Methods of Psychological Research Online 2001, Vol.6, No.2
Internet: http://www.mpr-online.de
Institute for Science Education
© 2001 IPN Kiel

Non-modeled item interactions lead to distorted discrimination parameters: A case study

Francis Tuerlinckx and Paul De Boeck[1]

Abstract

In this paper it is shown that differences in item discrimination may be explained in some situations by means of non-modeled interactions between the items. If two items have large estimated discrimination parameters, this may be the result of a positive interaction between them. Conversely, if two items have small estimated discrimination parameters, this may be the result of a negative interaction between them. The effect is shown to be present in two data sets.

Keywords: Item interactions; local item dependencies; local stochastic independence; two parameter logistic model.

---

[1] Author's addresses: Francis Tuerlinckx, Department of Psychology, University of Leuven, Tiensestraat 102, 3000 Leuven, Belgium (mailto:francis.tuerlinckx@psy.kuleuven.ac.be) and Paul De Boeck (mailto:paul.deboeck@psy.kuleuven.ac.be).
The first author was a Research Assistant of the Fund for Scientific Research - Flanders (Belgium). The research is funded by the GOA-2000/2 grant from the K.U.Leuven. We would like to thank Laurence Claes and Piet J. Janssen for the use of their data, collected in research project funded by the OBPWO grant 94.14 of the Ministery of the Flemish Community, Department of Education, awarded to Piet J. Janssen.

1. Introduction

A widely used model from item response theory (IRT; Embretson & Reise, 2000) is the two parameter logistic model (2PLM; Birnbaum, 1968):

$$\Pr(X_i = 1|\theta) = \frac{\exp\left(\alpha_i(\theta - \beta_i)\right)}{1 + \exp\left(\alpha_i(\theta - \beta_i)\right)} \tag{1}$$

where X_i is a random variable denoting the response on item i (i=1,...,n), θ is the ability of a person, β_i is the difficulty of the item and α_i is the item discrimination parameter. If α_i=1 then the model in Equation 1 reduces to the well-known Rasch model (Rasch, 1980). Although it will be assumed that we are dealing with ability items, our research is not restricted to those items only.

An important assumption of the 2PLM (and many related models) is local stochastic independence (LSI). This means that given the latent trait value θ, all association between the items should disappear. Stated otherwise, the latent trait is the only reason why items are associated. If the assumption of LSI is violated, some dependencies between items remain after controlling for the latent trait and these dependencies are called local item dependencies (LIDs). LIDs may be the result of undetected multidimensionality, but in this paper we focus on another cause of LIDs: non-modeled interactions between the items. In this paper, only interactions between pairs of items are considered for simplicity.

An interaction between two items occurs when the response on one item is directly related to the response to another item (above the common influence of the latent trait). An appropriate model for item interactions has been proposed by Hoskens and De Boeck (1995, 1997). For simplicity, it is assumed that there is an interaction between the first two items (items 1 and 2). The remaining items (items 3 to n) are considered to be local stochastic independent of each other and of the pair of interacting items 1 and 2. Next, we define the vector of random variables (X_1, X_2,..., X_n) denoting the responses on the n items. The probability for a realization (x_1, x_2,...,x_n) equals:

$$\begin{aligned}\Pr(X_1 = x_1, X_2 = x_2, ..., X_n = x_n|\theta) = \\ = c^{-1}\exp\left(x_1\alpha_1(\theta - \beta_1) + x_2\alpha_2(\theta - \beta_2) + x_1x_2(-\beta_{12}) + ... + x_n\alpha_n(\theta - \beta_n)\right)\end{aligned} \tag{2}$$

where β_{12} is the interaction parameter, θ is the person ability, β_i is the difficulty uniquely associated with item i, and α_i is the discrimination parameter of item i. More-

over, c is the proportionality constant that makes $\Pr(X_1 = x_1, X_2 = x_2, ..., X_n = x_n|\theta)$ sum to one over all response patterns.

The interaction parameter β_{12} can be interpreted by considering the following log odds ratio :

$$\begin{aligned}
&\log\left(\frac{\Pr(X_1=1,X_2=1,X_3=x_3,...,X_n=x_n|\theta)\Pr(X_1=0,X_2=0,X_3=x_3,...,X_n=x_n|\theta)}{\Pr(X_1=1,X_2=0,X_3=x_3,...,X_n=x_n|\theta)\Pr(X_1=0,X_2=1,X_3=x_3,...,X_n=x_n|\theta)}\right)\\
&=\log\left(\frac{\Pr(X_1=1,X_2=1|\theta)\Pr(X_1=0,X_2=0|\theta)}{\Pr(X_1=0,X_2=1|\theta)\Pr(X_1=1,X_2=0|\theta)}\right)\\
&=-\beta_{12},
\end{aligned} \tag{3}$$

for any response vector $(x_3,...,x_n)$. If $\beta_{12}<0$, then we have a positive interaction between items 1 and 2 because the probability of having both items correct (or incorrect) increases compared to the case where there is LSI ($\beta_{12}=0$). If $\beta_{12}>0$, then we have a negative interaction because the probility of having both items correct (or incorrect) decreases compared to the case of LSI ($\beta_{12}=0$). Note that if $\beta_{12}=0$, Equation 2 simplifies to the product of n 2PLM items (see Equation 1). In the following, we continue to work with an interesting special case of Equation 2, where all discrimination parameters α_i are equal to 1.

It is shown in many studies (see e.g., Thissen, Steinberg, & Mooney, 1989 and Yen, 1993) that the presence of LIDs may have a distorting influence on the estimated discrimination parameter if the LIDs are not taken into account. The main empirical finding is that the item discrimination is overestimated if positive LIDs are not considered. Extending the previous studies, Tuerlinckx and De Boeck (2001) showed that ignored item interactions between items (see Equation 2) can lead to distorted discrimination indices, and that this cannot always be detected with a test for global model fit. Specifically, Tuerlinckx and De Boeck (2001) showed that if $\beta_{12}<0$ in Equation 2 and if such data that contain interactions are modeled with a 2PLM, the discrimination parameter is overestimated (i.e., larger than its true value). The reverse happens for $\beta_{12}>0$: If such a negative interaction is neglected, the item discrimination is underestimated.

In this paper, we will show in two real data sets that ignored interactions can lead to biased discrimination parameters. Moreover, we will try to go one step further and show that a well-fitting 2PLM with differing discrimination parameters sometimes may hide the interactions in the data. In that case, the problem is approached from the opposite direction: Not the effect of ignored interactions on the estimated discrimination parameters is shown but some alternative explanations for high and low discrimination parame-

ters are given by taking into account interactions between items in a real data set. The applications are chosen to show that for real data that can be fitted with an equal discrimination model with interaction, also the 2PLM can fit the data, with differences in the item discriminations (and with no interactions).

In search for item interactions, we focused on sets of items that share a common reading passage since that is an obvious place for interactions between items to occur. It is plausible that for items related to the same text, interactions have a higher plausibility than LSI. However, it remains an empirical issue whether this is really the case.

The remainder of the paper is organized as follows. First, the data that will be analyzed are discussed. Next, the method of analysis is explained and thereafter the results are shown. Finally, the findings of our research are discussed.

2. Data

The data come from a Dutch test that is taken from 441 first-year law students. The test was part of a research project funded by the Flemish Department of Education (Claes, Janssen, & Wels, 1999). The purpose of the research project is to inform last-year high school students about how well their capacities match with the requirements of the chosen study (in this case, law).

The test consists of eleven texts from first and second year undergraduate law courses. Each text is followed by six or eight multiple-choice questions (with four alternatives) about the text[2]. The anwers of the subjects were coded as wrong or correct. Of the eleven subtests, two are chosen because interactions were found among the items: one about the illegal act as a source for contracts and one on the president and separation of powers in the US. Both these texts are followed by six questions.

[2] Although the items have a multiple choice format, we will not model the data with a three parameter logistic model (3PLM; Embretson & Reise, 2000) because of two reasons. First, if one tries to estimate the lower asymptote parameter, convergence problems occur very often (difficulty parameters go to plus or minus infinity or lower asymptotes go to zero). Second, even if the parameters are constrained to be equal (all lower asymptotes equal to each other, or all item discriminations equal to each other), the convergence problems do not always disappear. Moreover, if the problems disappear, the final model does not fit the data better than the other models we considered (see Results section).

3. Method

Our method of analysis for each set of questions differs on two points from what would be generally done with such data. First, this kind of data is mostly analyzed through a so-called score-based approach. In a score-based approach, the set of questions that follow the same reading passage is taken as the unit of analysis. Such a set of items is often called a testlet (Wainer & Kiely, 1987). Doing so, the sets of questions belonging to a text are analyzed together by modeling the sum score of the testlet with a polytomous model (the categories of the polytomous item are the possible realizations of the sum score). Modeling only the sum score results in an information loss since the response to the individual items cannot be recovered from the sum score. Contrary, in this paper we have chosen to model the full response patterns on a set of interacting items and not only the sum score. Taking this item-based approach, no information is lost.

Second, the data for each text are analyzed separately, because a joint analysis might require a multidimensional model, which may complicate things without resulting in a better understanding of the topic under study. Performing two within-text analyses allows us to give a more detailed account of the response process by considering interactions between individual items without the concern of multidimensionality.

The analysis of the data comprises three parts and these will be discussed in the subsequent subsections.

3.1. Part 1: The 2PLM part

In a first part of the analysis, each text is first analyzed with the Rasch model and then with the 2PLM. Both models are fitted by means of MULTILOG (Thissen, 1991) which yields (marginal) maximum likelihood estimates (MML; Thissen, 1982) for the item parameters of the 2PLM. As default, it is assumed in MULTILOG that the latent trait values θ are drawn randomly from a standard normal distribution. The output from MULTILOG contains also the likelihood ratio test statistic (denoted by G^2). Under the null hypothesis that the model is true, G^2 is asymptotically chi-square distributed with degrees of freedom the number of free response patterns minus the number of free parameters in the model.

A marginal parametric model (the model from Equation 1 together with a normal distribution for θ) is assumed in MULTILOG for the latent trait values, so that the test

of the model is also sensitive to wrong specifications of the latent trait distribution. The default density for latent trait values used by MULTILOG is the normal density. However, if the true latent trait density deviates from the normal density in skewness or tail heaviness, then one may shift to a density from a very general family, called the Johnson family (see Thissen, 1991). Working with the Johnson family allows for greater flexibility in finding an appropriate latent trait density. All reported analyses in this chapter are performed with the normal density as marginal density for the person parameters, unless indicated otherwise.

3.2. Part 2: The interaction part

In a second part of the analysis, loglinear IRT models are applied to fit the interaction models. As has been shown by Kelderman (1984) and Duncan (1984), the Rasch model and models with item interactions may be formulated as (quasi)-loglinear models (Agresti, 1990; Bishop et al., 1975). This derivation of a loglinear model from an IRT model is, however, only possible if a sufficient statistic exists for θ, so that by conditioning on that sufficient statistic, θ disappears from the probability formulas. For the Rasch model, this sufficient statistic is the sum score.

However, the translation of the general model presented in Equation 2 into a loglinear model is impossible because in this model there are no sufficient statistics for θ due to the presence of the discrimination parameters. Therefore, as said before, we restrict all discrimination parameters of the interaction model to be equal to 1. This restriction allows us to obtain item parameter estimates of the model in Equation 2 by applying a loglinear model on the incomplete [item 1 x item 2 x ...x item J x sum score]-table. A main effects model for this table corresponds to a loglinear Rasch model but by specifying interaction effects between items, a loglinear version of an interaction model is obtained. Details about how to estimate loglinear IRT models will not be given here. A theoretical account for deriving loglinear IRT models can be found in Kelderman (1984) and some practical guidelines for fitting IRT models with loglinear procedures are supplied by Ten Vergert, Gillespie, and Kingma (1993).

For the loglinear IRT analyses, the loglinear Rasch model will be fitted first and subsequently, an appropriate model with item interactions will be searched for. The loglinear interaction models are tested with the likelihood ratio test statistic (denoted by G^2). The inclusion of interaction effects is done by a forward stepwise selection procedure, starting with the Rasch model and in the next steps interactions that resulted in a sig-

nificant increase in the likelihood ratio test statistic were included. This is possible because we only consider hierarchical loglinear models so that previously formulated models are nested within subsequently formulated ones.

The aim of this subsequent loglinear analysis is to show whether differences in discrimination as found with MULTILOG can be explained as stemming from ignored interactions between items. If this is the case, then an equal discrimination model with interactions is also able to fit the data.

There are no nesting relations between the loglinear IRT models and the IRT models estimated with MULITLOG and therefore, the asymptotic distribution of the difference in the likelihood ratio test statistics is unknown. For assessing the relative fit of non-nested models, Akaike's Information Criterion (*AIC*; Akaike, 1977) can be used. The *AIC* is defined as minus two times the log likelihood plus two times the number of estimated parameters. The model with the smallest *AIC* is selected as the best fitting model.

3.3. Part 3: The LID index part

The third part of the analysis, called the *LID index part*, is intended to check which item pairs show LID, using indexes proposed in the literature. The indexes are calculated for the 2PLM that would fit the data, because, if the indexes can detect LID while the 2PLM seems to fit (based on a global test statistic), then it is possible that interactions may be the cause of differences in item discrimination. Chen and Thissen (1997) have proposed and evaluated several indexes for the detection of LIDs and thus these indexes may also be used for detecting item interactions. From the five indexes evaluated by Chen and Thissen (1997), two are used in this paper to check for item interactions. The first index is G^2_{LID} statistic, a likelihood ratio statistic, computed for every pair of items. For every possible pair of items, one can build a contingency table with the observed frequencies and another table with the expected frequencies from the estimated 2PLM. If the local stochastic independence assumption of the 2PLM is correct, the observed frequencies should not deviate much from the expected frequencies and this deviance can be measured with the G^2_{LID}. Using simulations Chen and Thissen (1997) show that G^2_{LID} is distributed chi-square with approximately one degree of freedom under the null hypothesis of no LIDs (the actual degrees of freedom is somewhat less that one).

The second statistic is the Q_3 *statistic*, originally proposed by Yen (1984). The Q_3 is the correlation between the residuals of two items. For every person, the expected probability of a correct response under the model is subtracted from the observed response on item i. The same is done for item j, and subsequently the correlations between these residuals on items i and j are correlated over persons. As a rule of thumb for the evaluation of Q_3, one can consider pairs of items with a Q_3 index larger than 0.2 or smaller than –0.2 as suspected (although this procedure is rather conservative, see Chen & Thissen, 1997). The sign of the Q_3 may also be used to infer whether the dependency between two items is negative or positive. The indexes of Chen and Thissen (1997) can be computed easily subsequently to the parameter estimation with MULTILOG using the program IRTNEW (Chen, 1993) that can be downloaded for free from http://www.unc.edu/~dthissen/dl.html.

4. Results

4.1. Results for Text 1

The results of the analysis for the first text (on illegal acts) are shown in Table 1. In the first column one can find the different models that are estimated. The three parts in the data analysis that were identified in the section on the Method will be discussed separately here.

4.1.1. Part 1: The 2PLM part

In the first part of the analysis, the Rasch model is estimated with MULTILOG (denoted by 'MULTILOG-Rasch' in Table 1) and this does not yield a satisfactory fit. Second, MULTILOG is used again but now the item discriminations are allowed to vary (referred to as 'MULTILOG-2PLM'). This model yields a satisfactory fit. In the last columns of Table 1 one can find the estimated item discriminations for the six items. These are all one for the Rasch model but differ for the 2PLM. The fifth column of Table 1 contains the *AIC* value for the different models. If only these MULTILOG analyses were performed, the 2PLM would be chosen as the best fitting model.

Table 1: Models fitted for Text 1.

MODEL	G^2	*df*	*p*	*AIC*	Item Discriminations					
MULTILOG - Rasch	88.3	56	.00	102.3	1	1	1	1	1	1
MULTILOG - 2PLM	60.6	52	.17	84.6	1.16	1.71	0.33	1.53	0.13	1.25
LOG - Rasch	65.8	52	.10	89.8	1	1	1	1	1	1
LOG - 1x2	46.2	49	.59	76.2	1	1	1	1	1	1
4x6										
1x5										

Note. LOG = loglinear model; *i*x*j* denotes an interaction between items *i* and *j*; G^2 = likelihood ratio chisquare; *df* = degrees of freedom; *p* = *p*-value; *AIC* = Akaike's Information Criterion.

4.1.2. Part 2: The interaction part

The next model that is fitted is again the Rasch model, but now by means of a loglinear IRT procedure (the term 'LOG' in the Table 1 refers to the fact that a loglinear model is estimated). The loglinear Rasch model fits the data but it has a larger *AIC* value than the 2PLM, which is the favorable model of the two. Finally, a loglinear IRT model with interactions between items is estimated and tested. The included interactions are indicated in the table after the word 'LOG'. For instance, '1x2' denotes the interaction between the first two items. The chosen interaction model contains 3 interactions between pairs of items: an interaction between items 1 and 2, between items 1 and 5, and between items 4 and 6. The model provides a good fit to the data. The AIC for the interaction model is the lowest among all estimated models; hence, this model is the preferred one for the first text.

The left part of Table 2 shows the parameter estimates for the best fitting loglinear model (with three two-item interactions) for the first text. It can be seen that two interactions are positive (because of the negative value of the β_{ij}-parameter) and that the interaction between items 1 and 5 is negative (because of β_{ij}=0.616 but notice the large *SE*). The items showing a positive interaction all have higher estimated discrimination parameters in the 2PLM. Item 4 and item 6 are involved in a positive interaction and their degrees of discrimination are 1.71 and 1.53 respectively. Item 1 and item 2 are also involved in a positive interaction, and they have estimated discrimination parameters of 1.16 and 1.71, respectively. However, the estimated item discrimination of item 1 is not as high as the item discrimination of items 2, 4 or 6, because it is also involved in a negative interaction with item 5. As a consequence, item 5 has a very low item dis-

crimination (0.13). Hence, the differences in item discriminations can be explained by taking into account the interactions between items.

4.1.3. Part 3: The LID index part

From the local dependence indexes (Chen & Thissen, 1997) that were computed, G^2_{LID} only flags a dependency between items 1 and 5 at the α=.05 level. The absolute value of the Q_3 never exceeds 0.2, even not for the negatively interacting items 1 and 5. The sign of Q_3 for items 1 and 5 is negative, indicating that there is a negative dependency between these two items. The significant interactions between items 1 and 2 and between items 4 and 6 are not detected by the LID indexes.

Table 2: Parameter estimates for the loglinear IRT models with interactions for two texts.

Text 1 LOG 1x2,4x6,1x5		Text 2 LOG 1x6,4x5	
Parameter	Estimate (SE)	Parameter	Estimate (SE)
β_1	0.513 (0.330)	β_1	–0.492 (0.191)
β_2	–0.805 (0.329)	β_2	0.663 (0.161)
β_3	1.273 (0.291)	β_3	–0.180 (0.158)
β_4	–1.099 (0.309)	β_4	–1.340 (0.197)
β_5	1.635 (0.297)	β_5	–1.504 (0.197)
β_6	0.000 (0.000)	β_6	0.000 (0.000)
β_{12}	–1.390 (0.421)	β_{16}	–1.510 (0.240)
β_{46}	–1.219 (0.532)	β_{45}	–0.736 (0.311)
β_{15}	0.616 (0.532)		

Note. The parameter estimates of β_6 are constrained to zero to obtain model identification.

4.2. Results for Text 2

Table 3 contains the fitted models for the second text (on the USA). Again, the three data analysis parts will be discussed separately.

4.2.1. Part 1: The 2PLM part

As can be seen in Table 3, the Rasch model does not fit the data. But, the 2PLM estimated with MULTILOG does not fit the data either. However, since the distribution

of the raw sum scores is bimodal for these data, this distribution could have caused the misfit. To check whether the misfit of MULTILOG can be attributed to a wrong specification of the latent trait density, more flexibility was allowed for the shape of the latent trait distribution by allowing MULTILOG to use any member of the Johnson family. When doing so, the fits for the Rasch model and the 2PLM increase somewhat, but not enough to accept one of the models.

As a further step in testing to what degree the misfit of the 2PLM is due to a wrong specification of the latent trait density, a loglinear analysis is conducted with different item weights. First, it will be explained why a loglinear analysis helps to avoid a wrong specification of the latent trait density, and then how we handled the differential weighting. First, by estimating and testing the model with a loglinear method, no information is used about the sum score distribution (i.e., the sum scores are considered fixed; Kelderman, 1984). Therefore, misfit of the model is not attributable to a wrong specification of the latent trait density. Second, the weights are defined as the estimated item discriminations taken from the MULTILOG analysis but rounded to the nearest number that is a multiplicative of 0.5. These imputed weights are shown in Table 3, for the model 'LOG - 2PLM'. The sufficient statistic for θ is now the weighted sum of the item scores. It can be seen from Table 3 that the loglinear 2PLM model (with imputed item discriminations) has improved the fit a lot and is by far the best fitting model up till now. Moreover, it could be argued that the frequency table is somewhat sparse (62.5% of the response patterns has a frequency lower than 5) and a way to overcome such sparseness is to add a small constant to each frequency. If for instance 0.5 is added to each frequency, the fit of the loglinear 2PLM becomes very good (G^2=49.6, df=42, p=0.20).

An analysis with another IRT program OPLM (One Parameter Logistic Model; Verhelst & Glas, 1995; Verhelst, Glas, & Verstralen, 1995) yields the same conclusion. OPLM uses the conditional maximum likelihood (CML) estimation procedure for estimating the parameters and also this approach guarantees a perfect fit of the sum score distribution. OPLM allows for different item discriminations by restricting them to integer values; therefore, a CML estimation procedure can be used. The fit index of OPLM, the R_{1c} which is a Pearson chi-square type of test statistic (Glas, 1988; Glas & Verhelst, 1995), yields an acceptable fit (R_{1c}=9.45, df=11, p=.51). This seems to confirm the hypothesis that the misfit of MULTILOG is due to the bimodality of the latent trait density.

Table 3: Models fitted for Text 2.

MODEL	G^2	*df*	*p*	*AIC*	Item Discriminations					
MULTILOG - Rasch (Normal distr.)	145.9	56	.00	159.9	1	1	1	1	1	1
MULTILOG - Rasch (Johnson curve)	112.8	52	.00	136.8	1	1	1	1	1	1
MULTILOG - 2PLM (Normal distr.)	109.1	52	.00	131.1	1.96	0.44	0.91	1.55	1.45	1.47
MULTILOG - 2PLM (Johnson curve)	101.1	47	.00	133.1	1.83	0.53	0.96	1.51	1.41	1.44
LOG - 2PLM	63.9	42	.02	107.9	2.0	0.5	1.0	1.5	1.5	1.5
LOG - Rasch	94.1	52	.00	118.1	1	1	1	1	1	1
LOG - 1x6 4x5	49.5	50	.49	77.5	1	1	1	1	1	1

Note. See Table 1.

4.2.2. Part 2: The interaction part

In the second part of the analysis, the loglinear Rasch model was estimated. As can be seen in Table 3, and as expected, this model does not fit the data. It is therefore extended by including interactions. The final model with two interactions does fit the data well. Compared to the other estimated models, it has the lowest *AIC* value (*AIC*=77.5). As was the case for Text 1, the model with the lowest *AIC* is the model that allows interactions between items but without varying item discriminations. The interactions between items 1 and 6 and between items 4 and 5 are included in the final model.

The parameter estimates for this model can be found in the right part of Table 2. For Text 2, all interactions are positive such that the items that interact with each other have higher estimated discrimination parameters. The items 1, 4, 5 and 6 show interactions and their item discriminations are 1.93, 1.65, 1.55 and 1.53 respectively, which is higher than 0.66 and 1.02 for the non-interacting items 2 and 3. The results of the analysis of the second text confirm what was already found for the first text: Large item discriminations may be explained by allowing interaction between items.

4.2.3. Part 3: The LID index part

Next, the two LID indexes, G^2_{LID} and Q_3, are computed. G^2_{LID} is significant at α=.05 level for the item pairs (2,5), (4,5) and (2,6), and significant at α=.01 level for the item pair (1,6). However, the item dependency for the latter pair is not detected by Q_3 since

the residual correlation is only –0.08. The Q_3 index flags a list of other item pairs showing LIDs: (1,3), (1,4), (1,5), (2,6), (3,6) and (4,6). Note that the LID indexes indicated more item pairs showing LIDs than the ones included in the final loglinear interaction model. With the G^2_{LID} index, in addition to the pairs (1,6) and (4,5) that were included in the model, also the pairs (2,5) and (2,6) were indicated as dependent. The Q_3 index does not detect any of the pairs that were included in the model, but it indicated six other pairs.

5. Conclusion and Discussion

For Text 1 it is possible to obtain a fitting 2PLM without interactions as well as a fitting interaction model with equal discriminations but the interaction model is the best fitting model. It is also more parsimonious than a 2PLM with different item discriminations. In this case it appears that non-modeled interactions in the data can hide in the discrimination parameters. Of course, this is an example of an extreme analysis, since the item discriminations could be explained completely away by the interactions but this will not always be the case. The situation for Text 2 is somewhat more complicated. Here it is shown that the non-modeled interactions can distort the estimated discrimination parameters but now it is not possible to conclude that the discriminations hide the interactions because the 2PLM does not fit the data quite well (although the misfit is probably only due to that part of the model related to the distribution of θ-values in the population).

One could argue that the interactions have to be included in the loglinear model to account for true differences in the discrimination parameters (so that interactions would hide differing discrimination parameters). This argument makes it clear that we have to rely on the assumption that the interaction model is the more plausible model and not the 2PLM. But that is not an unrealistic assumption for three reasons. First, the type of data that we have analyzed (items related to a common text paragraph) are a natural place to find item interactions. Second, for both texts, the interaction was according to several criteria the best fitting model (this is certainly the case for Text 2 since the 2PLM did not fit the data). Third, the specific LID indexes indicate, for Text 1 as well as for Text 2, at least for some pairs of items indicated that they are interacting.

As said before, in our case one could suspect item interactions, which is why the data sets were chosen. There are however many more situations in which item interactions

can appear. For instance, Knowles and Condon (2000) concluded that if similar items are grouped together, the intercorrelations among their responses may increase. Another example is presented by Hoskens and De Boeck (1997) where the solution of an ability item gives an indication of the solution of another item. But it is difficult to detect these interactions on first sight, and therefore it is advisable to check for them with possibly an appropriate parametric model or some LID indexes. The LID indexes flagged item pairs, which could not be included in the interaction model though (and vice versa), but at least the mere detection of dependencies can be considered as an indication that the correct model has not been chosen.

It is relatively simple to detect interactions between items in data sets, but it is a more complex task to find an explanation for these interactions. For the five significant item interactions that are found in the two texts, only one has a clear explanation. In the second text, items 1 and 6 are both very strongly related in content: Both questions ask to compare governmental and parliamentary power in the US and Belgium. Therefore, it is not unreasonable to find a positive interaction between items 1 and 6: Solving the first one will have an influence on the probability of solving the second one. The other interactions (also the negative one in Text 1) are much more difficult to interpret and therefore, they are not discussed.

With this paper we do not want to suggest that interactions always can explain all differences in discrimination. However, our two examples show that in cases where one might expect interactions, it is a good practice to check whether they really appear. And if that happens, the estimated item discrimination may be seriously biased.

References

[1] Agresti, A. (1990). *Categorical data analysis.* New York: John Wiley.

[2] Akaike, H. (1977). On entropy maximization principle. In P.R. Krishnaiah (Ed.), *Proceedings of the symposium on application of statistics* (pp. 27-47). Amsterdam: North-Holland.

[3] Birnbaum, A. (1968). Some latent trait models. In F.M. Lord , & M.R. Novick, *Statistical theories of mental test scores* (pp. 397-424). Reading: Addison-Wesley.

[4] Bishop, Y.M.M., Fienberg, S.E., & Holland, P.W. (1975). *Discrete multivariate analysis: Theory and practice.* Cambridge, MA: The MIT Press.

[5] Chen, W.H. (1993). *IRT LD: A computer program for the detection of pairwise local dependence between items* (Research Memorandum 93-2). Chapel Hill: L.L. Thurstone Laboratory, University of North Carolina at Chapel Hill.

[6] Chen, W.H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265-289.

[7] Claes, L., Janssen, P.J., & Wels, G. (1999). *Keuzebegeleiding van laatstejaars secundair onderwijs. Hoe jongeren helpen bij het maken en toetsen van hun studiekeuze? [Choice counseling of last year high school students. How to assist students in making and testing their study choice.]* Eindrapport OBPWO 95.09.

[8] Duncan, O.D. (1984). Rasch measurement: Further examples and discussion. In C.F. Turner & E. Martin (Eds.), *Surveying subjective phenomena* (Vol. 2, pp. 367-403). New York: Russell Sage Foundation.

[9] Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists.* New Jersey: Lawrence Erlbaum Associates.

[10] Glas, C.A.W. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika, 53*, 525-546.

[11] Glas, C.A.W., & Verhelst, N.D. (1995). Testing the Rasch model. In G.H. Fischer & I.W. Molenaar (Eds.), *Rasch models: Foundations, recent developments and applications* (pp. 69-96). New York: Springer-Verlag.

[12] Hoskens, M., & De Boeck, P. (1995). Componential IRT models for polytomous items. *Journal of Educational Measurement, 32,* 234-246.

[13] Hoskens M., & De Boeck, P. (1997). A parametric model for local item dependencies among test items. *Psychological Methods, 2*, 261-277.

[14] Kelderman, H. (1984). Loglinear Rasch model tests. *Psychometrika, 49,* 223-245.

[15] Knowles, E.S., & Condon, C.A. (2000). Does the rose still smell as sweet? Item variability across test forms and revisions. *Psychological Assessment, 12*, 245-252.

[16] Rasch, G. (1980). *Probabilistic models for intelligence and attainment tests.* Chicago: The University of Chicago Press.

[17] Ten Vergert, E., Gillespie, M., & Kingma, J. (1993). Testing the assumptions and interpreting the results of the Rasch model using loglinear procedures in SPSS. *Behavioral Research Methods, Instruments, & Computers, 25,* 350-359.

[18] Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika, 47*, 175-186.

[19] Thissen, D. (1991). *MULTILOG* [Computer software]. Mooresville, IN: Scientific Software.

[20] Thissen, D., Steinberg, L., & Mooney, J.A. (1989). Trace lines for testlets: a use of multilple-categorical-response models. *Journal of Educational Measurement, 26*, 247-260.

[21] Tuerlinckx, F., & De Boeck, P. (2001). The effect of ignoring item interactions on the estimated discrimination parameters in item response theory. *Psychological Methods, 6,* 181-195.

[22] Verhelst, N.D., & Glas, C.A.W. (1995). The one parameter logistic model. In G.H. Fischer & I.W. Molenaar (Eds.), *Rasch models: Foundations, recent developments and applications* (pp. 215-238). New York: Springer-Verlag.

[23] Verhelst, N.D., Glas, C.A.W., & Verstralen, H.H.F.M. (1995). *OPLM: Computer program and manual.* Arnhem: Cito.

[24] Wainer, H., & Kiely, G.L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185-201.

[25] Yen, W.M. (1984). Effect of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8,* 125-145.

[26] Yen, W.M. (1993). Scaling performance assessments: strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187-213.