Avoiding Unintentionally Difficult Test Items
for Immigrant Minority Students

Differentiation without DIF

Universität Duisburg-Essen
Avoiding Unintentionally Difficult Test Items
for Immigrant Minority Students
Differentiation without DIF
Avoiding Unintentionally Difficult Test Items for Immigrant Minority Students Differentiation without DIF

1. Introduction

Previous research into the Dutch ‘Final Test of Primary Education’ (an annually produced achievement test by the Netherlands National Institute of Test Development: CITO) has shown that this test contains a number of unintentionally and therefore unwanted, difficult test items for immigrant minority students whose parents speak Turkish or Arabic/Berber at home.

In order to find out the sources of these extra-difficult test items, known as Differential Item Functioning (DIF), CITO and the Babylon Centre at Tilburg University have been working together since 1986. Knowledge of sources of DIF is necessary for test constructors to avoid DIF in future tests as much as possible.

To identify DIF items in the ‘Final Test of Primary Education 1997’ for immigrant minority students two statistical procedures, the Mantel-Haenszel (MH) and the One Parameter Logistic Model (OPLM), have been used. Subsequently, five experiments have been conducted to detect possible sources of DIF. The results of these experiments revealed a number of hypotheses concerning possible (linguistic and cultural) sources of DIF that have been used in order to manipulate original DIF items into intentionally DIF-free items. These manipulated items were statistically analyzed with MH and OPLM again. Finally, a content analysis of the results of these statistical procedures took place. In this article the total research project will be presented and discussed. The main emphasis, however, will be on the discussion of possible sources of DIF and on formulating recommendations for avoiding DIF in future tests as much as possible.

2. Differential Item Functioning versus Differential Item Validity

In research into the sources of Differential Item Functioning (DIF), it is important to make a clear distinction between Differential Item Functioning, Differential Item Validity (also known as item bias) and difficulty level. DIF occurs when students of different subgroups, in this research native-Dutch students and students whose parents speak Turkish or Arabic/Berber at home, who are equally skilled in the construct to be measured, and therefore have a comparable achievement level in a certain subject matter area, do not have an equal chance on answering an item correctly. When the source of DIF is irrelevant to the construct to be measured -this is when answering the item correctly requires additional skills and knowledge other than those being measured- it can be concluded that there is a case of Differential Item Validity (DIV) or item bias. The implication of this is that not all
DIF items have to be DIV items as well.

It should be noted that DIF and DIV are not the same as difficulty level. The level of difficulty indicates the percentage of students that have answered an item correctly. Of course, it is possible here to make a distinction between different subgroups, but these subgroups do not necessarily have a comparable achievement level in a certain subject matter area. Because tests, and especially tests of educational progress, are meant to differentiate between high- and low-achieving students, one must be confident that scores of all student subgroups can be interpreted in a similar way. When differentiation occurs on basis of additional skills and knowledge that not all students have a thorough command of, students with a comparable level in a certain subject matter area do not have an equal chance to obtain the test result that is in line with their achievement level in the subject matter area being tested.

3. Research into Causes of Differential Item Functioning

3.1 Introduction

In item bias research, usually four steps are needed. First, statistical procedures have to be used to detect DIF items. Secondly, experiments have to be conducted in order to identify possible sources of DIF. For establishing which element(s) in a DIF item may be the cause of DIF, it is recommended that judgements about the possible sources of DIF are obtained from various relevant groups of informants in different ways, since it is very difficult to identify the element in the item that causes DIF with 100% certainty (Uiterwijk & Vallen, 2005). Thirdly, the possible sources of DIF have to be confirmed using the same statistical procedures as those that were used to identify the DIF items. Finally, the decision has to be made whether the possible source of DIF is relevant to the construct the test claims to measure. In other words, the discussion has to be made whether the concerning DIF item is a DIV item too.

As mentioned before, in this research the Dutch ‘Final Test of Primary Education’ is investigated with a focus on the identification of possible sources of DIF. This test of educational progress is produced by CITO and is administered at the end of primary school, when the students are about 12 years of age. A new form of this test is produced every year. An important objective of the test is to give a score-based indication as to which of the three levels of secondary education is most suitable for each student (Uiterwijk & Vallen, 2005). The majority of primary schools directors in the Netherlands (about 85 percent) offer the Final Test of Primary Education to their students. The Final Test of 1997 that is evaluated in this research consisted of 240 four-option multiple-choice questions, evenly distributed over the four subtests of Language, Mathematics, Information Processing and World Orientation. The latter subtest was the only not compulsory subtest. In the four areas the following academic skills are being measured (Cito, 1986):
- Language: spelling, writing, vocabulary
- Mathematics: understanding figures, mental arithmetic, percentages, fractions, calculations for measures, weights, money and periods of time
- Information Processing: reading comprehension, dealing with textbook material and sources of information such as the telephone directory; reading and understanding diagrams, schedules, tables and graphs
- World Orientation: applying knowledge acquired in the fields of geography, history, biology, physics, civics and religious movements.

In 1997 over 100,000 students of 4302 schools conducted World Orientation. In all over 122,000 students of 5246 schools participated in Final Test of Primary Education 1997. To be able to classify the students that participated in the Final Test of 1997 in ethnic subgroups and to subsequently investigate the differences between the test results of the different subgroups of students, a questionnaire is developed in 1997 and administered in 1998 with 5894 first-year secondary school pupils of which 5026 students participated in the Final Test of Primary Education 1997 (Derks & Vallen, 1999).

3.2 Detecting DIF Items

Empirical evidence has shown that the various DIF detection procedures do not necessarily lead to the same results (Camilli & Smith, 1990; Uiterwijk, 1994; Uiterwijk & Vallen, 2005). Therefore, in this research two statistical procedures have been applied in order to detect DIF items: the Mantel-Haenszel test (MH) and the One Parameter Logistic Model (OPLM). Item Response Theories like OPLM, claim that DIF items are those where the item characteristic curves of two subgroups do not coincide. In this research two random subgroups of native-Dutch students are compared. For these two subgroups the items did not show any DIF. In other words, the test items did fit a (DIF-free) unidimensional scale. The assumption is that such a DIF-free scale for native-Dutch students will not show any DIF either when Dutch students and Turkish and Moroccan students are compared. However, DIF will be detected in the items which exceed a given significance level in this comparison (Uiterwijk & Vallen, 2005).

The second statistical procedure that has been used, MH, is based on the classical test theory. This procedure claims that the test score is an adequate estimation of the skills that the test aims to measure. MH verifies the hypotheses that the difficulty level of an item is the same for students with a comparable ability. If the z-score of an item is > 1.95 or < -1.95, this item has DIF to the disadvantage and advantage of Turkish and Moroccan students respectively.

The results of the DIF analysis of the 240 items of the Final Test of Primary Education of 1997 show that 32 items contain DIF. From these DIF items 21 have DIF to the disadvantage of Turkish and Moroccan students and 11 have DIF to their advantage. Table 1 shows the number of DIF items that are detected in the statistical procedures on each subtest.
of the Final Test of Primary Education (i.e. Language, Mathematics, Information Processing and World Orientation).

Table 1. *Number of items of the Final Test of Primary Education of 1997 with DIF for Turkish and Moroccan students*

<table>
<thead>
<tr>
<th></th>
<th>MH</th>
<th>OPLM</th>
<th>MH + OPLM</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language disadvantage</td>
<td>5</td>
<td>-</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>Language advantage</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Mathematics disadvantage</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>Mathematics advantage</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>Information Processing</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>4</td>
</tr>
<tr>
<td>Information Processing</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>World Orientation disadvantage</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>World Orientation advantage</td>
<td>-</td>
<td>4</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>

### 3.3 Identification of Possible Sources of DIF

As mentioned before, it is important to obtain judgements of various relevant informants in different ways in order to identify the possible sources of DIF. Therefore, in this research five experiments have been conducted using three different research methods: (1) a Content Response Code method, (2) a Student Think-Aloud Method and (3) a Rewrite-Procedure, in which Dutch, Turkish and Moroccan students, teachers, teachers in training, test constructors and linguistics participated. In Table 2, an overview of the different experiments, research methods, and the number of informants is presented.

Table 2. *Conducted experiments*

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Research method</th>
<th>Informants</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment 1</td>
<td>Content Response Code</td>
<td>Students</td>
<td>16</td>
</tr>
<tr>
<td>Experiment 2</td>
<td>Content Response Code</td>
<td>Teachers in training</td>
<td>19</td>
</tr>
<tr>
<td>Experiment 3</td>
<td>Student Think-Aloud Method</td>
<td>Students</td>
<td>16</td>
</tr>
<tr>
<td>Experiment 4</td>
<td>Rewrite-Procedure</td>
<td>Teachers in training</td>
<td>18</td>
</tr>
<tr>
<td>Experiment 5</td>
<td>Combined Content Response Code</td>
<td>Teachers, Test constructors and</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>and Rewrite-Procedure</td>
<td>Linguistics experts</td>
<td>5</td>
</tr>
</tbody>
</table>

In the experiments 1 and 2 the Content Response Code is used. To apply this research method four steps are needed. First, the text the informants have to evaluate has to be divided by the researcher into small marked text units, called content units. These content units are not visible in the actual text that has to be evaluated. Subsequently, the informant
marks his/her positive and negative evaluations in the text margins using the plus and minus signs (+, -). The instruction the informants receive is rather broad and open in order to obtain a wide range of all sorts of problems. When the informant has read and marked the text, he is asked to explain the assigned plusses and minuses. Finally, the text is revised on the basis of the qualitatively incorporated data of the interviews (Vroom, 1985; De Jong, 1991; De Jong & Schellen, 1995). This research method is used to point out possible problems and their locations in both the text that has been evaluated and the multiple-choice questions to the text. It should be noted that the Content Response Code does not solve problems, but only points them out. The assigned minuses only indicate that, and where, there might be a problem. In addition, the informant can be asked to give suggestions to improve the text or the multiple-choice questions asked. Experiment 1 is conducted by 16 students, of which 6 were Turkish, 4 Moroccan and 6 native-Dutch students. All students were in primary education grade eight (11-12 year-old).

In experiment 5 the Content Response Code has been combined with the Rewrite-Procedure. Twenty-six teachers in primary grade eight, five test constructors of CITO and two linguists participated as experts in this experiment. They all were asked to evaluate all DIF items and to rewrite those items to which they had assigned one or more minuses. The instruction was to rewrite the items in such a way that they would be appropriate for the target group, i.e. all students in primary grade eight regardless of the language of their parents at home. The same Rewrite-Procedure is conducted in experiment 4, in which 18 students of a teacher training college participated.

The Student Think Aloud Method has been used in experiment 3. In this research procedure informants are asked to perform a task – i.e. a subtest of the Final Test of Primary Education - while thinking aloud. This method is used to gain an insight into the cognitive processes that influence the completion of an assigned task (Elshout, 1976; Ericsson & Simon, 1984). When using this method, it is very important to explicitly point out to the informants that they have to perform the task in the same way as they usually would do in order to prevent the cognitive processes to be influenced by the verbalization of the task. In this research the Student Think Aloud Method has been used in order to investigate whether and to which extent there are different answering strategies in students with different home languages in view of the questions asked in the subtests of the Final Test of Primary Education. Secondly, this procedure is used to find out whether possible sources of DIF are the reason for these differences. The Student Think Aloud Method is conducted by 16 students in primary grade eight (five Turkish, six Moroccan and five native-Dutch students).

3.4 First Confirmation of Possible Sources of DIF

On the basis of on the results of the five experiments, a number of hypotheses were formulated concerning the possible sources of DIF. Subsequently, the DIF items were manipulated in such a way that they were expected to be DIF-free because of the
manipulation. For the statistical confirmation whether the manipulations have had resulted into the intended effect, all items were manipulated as minimal as possible. These changed items have been divided over two tests, which have been presented to two groups of students in primary grade eight. Both groups consisted of native-Dutch, Turkish and Moroccan students. To operationalize the ethnicity of the students, all students were asked what language their parents speak at home. In addition to the manipulated items, also items of the Final Tests of Primary Education of 2002 and 2003 were included under the assumption that these items might also be DIF items on the basis of the formulated hypotheses on possible DIF sources. All items were, together with 80 additional DIF-free control-items, represented in the two tests.

Students of 142 schools have completed one of the test versions. These schools have been randomly divided into two subgroups; Test 1 has been conducted in 69 schools, test 2 in 73 schools. The tests have been completed on two consecutive days. The test structure of the two tests was as similar as possible as the structure of the Final Test of Primary Education. This means that the four subtests, i.e. Language, Mathematics, Information Processing and World Orientation, have been divided into multiple subtests themselves, which were spread over the tests. Table 3 presents an overview of the number of target students participating in the two tests.

Table 3. *Number of students in each test*

<table>
<thead>
<tr>
<th></th>
<th>test 1</th>
<th>test 2</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Native-Dutch students</td>
<td>640</td>
<td>588</td>
<td>1228</td>
</tr>
<tr>
<td>Turkish students</td>
<td>352</td>
<td>325</td>
<td>677</td>
</tr>
<tr>
<td>Moroccan students</td>
<td>363</td>
<td>297</td>
<td>660</td>
</tr>
<tr>
<td>Total</td>
<td>1355</td>
<td>1210</td>
<td>2565</td>
</tr>
</tbody>
</table>

The results of the two tests have been statistically analyzed by using the Mantel-Haenszel test (MH) and the One Parameter Logistic Model (OPLM) again. The analyses showed that most of the manipulated items did not show any DIF. This may lead to the conclusion that most of the assumptions about possible sources of DIF were relevant in fact. In the following part, only three hypotheses will be discussed:

1. The use of misleading illustrations can be a cause of DIF to the disadvantage of Turkish and Moroccan students.
2. The use of metaphorical language can be a cause of DIF to the disadvantage of Turkish and Moroccan students.
3. Asking questions concerning subjects related to religion can be the cause of DIF to the advantage of Turkish and Moroccan students.
Misleading Illustrations

The analyses of the results obtained from the five experiments mentioned before, revealed that Turkish and Moroccan students pay far more attention to illustrations in test items than native-Dutch students do. The main reason for Turkish and Moroccan students to opt for a specific answer out of four possibilities is often based on such an illustration. This is not only the case when the illustration actually is part of the multiple-choice question asked, but also when an illustration is added to a text for other reasons (illustrative, esthetical, etc) (Van Schilt-Mol, Vallen & Uiterwijk, 2005). The results of the experiments also show that Turkish and Moroccan students are more critical with respect to the illustrations used and that they notice discontinuities rather soon. In some cases this discontinuities leads to wrong answers in terms of the test as such, but in fact to correct answers when the students’ arguments are taken into consideration. An example of one of these items stems from the subtest Information Processing. This original, non-manipulated item has DIF to the disadvantage of Turkish and Moroccan students ($z = 1.95$). The question the students have to answer is the following question: *From which side the castle is drawn here?*

Final Test of Primary Education 1997: Information Processing Subtest 2, question 10

As can be derived from the legend [legenda], two different symbols are used for the ‘tower’ [toren] as referent. In addition, the symbol used for ‘well’ [put] shows much resemblance
with one of the symbols used for ‘tower’ [toren]. When a student (incorrectly) links the symbol for ‘well’ to the referent ‘tower’ [toren] (as is done in the Student Think Aloud fragment below), the right answer should be possibility D. It is remarkable that most of the Turkish and Moroccan students opt for this particular wrong answer possibility (28% and 31% respectively). Only 16% of the native-Dutch students opt this possibility.

Fragment of Student Think Aloud Method: Mohammed

Mohammed: Gornsteeren Castle. There is a legend. A stripe is a wall, a circle is a tower, a stripe with a … a kind of door, is a gate. A block with a cross in it is a tower. A circle with a kind of filling is a well.
Researcher: Yes.
Mohammed: From which... from which side the castle is drawn here?
(...)
Mohammed: C. Here there is a tower in every corner again. And then a gate. And to the left there is a tower and that is not right too.
Researcher: No.
Mohammed: And then you get to D again. At every corner there is a tower again, in the middle there is a gate. And to the right there is the tower, so that is the right answer.

To investigate whether the use of different symbols for the same referents and similar symbols for different referents is the source of DIF in this particular item, the item has been manipulated. In this manipulation the symbols for the referent ‘tower’ are similar and clearly different from the symbol used for the referent ‘well’. The item changed in this way has been included in Test 1 and is statistically analyzed with MH and OPLM again. These analyses reveal that the manipulated item does not show any DIF to the (dis-)advantage of Turkish and Moroccan students (z = .51 and z = 1.00 respectively). Because the item is manipulated as minimal as possible, i.e. only one aspect of the item has been changed, it can be assumed that in this item the cause of DIF has been the use of the symbols in the legend.

A second example of unclear illustrations leading to DIF to the disadvantage of Turkish and Moroccan students is the following item (z = 3.42). The answer possibilities are translated from Dutch in such a way that the wording is as comparable as possible to the original Dutch item.
Translation

Sign: Soft asphalt. Temporarily closed.

The local authorities sometimes put this sign on this country lane. When will they do so?

A when there is an earthquake    C when thaw has set in
B when there is a heat wave     D after a long period of rain

Although the right answer should be ‘when there is a heat wave’ most Turkish and Moroccan students opt for answer D, ‘after a long period of rain’ (37% and 43%, respectively, compared to 23% of the native-Dutch students). The underlying motivation of most of the students to opt for D is fully based on the drawings of the country lane, an canal and a windmill, that from the perspective of the students can only be in the Netherlands.

Fragment of Student Think Aloud Method: Mohammed

Mohammed: Question 30. The local authorities sometimes put this sign on this country lane. When will that be? Here it says: soft asphalt. ‘Tempor... temporarily closed’, it says on that sign. A when there is an earthquake, B when there is a heat wave, C when thaw has set in, D after a long period of rain. Well, here it says soft asphalt, tempor... temporarily closed. You can see a river or a canal there. So it has to be D: after a long period of rain.

On the basis of the illustration attached to this item students might select the wrong answer. Therefore, the original illustration is replaced by another one in the manipulated item.
this new item the drawing of the country lane is replaced by a motorway in a city and the question has been adjusted as follows: ‘The local authorities sometimes put a sign on this motorway with the following text: Soft asphalt. Temporarily closed.’ When will the local authorities do so?’

Final Test of Primary Education 1997: manipulated item
(World Orientation Subtest 3, question 3)

Statistical analyses show that the manipulated item does not have DIF. Because of the fact that the item was manipulated as minimal as possible, it can be assumed that the illustration has been the cause of DIF in this item.

**Metaphorical Language Use**

A lot of research has paid attention to the use of metaphorical language and idiomatic expressions as possible sources of DIF to the disadvantage of Turkish and Moroccan students (Coenen & Vallen, 1991; Uiterwijk, 1994; Vallen & Uiterwijk, 1994; Klatter-Folmer & Vallen, 1995; Van Schilt-Mol, Vallen & Uiterwijk, 2003; Uiterwijk & Vallen, 2005, Van Schilt-Mol, Vallen & Uiterwijk, 2005). The analyses of the five experiments conducted in this research project confirm that metaphorical language use might cause DIF in test items with questions that refer to or include this kind of language. An example is the following item. Again, text and answer possibilities are translated into English in such a way that the wording is as comparable as possible to the original Dutch item.
When they enter the playground after some time, they see Dick bent forward at a rubbish bin. ‘Of course! Hungry again!’, Koen says. ‘No way, my guilder has fallen into it.’ He pushes the rubbish aside. Suddenly he takes out a plastic bag with sandwiches and lifts it up. His eyes enlarge. With a leering smile he shows it to Koen and Katlijn. But when he sees the wry expression on their faces, he changes his mind and continues searching.

Read: When … searching. (l. 23-32)
What does this fragment make clear about Dick. That Dick is known as …

A a curious boy
B a slob
C a glutton
D a forgetful boy

The analyses of the Student Think Aloud Protocols and the Content Response Code show that the word ‘veelvraat’ [glutton], which is metaphorically used in Dutch, is leading to problems for a lot of students. The students indicate to be unfamiliar with the meaning of this word as is illustrated in the fragment below.

Fragment of Student Think Aloud Method: Nyazi

Nyazi: C a glutton ... well yes ... I pick C a glutton.
Researcher: Why do you think it has to be C?
Nyazi: Because, A a curious little boy, he is not that. D a forgetful little boy, he is not that either. B ... yes ... a slob, he is not that either, so ... and ... yes ... C, a glutton, I don’t really know what that means, so... Yes ... I think that when all the others are not right, it must be that one.

Another conclusion that can be drawn from the analyses of the difficulty level percentages is that almost all students agree that answer possibility B is right in fact too: 23% of the native-Dutch, 45% of the Turkish, and 37% of the Moroccan students opt for this possibility. Although one could argue that the source of DIF in this item might not be a linguistic but a cultural one, the manipulation of the item shows something else. In this manipulation the answer possibilities B and C are rewritten in such a way that no metaphorical language is used.

Manipulated item

Read: When … searching. (l. 23-32)
What does this fragment make clear about Dick?

A That Dick is always curious.
B That Dick is always dirty.
C That Dick is always in the mood for food.
D That Dick is always forgetful.

Analyses show that less students pick up answer possibility C in the manipulated item: 17% of the native-Dutch, 18% of the Turkish and 20% of the Moroccan students. Answer possibility C is now chosen most frequently by all students (69% of the native-Dutch, 66% of the Turkish, and 63% of the Moroccan students). Because the item was manipulated as minimal as possible, it can be assumed that the cause of DIF in the original item has been the metaphorical use of language in the answer possibilities B and C.

Another indication for the fact that the DIF source is in fact a linguistic and not cultural one, can be found in one of the items of the Final Test of Primary Education of 2002. This item was submitted to the students under the assumption that it might have DIF to the disadvantage of Turkish and Moroccan students on the basis of the use of metaphorical language both in the question and the answer possibilities. Statistical analyses confirm that this indeed is the case for Turkish student ($z = 2.27$). Again, it should be noted here that text, question and answer possibilities have been translated into English in such a way that the wording is as comparable as possible to the original Dutch item.

_Final Test of Primary Education 2002: Language Subtest 2, question 11_

_‘It was pitch-dark [pikdonker] in my room.’_
Read: It … room (l. 4)

What could Marieke have written in stead of: pitch-dark [pikdonker]?

A pitch-dark [aardedonker]
B deep-dark [diepdonker]
C ultra-dark [oerdonker]
D stone-dark [steendonker]

_Religion_

The third hypothesis to be discussed in the rest of this article is that questions concerning subjects related to religion can be the cause of DIF to the advantage of Turkish and Moroccan students. In the Final Test of Primary Education 1997 four questions refers to religious issues. Twice it concerns Islam questions and once Christianity and Judaism themes. All these questions have DIF to the advantage of Turkish and Moroccan students.

In order to investigate whether questions related to religion are the cause of DIF to the advantage of Turkish and Moroccan students, one of the items of the Final Test of Primary Education 1997 (World Orientation, subtest 1, question 19) has been manipulated and one item of the Final Test of Primary Education 2003 (World Orientation, subtest 3, question 4) has been submitted to the students.
The original item of 1997 \((z = -3.35)\) shows a photograph of the Wailing Wall on which six Israeli flags are visible. The question asked is to what the religious group the people belong that visit the Wailing Wall to pray. Students have a choice from (A) Islam, (B) Buddhism, (C) Christianity and (D) Judaism. Analyses show that a lot of native-Dutch students opt for possibility A (22% compared to 15% of the Turkish students and 10% of the Moroccan students). The Student Think Aloud Protocols show that most Turkish and Moroccan students opt for answer D based on the photograph added to the question.

_Fragment of Student Think Aloud Method: Nyazi and Mohamed_

Nyazi: _the Wailing wall in Je – ru – sa – lem. From all over the world people come to Jerusalem to pray at the Wailing Wall. This wall is a remain of an old temple. What is the religion of the people that come to pray here? A Islam, B Buddhism, C Christianity, D Judaism. Well ... I think ... uh yes ... I opt for D, Judaism, because, ... yes ... because yes ... if you look at the picture ... you see a flag. And that flag did the Jews have too._

Mohamed: _From all over ... all over the world people come to Jer... Jerusalem to pray at the Wailing Wall. This wall is a remain of an old temple. What is the religion of the people that come to pray here? Islam, Islam, uh ... no. Well they are going, they are going to Mekka. B Buddhism. No, they have, in their country, China I believe, they have their own statues there. C Christianity. They are not going ... they are going to Jerusalem, but uh ... but they are going to a river or something. Judaism, yes, they are coming here. Because here you see ... here you see the Jewish flag._

To investigate whether this item still has DIF to the advantage of Turkish and Moroccan students when they are not able to make there choice on the basis of the photograph added to the question, the picture is removed in the manipulated item. This new item was submitted to the students and statistically analyzed with MH and OPLM. It turned out that the manipulated item had DIF to the advantage of both Turkish (\(z = -2.11\)) and Moroccan students (\(z = -3.51\)) too.

As mentioned before, one item of the Final Test of Primary Education 2003 has been included in the experiment to test the hypothesis that questions concerning subjects related to religion can cause DIF to the advantage of Turkish and Moroccan students. Again, it should be noted that the question and answer possibilities are translated into English in such a way that the wording is as comparable as possible to the original Dutch item.

_Final Test of Primary Education 2003: World Orientation, Subtest 3, question 4_

**Bianca’s religion**

Bianca talks about her religion in her classroom. She tells that she prays at home in front of her home altar and that she burns incense for many gods. Some animals are holy in her religion. She believes that she will return on earth in a new live after her death.
What religion does Bianca believe in?
A Islam
B Christianity
C Hinduism
D Judaism

According to the hypothesis formulated in this respect, it was expected that this item should have DIF to the advantage of Turkish and Moroccan students. Statistical analyses show that this is the case for Moroccan students \( z = -2.58 \) indeed.

3.5 DIF or DIV?

When statistical analyses of the manipulated items show that the hypotheses about the possible sources of DIF do not have to be rejected, the decision has to be made whether the actual source of DIF is relevant or irrelevant to the construct to be measured. When the source is relevant, the item only has DIF. In this case it is not necessary to change the item or remove it from the test. The difference in the test scores of the different subgroups is, in this case, a result of a difference in the mastery level of the construct being measured.

When the source of DIF is irrelevant to this construct and when, in other words, answering the item correctly requires additional skills and knowledge other than being measured, it can be concluded that there is a case of Differential Item Validity (DIV) or item bias. When it can be assumed that not all students have an equal command of these additional skills and knowledge, these items are not suitable to differentiate between students. Tests, and especially tests of educational progress, are meant to differentiate between high- and low-achieving students. This means that one has to be confident that scores of all student subgroups can be interpreted in a similar way. This is only possible when all students with a comparable proficiency or achievement level stand the same chance to have an equal chance to obtain a test result that is in line with their achievement level in the subject matter area being tested. When differentiation occurs on basis of additional skills and knowledge that not all students have a thorough command of, students with a comparable level in a certain subject matter area do not this equal chance. For that reason, items that have DIV should be adapted or removed from the test.

To determine whether the DIF items detected in this research only have DIF or have DIV too, a final experiment is now conducted. In this experiment several experts - test constructors, textbook developers, teaching methods developers, teachers of primary grade eight and linguistic experts -, will be asked to indicated whether the source of DIF is relevant to the construct to be measured according to their professional knowledge. On the basis of the results of this latter experiment and the other experiments conducted, it will be possible to make a clear distinction between those sources of DIF that are connected with the items through the construct being measured and those sources that cause DIF not all
students having a thorough command of. This latter type of source will make it possible to formulate recommendations for avoiding DIV in future tests.
**Literature**


