To cite this article:

Kahraman, N., De Boeck, P., & Janssen, R. (2009). Modeling DIF in complex response data using test design strategies. *International Journal of Testing*, *9* (2), 151-166. https://doi.org/10.1080/15305050902880744

RUNNING HEAD: Modeling DIF

Modeling DIF in Complex Response Data Using Test Design Strategies

Nilufer Kahraman

National Board of Medical Examiners

University of Leuven, Belgium

Paul De Boeck

Rianne Janssen

University of Leuven, Belgium

CORRESPONDING AUTHOR:
Nilufer Kahraman
NBME MCS
3750 Market Street
Philadelphia, PA 19104
Phone: 215 590 9728
E-Mail: nkahraman@nbme.org

Abstract

This study introduces an approach for modeling multidimensional response data with construct-relevant group and domain factors. The item level parameter estimation process is extended to incorporate the refined effects of test dimension and group factors. Differences in item performances over groups are evaluated distinguishing two levels of differential item functioning (DIF): a domain level and an item level.

An illustration is presented using a Dutch spelling proficiency scale administered to two subgroups. DIF is modeled by the interaction between group and item domain (domain level DIF), and by the interaction between groups and items within each domain (Item level DIF). A set of IRT Models was estimated using an adaptation of the logistic regression approach. The model with domain specific item-by-group interactions or DIF performed better than the other models neglecting domain or group differences.

The method appears to be promising in that explicit domain factors can be implemented into model estimation procedure to better understand why items favor a specific language group over another.

Keywords: DIF, explaining DIF, multilevel models, spelling ability, test design strategies, multidimensionality.

Modeling DIF in Complex Response Data Using Test Design Strategies

An important issue in test adaptation across languages and cultures is whether inferences made on the basis of a test are comparable for different groups. Differential item functioning (DIF) is a threat to comparability, and occurs if an item is easier for one group of test takers than for another after controlling for overall ability. When checking for DIF, individual test items are evaluated over subpopulations, which are presumed to have an equal possession of ability, to assure that the conditional probability of endorsing an item is independent from group membership (Lord, 1980). For example, serious complications would arise in the interpretation of test scores when there is evidence that scores of a particular language group of test takers are systematically underestimated or overestimated. Ideally, DIF items would be identified as "biased" only after a logical analysis to why they are easier (or more difficult) for a focal group of test takers in comparison with a reference group (Hambleton & Swaminathan, 1985).

The DIF literature is quite extensive and includes a wide range of applications of the procedures developed to detect DIF items under various conditions. To date, however, the focus in DIF research is largely on manifest group factors interacting with item effects, and not on relevant group factors explaining DIF. This makes it rather difficult to understand why some test takers are disadvantaged by particular test items (e.g., Cohen & Bolt, 2005, Johanson & Asmaldi, 2002). In fact, in quite a number of studies, testing for DIF is merely done from the perspective of 'test cleaning', i.e., omitting the items showing DIF from the test. However, the presence of DIF can also be interesting from a substantive point of view.

One of the possible explanations of DIF is that an item measuring a composite of multiple dimensions would function differently over subgroups. The composite nature of test constructs are expected to be amplified when items measure multiple dimensions, and groups differ with respect to one or more these dimensions (e.g. De Ayala, et. al. 2002, Lubke & Muthen, 2005). In this case, evidence for DIF may be a reflection of model misspecification to the degree that the observed variations in the data are in conflict with the presumed test dimensionality. For example, a spelling test can be comprised of items measuring a different composite of a number of spelling rules. If examinees are from different language subpopulations, the overall group will be heterogeneous with respect to the required level of spelling proficiency. Hence, investigating the dimensional structure of the test data can provide a basis to understand why some items perform differently over subpopulations that are known to differ in their primary language, developmental level, etc.

Second, DIF is mostly studied for demographically distinct subpopulations defined by gender or ethnic background. However, from a substantive point of view, it can also be interesting to look for DIF in items that are administered across different time points or across different grade levels (Mislevy, Johnson, & Muraki, 1992). In developmental studies, the differences in item performances across grade groups can be attributed to impact only if the differing item characteristic curves (ICC) relate to real ability differences. Then, this impact refers to the difference in general level of proficiency, and can be modeled either through a change in the item parameters or in the ability parameters of the different groups. For example, an item reflecting a fourth-grade spelling curriculum can be expected to show impact in the form of a difference in performance between a reference group of fourth

graders and a focal group of third graders. The item, however, would be considered to be a DIF item if third and fourth graders with the same overall spelling score perform differently on this item.

In the present study, we applied a multidimensional perspective on DIF to investigate how item properties relate to complexity factors of tests. The procedure is demonstrated for a developmental scale for Dutch spelling, considering two factors: (1) test dimensions referring to two spelling subdomains, and (2) subpopulations referring to students from two adjacent grades in primary education. The aim of the study is twofold. First, the spelling performance and progress through grades is portrayed in a way that lends itself to a substantive understanding, which is interesting from a psychological point of view. Second, domain ties of DIF is investigated by modeling within and between group variations of items from two spelling subdomains, which is interesting from a psychometric point of view. After the presentation of this data set, the approach is explained.

The spelling data

A particular problem in Dutch spelling is the writing of short and long vowels in open and closed syllables. Given the dominant principle of phoneme-grapheme correspondence, a simple rule might have been to write a single letter representing the corresponding phoneme for a short vowel, like the o in ‘bot’ (bone), and to write a doublet of this letter for the corresponding long vowel, like in ‘boot’ (boat). As is more often the case in spelling, the simple rule is more complicated. For long vowels, a doublet is indeed written, but only in closed syllables, as in ‘boot’. In open syllables, however, the doublet

representing the long vowel is halved, like in ‘boten’ (boats). The halving of a doublet of vowels is one spelling rule children have to acquire when learning to write vowels. The rule is called ‘klinkerverenkeling’, which literally means to make the vowel single.

Because of the vowel rule, the pronunciation of a single letter representing a vowel becomes, in a sense, ambiguous. In open syllables, a single letter represents a long vowel, like in ‘boten’ (boats), while in closed syllables, it represents a short vowel, like in ‘botten’ (bones). Hence, to guarantee that a single letter representing a vowel is read as a short vowel, it should be contained in a closed syllable. The latter can be ensured by doubling the consonant following it. This spelling rule is called ‘medeklinkerverdubbeling’, which literally means to double the consonant. The consonant rule is applied in different situations, for example, in the plural of some nouns, e.g. bot – botten (bone – bones), in derivative words, e.g. pret – prettig (pleasure –pleasant), or in words as such, e.g. ladder (ladder) or terras (terrace). Although dominantly applicable, the vowel and consonant rule also have their exceptions, making it a difficult developmental task for children to acquire the correct writing of short and long vowels in open and closed syllables in Dutch. Both rules are taught from the second grade in primary education onwards, but full mastery of the rules and their exceptions is only expected around the sixth grade, which is at the end of primary education.

In the present study a collective, paper and pencil test for this particular spelling problem (Vangeneugden, 2004) was administered to a random sample of 269 children from the third grade of primary education in Flanders and of 266 children from the fourth grade. A one-step sampling procedure was used, with a sampling of schools of a particular region

at the first step. All the children of the intended grade from a sample school participated in the study. The data studied included 13 open and 13 closed syllables, with as dominant difficulty the vowel rule and the consonant rule, respectively. The student had to write each word down, after hearing a sentence illustrating the meaning of the word. Each word was scored on correctness, but only taking into account the number of vowels and consonants.

Modeling Strategy

Two levels of DIF

The design for the present study consists of two complexity factors: domain and group. The domain is an item feature and it refers to the two spelling rules for open and closed syllables in Dutch (the vowel rule for open syllables and the consonant rule for closed syllables). The group factor is a feature of the respondents and it refers to the two groups (third and fourth grades of primary school). Four types of effects were considered: (1) the main effects of the two complexity factors (domain and group), (2) their interaction, (3) the item main effects, and (4) the item-by-group interaction effects.

Two levels of DIF can be distinguished. First, the interaction effect between item domain and group is a domain level of DIF. This is also called Differential Feature Functioning (DFF), which occurs when DIF can be reduced to item features functioning differently for the groups (Engelhard, 1992, Meulders & Xie, 2004; Wang & Wilson, 2005). In the case where the item feature refers to different domains, one would no longer find DIF if the domains were scored separately, instead of using a composite score. Investigating a domain level of DIF is interesting from a developmental point of view, because the

interaction effect between item domain and group indicates that the progress made from one grade to the next is different for both domains. The absence of an interaction effect implies parallel development.

Second, there is the item level of DIF, which refers to the differential performance of individual items within a domain across the groups. The inclusion of an item-specific level of DIF is independent of the dimensionality of the test, and can occur in a unidimensional as well as in a multidimensional model. The item-specific level of DIF can be modeled by including within-group nested item effects with or without including a domain level of DIF.

The proposed approach is primarily a modeling approach rather than a DIF detection approach, and aims at explaining item performances on two levels of main effects and interaction effects rather than at detecting biased items.

Proposed models

Given the four types of effects and the two levels of DIF, the following models can be inferred, all with item discriminations: (1) a model with item main effects but no DIF (the regular 2PL Model), (2) a model with DIF at the item level, (3) a model with DIF at the domain level, and (4) a model with DIF at both the item and the domain levels. The four models presented differ with respect to the dimensionality of θ and the modeling of the item location parameter.

The reasoning in the selection of these models is as follows. The *2PL* model is a reference model being a simple unidimensional model. The model with DIF at the item

level is an evident extension for the 2PL model when confronted with the possibility of DIF. When DIF is found at the item level, a logical next step is to investigate whether DIF can be summarized or explained by a feature of items, more in particular by the item domains. This can be modeled by including a second dimension in the model to explain group differences. One may as well extend this second dimension into one that also allows for individual differences within the groups (implying individual differences in DIF). The resulting model is called DIF at the domain level. In this model there are no longer item-by-group interactions but only domain-by-group interactions instead. Finally, in the model with DIF at both levels, the item-by-group interactions are added to the previous model in the form of item effects nested within groups. The final model with two levels is referred to as the complete model in the following text. The parameterizations of these models are presented below, after the presentation of the general modeling approach, which is based on item response theory (IRT).

Model Estimation

The logistic regression model

The logistic regression procedure is commonly used to detect DIF items (see Borsboom, Mellenbergh & van Heerden, 2002; Gelin & Zumbo, 2003; Miller & Spray, 1993; Raju, Laffitte & Byrne, 2002). A logistic regression model defines the probability of observing a correct response, u, as a function of two explanatory variables: the observed test score t, and a group indicator variable $\underline{g}$ (Swaminathan & Rogers, 1990, Rogers & Swaminathan, 1993). The model testing for DIF at the level of individual items can be written as

$$P(u=1|t,g)=\frac{\exp(\beta_0+\beta_1 t+\beta_2 g+\beta_3 t\times g)}{1+\exp(\beta_0+\beta_1 t+\beta_2 g+\beta_3 t\times g)}, \quad (1)$$

where β_0 refers to the intercept, β_1 to the association between the ability and the score on the item, β_2 to the impact, and β_3 to the DIF. Individual parameters in the model are tested using a Wald statistic, which is the ratio of an estimate to its standard error and which has an asymptotic χ^2-distribution. Model improvement by including the group and the group-by-total score effects are also evaluated by testing the differences in the log likelihood functions of full and reduced models.

An IRT approach

Equation 1 can be easily extended from a logistic regression model on the total test score to an IRT model with as predictor the latent trait θ instead of the score *t*. In IRT models, the characteristics of the test items are described using a set of item parameters and a functional form that relates a vector of person parameters to the probability of a correct response to each item. As a general model, we used the compensatory two-parameter logistic (2PL) multi-dimensional model (MIRT) (Reckase and McKinley, 1991), which is a multivariate extension of the 2PL logistic model. In this model, the probability that a person *j* answers item *i* correctly is written as:

$$P(y_{ij}=1|\theta_j)=\frac{\exp(\sum_{k=1}^{K} a_{ik}\theta_{jk}-d_i)}{1+\exp(\sum_{k=1}^{K} a_{ik}\theta_{jk}-d_i)}, \quad (2)$$

where y_{ij} is the score on item i (i=1,…,I) by person j (j=1,…,J), a_i is a vector of item discrimination parameters a_i = (a_{i1}, a_{i2}, …, a_{ik})', d_i is a scalar parameter related to item location, and θ_j is a vector of ability parameters for person j on k dimensions (k=1,…,K), θ_j =(θ_{j1} , θ_{j2},…., θ_{jk}). The θ follow a multivariate normal distribution, with zero mean and variances restricted to one, and with covariances as free parameters. The model in Equation 2 reduces to a simple structure MIRT model when item domains do not overlap. For a two-dimensional test, a simple dimensional structure would imply that each item has one location parameter (d_i), but two discrimination parameters (a_{i1} and a_{i2}) one of which is zero. The model further simplifies and becomes a unidimensional 2PL model when k = 1.

In the present paper, the procedure NLMIXED from the general statistical package SAS was used to estimate the different IRT models (De Boeck & Wilson, 2004). For the analysis we used a quasi-Newton-Raphson optimization technique and a nonadaptive Gauss-Hermite approximation with 10 quadrature points for each dimension (SAS V8, 1999).

Item effects

For the modeling of item location parameter (d_i), two types of item effects were used depending on the model design: (1) item main effects, τ_i, and (2) item-by-grade interaction effects, τ_{ic} (items nested within groups), grade groups being denoted by c = 1,…,C.

Item main effects were estimated as item-specific weights:

$$\tau_i = \sum_{h=1}^{H} \tau_h X_{ih}, \quad (3)$$

where design facets are denoted by h = 1,…,H, so that H=I, and X is an identity matrix with an indicator vector for each item, so that X_{ih}=1 if i=h, and X_{ih}=0 otherwise.

Item-by-grade interaction effects were estimated as an item-specific weight within grades:

$$\tau_{ic} = \sum_{h=1}^{H} \tau_{hc} X_{ih}, \quad (4)$$

where X is again an identity matrix.

In fact, for the purpose of the present manuscript, X is transformed into an effect-coded matrix (Cohen & Cohen, 1983) minus one column, because the mean is taken care of by the rest of the model (see below in Table 1). The effect of item i on the dependent variable is parameterized as the deviation of item i from the expectation based on the other terms of the model. This parameterization nicely separates the two intended levels of DIF (domain level and item level).

Formulation of the model with DIF at both levels: the complete model

The equation for the location of an item from domain k presented to grade g is

$$d_{i(k)c} = \beta_0 + \beta_1 \nu_k + \beta_2 g + \beta_3 \nu_k \times g + \tau_{ic}\ , \quad (5)$$

where β_0 is the overall mean, β_1 is the domain main effect, β_2 is the grade main effect, β_3 is the grade-by-domain interaction effect, and τ_{ic} is the item effect nested within grade. In the context of a DIF analysis, β_3 refers to domain-related DIF and τ_{ic} refers to item-specific DIF

within grades. In Equation 5 the group (Grade 3 vs. Grade 4), and item domain (open versus closed syllables) are contrast coded with

$$g = \begin{cases} 1 & \text{if the examinee is from grade 3,} \\ -1 & \text{if the examinee is from grade 4,} \end{cases}$$

and,

$$\nu = \begin{cases} 1 & \text{if the item is from the open syllables domain,} \\ -1 & \text{if the item is from the closed syllables domain.} \end{cases}$$

Table 1 shows a schematic representation of the overall frame of the design matrix for the complete model with domain-by-grade interaction effect, and grade-nested item effects. The first column refers to the intercept, followed by the grade factor, the domain factor, and their interaction, and finally the effect coded matrix *X* for the item effects within grades. A simple structure two-dimensional IRT model was estimated, consistent with the test design (each item loaded into either the open syllables factor or the closed syllables factor, not on both factors).

Insert Table 1 about here

Formulation of the other models

Similar to the design for the complete model with DIF at both levels, *the model with DIF only at the domain level* is defined by replacing the nested item effects in the complete model by item main effects:

$$d_{i(k)} = \beta_0 + \beta_1 \nu + \beta_2 g + \beta_3 \nu \times g + \tau_i . \quad (6)$$

This means that, the domain-related DIF (β_3) is retained in the model, while item-specific DIF within grades (τ_{ic}) are replaced by item main effects (τ_i), and again two ability parameters are used for each person.

The model with DIF at the item level can be estimated by omitting the domain main effect and the domain-by-grade interaction effect from the complete model:

$$d_i = \beta_0 + \beta_2 g + \tau_{ic}. \quad (7)$$

This means that the model with DIF at item level includes the grade main effect (β_2) along with item-specific DIF within grades (τ_{ic}), and only one θ is used.

Finally, a base model was estimated using the 2PL model, with

$$d_i = \beta_0 + \tau_i, \quad (8)$$

retaining item main effects (τ_i) only.

The different models were compared using the negative log likelihood and Akaike's Information Criterion (AIC; Akaike, 1977; Bozdogan, 1987).

Results

Exploring the spelling data

Figure 1 plots the means of item proportions correct (MPC) over the 13 items of each subdomain conditioned on the observed number correct items on the total test. The higher values of the MPC for the open syllables items in Figure 1a reflect the domain main effect: applying the vowel rule in open syllables is on the average easier than applying the consonant rule in closed syllables. The difference between the two domains is more pronounced for low ability students and diminishes toward the end of the scale. The

difference between the two subdomains would disappear towards the lower end as well if there had been low scores observed. It is a necessary consequence of using the number correct score on the absis. Figure 1a suggests that inferences on the level of proficiency of the spelling domain based on the total test score can be misleading because the total score is not a good indicator for the level of proficiency in both subdomains. A solution, therefore, could be to derive two scores, one for the open syllables domain and one for the open syllables domain. However, Figure 1b shows that the conditional MPC lines are not parallel for the two groups, that is, there is an interaction effect between domain and group factors. While the students of the fourth grade are performing better on closed syllables on the average, the students of the third grade are performing better on the open syllables, when one controls for the total number of correct items.

The MPC averaged over all items from the two subscales in Figure 1 illustrate that there are domain-specific within-group variations. In fact, item-specific within-group variations can also be shown to be present in the data, which is in agreement with the item-specific interaction model.

Insert Figure 1about here

Estimated Models

Table 2 presents the goodness of fit and the parameter estimates for the four models, except those of the individual item effects. The 2PL model was estimated to function as a base model. The goodness of fit improves considerably when a grade main effect and item-specific interactions are included to the base model, forming the item level DIF model. The effect of grade is large and positive, pointing to the progress in spelling made from grade 3

to grade 4. This effect alone cannot explain the difference in the goodness of fit, because for a model with the main effect and without the item-specific interactions, the values of -2* Log likelihood and the AIC are 12411 and 12517, respectively.

Insert Table 2 about here

The likelihood of the (two-dimensional) model with DIF at the domain level is only slightly better than that of the model with DIF at the item level. It is again clear from the results that grade 4 performs better than grade 3 (β_3=0.569, p<0.001), but the domain seems to matter too (β_3=0.336, p<0.001), indicating that the progress from grade 3 to Grade 4 is larger for the closed syllables domain than for the open syllables domain. Overall, open syllables items are easier than the closed syllables items (β_2=0.806, p<0.001). The interaction can be understood as grade 3 making a leap forwards in the closed syllables domain while for the open syllables domain a high level of proficiency was already reached in grade 3, so that not much further improvement is required. The fact that the likelihood of the domain level model is about as good as the item level model indicates that the DIF added to the unidimensional model does not really improve goodness of fit if the domain-by-group interaction is included, and the model is made into a two-dimensional model. The two dimensions are rather highly correlated (r=0.625, p<0.001), but it pays off to distinguish between the two.

Finally, when the item-by-group effects are added to the domain level model forming the complete model, a much better goodness of fit is obtained, both for the

likelihood and the AIC. This implies that both item main and item-by-group interaction effects are evident, one at the domain level and another at the item level within domains. The estimates of the other effects of the complete model are in line with their estimated values in the other models: grade 4 performs better than grade 3 ($\beta_1 = 0.777$, $p < 0.001$), open syllables items are easier ($\beta_2 = 0.827$, $p < 0.001$), and the progress from grade 3 to grade 4 is larger for the closed syllables domain ($\beta_3 = 0.397$, $p < 0.001$).

Grade effects

Table 3 summarizes the expected group means of the item location parameters as estimated by the four models on which individual item location parameters were anchored. The 2PL model predicts the mean of the item location parameters to be approximately -0.84 over grades, which means that the test is rather easy on the average. The item level model including the grade effect alone, and allowing item effects to be estimated within grade levels, predicts the mean item location parameters of items as -0.17 and -1.65 for grade 3 and grade 4, respectively. The expected grade impact with respect to the total test score (the difference between these two means) is, therefore, 1.48. The domain level model simultaneously including both grade and domain effects, and allowing item effects to be estimated within domain items, predicts the grade impact to be 1.81 for the closed syllables and 0.47 for the open syllables domains, taking the domain-by-grade interaction effect of 1.35 into account. The grade impact clearly varies when domains are taken into account.

The complete model including both grade and domain effects as well, but also allowing item effects to be estimated for each domain within grades, predicts the grade

impact to be 2.35 for the closed syllables and 0.33 for the open syllables domains. The interaction effect estimated by the complete model is 1.59, and greater than that of the domain level model, showing that there is a greater grade contrast for the closed syllables domain than there is for the open syllables domain when both grade and domain factors are considered simultaneously.

Insert Table 3 about here

Discussion

A test design approach, which is commonly used to infer diagnostics on *relevant aspects* used in the process of solving items (Embretson, 1985), was extended to a DIF study, which is typically done to infer diagnostics on *irrelevant aspects* used in this process. We specifically investigated whether spelling ability can be conceptualized as a multi-dimensional construct to explore domain-related aspects of differential item functioning over grade groups.

The results from the spelling proficiency scale confirm that spelling ability can be conceived as a composite construct in the onset of development. The spelling ability measure modeled in this study was a composite from two subdomains: the open syllables domain and the closed syllables domain. Items in the two domains functioned differently in the two adjunct grade groups leading to the investigation of DIF at two levels: at the item level and at the domain level. There was a grade impact, as expected. However, the impact effect was different for the two subdomains, such that, if overall scores were computed

neglecting the domain factor, impact would be overestimated for items in the closed syllables domain and underestimated for the items in the open syllables domain.

When the total test score seems to measure a different composite over groups, neglecting domain-related factors would imply DIF that is *domain-specific*. The situation becomes even more complicated when there are also item-by-group effects. Our results show that, when evident, item-by-group interactions must be fine-tuned further within domains.

Comparing models with domain-related and domain-free item effects would provide reliable diagnostic information about DIF. For example, children in primary schools in Flanders do not only come from homes where Dutch is the primary language, but also from homes where another language is the primary spoken language, such as Moroccan or Turkish. As a further study, models with item-specific effects can be formed by implementing other language subgroups to investigate potential threats for the validity of test scores.

The proposed approach of estimating item and domain level effects simultaneously over subpopulations appears to be promising as an explanatory tool relating psychometric and substantive aspects of language assessment. Besides the gain in the goodness of fit, interaction terms extracting auxiliary information might also purify item functions, leading to a better understanding why some items function differently across languages and cultures.

References

Akaike, H. (1977). Factor analysis and AIC. *Psychometrika, 52,* 317-332.

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29,* 67-01.

Borsboom, D., Mellenbergh, G. D. & van Heerden, J. (2002). Different kinds of DIF: A distinction between absolute and relative forms of measurement invariance and bias. *Applied Psychological Measurement, 26,* 433-450.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52,* 345-370.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement, 42,* 133-148.

Cohen, J. & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences.* London: Lawrence Erlbaum.

De Ayala, R. J., Kim, S., Stapleton, L. M., & Dayton, C. M. (2002). Differential item functioning: a mixture distribution conceptualization. *International Journal of Testing, 2*, 243-276.

De Boeck and Wilson (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer-Verlag.

Engelhard, G. (1992). The measurement of writing ability with many faceted Rasch model. *Applied measurement in education, 5,* 171-191.

Embretson, S. E. (1985). Introduction to the problem of test design. In S. E. Embretson (Ed.), *Test Design: Developments in psychology and psychometrics* (pp. 3-17). New York: Academic Press.

Gelin, M. N & Zumbo, B. D. (2003). Differential item functioning results may change depending on how an item is scored: An illustration with the center for epidemiologic studies depression scale. *Educational and Psychological Measurement, 63*, 65-47.

Hambleton, R. K., & Swaminathan, H. (1985*). Item response theory. principles and application*. Boston: Kluwer-Nijhoff.

Johanson, G. & Asmaldi, A. (2002). Differential person functioning. *Educational and Psychological Measurement, 62*, 435-443.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lubke, G. H. & Muthen, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological methods, 10*, 21-39.

Meulders, M. & Xie, Y. (2004) Person-by-item predictors. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp.213-240). New York: Springer-Verlag.

Miller, T. R. & Spray, J. A. (1993). Logistic discriminant function analysis for dif identification of polytomously scored items. *Journal of Educational Measurement, 30*, 107-122.

Mislevy, R. J., & Bock, R. D. (1984). *BILOG Version 2.2*: Item analysis and test scoring with binary logistic models [Computer program]. Mooresville, IN: Scientific Software.

Mislevy, R.J., Johnson, E.G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics, 17*, 131-154.

Raju, N. S., Laffitte, L. J. & Byrne, B. M., (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87*, 517-529.

Reckase, M. D. & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 15*, 361-373.

Rogers, H. J., Swaminathan, H. (1993). A comparison of logistic regression and Mantel Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement. 17*, 105-116.

SAS Institute (1999) SAS *Online Doc* (Version 8). Cary: SAS Institute Inc.

Swaminathan, H. & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361-370.

Vangeneugden, E. (2004). *De verdere ontwikkeling van de criteriumtoets SPOMEK voor het lager onderwijs* [The further development of the criterion test SPOMEK for primary education]. Unpublished master's thesis. Department of Educational Sciences, Katholieke Universiteit Leuven, België.

Wang, W., & Wilson, M. (2005). Assessment of differential item functioning in testlet-based items using the Rasch testlet model. *Educational and Psychological Measurement, 65*, 549-579.

Acknowledgement
We wish to thank Evi Vangeneugden and her promoter Pol Ghesquière for the use of their data set. The present study was supported by grant GOA/00/02 (ZKA4511) and by a postdoctoral fellowship of the Fund for Scientific Research – Flanders (Belgium) to the last author.

Address correspondence to Nilufer Kahraman,
NBME 3750 Market Street, Philadelphia PA 19104
Phone: 215 590 9728; e-Mail: nkahraman@nbme.org

Table 1.
Design matrix of the complete model for the 13 closed syllables and 13 open syllables items

Items	1	Contrast Codings				Item Effect Codings			
		g	v		gxv	X_1 X_2 .... X_{12}	X_{13} X_{14} .... X_{24}	X_{25} X_{26} ... X_{36}	X_{37} X_{38} … X_{48}
i_1	*1*	*1*	**closed syllables**	*1*	*1*	*1 0 0 …. 0*			
i_2	*1*	*1*				*0 1 0 …. .*			
.	.			.	.	*0 0 1*			
.	.			.	.	. *1 …. .*	*0*	*0*	*0*
		.				. *1 0*			
.		.				. *1*			
i_{13}		.		*1*	*1*	*-1 -1 . . . . . -1 -1*			
i_{14}		**G3**	**open syllables**	*-1*	*-1*		*1 0 0 …. 0*		
i_{15}							*0 1 0 …. .*		
.				.	.		*0 0 1*		
.				.	.	*0*	. *1 …. .*	*0*	*0*
.	.						. *1 0*		
	.	*1*					. *1*		
i_{26}	*1*	*1*		*-1*	*-1*		*-1 -1 . . . . . -1 -1*		
i_1	*1*	*-1*	**closed syllables**	*1*	*-1*			*1 0 0 …. 0*	
i_2	*1*	*-1*						*0 1 0 …. .*	
.	.			.	.			*0 0 1*	
.	.			.		*0*	*0*	. *1 …. .*	*0*
		.			.			. *1 0*	
.		.						. *1*	
i_{13}		.		*1*	*-1*			*-1 -1 . . . . -1 -1*	
i_{14}		**G4**	**open syllables**	*-1*	*1*				*1 0 0 …. 0*
i_{15}									*0 1 0 …. .*
.				.	.				*0 0 1*
.				.	.	*0*	*0*	*0*	. *1 …. .*
.	.								. *1 0*
	.	*-1*							. *1*
i_{26}	*1*	*-1*		*-1*	*1*				*-1 -1 . . . . -1 -1*

* Each cell represents a domain with 13 items (rows) and 13-1 coded item effects (columns).

** Italics indicate the entries in the operational design matrix.

Table 2.
The Four Models with Fit Indices and Individual Test Statistics

Models	**Model Fit Indices** *-2LogL AIC*	**Number of parameters in the model**			**Test-Level Parameters (Std. Dev.)**				**Correlation**
					Intercept	*Grade Contrast* g3 → 1 g4 → -1	*Dimension Contrast* MV → 1 KV → -1	*G-by-D Contrasts* g3 MV → 1 g3 KV → -1 g4 MV → -1 g4 KV → 1	
		Test Level	*Item Level*	*Total*	β_0	β_1	β_2	β_3	
1. 2PL	12426 12530	1	51	52	-0.835* (0.132)				
2. Item-level	12143 12299	2	76	78	-0.912* (0.057)	0.742* (0.057)			
3. Domain-level	12131 12237	3	50	53	-0.940* (0.064)	0.569* (0.062)	0.806* (0.044)	0.335* (0.039)	0.625* (0.015)
4. Complete	*11951* *12109*	5	74	79	-0.979* (0.065)	0.777* (0.046)	0.827* (0.064)	0.397* (0.043)	0.624* (0.041)

* $p<0.0001$

** Individual item main effects, τ_i's of the 2PL and the item level models, or item-by-grade effects τ_{ic}'s of the domain level and the complete models are not shown here.

*** We also estimated the complete model allowing item discrimination parameters vary over grade levels, however, the improvement observed was not of practical importance given the parsimony of the model presented here (the values of -2* Log likelihood and the AIC were 11923 and 12133, respectively).

Table 3.
Group means of the item location parameters estimated by the four design models

Models	**Complexity factors -item effects**	**Domain Scale**	**Means**		**Mean difference between grades**	**Interaction effect**
			Grade 3	**Grade 4**		
1. 2PL	-Items	Closed Syll. Open Syll.	-0.835			
2. Item-level	Grade -Items within grades	Closed Syll. Open Syll.	-0.170	-1.654	1.484	
3. Domain-level	Grade Domain Domain by Grade -Items	Closed Syll. Open Syll.	0.771 -1.513	-1.040 -1.978	1.811 0.465	1.346
4. Complete with two levels	Grade Domain Domain by Grade -Items within grades	Closed Syll. Open Syll.	1.021 -1.427	-1.327 -2.186	2.348 0.329	1.588

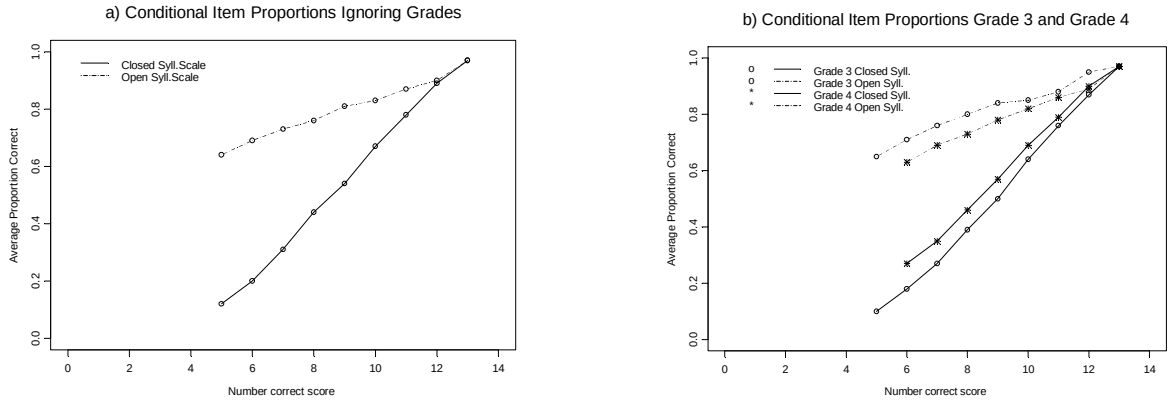

Figure 1. Conditional mean item proportions for the Closed Syllables and the Open Syllables Scales (Smoothed)