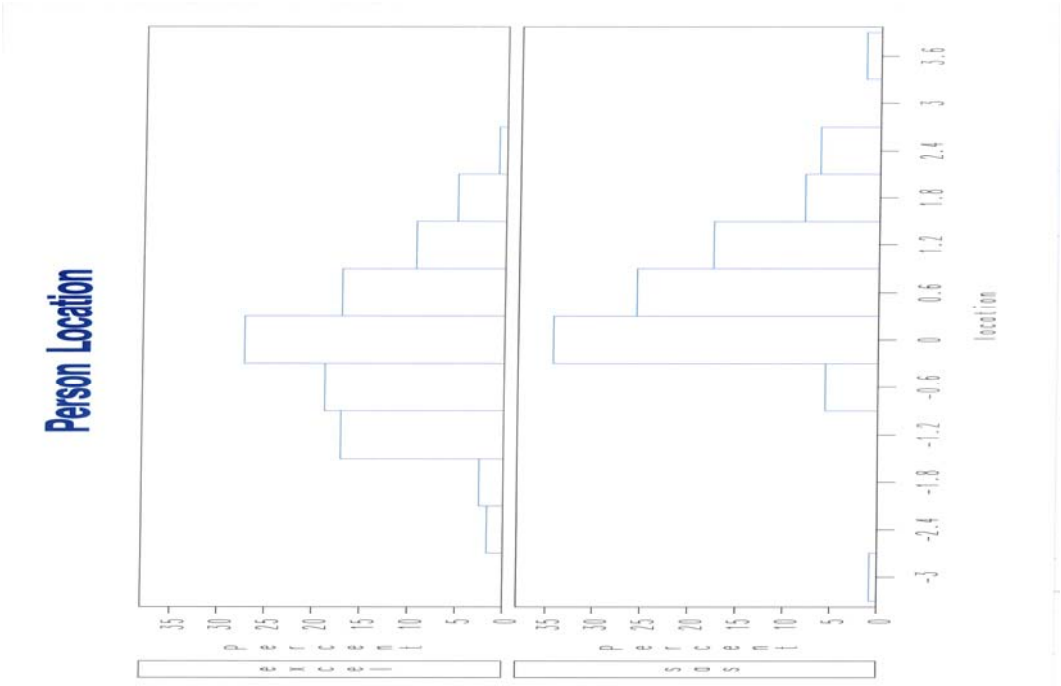
Selve fittet af Raschmodellen er gennemført ved hjælp af RUMM 2020, det samme program som anvendes ved analyserne af de nationale tests. Resultatet af test of fit er vedlagt og viser ikke alvorlige fejl! (det er udmærket, at man kan finde svagere tilpasning, hvis man anvender stærkere metoder, men nu har vi valgt at gennemføre 'standard' på området= de programmer, der anvendes fx i de nationale tests)

Der er helt sikkert en vis grad af lokal afhængighed – ikke overraskende – på grund af den mediering, som finder sted under det alternative regime, men det er ikke så markant, at RUMM – analysen falder på gulvet af den grund!
Kritikernes mening om lokalafhængighedens (punkt 3) indflydelse på opgavernes validitet må skyldes en misforståelse .- validitet har ikke noget at gøre med 'forvaltningen' af responser, herunder muligheden for 'afsmitning' eller lokal afhængighed.

Bilag 6.1 (fortsat)



Bilag 6.1 (fortsat)



Bilag 6.1 (fortsat) Vedhæftet PDF-dokument med test-of-fit data indsat her.

RUMM2020 Project: SAS Analysis: RUNALL
Title: ALL 8 V-SAS ITEMS INCLUDED.
Display: SUMMARY TEST-OF-FIT STATISTICS Date: 4 jun 2008 01:07:26

ITEM-PERSON INTERACTION

	ITEMS		PERSONS	
	Location	Fit Residual	Location	Fit Residual
Mean	0.000	-0.406	0.649	-0.291
SD	1.148	0.835	0.902	0.664
Skewness		0.448		1.277
Kurtosis		-1.190		0.812
Correlation		0.000		-0.258
Complete data DF =		0.827		

ITEM-TRAIT INTERACTION		RELIABILITY INDICES	
Total Item Chi Squ	54.479	Separation Index	0.46699
Total Deg of Freedom	32.000	Cronbach Alpha	N/A
Total Chi Squ Prob	0.007880		

LIKELIHOOD-RATIO TEST	POWER OF TEST-OF-FIT
Chi Squ	Power is LOW
Degrees of Freedom	[Based on SepIndex of 0.46699]
Probability	

Bilag 6.2.

”Afsnit til Dolin’s papir om Rasch

Peter Allerup 22. juni 2009

Raschanalyser af opgaver med ændret format.

Når man sammenligner besvarelser fra opgaver der er stillet under forskellige betingelser kan gøre det på flere måder. I nærværende studie drejer ’forskelligheden’ sig om, at opgaverne, hvis besvarelser skal sammenlignes, er stillet under benyttelse af forskellige opgaveformater. Opgaverne kan derfor, groft sagt sammenholdes to og to, den ene stillet i det sædvanlige PISA format, den anden i det såkaldte VAP format. Der er andre stede redegjort for hvilke opgaver der er tale om.

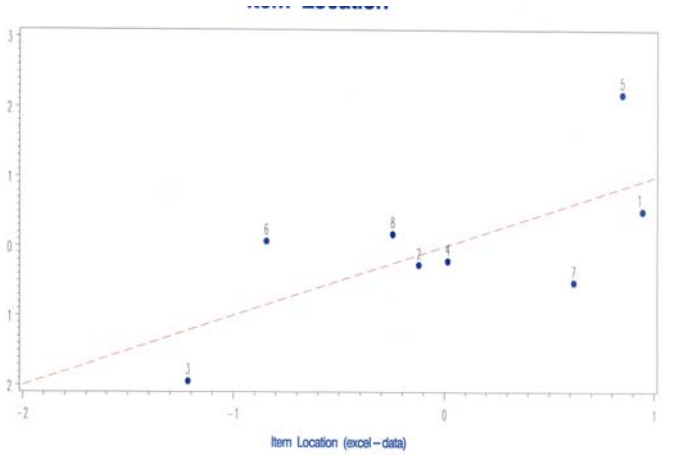
Sammenligningen imellem besvarelser afgivet under de to formater kan i princippet udføres opgavepar for par og man kan efterfølgende danne en overordnet konklusion vedrørende VAP-formatet. Strategien er ikke optimal af flere grunde. Dels ser man væk fra, at PISA opgaverne er blevet underkastet psykometriske (Rasch) analyser og blevet ’godkendt’ mht. til at være ’dele’ af en fælles PISA-skala, dels ville man være uvidende om VAP-opgavernes modsvarende skala egenskaber. De sidste er lige så nødvendige som de originale PISA – opgavers skalaegenskaber mht. til at kunne bruges som grundlag for sammenligninger mellem elevbesvarelser. Et hovedargument bag ved ønsket om at opgaverne skal have passende skalaegenskaber er – ud over muligheden for at måle én fælles dimension – at gøre det så statistisk præcist som muligt. Minimalisering af den såkaldte ’error of measurement’ som i PISA designet fører til at man netop anvender mange opgaver i hvert hefte.

Det er en styrke ved opgaver, der allerede er ’godkendt’ ved de psykometriske analyser, at man kan anvende et hvilket som helt delsæt af opgaverne og stadigvæk opnå, at man måler samme ’dimension’ som alle opgaverne. Det anvendte delsæt af PISA opgaver nyder godt af disse egenskaber mht. til at ’måle’ PISA-relevante aspekter. Det er hensigten, at besvarelsen af VAP opgaverne skal sammenlignes med besvarelsen af de originale PISA opgaver under udnyttelse af de samme psykometriske fordele , som gælder ved PISA . Det vil derfor sige, at VAP opgaverne underkastes en Rasch analyse og derefter kan man sammenligne med PISA opgaver med VAP opgaver effektivt ud fra skala-scores fra de to opgavescenarier.

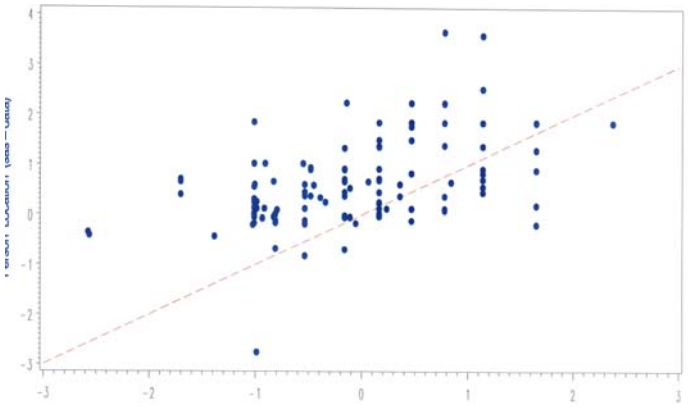
Detaljerne bag ved Rasch analyserne gengives ikke, men de teststørrelser, som normalt anvendes ved kontrollen af den statistiske Rasch Model peger på flere problemer mht. til at indplacere de 8 VAP opgaver på én dimension. Fx er det ikke helt klart, at den relative sværhedsgrad af enkelte opgaver er uafhængig af, om det er en ’dygtig’ eller ’svag’ elev, der besvarer opgaven. Noget som skal være tilfældet, hvis Rasch analysen skal godkende alle 8 opgaver. Der er ikke plads til at sortere opgaver fra blandt de 8 opgaver, sådan som man normalt gør det (forud for indplaceringen af de originale PISA opgaver, går en fase, hvor ca. 50% af opgaverne tages ud pgr. af psykometriske problemer). Problemerne er samlet vurderet til at være af et sådant omfang, at man kan gå videre med alle 8 opgaver, som ’godkendt’ af Rasch analysen.

Bilag 6.2 (fortsat)

For at sammenligningen derefter kan gennemføres på skala-niveau kræves, at PISA opgaverne og VAP opgaverne for, i den mindste nogle få opgavers vedkommende, lapper over. Dvs. har samme relative sværhedsgrad. Dette er undersøgt og fundet acceptabelt, lidt overraskende, for de fleste af opgaverne. Tegningen herunder viser de 8 items relative sværhedsgrader, vurderet i VAP skala og i PISA skala. Ens relative sværhedsgrader opnås, hvis punkterne ligger på den stiplede linje.

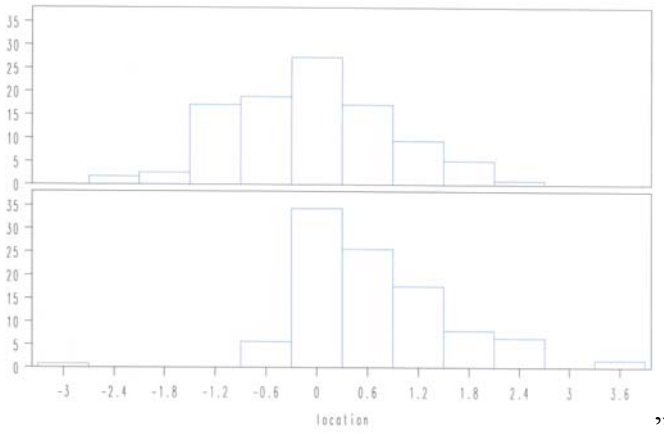


Endelig kan elevernes præstationer sammenlignes via de to godkendte Rasch skalaer. Det er gjort i figuren herunder. Som det fremgår, ligger majoriteten, en signifikant stor del, af eleverne *over* den indlagte stiplede identitetslinje, hvilket betyder, at eleverne vurderet med VAP skalaen er 'dygtigere' end elevpræstationerne beregnet ved hjælp af de originale PISA opgaver



Fordelingsmæssigt kan forskellen mellem de to sæt scoreværdier anskueliggøres ved at sammenligne to histogrammer over samlingen af elevpræstationer fra de to svarskalaer. Dette er gjort neden for, og det fremgår at VAP scorefordelingen (den nederste) ligger tydeligt til højre i forhold til fordelingen af PISA scores (den øverste).

Bilag 6.2 (fortsat)



”

Bilag 6.3:

Date: Wed, 26 Aug 2009 23:12:19 +0200 [08/26/2009 11:12:19 PM CEST]
From: Peter Allerup <nimmo@dpu.dk>
To: Lars Brian Krogh <lars.krogh@ivs.au.dk>, dolin@ind.ku.dk **Cc:** Peter Allerup <nimmo@dpu.dk>
Subject: Hjælp til PISA gentestningstolkning

Kære Lars og Jens,
Så fandt jeg tallene frem fra gemmerne. Det var lidt støvet, og jeg skulle lige igennem en lille renselsesproces, før jeg kunne genkende *hvad* det hele egentlig gik ud på.

Jeg er nu ret sikker på, at

- den procentvise forskel mellem de to gruppe ligger på ca 25% rigtige

- at denne 'rå' forskel udmålt på en PISA målestok (med de sædvanlige internationale 500 i midten og en standardafvigelse på 100) beløber sig til ca 125 point

Håber at I kan bruge det
mange hilsner fra

Peter

Bilag 6.4

Date: Thu, 12 Jun 2008 08:20:51 +0100 [06/12/2008 09:20:51 AM CEST]
From: Peter Allerup <nimmo@dpu.dk>
To: Peter Allerup <nimmo@dpu.dk>, lkrogh@phys.au.dk
Cc: Jens Dolin <dolin@ind.ku.dk>, Lars Krogh <lars.krogh@si.au.dk>
Subject: SV: SV: PISA items

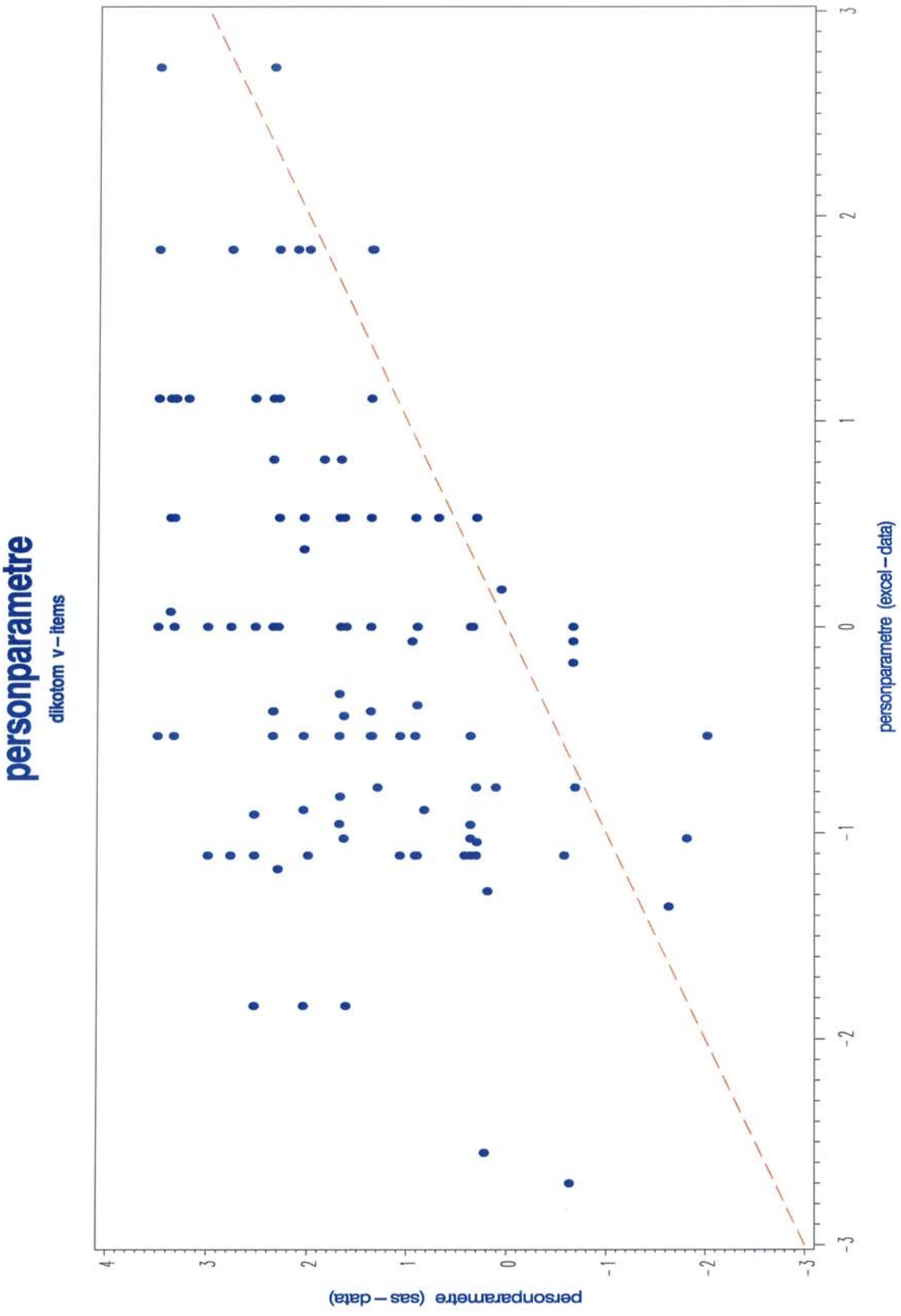
Show this HTML in a new window?

Kære Jens og Lars,
Jeg har lige fået tilsendt sidste 'sending' (efter tlf mødet) fra Andre - det er lift kortfattet og uden ord, men forhåbentlig fik i nedenstående mail fra mig i går med de vigtigste opsummeringer.
Lidt uddybende i forhold til de to sidste tegninger: Der er et plot nr 1, hvor elevernes parametre er estimeret under de to seancer og er holdt op mod hinanden - tydeligvis med fordel til 'medieringsseancen' selv om der findes undtagelser (dem kan vi selvfølgelig identificere). Estimationen er foretaget under den dikotomisering, som vi talte om i telefonen.

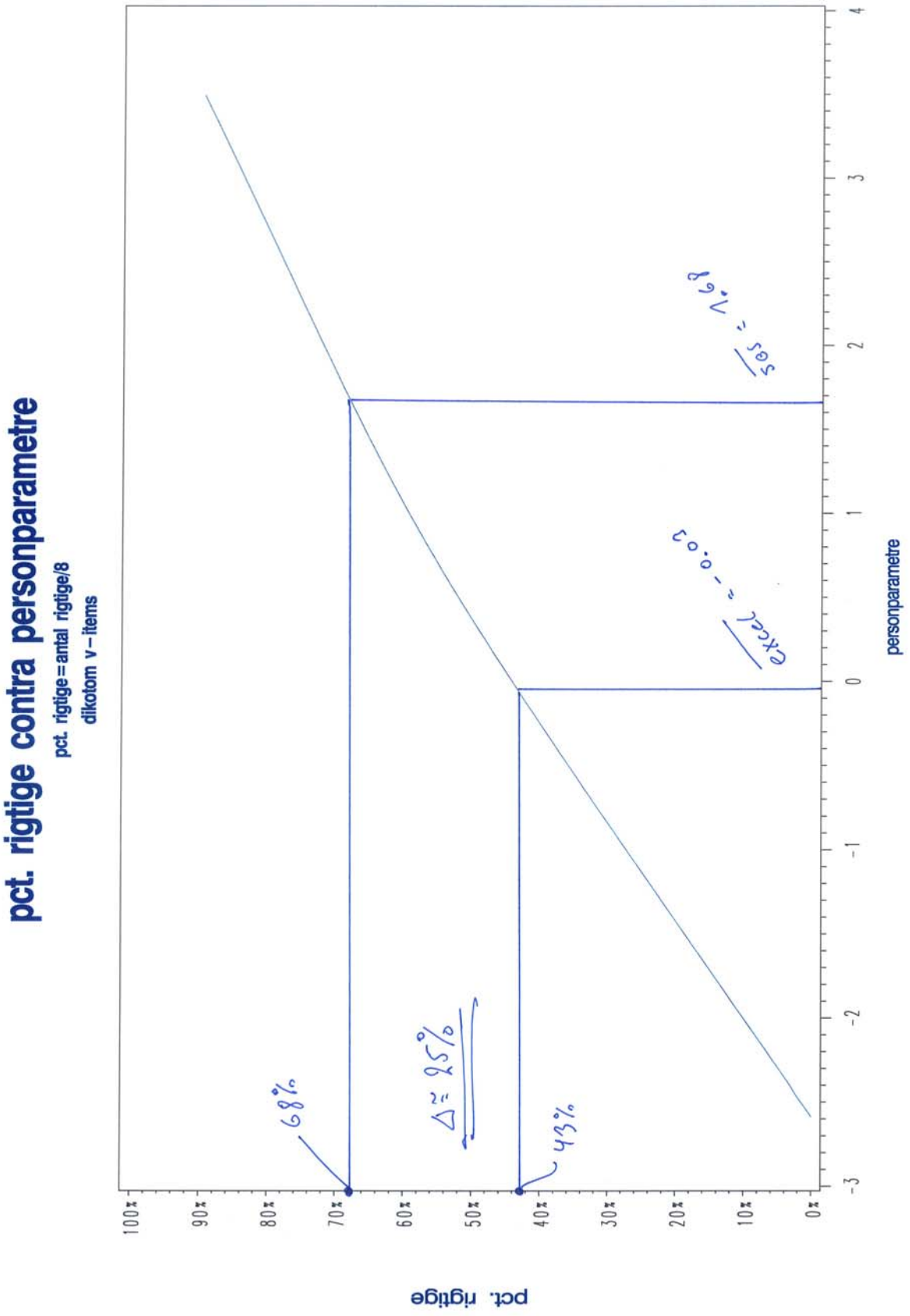
Plot nr 2 angiver kort sammenhængen mellem Rasch-score og pct rigtigt-score og de 25% forskel opstår som 'tilbageregnet' forskel fra den beregnede gennemsnitlige forskel på x-aksen over Rasch scores i de to grupper ---

håber at det er klart, ellers skriv - jeg sider på min øde ø med et meget smart USB modem i siden der tillader mig at køre ret hurtigt -- - og ringe næsten gratis til hele verden. Så hvis det virkeligt kniber, så skriv et telefonnummer (med alle internationale koder mv), så ringer jeg
mange hilsner fra
Peter

Plot 1:


personparametre
dikotom v – items
personparametre (excel – data)
personparametre (sas – data)
4
3
2
1
0
-1
-2
-3
3
2
1
0
-1
-2
-3

Plot 2:


pct. rigtige contra personparametre
pct. rigtige = antal rigtige/8
dikotom v – items
pct. rigtige
personparametre
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
-3
-2
-1
0
1
2
3
4
68%
43%
Δ ≈ 25%
excel = -0.02
sas = -1.98