Dr. Are Turmo is associate professor in the Department of Teacher Education and School Development at the University of Oslo.

Dr. Eyvind Elstad is full professor in the Department of Teacher Education and School Development at the University of Oslo.

ARE TURMO
Department of Teacher Education and School Development, University of Oslo, Norway.
are.turmo@ils.uio.no

EYVIND ELSTAD
Department of Teacher Education and School Development, University of Oslo, Norway.
eyvind.elstad@ils.uio.no.

What factors make science test items especially difficult for students from minority groups?

Abstract

Substantial gaps in science performance between majority and minority students are often found in standardized tests used in primary school. But at the item level, the gaps may vary significantly. The aims of this study are: (1) to identify features of the test items in science (grade 5 and grade 8 students) that can potentially explain group differences; and (2) to analyze what factors make test items especially difficult for minority students. Explanatory variables such as reading load, item difficulty, item writing load, and use of the multiple-choice format are found to be major factors. The analysis reveals no empirical relationships between performance gap and either item subject domain, item test location, or the number of illustrations used in the item. Subtle issues regarding the design of items may influence the size of the performance gap at item level over and above the main explanatory variables. The gap can be reduced significantly by choosing "minority friendly" items.

Introduction

Research in many countries indicate that minority students, as a group, are outperformed by majority students in school subjects (Heath, Rothon & Kilpi, 2008; Heath & Brinbaum, 2007), even though substantial differences exist between different groups of minorities, even in Norway (Støren, 2006). Part of the picture is that a few sub-groups of minority students outperform majority students. For instance, the second-generation youths of Vietnamese ancestry in Norway outperform the majority population of youths (Støren, 2006). The same is also often found regarding minority students from western backgrounds (EU/EEC countries, North America, Australia, and New Zealand). In general, comparisons from Norway show that non-western minority students as a group achieve significantly less in school, compared with majority students (Engen et al., 1996; Bakken, 2003; Bakken 2008; Hvistendahl & Roe, 2004; Fekjær, 2007).

The catalog of suggested measures to close the achievement gaps between majority and minority students does not comprise the construction of test items (Uline & Johnson, 2004). This article analyses what factors make science test items especially difficult for students from minority groups. We define student minority status based on the language predominantly spoken at home. A minority student is defined as a student who predominantly speaks a non-western language at home, while a majority student predominantly speaks Norwegian at home. A justification for this choice of definition is given in the following section.

The school governing body in Oslo uses standardized achievement tests in science to measure how much students learn in science at different grade levels. Data from the Oslo 2007 standardized tests in science have been collected at both grade 5 and grade 8, and also at the individual item level. These science tests are compulsory for all grade 5 and grade 8 students. The 2007 Oslo tests in science for grade 5 and grade 8 show substantial gaps in overall science performance between the majority and the minority students. However, at the item level, the performance gaps between the majority and the minority students vary significantly. At grade 5, the mean differences of each item vary from 0.10 to 0.80, expressed as Cohen's d (Cohen, 1992). Cohen's d is defined as the difference between two means divided by a standard deviation for the data. Corresponding Cohen's d-values are 0.00 and 0.53 at the grade 8 level. In this article, we explore how this variation between individual test items can be explained.

Standardized tests (i.e. measurement instruments that are used to evaluate levels of specific proficiencies, aptitudes or skills, such as knowledge of science) are used in education for several purposes, including charting and measuring changes in student attainment, as well as for diagnostic and aptitude purposes (Shephard, 2001; Black & Wiliam, 1998). When tests are also used as part of the basis for awarding student grades, they can contribute towards assessing which schools the students may attend in the future, the students' self-efficacy (Bandura, 1997), and other such considerations. Norwegian educational policy dictates that "all students are to have equal opportunity to use their abilities and achieve their goals, regardless of social background" (Parliamentary White Paper 16, 2006-2007). This article discusses the question of whether the composition of standardized tests in the natural sciences (e.g. the Oslo tests) can contribute to exaggerating the disparity between students from majority and minority backgrounds.

Standardized tests such as the Oslo tests measure the reading ability of minority-group students in their second language in tandem with their knowledge in the subject in question. After presenting a theoretical evaluation of what elements are significant for how students with Norwegian as a second language (students whose primary language is not Norwegian) appreciate and understand test instructions in comparison with first-language Norwegian students, we will analyze connections between differences in performance at item level via statistical analysis. On this basis, we will identify science test items that are particularly unfavorable towards minority groups and those that are favorable, and we discuss plausible reasons for these differences.

Definitions of majority and minority students

Statistics Norway uses two main groups of countries in its official statistics on immigration (see www.ssb.no):

1) EU/EEC-countries, USA, Canada, Australia and New Zealand.
2) Asia, Africa, Latin-America, Oceania, excluding Australia and New Zealand, and Europe, excluding EU/EEC.

These two groups were until recently labeled "western" and "non-western", respectively. Also in the research literature on minority students, the distinction between "western"/"European" and "non-western"/"non-European" immigrants is frequently used (for instance Heath et al., 2008). Often, the minority status of the students is defined based on the students' and the parents' country of birth; for example, in cases where both the students and the parents are born abroad, the students have been labeled first-generation immigrants. Students born in Norway with parents born abroad have been labeled second-generation immigrants. Other terms used are immigrants and descendants.

In the data from the 2007 Oslo tests, information is included about both the students' and the parents' country of birth. Additionally, the data contains information about the language the students predominantly speak at home. Finally, the data also contains information on how long students born abroad have lived in Norway. Our theoretical framework, which will be elaborated upon in the next section, underlines the importance of language in the process of learning science, as well as understanding science test items. Furthermore, as stated above, the distinction between western and non-western immigrants are frequently used in the research literature on minority students. We have therefore chosen to define minority/majority status based on the language predominantly spoken at home; and, following the Statistics Norway official definitions, we initially distinguish between the following three categories of students:

1) Majority students: Students who predominantly speak the Norwegian language at home.
2) Western minority students: Students who predominantly speak a western language other than Norwegian at home.
3) Non-western minority students: Students who predominantly speak a non-western language at home.

The data contains only very few students in category 2 above. In this article, we therefore focus on the distinction between categories 1 and 3, and they are labeled as majority and minority students, respectively. When asked which language they predominantly speak at home, a few students responded both Norwegian and a non-western language. These few borderline students have been classified as non-western minority students.

At grade 5, the results show that two out of five minority students have lived in Norway for a shorter period than one year. Four out of five students have lived in the country five years or shorter (the modal age of the students is 10 years at the time of the test). At grade 8, one out of three students has lived in Norway for a shorter period than one year. Furthermore, three out of five students have lived in the country shorter than five years (the modal age of the students is 13 years at the time of the test). This information is based on the students' own statements in the questionnaire.

The group of minority students predominantly speaks many different non-western languages at home. Based on the students' responses in the questionnaire, the most frequent language at both grade levels is Urdu, a language mainly spoken in Pakistan and India. At grade 5, the second most frequent language is Somali, a language mainly spoken in Somalia, Djibouti, Ethiopia, Yemen, and Kenya. At grade 8, Arabic is the second most frequent language, followed by Somali. Even though the minority students predominantly speak different languages at home, these students definitely have some distinct features in common. In general, the students in this group are relatively new to the Norwegian culture, and their cultural backgrounds are also rather distant from the Norwegian majority culture along several dimensions. Accordingly, they are relatively new to the main language of instruction in school, as well as the language used in the Oslo science tests. Students for whom the main language of the school and the Oslo science test is their second language have a different basis for understanding test instructions to students for whom the school's language is their native tongue. We elaborate on this further in the next section.

Theoretical framework

Readers of test items integrate the tasks they have read into their existing schemata (Brewer & Treyens, 1981). Good comprehension of a composite test item involves abstracting a macrostructure in the written words and the illustrations related to the test item (Kintsch & van Dijk, 1978). Students for whom the main language of the school is their second language have a different basis for understanding test instructions that may contain words with which some students are more unfamiliar than students for whom the school's language is their native tongue. Some theories point to a lack of knowledge and skills in the majority language (e.g. van de Werfhorst & van Tubergen, 2007). Kulbrandstad (1996) provides an overview of research on the relationship between first-language and second-language reading. The second group has a smaller vocabulary in the relevant language than the first. Furthermore, Kulbrandstad asserts that reading in a second language is slower with less fluency and involves a lower degree of comprehension than reading in a first language.

Distinct forms of knowledge are involved in science learning (Mayer, 1992). We are especially interested in distinguishing between students' semantic knowledge (i.e. familiarity with technical terms necessary for expressing scientific phenomena), their factual knowledge (i.e. their ability to recognize technical issues after working with the learning materials), and their logical knowledge. The semantic knowledge is often conceptual: knowledge that is acquired inductively. Minority students' semantic knowledge is often related to their primary language, which is not Norwegian. Logical knowledge arises from the exercise of the students' thinking. This type of knowledge is distinct from semantic knowledge and factual knowledge. Factual knowledge comprises facts a student can declare to others and exist in the memory in tabular form, something that the student can and is in a position to reproduce. Such tables are regarded as being organized as a network of nodes. The associative connections of a network structure must be regarded as significant for a student's ability to reason.

When we speak of comprehension in connection with the design of test items, we assume that a table contains "slots" that are filled with specific information relevant to the test (Rumelhart & McClelland, 1986; Rumelhart & Norman, 1978). These "slots" are network structures of tables. The skills in science and the thinking expertise are at least heavily dependent on knowledge held in long-term memory (Sweller, 2004). Long-term memory is described as a node-linked network of neurons (Anderson, 2005). If the student has an abundance of tables with clear associative connections between the nodes, the student will be able more quickly and effectively to reason and to solve the problem posed in the test question. The student assimilates, via his/her construction of the information contained in the test question and its illustrations, with the existing knowledge. Working memory is characterized by a limited and immediate work span. The processing of facts and concepts are interlocked in the functioning of working memory. The cognitive load theory (Paas, Renkl & Sweller, 2003) predicts that students with limited network structure have more difficulties with connections of their network structures to the target node, the test item, than students with a richer network structure. The minority students' reading efforts consume the constrained capacity of the working memory. We expect that minority students have a lower degree of comprehension of the test items than if they were reading a similar text in their first language, and also lower performance on constructed response items (items requiring the students to construct their own written response).

The students' competence in the main language has considerable significance for their ability to read and understand test items in the Norwegian language that require reasoning, especially test items that extend beyond recalling factual knowledge. Tasks that require much reading are therefore significant in terms of the performance of minority-group students in these types of tests. Students whose primary language is not Norwegian are sometimes overwhelmed by the quantity of the information elements that need to be processed before comprehension can commence. Our hypothesis 1 is that the number of words (reading load) is not favorable to minority-group students in their second language (alongside their ability in the subject in question). If the test item contains a large amount of text, we can assume that much of the student's capacity to relate to and process new information will be used up. A corollary hypothesis 2 is that the use of visual illustrations may be favorable to minority students. The illustrations could clarify the item content as a supplement to the written text itself.

Writing ability is another aspect of language competence. Some test items may require the students to formulate their own written response (often called constructed response items); while other formats allow the students to choose the correct answer from a predefined list of alternatives, i.e. the traditional multiple choice format. The process of carrying out thinking and ideas by generating written text as a response to a test item requires a great amount of information-processing capacity (Bereiter & Scardamalia, 1987). Learners with low automaticity of spelling, grammar, and vocabulary are expected to perform worse than learners with high automaticity (ceteris paribus). Our hypothesis 3 is that test items that require the students to formulate their own written responses via the item writing load are not favorable to minority students in their second language.

Studies indicate that, on average, minority students tend to be motivated to learn science and more interested in the process than majority students (i.e. Elstad & Turmo, 2007). Minority students also report using different learning strategies in science more intensively than majority students. We expect (hypothesis 4) that the achievement gap between minority and majority students decrease through the test, as a consequence of the minority students' favorable motivational orientation. However, specific motivational patterns related to the test situation may also come into play. Accordingly, it may be predicted that the minority students have a relative advantage as the overall difficulty of the items increase (hypothesis 5). When the cognitive challenges in the test item are also large for the majority students, the negative influences of these students lower motivation, and less strategic behavior might also be expected to be more significant.

Furthermore, it is also of interest to study whether or not the performance gap between minority and majority students varies according to science domains (physics, biology etc.). International comparative studies have shown that the Norwegian students' science performance in general varies significantly by science sub-domains. For example, the Trends in International Mathematics and Science Study conducted in 2007 shows that Norwegian students score relatively lower in the domain of physics than in biology (Martin et. al, 2008). Furthermore, the study also reveals that the relative strengths vary significantly according to the participating country. Based on this, it is of interest to explore whether or not the majority-minority gap also varies by science sub-domain.

Based on the theoretical considerations above, and investigating group differences at item level as dependent variable, the following explanatory variables are explored:

1) Reading load (number of words)
2) Illustrations (number of illustrations)
3) Format (constructed response/multiple choice)
4) Writing load (constructed response format)
5) Difficulty (p-value)
6) Test location (early/late in test)
7) Subject domain (biology, chemistry, physics, geology)

Research questions

As stated in the introduction, the 2007 Oslo tests in science for grade 5 and grade 8 show a substantial gap in overall science performance between the majority and the minority students. At the item level, the performance gaps between the majority and the minority vary significantly. At grade 5, the differences in score between the majority and the minority vary from 0.10 to 0.80, expressed as Cohen's d. Corresponding values are 0.00 and 0.53 at the grade 8 level. Based on the observed differences in Cohen's d, the following research questions will be explored:

1) What factors make science test items especially difficult for students from minority groups?

2) Can the performance gap be reduced significantly by choosing the most minority "friendly" items?

In the empirical investigation, we will explicitly compare the power of the explanatory variables at grade 5 and grade 8, respectively. Based on the variables that explain the most variance in d, we will generate a prediction model for the size of d. In the end, we make a qualitative analysis of the items that deviates the most from this prediction model generated by linear regression analysis. The aim of this study is to identify additional features of the items that can potentially explain the variation in d.

Methods

Instruments

The school governing body in Oslo uses standardized achievement tests in science, measuring how much students have learned in science at different grade levels. These science tests are compulsory for all grade 5 and grade 8 students. The tests are implemented at the beginning of the school years 5 and 8, and are constructed to measure the curriculum competence aims in science by the end of grade 4 and grade 7. The Norwegian science curriculum focuses on the following main areas: "The budding researcher", "Diversity in nature", "Body and health", "The universe", "Phenomena and substances", and "Technology and design". Along with the science tests, a short questionnaire has been administered. This makes it possible to identify student minority status by using the questions presented in the minority definition section earlier in this article.

Sample

Random samples of students have been provided at both grade 5 and grade 8. The numbers of majority and minority students at the two grade levels, according to the applied definitions outlined earlier, are given in Table 1.

Table 1: Number of majority and minority students based on the applied definitions.

Grade	Number of majority students	Number of minority students
5	358	126
8	254	103

Analysis

A secondary analysis of the data from the Oslo tests has been conducted using the statistical software SPSS 16. Mean p-values have been calculated for the groups of minority and majority students according to the applied definitions (N values underlying these means are give in Table 1). The differences between the majority and minority means for each item have been calculated, and the effect sizes expressed as Cohen's d have been estimated. Cohen's d is defined as follows:

$$\text{Cohen's d} = M_1 - M_2 / \sigma_{\text{pooled}}$$

$$\text{where } \sigma_{\text{pooled}} = \sqrt{[(\sigma_1^2 + \sigma_2^2) / 2]}$$

M and σ denote the means and standard deviations of two groups (which are compared).

Accordingly, the number of items in the two tests constitutes the N values in the quantitative analyses that are presented. The d values are analyzed in relation to potential explanatory variables, using correlation and regression analysis. The potential explanatory variables are operationalised as follows: reading load is the number of words in the item; item difficulty is the overall p-values; test location is the successive numbering of items throughout the booklets; number of illustrations is a count of the illustrations included in the item; and item writing load is defined as a dichotomous variable differentiating between constructed response items (numerical value=1) and other item formats (numerical value=0). Correspondingly, we use a dichotomous variable differentiating between multiple choice items (numerical value=1) and all other item formats (numerical value=0).

Empirical results

Variations in d at grade 5 and grade 8

Table 2 presents descriptive data for Cohen's d at the two grade levels. As shown in the table, the test at grade 5 contained 40 items, while the grade 8 test contained 51 items in total. The table shows a substantial variation in d at both grade levels. However, the range of d values is clearly larger at grade 5. The table also shows that the mean performance gap between minority and majority students is larger at grade 5 than at grade 8. This finding is as expected, taking the grade 5 minority students' shorter average living time in Norway into account (as analyzed earlier in the article).

Table 2: Descriptive data for d in the two tests at grade 5 and grade 8.

	Number of items	Min. d	Max. d	Mean d	SD d
Grade 5	40	0.10	0.80	0.45	0.18
Grade 8	51	0.00	0.53	0.26	0.13

Potential explanatory variables

Table 3 shows the relationship between Cohen's d and the potential explanatory variables at item level (for operational definitions, see Methods). Regression analysis shows that the variables in Table 3 can explain 25 percent of the variance in d at grade 5, while 15 percent of the variance at grade 8 can be explained.

Table 3 shows some striking similarities and differences between the results for the two tests at grade 5 and grade 8. Firstly, the table shows only very weak relationships for the item test location and number of illustrations at both grades. In other words, these two variables do not seem to have any important explanatory power regarding the variance in d. However, the item writing load can explain 4-5 percent of the variance in d at both grade levels. The values of d tend to be higher for the constructed response items where the students are required to write or construct their responses. This result indicates that constructing responses is relatively more challenging for the minority students. Furthermore, Table 3 also shows that the use of the multiple-choice format can predict the variance in d, especially at the grade 5 level (correlation -0.39), but also at grade 8 (correlation -0.25). These correlations mean that the multiple-choice items tend to have smaller d values than the group of all other item formats. Table 3 also shows that as the items become more easy, the performance gap between the majority students and the minority students increases. However, this effect is clearly stronger at grade 8.

Regarding reading load, the table shows a complex picture. We find a clear positive relationship at grade 5 level, while there is a weaker negative relationship at grade 8. In other words, the more reading an item requires at grade 5, the larger the performance gap between the minority and the majority students. However, at grade 8 there is a tendency that the gap is smaller when the item requires more reading. The first of these findings is in line with the hypothesis, while the second is more challenging to interpret. We therefore study this in more detail in the following portion of this paper, especially with regard to co-variations between reading load and other explanatory variables in the two tests.

Table 3: Empirical relationships between Cohen's d and potential explanatory variables at item level. Grade 5 and grade 8. Number of items: N=40, N=51.

Explanatory variables	Grade 5	Grade 8
Reading load	0.27	-0.15
Item difficulty (p-value)	0.11	0.30
Item test location	-0.01	-0.07
Number of illustrations	-0.04	0.03
Item writing load	0.21	0.22
Multiple-choice format	-0.39	-0.25

Exploring the differential effect of reading load

To better understand the differential effects of reading load at the two grade levels, we studied the co-variation between reading load and other explanatory variables at item level. The interaction between reading load and item format in the two tests turns out to be important. Table 4 shows the mean d for the two main categories of item formats (multiple choice and constructed response) in the two tests, as well as the average reading load. The table shows that the constructed response items have larger d on average than the multiple-choice items, and the difference is equally large in the two tests. However, the table reveals an important difference regarding reading load. At grade 5, the two item format groups have approximately the same mean reading load, while there is a large difference at grade 8, where the multiple-choice items require significantly more reading than the constructed response items.

Table 4: Mean reading load and d for the two main categories of item formats in the two tests.

	Multiple-choice items	**Constructed-response items**
Grade 5	31 words, d=0.38	28 words, d=0.51
Grade 8	44 words, d=0.23	25 words, d=0.30

Based on the results in Table 4, it is relevant to study the correlations between reading load and d within the groups of different item formats. These results are displayed in Table 5. The results in Table 5 show the empirical relationships are strikingly parallel at the two grade levels. Among the multiple-choice items, the tendency is that the more reading the item requires, the larger the performance gap between the minority and the majority. Furthermore, among the constructed-response items there are very weak negative relationships between reading load and d. However, we do not regard these very weak relationships as substantially important.

Table 5: Correlations between d and reading load within item formats at the two grade levels.

	Grade 5	**Grade 8**
Multiple-choice format	0.19	0.18
Constructed response format	-0.09	-0.04

Variation in d between subject domain

The items in the grade 8 and grade 5 tests can also be classified according to subject domains. A classification shows that the majority of items in the grade 5 test are within the biology domain. The mean performance gap in this category corresponds to the mean gap found for the geology items. Furthermore, the physics items and the chemistry items on average show slightly smaller performance gaps. However, the results show that the variation in mean d between subject domains can in large part be explained by the variation in explanatory variables as displayed in Table 3. For example, the items in the category environmental science, showing the largest mean performance gap, have by far the highest reading load, and also the lowest percentage of multiple-choice items. Furthermore, these items are also the most difficult on average. On the other hand, the items in physics, showing a relatively small performance gap, have the lowest reading load in combination with an above average percentage of multiple choice items. At grade 8, most of the items can be classified within the domains of physics and biology. The mean d for these two categories of items is rather similar. In conclusion, based on the relative small number of items in the different subject domain categories, as well as the co-variation between mean d and other explanatory variables, it cannot be concluded that there are systematic differences in d by content subject domain based on the present science tests for grade 5 and grade 8.

Prediction models of d

Based on the coefficients from a linear regression model, a predictor construct of d has been calculated at both grade 5 and grade 8 level as the linear combination of the variables multiple choice format, reading load, item writing load and item difficulty, as defined in Table 3. As shown in Table 6, these models explain 22 percent of the variance in d at grade 5 and 15 percent at grade 8. However, as shown from the table, the stability of these models should not be over-emphasized.

Table 6: Enter regression models testing the relationships between d and four explaining variables

Dependent variable	Explaining variable	B	SE	Beta	t	p
d; grade 5	Multiple-choice	-0.13	0.08	-0.35	-1.57	0.12
	Reading load	0.00	0.00	0.24	1.45	0.16
	Writing load	0.00	0.09	0.00	0.01	0.99
	Difficulty	0.18	0.19	0.15	0.97	0.34
	Constant	0.31	0.17		1.84	0.07
d; grade 8	Multiple-choice	-0.03	0.05	-0.11	-0.54	0.59
	Reading load	0.00	0.00	0.06	0.37	0.72
	Writing load	0.04	0.06	0.16	0.71	0.48
	Difficulty	0.21	0.11	0.31	2.01	0.05
	Constant	0.12	0.11		1.14	0.26

Notes: *R square=0.22 at grade 5, R square=0.15 at grade 8.*

at grade 5 and grade 8.

"Minority friendly" sub-tests

Based on the two strongest predictors of d at grade 5 and grade 8, we have constructed sub-tests that are especially favorable for the minority students. At grade 5, we have chosen items with below average reading load and multiple-choice formats. This sub-test then consists of 14 items. The results show that the average value for d is 0.36 in this sub-test, compared to 0.45 for the test as a whole (see Table 2). At grade 8, we made a sub-test consisting of the items with below average difficulty and the multiple-choice format (13 items). The mean value of d is here 0.18, while the value for the test as a whole is 0.26 (see Table 2). This illustrates that the performance difference between the majority students and the minority students can be reduced significantly by choosing "minority friendly" items. However, there is still a significant performance gap between the two groups.

Items with the largest deviations from the prediction models

Table 7 shows the items with the largest deviations in d from the predictor construct values (i.e. values calculated as linear combinations based on the unstandardized beta coefficients in Table 6). Negative deviation values here mean that the item shows a smaller minority-majority performance gap than expected based on the prediction model. Accordingly, positive values mean that the item is relatively difficult for the minority students.

Item A in Table 7 turns out to be significantly easier than predicted for the minority students relative to the majority. In this item, the students are given pictures of sugar in two different forms: pieces of sugar (20 grams) and loose sugar (20 grams). The students then state which form of sugar they believe will dissolve the fastest in water, and then explain their answer. According to the applied coding guide, one score point is to be given to responses stating that loose sugar is already in smaller pieces, or that the pieces of sugar are more compact. Students who merely

Table 7: Items that are relatively easier or more difficult than expected for the minority students.

	d	Reading load	Item format	Difficulty	Deviation
Grade 5					
Item A	0.10	41 words	Constructed response	0.59	-0.44
Item B	0.72	14 words	Constructed response	0.62	0.25
Grade 8					
Item C	0.00	21 words	Multiple choice	0.99	-0.30
Item D	0.44	65 words	Multiple choice	0.45	0.26

state loose sugar without any explanations are given zero points. This item is obviously closely related to everyday experiences, and the explanation required also requires common sense. It may be argued that this item to a very limited extent requires knowledge typically learned in school science. Furthermore, the item does not require much reasoning. Rather, the students can receive one score point based on concrete everyday experiences and a rather common sense explanation. Our interpretation is that the "minority friendly" aspect of this item may be largely explained by these features of the item.

Item B is also a constructed-response item that reads as follows: the human body is covered with skin. Write down one task (in Norwegian: oppgave) the skin has. One score point is given to responses stating that the skin protects the body, or prevents the body from dehydrating, or that it regulates the temperature (by sweating). This item is far more difficult for the minority students than predicted by the model. The specific use of language in this item requires special attention. Here, the skin is treated as a "subject", which can be responsible for certain tasks. Our hypothesis is that this abstract use of language (Kulbrandstad, 1996) is a main reason why this item is far more difficult for the minority students than predicted.

Item C is a very easy multiple-choice item, both for the majority and the minority students. It asks which of these four organisms you need a microscope to see: worms, bacteria, spiders, or butterflies. This item uses only terms that are either relatively frequent in everyday language, and/or typically explained in science textbooks. Furthermore, the item is closely related to everyday experience; it is rather unlikely that students would have had no concrete experiences with worms, spiders, and butterflies.

Item D is another multiple-choice item that turns out to be far more difficult for the minority students than predicted. The item stem states that one often builds high fences along motor roads in highly populated areas. The students are then asked to choose the most likely (in Norwegian: mest sannsynlige) reason why the fences are built among four given alternatives, the correct alternative being "to reduce sound from the traffic". The expression "most likely" is obviously low-frequent in the everyday language of 13-year-olds. Neither is it a concept that is part of the science curriculum, and thus not explicitly taught in science classes. Our hypothesis is that the use of this kind of low-frequent and formal language (Kulbrandstad, 1996) is a major reason why the minority students score lower than predicted on this item.

Discussion and conclusion

Our analysis shows that a predictions model based on four explanatory variables; reading load, item difficulty, item writing load, and use of multiple-choice format, can explain 22 percent of the variance in Cohen's d at grade 5. The explanatory power of this model is however significantly weaker at grade 8 (explains 15 percent of the variance). It is also interesting that we do not find any empirical relationships between d and either item subject domain, item test location or the number of illustrations used in the item.

As expected, we find lower Cohen's d for multiple-choice items than for constructed-response items. Multiple-choice items are thus to be regarded as minority-favorable. Initially, it is surprising to find a negative connection between Cohen's d and reading load for grade 8 students. However, this finding is explained by interactions between item format and reading load. The multiple-choice items in the grade 8 test have on average larger reading load than the constructed response items.

Item difficulty can predict variations in d at both grade levels. Our initial hypothesis was that the minority students may have a relative advantage as the items become more difficult, based on their reported higher motivation and increased use of learning strategies in science. The empirical results show positive correlations between p-value and d (the higher p-value, the easier item).

The analysis indicate that subtle issues regarding the use of language in the items may influence the size of Cohen's d over and above the four explanatory variables in the prediction model. The use of abstract and low frequent language (Kulbrandstad, 1996) seems to influence the size of d significantly, over and above the quantity of reading (number of words) in itself. Thus, we need to study more closely the choice of formulations made by item writers. Furthermore, the results also indicate that minority-group students score relatively better when the items can be answered correctly based on common everyday experiences.

In general, the results show that the item design may significantly influence the size of the majority-minority gap in science. However, also choosing the most "minority friendly" items leaves us with a substantial performance gap. When discussing item design, one must also consider carefully which features are construct relevant, as opposed to construct irrelevant. The item format (e.g. multiple choice, constructed response etc.) is an example of an obviously construct irrelevant aspect. In contrast, it may be argued that language competency (reading and writing) is an integrated part of the construct of scientific literacy (e.g. Fang, 2005; Fang, 2006; Norris & Phillips, 2003).

The conclusion is that if society is concerned that "all students are to have equal opportunity to use their abilities and achieve their goals, regardless of social background" (Parliamentary White Paper 16, 2006-2007), the design of test items influences the differences in performance between minority and majority students. We are of the opinion that this question should be awarded greater attention in the debate about how test items are written, simply in order to avoid questions that are particularly unfavorable to minority groups. Minority students will benefit from test conditions that reduce the cognitive load to more manageable levels: the task challenges should be at the correct grain size (van Merriënboer, 1997). What is important is that the test items afford enough scaffolds to the missing interactions between the minority students' prior knowledge and their central executive (how test elements are coordinated with their prior knowledge) when dealing with information in the test item. We suggest the study of the characteristics of tests for future research.

Note

The Oslo 2007 science tests were developed and implemented by Rolf V. Olsen, Anubha Rohatgi, Sonja M. Mork and Svein Lie, University of Oslo. We thank them for making the data available for analysis. We also want to thank two anonymous referees for their useful comments.

References

Anderson, J. R. (2005). *Cognitive Psychology and Its Implications: Sixth Edition*. New York: Worth Publishing.

Bakken, A. (2003). *Minoritetsspråklig ungdom i skolen. Reproduksjon av ulikhet eller sosial mobilitet?* Rapport nr.15. Oslo: Norsk institutt for forskning om oppvekst, velferd og aldring.

Bakken, A. (2008) Er kjønnsforskjeller i skoleprestasjoner avhengig av klassebakgrunn og minoritetsstatus? *Tidsskrift for ungdomsforskning*, 8 (1):85-93

Bandura, A. (1997). *Self-efficacy: The exercise of control.* New York: W. H. Freeman.

Bereiter, C. & Scardamalia, M. (1987). *The psychology of written composition.* Hillsdale, New Jersey: Erlbaum.

Black, P.J. & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, 5, 7-77.

Brewer, W. F., & Treyens, J. C. (1981). Role of schemata in memory for places. *Cognitive Psychology*, 13 (2), 207-230.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155-159.

Elstad, E. og Turmo, A.(2007) Strategibruk, motivasjon og interesse for naturfag: forskjeller mellom minoritetselever og majoritetselever. *Tidsskrift for ungdomsforskning*, 7 (2), 23-44.

Engen, T.O., Kulbrandstad, L.A., & Sand, S. (1996). *Til keiseren hva keiserens er? Om minoritetselevenes utdanningsstrategier og skoleprestasjoner.* Sluttrapport fra “Minoritetselevers skoleprestasjoner”. Vallset: Oplandske Bokforlag.

Fang, Z. (2005). Scientific Literacy: A Systemic Functional Linguistic Perspective. *Science Education*, 89, 335-347.

Fang, Z. (2006). The Language Demands of Science Reading in Middle School. *International Journal of Science Education*, 28, 491-520.

Fekjær, S.N. (2007). New differences, old explanations: Can educational differences between ethnic groups in Norway be explained by social background? *Ethnicities*, 7, 367-389

Heath, A.F. & Y. Brinbaum (2007). Guest editorial: Explaining ethnic inequalities in educational attainment. *Ethnicities*, 7, 291-304.

Heath, A., Rothon, C. & Kilpi, E. (2008). The second generation in the Western Europe: education, unemployment and occupational attainment. *Annual Review of Sociology*, 34, 211-235.

Hvistendahl, R.E. & Roe, A. (2004). The Literacy Achievement of Norwegian Minority Students, *Scandinavian Journal of Educational Research*, 48 (3), 307-324.

Kintsch, W. & van Dijk, T.A. (1978). Toward a model of text comprehension and reproduction. *Psychological Review*, 85, 363-394.

Kulbrandstad, L.A. (1996). *Lesing på et andrespråk. En studie av fire innvandrerungdommers lesing av læreboktekster på norsk*. Oslo: HF-fak., Universitetet i Oslo

Martin, M.O., Mullis, I.V.S., & Foy, P. (with Olson, J.F., Erberber, E., Preuschoff, C., & Galia, J.). (2008). *TIMSS 2007 International Science Report: Findings from IEA’s Trends in International Mathematics and Science Study at the Fourth and Eighth Grades.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mayer, R.E. (1992). *Thinking, problem solving, cognition.* New York: Freeman.

Norris, S. P. & Phillips, L. M. (2003). How Literacy in Its Fundamental Sense Is Central to Scientific Literacy. *Science Education*, 87, 224-240.

Paas, F., Renkl, A. & Sweller, J. (2003). Cognitive load theory and the instructional design: Recent developments. *Educational Psychologist*, 38, 1-4.

Parliamentary White paper no. 16 (2006-2007) … *og ingen stod igjen*. Oslo: Ministry of Education

Rumelhart, D.E. & McClelland, J.L. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition.* Cambridge: MIT Press.

Rumelhart, D.E. & Norman, D.A. (1978). Accretion, tuning and restructuring: Three modes of learning. In J.W. Cotton & R. Kaltsky (Eds.), *Semantic factors in cognition* (pp. 37-53). Hillsdale: Erlbaum.

Shepard, L. (2001). The role of classroom assessment in teaching and learning. In V. Richardson (Ed.), *Handbook of research on teaching* (pp.1066-1101). Washington DC: AERA.

Støren, L.A. (2006). Nasjonalitetsforskjeller i karakter i videregående opplæring. *Tidsskrift for ungdomsforskning* 6(2), 59-86.

Sweller, J. (2004). Instructional Design Consequences of an Analogy between Evolution by Natural Selection and Human Cognitive Architecture. *Instructional Science*, 32, 9-31.

Uline, C. L. & Johnson, J. F. (2004). Closing the achievement gap: What will it take? *Theory into practice*, 44, 1-3.

Van De Werfhorst, H.G. & Van Tubergen, V. (2007). Ethnicity, schooling, and merit in the Netherlands. *Ethnicities,* 7, 416-444.

Van Merriënboer, J. J. G. (1997). *Training Complex Cognitive Skills: A Four-Component Instructional Design Model for technical Training.* Englewood Cliffs, NJ: Educational Technology Publications.