Psychology Science Quarterly, Volume 50, 2008 (1), p. 64-74

Estimation of item location effects by means of the generalized logistic regression model: a simulation study and an application

RAINER ALEXANDROWICZ[1] & HERBERT MATSCHINGER[2]

Abstract

The present paper deals with the application of the generalized logistic regression model to the estimation of item location effects. A Monte Carlo study demonstrates that item difficulties and item location effects show excellent parameter recovery when distributional assumptions of the marginal maximum likelihood method are not met. A practical application to a reasoning test revealed the existence of item location effects and an effect of total test taking time. This model allows for the flexible utilization to a wide range of problems and thus provides a powerful tool for item analysis.

Key words: Item location effect, position effect, Generalized Logistic Regression Model, Generalized Linear Mixed Model, Linear Logistic Test Model

[1] Univ.Ass. Mag.Dr. Rainer Alexandrowicz, University of Klagenfurt, Institute of Psychology, Department of Applied Psychology and Methods Research, Universitätsstraße 65-67, 9020 Klagenfurt, Austria; Phone: +43 (0) 463 2700 1627, email: rainer.alexandrowicz@uni-klu.ac.at

[2] University of Leipzig, Germany

1. Introduction

This paper deals with the estimation of item location effects (Kingston &Dorans, 1984; for an overview see Leary & Dorans, 1985), which alter the item difficulty depending on its very position (1st, 2nd, ... , last) within a test (throughout this article we will refer to tests; of course, other applications can be thought of as well). Several factors causing item location effects can be identified, some of which are warm-up, learning or practice, fatigue or time shortage. If *e.g.* fatigue takes place, items of basically equal difficulty will appear more difficult towards the end of the test – mere effects of the position of an item will be ascribed to its characteristics. Such effects violate the unidimensionality assumption (Whitely & Dawis, 1976; Yen, 1980). Hence item location effects will distort the estimation of item difficulties Thus they invalidate conclusions concerning psychometric properties of an item or the level of performance of a respondent, respectively. This is of particular importance when the sequence of items differs between calibration phase and the administration of a test. This occurs inevitably in the context of adaptive testing: based on a pool of items considered to be in accordance with the assumptions of the Rasch-model, an adaptive testing strategy chooses the most informative item(s) with respect to the estimated ability level. In such a context an item can occur at virtually any position, therefore it is of indispensable importance that item characteristics obtained from a calibration study are maintained in test administration.

One model that allows the estimation of position effects is the Linear Logistic Test Model (LLTM; Fischer, 1972, 1973, 1995). Here, the item difficulty parameter β_i is being split up into a sum of several basic parameters, η_j, weighted according to the entries of a design matrix **W**: $\beta_i = \sum_{\mathrm{j}} w_{\mathrm{ij}}\eta_j$, (j = 1...m indexing the vector of basic parameters, i = 1...k indexing the item difficulty parameters). If a set of items had been administered in at least two (favorably more) different compilations of the test (*i.e.* test versions with items arranged at different positions), two sets of basic parameters are used for each item: one set describes the difficulty of an item if there were no position effects present (subsequently these will be termed *conditional difficulties*) and the second one is used for the estimation of the effect the position has on this item (of course, both sets can consist of one single η_j each). Different parametrizations of the second set can be thought of. If *e.g.* continuous learning or tiring is assumed, a linear effect might be chosen: with each item a respondent has worked on, the same amount of learning or tiring occurs. This would require one parameter η_j that estimates the in-/decrease of difficulty per each item. Then, the entries in the corresponding column of **W** are the position of the item. Another possibility would be to introduce one parameter to each position, each of which describes the effect of this very position (the matrix **W** would have to be extended by the appropriate number of positions, which in this case equals the number of items). Such a parametrization not only allows for the detection of position effects as such, but also for a structural description of the kind of position effect that can be derived from the data, because no structure of this effect (*e.g.* linear) is assumed *a priori.*

In the present paper we want to take another approach towards the estimation of position effects. The problem can be framed by a *mixed logistic regression model*, which allows for a reformulation of the LLTM (Rijmen et al., 2003).

The model equation of a *generalized logistic regression model*, can be expressed as follows:

$$p\left(y_{vi}=1\,|\,\boldsymbol{x}_{vi},\boldsymbol{z}_{vi},\boldsymbol{\beta},\boldsymbol{\theta}_v\right)=\frac{\exp\left(\boldsymbol{x}'_{vi}\boldsymbol{\beta}+\boldsymbol{z}'_{vi}\boldsymbol{\theta}_v\right)}{1+\exp\left(\boldsymbol{x}'_{vi}\boldsymbol{\beta}+\boldsymbol{z}'_{vi}\boldsymbol{\theta}_v\right)}, \qquad (1)$$

with
y_{vi} the binary response of level-2-unit v to level-1-subunit i (v = 1..n; i = 1..k)
$\boldsymbol{x}_{vi}$ the user provided covariate vector for the fixed effects ($1\times p$)
$\mathbf{z}_{\mathrm{vi}}$ the user provided covariate vector for the random effects ($1\times q$)
$\boldsymbol{\beta}$ the p-dimensional parameter vector for the fixed covariates
$\boldsymbol{\theta}_{\mathrm{v}}$ the q-dimensional parameter vector for random covariates of unit v

Basically, this is a two level model (Goldstein, 1995; Langer, 2004; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999) and responses y_{vi} are stacked one below the other, resulting in a response vector **y** with dimensions ($nk \times 1$). By stacking all row vectors $\mathbf{x}_{vi}$ and $\mathbf{z}_{vi}$ across the items a person v has responded to, one obtains the design matrices $\mathbf{X}_v$ and $\mathbf{Z}_v$, respectively. If we further stack all design matrices $\mathbf{X}_v$ and $\mathbf{Z}_v$ over persons (units), we obtain the supermatrices **X** ($nk \times p$) and **Z** ($nk \times q$). The distribution of the error term of the random parameter is usually assumed to be normal. In the present application an *items-within-respondents* design is applied, respondents are assigned to level-2-units and items to level-1-units. Item difficulties and position effects are modelled through the fixed effects, and the ability of the respondent is covered by one random parameter (therefore, $\mathbf{Z}_v$ is in this case a column of ones). Note that in contrast to the definition above, the fixed parameters β_j used in (1) correspond with the η_j of the LLTM, hence the same subscript will be used. This model is extensible as it allows for the inclusion of predictors for both fixed and random effects. These predictors can be both quantitative and qualitative and interactions within and across levels can be estimated (De Boeck & Wilson, 2004; Skrondal & Rabe-Hesketh, 2004).

The generalized logistic regression model is a special case of the generalized linear mixed model (GLMM; McCulloch & Searle, 2001). One characteristic of GLMMs is that they require the application of the marginal maximum likelihood estimation method (MML) to be applied, entailing distributional assumptions. Usually – but not necessarily – normal distribution is assumed. This is a basic difference to the LLTM, where a conditional ML estimation method (CML) is applicable (Andersen, 1970). Therefore, the present study investigates the parameter recovery of item difficulties and position effects using the generalized logistic regression model, when the distributional assumptions are not met. Furthermore, some of the extended possibilities of the generalized logistic regression model compared to the LLTM shall be demonstrated in a practical application.

2. Methods and Material

In the first part, we want to evaluate by means of a Monte Carlo study, to what extent estimates will differ if the underlying ability distribution does not fit the assumed normal distribution. Data sets were generated according to the LLTM, with a parameter η_i (describing the conditional difficulty of item i) and a position parameter η_{k+p} for each item (note that subsequently p will be used for denoting the position of an item). Each data set contains five items and 1000 observations, the latter being split into five subgroups of equal size, displaying different item arrangements. For the item arrangements across the groups a latin square design was chosen, *i.e.* each item occurred at each position once (subgroup 1: item sequence: 1,2,3,4,5; subgroup 2: item sequence 2,3,4,5,1; ...). Details of the simulation parameters are given in Table 1.

Table 1:
Simulation parameters of the Monte Carlo study (Note: In the generalized logistic regression model additive parameterisation is applied, so item easiness parameters were used for compatibility reasons)

	n	k	β_i	β_{k+i}	θ	Samples
Design 1	1000	5	2, 1, 0, –1, –2	–2, –1, 0, 1, 2	N(0;1)	500
Design 2	1000	5	2, 1, 0, –1, –2	–2, –1, 0, 1, 2	$\chi^2_{[1]} - 2$	500
Design 3	1000	5	–1.38, 0.11, 0.54, –0.27, 1.02	1, 1, 1, –1, –2	N(0;1)	500
Design 4	1000	5	–1.38, 0.11, 0.54, –0.27, 1.02	1, 1, 1, –1, –2	$\chi^2_{[1]} - 2$	500

The first two designs contain linearly de- and increasing item difficulties β_i and position effects β_{k+i} respectively. Such a structure might be observable when learning takes place: an item decreases in difficulty (increasing β_{k+i}) when presented at a later position. While in designs 1 and 3 the normal ability assumption was met, for designs two and four a skewed distribution was chosen (as the $\chi^2_{[1]}$)distribution only realizes positive values, its theoretical mean of two was subtracted for a better coverage of both positive and negative values). In designs 1 and 2 a linear learning effect was assumed, in designs 3 and 4, a non-linear position effect, reflecting a sudden increase of item difficulty on the last two positions, was superimposed. This effect describes a situation in which the last two items show a sudden increase in difficulty, possibly a result of time shortage or fatigue. Comparing estimates of designs 1 and 2 (or 3 and 4, respectively) allows for an assessment of the effect of the violation of the distributional assumption. As this was the primary concern of this simulation study, we tried to warrant estimability of models by choosing the remaining simulation parameters in a fail-safe manner: in a latin-square comparably marked design position effects and large samples were chosen and the number of items rather low. Of course, such conditions will seldom be met in real life applications.

For the estimation of the generalized logistic regression model, effect coding for the item difficulty parameters and the parameters of the position effect was applied, as this method yields centered estimates matching the sum-zero norming of item parameters usually applied in Rasch measurement. Efron & Tibshirani (1993, p. 188) propose between 200 (p. 52) and 1000 replications for different designs. So we decided to repeat each design 500 times, which was a reasonable trade off between precision and time consumption.

In order to demonstrate the capabilities of the proposed approach, the generalized logistic regression model was applied to a reasoning test involving 78 items (BBT, Begriffsbildungstest [*concept forming test*]; Fischer, 1991; Kubinger, Fischer, & Schuhfried, 1993). An earlier study on this test (Alexandrowicz, 1999) found no violations of the assumptions of the Rasch-model (under certain prerequisites). Due to the workload of each item no person would be able to work on all 78 items. Therefore, 13 groups of 12 items each had been generated. Across these groups each item appeared on two different positions. This procedure led to *missings-by-design*. It is one of the merits of multilevel models that no further steps regarding the handling of *missings-by-design* are required. When the item groups were developed, estimates from Fischer (1991) were available. These had been utilised to form groups of approximately equal average difficulties. Furthermore, items within each group were arranged in a special way: the first item was always a rather easy one, giving respondents the opportunity to become acquainted with the handling of the test. The next five items became increasingly difficult. The sixth item was again a rather easy one in order to give weaker respondents a feeling of success (*cf.* Häusler, 2006). Then items became increasingly difficult again. This principle was maintained throughout all the 13 groups.

In this study we want to reanalyse the original data by means of the generalized logistic regression model in order to test whether position effects occur. Further predictors were included to test for gender or age specific interactions and for cross-level interactions concerning test duration and test solving strategy. Data simulation was performed by means of a program written by the first author of this article, the generalized logistic regression model was estimated using the procedure `xtlogit` of Stata (StataCorp, 2005). The risk of a type-I-error was chosen at 5% throughout all analyses.

3. Results

The Simulation Study

The parameter estimates of the four designs are given in Table 2. The values represent the means and standard deviations of each parameter over the 500 replications. Parameter distributions were inspected through histograms and Q-Q-Plots, both of which did not reveal any striking aberration from normality. Furthermore, the Kolmogorov-Smirnov-Test of normality did not show any significant results for any of the parameter distributions under consideration.

The linear position effects in designs 1 and 2 were discernible as was the training effect, making items gradually easier the later they are presented. The same is true for designs 3 and 4, where the positions of the conditional item difficulties were identifiable as was the "sudden" increase of item difficulty at positions 4 or 5.

All estimates are close to the true values, especially when located close to zero. For designs 1 and 2, the (absolute) larger parameters seem to slightly underestimate true values, but mostly the difference stays within the quartiles. This effect is somewhat more pronounced for the skewed distributions (designs 2 and 4), though maintainable. In general, distributions of estimates are very narrow, no heavy outliers are observable.

Table 2:
Parameter recovery for the four simulation designs (TP = true parameter)

	Design 1			**Design 2**			**Design 3**			**Design 4**		
	TP	**Mean**	**SD**	**TP**	**Mean**	**SD**	**TP**	**Mean**	**SD**	**TP**	**Mean**	**SD**
η_1	2	1.91	*0.09*	2	1.87	*0.09*	–1.38	–1.38	*0.08*	–1.38	–1.35	*0.10*
η_2	1	0.94	*0.08*	1	0.88	*0.08*	0.11	0.11	*0.07*	0.11	0.10	*0.08*
η_3	0	0.00	*0.08*	0	–0.02	*0.08*	0.54	0.53	*0.07*	0.54	0.52	*0.08*
η_4	–1	–0.94	*0.08*	–1	–0.91	*0.09*	–0.27	–0.30	*0.07*	–0.27	–0.33	*0.07*
η_5	–2	–1.91	*0.09*	–2	–1.83	*0.11*	1.02	1.03	*0.08*	1.02	1.06	*0.09*
η_6	–2	–1.91	*0.09*	–2	–1.83	*0.10*	1	0.98	*0.07*	1	0.95	*0.07*
η_7	–1	–0.94	*0.07*	–1	–0.91	*0.09*	1	1.01	*0.07*	1	0.99	*0.08*
η_8	0	0.00	*0.08*	0	–0.03	*0.08*	1	0.99	*0.07*	1	0.97	*0.08*
η_9	1	0.94	*0.08*	1	0.89	*0.08*	–1	–0.98	*0.07*	–1	–0.98	*0.09*
η_{10}	2	1.91	*0.09*	2	1.88	*0.09*	–2	–2.00	*0.10*	–2	–1.93	*0.11*

The Practical Application

A total sample of 552 was obtained, two respondents had not solved a single item, therefore they were deleted ; no respondent solved all of the items, so an effective sample size of 550 was available. The majority of respondents were advanced students of psychology at the University of Vienna. Seventy-five percent of respondents were female, the average age was 25 years with a standard deviation of five years.

First of all, estimates according to the Rasch-model without considering position effects were compared to those from the original study (Alexandrowicz, 1999). Item parameter estimates were satisfyingly similar (r = .93) with a maximum difference of 0.71 logits. The linear regression line of the current estimates (MML) on those according to Alexandrowicz (1999), which were obtained by means of CML estimation, had an intercept of 0.01 and a slope of 1.05, *i.e.* it closely resembled the identity line. Therefore, no systematic differences were detectable. This analysis was undertaken to ascertain equivalence of the two different methods of estimation.

In order to test for position effects, the design matrix $\mathbf{X}_v$ and the fixed parameter vector $\boldsymbol{\beta}$ were extended, so that a conditional difficulty $\boldsymbol{\beta}_i$ and a position effect $\boldsymbol{\beta}_{k+p}$ were estimated. According to the results of Alexandrowicz (1999) the first item each respondent had worked on was treated as warm-up item. For better readability, position effects were dummy coded, the first item under consideration (*i.e.* the second item each respondent had worked on) was taken as reference category (*i.e.* position) and therefore was omitted from estimation. Of course, effect coding might have been applied as well; estimates obtained from different coding schemes can be transformed into one another (Bock, 1963, 1975). The estimates obtained are depicted in Figure 1.

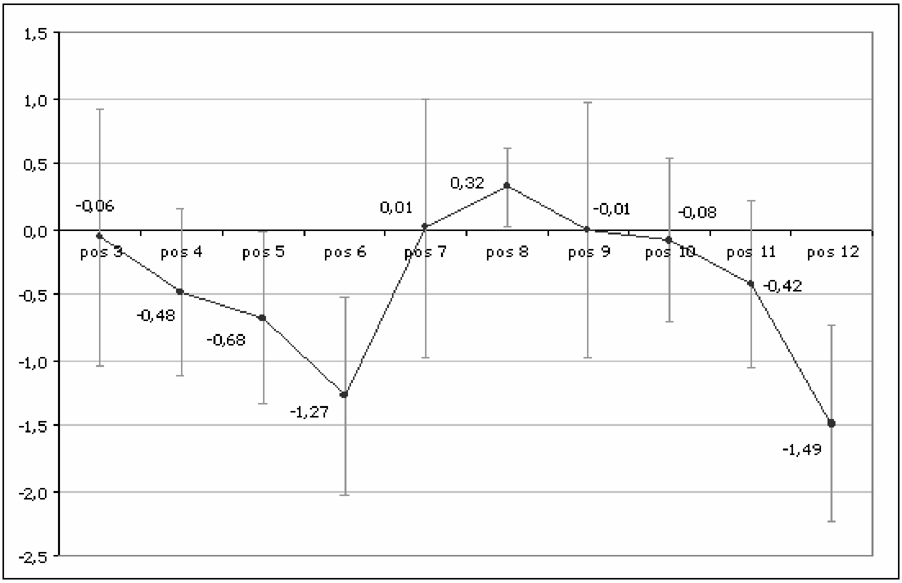

Figure 1:
Position effects β_{k+p} occuring in the reasoning test
(note: position 1 was not scored and position 2 is the omitted reference category; antennas represent the 95% confidence interval of the estimates)

A clear-cut position effect is discernible: the values show the change in the probability of solving an item depending on its position compared with the reference category. Item difficulty increases at positions three to six, the next item is again much easier, followed by another increase in difficulty. This exactly reflects the structure of item difficulties chosen within groups. Here a noteworthy effect of the study design becomes visible: a certain succession of item difficulties was chosen *a priori* in the same manner for all groups, a structure that arises in the position parameters. This is attributable to the fact that possible position effects are perfectly confounded with the chosen item difficulties, because no variation of the sequence of item difficulties between groups was introduced. Nevertheless, the estimation revealed correct results within the frame of reference.

In order to test for a gender effect the fixed part of the model (i.e. design matrix $\mathbf{X}_v$) was augmented by a column for gender (thus introducing another parameter β_{gender}). The estimate of the global gender effect parameter β_{gender} was 0.23 ($p_{[H_0:\beta=0]} = 0.01$). The reference (omitted) category was female, therefore the positive parameter globally indicates a (significantly) higher probability of solving an item for male respondents. If we introduced an interaction term for each item with gender and if these terms were not of equal size, *i.e.* the gender effect turned out to be item specific, which would indicate *differential item functioning* (DIF; Holland & Wainer, 1993). The practical realization would be testing the model without interaction effects against the model including interaction effects by means of a likelihood ratio test. If the test becomes significant, this indicates that the model with interaction effects describes data better, hence DIF occurs. In our case this test revealed a non-significant result ($\chi^2 = 77.0$; $df = 77$; $p = .48$), so no DIF with respect to gender could be evidenced.

Furthermore, we tested for a cross-level interaction of gender and position effects, *i.e.* whether the latter are taking place gender specifically. For that purpose an interaction parameter for each position and gender was estimated (*cf.* fig. 2, dot-dashed line). As can be seen, differences compared with the global effect of gender seem rather small (from a descriptive point of view). Again models are nested, therefore the gain of the additional position specific interaction terms can be evaluated by means of a Likelihood Ratio Test. Results were unambiguous (LR-$\chi^2 = 7.73$, $df = 10$, $p = .66$), so there are no clues that gender groups behave differently on certain positions.

The total test taking time of each respondent was available. Its effect on the probability of solving an item can be assessed by means of a cross level interaction. We therefore introduce one parameter, β_{time}, and insert one more column in the design matrix $\mathbf{X}_n$, containing each respondent's test duration. This complies with assuming a linear effect of test taking time (of course, a more complex effect could be considered as well, but in this case we want to demonstrate how quantitative covariates can be introduced). The estimate of this parameter was extremely small but significant ($\beta_{\text{time}} = 0.0002$, $p < .001$) – which at first sight may seem suprising, but makes perfect sense: The effect of test duration is assumed linear, so the parameter reflects the change in logits *per unit of the predictor* – which in this case were seconds. So for each hour a person worked on the test, an effect of 3600 × 0.0002 = 0.72 would be obtained. This denotes an increase in odds of exp(0.72) = 2.05, This means that the longer a respondent worked on the test, the better his or her result would become, so accurateness payed.

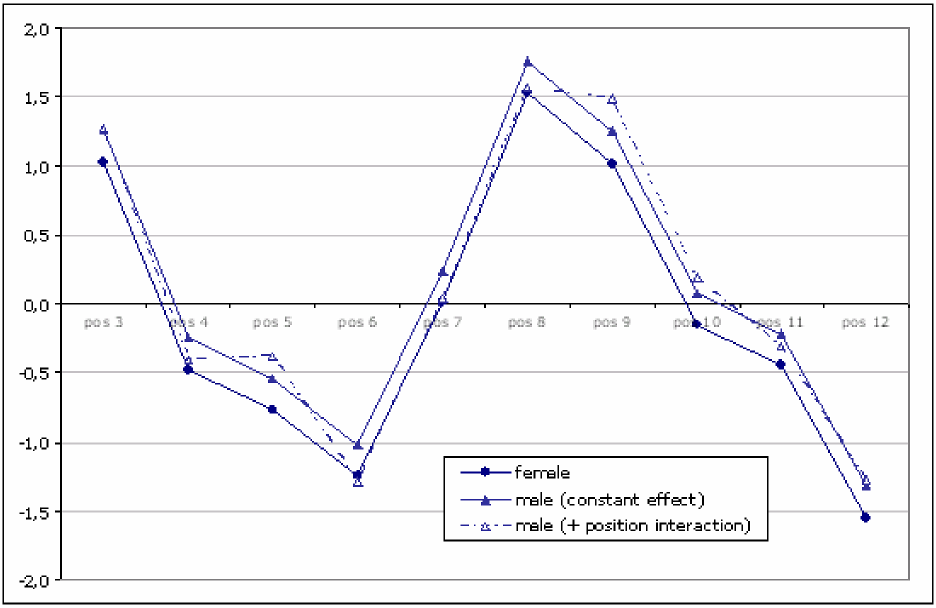

Figure 2:
Gender specific position effects
(note: the reference category was male; full triangles depict a gender effect assumed constant over all positions; empty triangles depict a gender effect assumed to be position specific)

4. Discussion

A simulation study was performed in order to investigate the estimability of position by means of a generalized logistic regression model, when distributional assumptions are not met. Four designs were analysed, exhibiting different combinations of item difficulties, position effects, and ability distribution. For all four designs under consideration parameter recovery can be seen as excellent, the initial structure is clearly identifiable and numerical differences were not larger than 0.2 logits. We found that (absolute) large parameters were somewhat underestimated. This seems to be an effect of the MML estimation, because this phenomenon was slightly more pronounced for the skewed conditions. Carstensen (2000, p. 137) reported a similar phenomenon: in a comparison of MML and CML estimates the former also showed the same tendency while the latter did not, which further evidences our explanation. Parameter estimates did not reveal any deviances from normality, which is in line with maximum likelihood estimation theory. Marked position effects and large samples have been chosen and only five items were taken in order to warrant detectability and estimability. It remains to be investigated systematically which sample size is required to ensure detectability of relevant effects and to what extent distributional assumptions play a role then.

In an application to a reasoning test a clear cut position effect could be detected. This application is a distinctive indication for the importance of carefully choosing an appropriate design regarding the succession of items in different compilations. In the present case, items were arranged according to *a priori* known item difficulties, therefore, position effects attributable to change of item difficulty during test taking were confounded with this predefined structure of item difficulties. A more meaningful design would be to put items in as many different positions as possible; of course, the chosen latin square design of the simulation study is not realistic for longer scales.

We could have estimated position effects by means of the LLTM, but by using the generalized logistic regression model quantitative covariates and its polynomials could be included in a straightforward manner. In the present case, total test taking time was analyzed and did reveal a significant influence on the probability of solving an item. When applying the LLTM for such a problem, test duration would have to be split into groups, which on the one hand would cause loss of information and on the other hand the outcome of such an analysis stands or falls with the appropriateness of the cut-off(s) chosen. The generalized logistic regression model is a special case of the generalized mixed model. This allows for further extensions, such as introducing a weight λ_i to model the slope, *i.e.* to adopt the idea of the 2PL (Birnbaum, 1968; Rabe-Hesketh, Skrondal, & Pickles, 2004, p. 72). This would further allow to test for the underpinning assumption of equal item discrimination of the Rasch family of models within the same model framework.

One crucial aspect essential for applying the Rasch-model must not be missed, *viz.* the fact that this model allows for specific objective comparisons, which in turn are the foundation for testing model validity. Andersen (1973) showed a way of testing the model by means of evaluating the assumption of indifferent parameter estimates across subgroups (formed by a split variable S) through a likelihood ratio test. This test could be approximated in the generalized mixed model framework by checking whether interaction effects of item parameters and β_S occur, *i.e.* whether the effect of the split criterion is item specific (Verhelst & Verstralen, 2001, p. 99).

As a disadvantage of this approach it might be seen that parameter estimation is based upon the marginal maximum likelihood estimation technique, which introduces distributional assumptions concerning the trait under consideration. Mostly normal distribution is assumed, but Micceri (1989) has bad news in finding that normality can seldom be assumed. On the other hand, Rost (2004, p. 310) states that the procedure is rather robust for misspecification of the distribution, so estimates at least approximately allow for the right conclusions. Our results support the latter view, differences of parameter estimates compared over the two distributions chosen seemed tolerable – although the $\chi^2_{[1]}$ is far from normality ($\gamma_1 = 2^{3/2} df^{-1/2} \sim 2.86$ and $\gamma_2 = 3 + {}^{12}/_{df} = 15$; Evans, Hastings & Peacock, 2000, p. 53). So due to its generality, the generalized mixed model framework might prove a valuable tool for a differentiated item analysis.

5. References

Alexandrowicz, R. (1999). *Normierung und Validierung des Begriffsbildungstests.* [Calibration and Validation of the Concept Forming Test]. Unpublished Master Thesis. University of Vienna.

Andersen, E. B. (1970). Asymptotic Properties of Conditional Maximum Likelihood Estimators. *Journal of the Royal Statistical Society B, 32*, 283–301.

Andersen, E. B. (1973). A Goodness of Fit Test for the Rasch Model. *Psychometrika*, *38*, 123–140.

Birnbaum, A. (1968). Some Latent Trait Models and Their Use in Inferring an Examinee's Ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical Theories of Mental Test Scores* (pp. 395-479). Reading, MA: Addison-Wesley.

Bock, R. D. (1963). Programming Univariate and Multivariate Analysis of Variance. *Technometrics*, *5*, 95-117.

Bock, R. D. (1975). *Multivariate Statistical Methods in Behavioral Research*. New York: McGraw-Hill.

Carstensen, C. H. (2000). *Ein Mehrdimensionales Testmodell mit Anwendungen in der pädagogisch-psychologischen Diagnostik.* [A Multidimensional Test Model with Applications in Educational and Psychological Diagnostics]. IPN-Schriftenreihe 171. Kiel: IPN.

De Boeck, P. & Wilson, M. (2004). (Eds.), *Explanatory Item Response Models: A General Linear and Nonlinear Approach.* New York: Springer.

Efron, B. & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* Boca Raton: Chapman & Hall.

Evans, M., Hastings, N., & Peacock, B. (2000). *Statistical Distributions* (3rd ed.). New York: Wiley.

Fischer, D. (1991). *Begriffsbildungstest – BBT.* [The Concept Forming Test]. Unpublished Master Thesis, University of Vienna.

Fischer, G. H. (1972). A Measurement Model for the Effect of Mass-Media. *Acta Psychologica*, *36*, 207-220.

Fischer, G. H. (1973). The Linear Logistic Test Model as an Instrument in Educational Research. *Acta Psychologica*, *37*, 359-374.

Fischer, G. H. (1995). The Linear Logistic Test Model. In: G. H. Fischer & I. Molenaar (Eds.), *Rasch Models. Foundations, Recent Developments, and Applications* (pp. 131-156). New York: Springer.

Goldstein, H. (1995). *Multilevel Statistical Models* (2nd ed.). London: Edward Arnold.

Häusler, J. (2006). Adaptive Success Control in Computerized Adaptive Testing. *Psychology Science*, *48*, 436-450.

Holland, P. W. & Wainer, H. (Eds.), (1993). *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Kingston, N. M. & Dorans, N. J. (1984). Item Location Effects and Their Implications for IRT Equating and Adaptive Testing. *Applied Psychological Measurement*, *8*, 147-154.

Kubinger, K. D., Fischer, D., & Schuhfried, G. (1993). *Begriffs-Bildungs-Test (BBT)*. [Concept Forming Test]. Mödling: Schuhfried.

Langer, W. (2004). *Mehrebenenanalyse. Eine Einführung für Forschung und Praxis*. [Multilevel-Analysis. An Introduction to Research and Practice]. Wiesbaden: VS Verlag für Sozialwissenschaften.

Leary, L. F. & Dorans, N. J. (1985). Implications for Altering the Context in Which Test Items Appear: A Historical Perspective on an Immediate Concern. *Review of Educational Research*, *55*, 387-413.

McCulloch, C. E. & Searle, S. R. (2001). *Generalized, Linear and Mixed Models (Wiley Series in Probability & Statistics)*. New York: Wiley.

Micceri, T. (1989). The Unicorn, The Normal Curve and Other Improbable Creatures. *Psychological Bulletin. 105*, 156-166.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). *GLLAMM Manual*. Working Paper 160. Berkeley, CA: U.C. Berkeley Division of Biostatistics Working Paper Series.

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical linear models: Applications and Data Analysis methods* (2nd ed.). NewburyPark, CA: Sage.

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A Nonlinear Mixed Model Framework for Item Response Theory. *Psychological Methods*, *8*, 185-205.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion*. [Textbook Test Theory and Test Construction] Second, revised and extended edition. Bern: Huber.

Skrondal, A. & Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling; Multilevel, Longitudinal, and Structural Equation Models*. London, New York: Chapman &Hall/CRC

Snijders, T. & Bosker, R. (1999). *Multilevel Analysis: An introduction to basic and advanced multilevel modeling*. NewburyPark, CA: Sage.

StataCorp. (2005). *Stata Statistical Software: Release 9*. Edited by StataCorp LP. TX: College Station.

Verhelst, N. D. & Verstralen, H. H. F. M. (2001). An IRT Model for Multiple Raters. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on Item Response Theory*. Lecture Notes in Statistics 157 (pp. 89-108). New York: Springer.

Whitely, S. E. & Dawis, R. V. (1976). The Influence of Test Context on Item Difficulty. *Educational and Psychological Measurement, 36*, 329-337.

Yen, W. M. (1980). The Extent, Causes, and Importance of Context Effects on Item Parameter for Two Latent Trait Models. *Journal of Educational Measurement, 17*, 297-311.