Spelling in primary school: Mixture IRT-models as a diagnostic tool

Daniël Van Nijlen and Rianne Janssen

Katholieke Universiteit Leuven, Belgium

Author’s address: Daniël Van Nijlen, Faculty of Psychology and Educational Sciences, Dekenstraat 102 (PO box 3773), 3000 Leuven, Belgium. E-mail: daniel.vannijlen@ped.kuleuven.be. Phone: ++32-16-32.59.05
Spelling in primary school: Mixture IRT-models as a diagnostic tool

Abstract

In the current study it is investigated whether mastery of two crucial spelling rules in Dutch can be considered to be purely dimension-like or that there are more category-like elements to be distinguished. If merely quantitative differences in mastery appear the mastery can be considered to be dimension-like. However, if qualitative differences appear, mastery will have some category-like features. Moreover, the issue will be addressed that, if qualitative differences appear, these could be linked to manifest features of the pupils or that the use of latent groups can be more informative. Results indicate that mastery of the spelling rules cannot be considered to be strictly dimension-like. Three latent groups of students can be distinguished that show qualitative differences in the mastery of one of these spelling rules and the related rules. The educational importance of these results and the applied methodology will be illustrated.
Spelling in primary school: Mixture IRT-models as a diagnostic tool

A crucial aspect in measuring a given psychological phenomenon is the issue whether it should be considered to be categorical or dimensional. Based on item response theory (IRT), the Dimcat-framework (De Boeck, Wilson & Acton, 2005) addresses the question whether, from a latent perspective, a given phenomenon is more category-like or more dimension-like.

Central in the Dimcat-framework are two distinctions. The first refers to the distinction between quantitative and qualitative differences, where qualitative differences can be considered to be more category-like and quantitative differences more dimension-like. Quantitative means that (manifest) categories are different by a degree of something identical (the latent dimension is the same), whereas qualitative means that the differences cannot be related to a degree of something identical, but that something of a different kind is involved (different latent dimensions). The second distinction is that of what De Boeck et al. refer to as within-category heterogeneity and homogeneity, where the former refers to systematic differences in degree within a category and the latter to lack of systematic differences within a category. Within-category heterogeneity is considered to be more dimension-like, while homogeneity of a category is more category-like. Although originally presented for manifest categories like gender, this framework can be easily extended to latent categories, which in essence corresponds to using a mixture IRT model when analyzing data (Hidegkuti & De Boeck, 2006).

The distinction between qualitative and quantitative differences within a certain population can be clearly linked to the issue of differential item functioning (DIF) in a measurement situation. If different subpopulations exist with respect to the variable that is being measured this will be reflected in a differential performance of the subpopulations on the tasks, items used. Purely quantitative differences in the performance will not be considered to be DIF, because when the overall ability has been taken into account no
performance difference between the subpopulations occurs. Qualitative differences, however, can be considered as DIF.

**Differential Item Functioning**

In the theorizing about DIF and the praxis of DIF there has been a shift throughout the years (Zumbo, 2007). Initially, DIF was merely considered to be a nuisance to the measurement properties of the test and, hence, focus lay on the detection of DIF items so they could be excluded from the test (‘test cleaning’). Later on there has been a transition to not only a focus on DIF detection but also on why DIF was occurring and what can be learned about the item response process from the DIF present in the test. DIF turned out to be possibly interesting from a substantive point of view.

DIF may be viewed as multidimensionality in the data (Ackerman, 1992), where the different groups perform in a different way on the additional dimension(s) in the data. Hence, a first step when analyzing data with regard to possible DIF can be to check whether it is possible to achieve satisfactory data-model fit using a unidimensional IRT-model. If this is not the case, one can allow for multidimensionality in the form of DIF for some manifest, preexisting groups (e.g. gender, ethnicity,…).

However, Cohen and Bolt (2005) found a weak relationship between the manifest characteristic associated with DIF and the actual advantaged or disadvantaged groups. Manifest groups are not as homogeneous as often assumed (De Ayala, Kim, Stapleton & Dayton, 2002; Samuelsen, 2005); which makes them not always the most interesting variable to investigate. As early as 1988, Tatsuoka, Linn, Tatsuoka and Yamamoto pointed to the fact that group membership often is only a weak proxy for the aspects that are of actual instructional relevance, thus, resulting in a lack of interpretability of the results from the DIF analysis.
A logical response to this is to no longer only focus on manifest, preexisting groups in the DIF analysis, but look for latent groups that show DIF. This idea of looking for latent groups in the context of DIF has recently gained considerable attention (Cohen & Bolt, 2005; De Ayala et al., 2002; Samuelsen, 2005; Webb, Cohen & Schwanenflugel, in press), also in the context of modeling spelling ability (Hoijtink & Notenboom, 2004; Notenboom & Reitsma, 2007).

The use of mixture IRT models (Fieuws, Spiessens & Draney, 2004) can circumvent the problems that rise with the use of manifest groups, as these models make it possible to detect latent groups for which the items function differently. These mixture IRT models, that combine a Latent Class-approach with an IRT-approach, can both detect subpopulations that show qualitative differences and at the same time quantify the differences in ability within the groups (e.g. Rost, 1990; Mislevy & Verhelst, 1990). Using these models one can determine differential item functioning by checking the differences in item parameters between the latent classes.

Spelling of short and long vowels in Dutch

**Relevant spelling rules**

A specific problem in Dutch spelling is the writing of long and short vowels. The basic principle applied is that of phoneme-grapheme correspondence (‘what you hear is what you write’). If you hear a long vowel you write a double letter (e.g. 'laat'; late); if you hear a short vowel you write a single one (e.g. 'lat', ruler). However, this principle is not sufficient for the writing of long and short vowels in words consisting of more phonological parts.

If a long vowel is followed by another phonological part it is represented with a single vowel (e.g. 'later'; later). The rule is called ‘klinkerverenkeling’ (KV), which literally means to make the vowel single (referred to as the KV rule). For short vowels, the consonant
following a short vowel is doubled (e.g. 'ladder'; ladder). This spelling rule is called ‘medeklinkerverdubbeling’ (MV), which literally means to double the consonant (MV rule).

As is often the case in language spelling, there are some additional rules and exceptions to this general MV rule, in which case the consonant is not doubled. First, the consonant is not doubled when it is preceded by a silent vowel (referred to as DK, e.g. monniken, monks). A second additional rule is that the consonant is not doubled after a diphthong (TK, e.g. houten, wooden). A third additional rule states that a single consonant is written after eu, i(e) and oe (EIO, e.g. kleuter, toddler). Finally, there are some words that are merely exceptions to the MV rule (e.g. papegaai, parrot).

**Spelling and the Dimcat-framework**

The issue of making a distinction between qualitative and quantitative differences in performance between groups is extremely relevant in the context of learning difficulties, also with regard to the writing of short and long vowels. It is a crucial distinction with regard to the research on the appropriate instructional interventions for certain learning disabilities or difficulties (Yang, Shaftel, Glasnapp & Poggio, 2005).

It seems plausible that qualitative differences are present, e.g. due to different approaches in the spelling instruction by teachers. Since the rules under consideration are only taught throughout the different grades of primary school, different strategies might be used by pupils in different grades. This may cause DIF in the items across the grades leading to a non-fitting unidimensional IRT scale. So, first it will be checked whether it is possible to achieve satisfactory data-model fit using a unidimensional IRT-model. If this is not the case, multidimensionality in the form of DIF across grades will be allowed for.

However, heterogeneity of a population, heterogeneity within a certain group in a qualitative sense may be attributed to known sources (manifest categories, e.g. grade) but also may be due to unknown sources (latent categories). There may be unknown subpopulations
that may respond in a different way to relevant indicators. And, as has been pointed out before, the manifest grouping might not always be the most relevant variable to investigate. Therefore, a next step in analyzing the data will be to check for the presence of latent groups by using mixture IRT models.

In the current study it is investigated whether mastery of the above-mentioned spelling rules can be considered to be purely dimension-like or that there are more category-like elements to be distinguished. Moreover, the issue will be addressed that, if qualitative differences appear, these could be linked to manifest features of the pupils or that the use of latent groups can be more informative.

Method

Test

In order to detect children that have problems with the spelling rules for the writing of short and long vowels, a specific spelling test was constructed assessing the ability of pupils in applying these rules (Vangeneugden, 2004). The final goal of this criterion test was to work with two parallel versions of the tests. In this way one version could be used for the detection of spelling difficulties with regard to these rules, while the second version could be administered to evaluate the progress a pupil made. To achieve this goal an overlap in test items was included so that it would be possible to fit both tests on a common IRT-scale. In this way comparability of results could be ensured and progress could be evaluated in a meaningful way.

In the current paper analyses will focus on one of the two versions that were constructed for the criterion test. The test consisted of 45 words imbedded in a sentence that clarified the meaning of the word that had to be spelled by the pupils. Of the test 39 words were common to all grades. The spelling was only scored for the application of the rules
under consideration. Scoring could be done at the word level or the item level within a word
when more than one rule was applied.

Participants

The test was administered to 1078 children coming from Grade 3 through Grade 6 of
primary education in Flanders. The majority of these pupils was Dutch-speaking (n=1005)
and 538 of them were girls. In Table 1 a more detailed description of the participants is given.

Table 1
Cross-tabulation of participants for grade and gender

<table>
<thead>
<tr>
<th>Grade</th>
<th>Boys</th>
<th>Girls</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>135</td>
<td>134</td>
<td>269</td>
</tr>
<tr>
<td>4</td>
<td>122</td>
<td>144</td>
<td>266</td>
</tr>
<tr>
<td>5</td>
<td>154</td>
<td>132</td>
<td>286</td>
</tr>
<tr>
<td>6</td>
<td>127</td>
<td>130</td>
<td>257</td>
</tr>
<tr>
<td>Total</td>
<td>538</td>
<td>540</td>
<td>1078</td>
</tr>
</tbody>
</table>

Analyses

Only the 39 words common to all grades were included in the analyses. A scoring at
the item level was used, which resulted in a total of 50 items. Of these items, 13 referred to
the writing of long vowels (KV rule), 13 to the doubling of the consonant after a short vowel
(MV rule) and 24 to the application of additional rules to the latter rule. Only pupils with a
non-perfect response pattern were included (n=1029) in the analyses.

First, a more standard IRT approach was taken in analyzing the data. A
unidimensional IRT model (Rasch and 2PL) was fit to the data. Second, to check for potential
DIF across grades a Rasch scale DIF analysis was done using BILOG-MG (Zimowski,
Muraki, Mislevy & Bock, 2003). Finally, mixture IRT models with one to four latent classes
were estimated. More detail on this model is provided in the next section.
Mixture IRT model

In the following the model is formulated in a general way so it can accommodate both the mixture Rasch model (Rost, 1990) and the mixture LLTM model (Fischer, 1973; Mislevy & Verhelst, 1990). Formally, according to the mixture IRT model, the data are assumed to be independent Bernoulli realizations conditional on the latent variables, with the probability of success modeled as

$$\Pr(Y_{ni} = 1|\theta_{nt}, \xi_n = t) = g(\theta_{nt} + x_i \eta_t),$$

where $Y_{ni}$ is the binary response variable for person $n$ on item $i$, $\xi_n$ is a categorical latent variable indicating class-membership for person $n$, $\theta_n$ is the person specific intercept (continuous latent variable, ability) of person $n$ in class $t$, the vector $x_i$ contains the values of item $i$ on the item covariates, $\eta_t$ is the vector of regression weights for class $t$, and $g$ is the response function (either the cumulative logistic or normal ogive function). Furthermore, we assume $\theta_{nt} \sim N(0, \sigma^2)$ for all $n$ and $t$.

Averaging over the latent space, the marginal probability of a response pattern $y_n$ is

$$\Pr(y_n) = \sum_t p_t \int \left( \prod_i \Pr(y_{ni} | \theta_{nt}, \xi_n = t) \right) \mathcal{N}(\theta_{nt}, 0, \sigma^2) d\theta_{nt},$$

where $p_t$ is the class weight or probability of belonging to class $t$, also known as the mixing parameter. Since the class weights have to sum to one, $p_T$ is set to $1 - \sum_{t=1}^{T-1} p_t$.

The models were estimated using an adaptation of the EM-algorithm introduced in Rijmen and De Boeck (2003). In the current paper the mixture Rasch model (Rost, 1990) was used to perform the analyses.

Results

Standard IRT models
Using a unidimensional IRT model no model-data fit could be achieved, neither for the Rasch model or the less restrictive 2PL. As mentioned before, this might be caused by the fact that children acquire these rules throughout the different grades of primary school and, as a consequence, that children in different grades apply different strategies to solve the items. A DIF-analysis using BILOG-MG indeed showed that a considerable number of items suffered from DIF across grades. However, it was fairly difficult to detect a clear, interpretable pattern in the DIF.

In the present case, in Flanders different schools start teaching the rules at different times in primary school, and, children might acquire the rules at different points during primary school. This makes it plausible that grade might not be the most relevant variable to study. As stated before, mixture IRT models might provide a solution to this issue.

Mixture IRT models

Table 2 describes the results for the estimated models. For each model the loglikelihood and the number of estimated parameters is reported. The total number of parameters equals the number of items for each class, the variance of the ability parameter theta (which is constant over classes), and the mixing parameters for models with more than one latent class. Based on this information both the AIC and the BIC of the model are calculated so these can be used in the model selection. For the models with two or more latent classes the mixing parameter for each of the latent classes is reported. Only t-1 mixing parameters are estimated because they are set to sum to one.

<table>
<thead>
<tr>
<th>Classes</th>
<th>LL</th>
<th>npar</th>
<th>AIC</th>
<th>BIC</th>
<th>p cl1</th>
<th>p cl2</th>
<th>p cl3</th>
<th>p cl4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-16741</td>
<td>51</td>
<td>33584</td>
<td>33836</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-15709</td>
<td>102</td>
<td>31622</td>
<td>32126</td>
<td>.36</td>
<td>.64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-15547</td>
<td>153</td>
<td>31400</td>
<td>32155</td>
<td>.22</td>
<td>.39</td>
<td>.39</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>-15498</td>
<td>204</td>
<td>31404</td>
<td>32411</td>
<td>.23</td>
<td>.01</td>
<td>.37</td>
<td>.39</td>
</tr>
</tbody>
</table>
Both the AIC and BIC indicate that a solution that assumes that all pupils belong to the same latent class is not desirable to describe the data. Also the solution assuming that four latent classes could be distinguished showed to be not the most suited for the data. The additional latent class only contained one percent of the pupils and both the AIC and the BIC indicated that this was not the preferable solution. The lowest value for the AIC indicates that a solution with three latent classes describes the data adequately, while the BIC refers to a solution with two latent classes as the preferable one.

To decide on the number of latent classes the shift from the pupils between the solutions with two or three latent classes was investigated (Table 3). For each pupil a probability was estimated for belonging to one of the latent classes. Each pupil was allocated to the latent class with the highest probability. Compared to the two class solution an additional middle group is created in the three class solution. As the latent classes were ordered based on the average total score of pupils within a class, one can really consider this additional group as a middle group. None of the pupils in the low-performing class of the two-class solution makes a shift to the highest-performing group in the three-class solution, but about one third of them makes the shift to this middle group. A similar, but opposite pattern shows for the pupils in the high-performing group of the two-class solution.

<p>| Table 3 |
| Cross-tabulation of class membership for a solution with two and three latent classes |</p>
<table>
<thead>
<tr>
<th>Two latent classes</th>
<th>Three latent classes</th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>235</td>
<td>131</td>
<td></td>
<td>366</td>
</tr>
<tr>
<td>2</td>
<td>250</td>
<td>413</td>
<td></td>
<td>663</td>
</tr>
<tr>
<td>Total</td>
<td>235</td>
<td>381</td>
<td>413</td>
<td>1029</td>
</tr>
</tbody>
</table>
Because the AIC clearly indicated to a three class solution and the difference in BIC between the two and three class solution was not that big, and the information presented in Table 3, it was decided to continue based on the three class solution.

Interpretation

The three latent classes were ordered based on the average total score on the test for each class. As a consequence the first latent class can be referred to as the lowest performing group, the second latent class as the medium performing group, and, finally, the third latent class as the highest performing group. However, a more substantive interpretation to the results is desirable. To achieve this, two approaches are used. First, the results for the latent classes are linked to the manifest categories of grade and gender. This mapping of manifest groups onto the latent classes, may hint at a first possible interpretation of why it was not possible to analyze the data using a unidimensional IRT model and what the source of DIF is (Samuelsen, 2005). Second, the results are interpreted based on a comparison of the item parameters for the three latent classes.

Interpretation based on classification of students

When we look at the distribution for gender across the different latent classes (Table 4), we find only minor differences between boys and girls, with slightly more girls in the highest performing group and slightly less in the lowest performing. Class membership does not seem to be very much related to the gender of the pupil.

Table 4
Cross-tabulation of gender and class membership

<table>
<thead>
<tr>
<th>Gender</th>
<th>class</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>Boy</td>
<td>.25</td>
<td>.37</td>
<td>.39</td>
<td>519</td>
<td></td>
</tr>
<tr>
<td>Girl</td>
<td>.21</td>
<td>.37</td>
<td>.42</td>
<td>510</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>235</td>
<td>381</td>
<td>413</td>
<td>1029</td>
<td></td>
</tr>
</tbody>
</table>
The latent classes clearly differed in their composition when grade is considered (Table 5). The majority of the grade 3 pupils (60%) belonged to the first, lowest performing, latent class, while the majority of grade 6 (68%) belonged to the third, best performing, latent class. So there appears to be some relation between grade and class membership.

Table 5
Cross-tabulation of grade and class membership

<table>
<thead>
<tr>
<th>Grade</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>.60</td>
<td>.32</td>
<td>.08</td>
<td>267</td>
</tr>
<tr>
<td>4</td>
<td>.18</td>
<td>.46</td>
<td>.36</td>
<td>262</td>
</tr>
<tr>
<td>5</td>
<td>.09</td>
<td>.39</td>
<td>.52</td>
<td>270</td>
</tr>
<tr>
<td>6</td>
<td>.02</td>
<td>.30</td>
<td>.68</td>
<td>230</td>
</tr>
<tr>
<td>Total</td>
<td>235</td>
<td>381</td>
<td>413</td>
<td>1029</td>
</tr>
</tbody>
</table>

However, grade at best can be considered as a partial proxy for class membership, but not a very accurate one. For instance, 8% of the grade 3 pupils in fact belonged to the third, best-performing, latent class and for each grade at least about one third of the pupils is classified within the second, medium-performing class. This suggests that acquiring these rules is a more subtle process than one that can be captured by grade. An interpretation based on the item parameters for the latent classes might provide more insight in what actually constitutes the typicality of each latent class.

Interpretation based on item parameters

The item parameters for the three latent classes were compared using one reference class (i.e. the lowest performing class) and correcting the item parameters of the other classes for the average performance difference with this reference class (DuToit, 2003). To compare the corrected item parameters the standard errors for these parameters were taken into account to check for significant differences. Before looking at the significance of differences in item parameters across classes, a visual representation of the corrected item parameters can
provide crucial information on the substantive interpretation (Figure 1). Note that for each item an item easiness parameter has been estimated. As a consequence the difficult items are represented at the left hand side of the scale and the easier items on the right hand side.

The first pane in the top left hand corner plots the item parameters for all 50 items. The top line represents the item parameters for the first, low performing class. On the middle line the item parameters for the second, medium performing latent class are plotted. On the bottom line the item parameters for the third, high performing latent class are represented. Vertical lines connect the parameters for the same items in the different classes. If there would be no DIF in any of the items (and there would be no error in the parameter estimates) these would be straight, vertical lines. This is clearly not the case. However, it is also fairly difficult to detect a clear pattern in the vertical lines.

Therefore, it might be interesting to make separate plots for the items pertaining to the different rules. A separate plot was made for the KV rule items, the MV items and the items pertaining to the additional rules to the MV rule. These plots are represented in the remaining three panes of Figure 1.
Figure 1: Graphical illustration of the corrected item parameters for the latent classes

For the items pertaining to the KV rule (pane b) the lines do go down in a rather straight vertical way, suggesting that there is not really an issue with DIF for these items. In pane c the MV items are plotted and for these items there is a clear tendency in the vertical lines. The MV items are very difficult for pupils in the first latent class, become a bit easier for the ones in the second, medium performing latent class and are clearly easier for the pupils in the high performing latent class. For the items pertaining to the additional rules (pane d) we see an opposite, albeit less clear, pattern in the vertical lines. There is a tendency for these items to be relatively easy for the pupils in the first, low performing latent class and they become gradually more difficult for the higher performing latent classes.

Table 6 gives an overview of the significance tests that take into account the estimation error in the item parameters. Several situations can occur for each item. The simplest case is if none of the comparisons between classes are significant, pointing to the fact that there is no DIF for the item. Other clear situations are when the items, going from the low performing to the high performing group, become gradually easier (HP>MP>LP) or more difficult (LP>MP>HP). Sometimes one of the steps from one group to the other is not significant but they still follow a gradual pattern (e.g. LP=MP>HP). For a few items the results for the significance tests were not in accordance with some logical pattern. These are combined under the label ‘other’. The results are presented in separate columns for the different types of items.
Table 6

Results significance tests pairwise comparison of item parameters for the latent classes

<table>
<thead>
<tr>
<th>KV</th>
<th>MV</th>
<th>Add</th>
</tr>
</thead>
<tbody>
<tr>
<td>HP&gt;MP&gt;LP</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>HP=MP&gt;LP</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>HP&gt;MP=LP</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>LP&gt;MP&gt;HP</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LP&gt;MP=HP</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LP=MP&gt;HP</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Other</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>No DIF</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>13</td>
<td>13</td>
</tr>
</tbody>
</table>

The results in Table 6 confirm the patterns that were found in the visual representation of the item parameters. The items concerning the writing of long vowels (KV rule) show almost no DIF. Almost all of the items measuring the MV rule turn out to be relatively more difficult for the lowest performing class compared to the medium-performing class and in its turn for this class compared to the high-performing class. The results for the additional rules to the MV rule seem to be less straightforward to interpret. Six of the items pertaining to the additional rules show no DIF at all between the three groups. Five out of these six items refer to one of the specific additional rules (EIO). If we disregard these items, almost all of the remaining additional rule items are relatively easy for the low performing class compared to the medium performing class and for that class compared to the high performing one.

Conclusions and educational importance

Data with regard to two crucial spelling rules in Dutch were analyzed, applying mixture IRT models. It was shown that pupils from grade 3 to grade 6 could not be considered to be one homogeneous group with regard to the mastery of these rules. Not only
were there quantitative differences in their performance on the test, but it was also necessary to take into account qualitative differences in the modeling.

Three latent classes could be distinguished, primarily based on the mastery of the rule for doubling consonants following short vowels. The first class consisted of low-performing spellers that did not yet master the 'double consonant' rule but, as a consequence, have relatively little problems with the additional rules and exceptions to that rule. The second class contained medium-performing pupils who are still acquiring the 'double consonant' rule. And finally, the third group was made up of high-performing spellers for whom the 'double consonant' rule is relatively easy but who overgeneralize this rule and apply it in cases they should not use it.

The application of mixture models to these educational data made it possible to let a pattern surface that could not be detected using a classic DIF-procedure. This illustrates that the application of this kind of models might have great value in an educational context to find out where specific difficulties in acquiring a domain can be found and what the educational practice should focus on. At the same time, from a diagnostic point of view, the use of these models enables one to not only detect quantitative differences in mastery of specific educational issues, but also to find out in a qualitative way where the specific problems of a pupil lie.
References


Zumbo, B.D. (2007). Three generations of DIF analysis: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*, 223-233.