Psychology Science Quarterly, Volume 50, 2008 (3), pp. 379-390

Analyzing position effects within reasoning items using the LLTM for structurally incomplete data

JULIA HAHNE[1]

Abstract

Position or transfer effects on an individual's ability while processing a series of test items are often ignored when tests are created. It is often implicitly assumed that such effects, if they occur, are a) the same for all persons and b) for all items and thus do not contribute to information about person ability or item difficulty. Rasch model analyses cannot quantify position effects because they are invariably confounded with the item difficulty parameters. In case of adaptive testing, where the examinees are administered the same items at different positions, effects of the position of item presentation lead to unfair estimations of item (and, consequently, person) parameters, and are therefore absolutely unwarranted. This study applies the Linear Logistic Test Model (LLTM, Fischer, 1973) for structurally incomplete data to illustrate how a series of test items can be evaluated for position effects. The test material consists of the Viennese Matrices (WMT, Formann & Piswanger, 1979) presented in varying item order to six groups of examinees. The study sample group consisted of 405 high school students. The concept of virtual items is introduced and applied to different models. Several hypotheses are tested by means of hierarchically applied Andersen's Likelihood Ratio tests. As a result of these analyses, no significant position effect can be found.

Key words: LLTM, position effects, Likelihood Ratio tests, hierarchical testing

[1] Mag. Julia Hahne, CEOPS - Center of Excellence for Orthopaedic Pain management Speising, Orthopaedical Hospital Speising, Speisinger Straße 109, A-1130 Vienna, Austria, Europe; email: julia.hahne@ceops.at

Introduction

Position effects are generally ignored when test items are presented to examinees in a conventional fashion with all items being presented to each examinee in the same order. In principle, there are two possibilities how position can influence item difficulty in a rather general way: If items become easier the later they appear in the test, we speak of learning (or practice) effects. In contrast, fatigue effects lead to increasing item difficulty the later an item appears within the test occasion. If learning takes place during a test session, item difficulty will be underestimated with progressive item position. In accordance with Item Response Theory (IRT), particularly the Rasch model, it is generally assumed that learning or position effects are the same for each person being examined. As long as the Rasch model holds true, the latent ability of the examinees can be concluded directly from the number of solved items. However, position effects are inseparably connected with item difficulty parameters and can thus neither be confirmed nor quantified.

Whereas position effects may not be a problem in conventional testing, they are certainly not acceptable in adaptive testing, where items are presented to the examinees in varying orders. Items presented at the beginning or end of a training session no longer have the same difficulty and 'fair' comparisons between examinees are no longer possible.

The Linear Logistic Test Model

Fischer (1973) developed the Linear Logistic Test Model (LLTM), in which the item parameters σ_i , i=1,…, k, of the Rasch model are composed in a linear combination of elementary (or basic) parameters η_l, l=1, …, p, representing cognitive processes necessary to solve test items. Item difficulty is explained as the weighted sum of the elementary parameters:

$$\sigma_i = \sum_{l=1}^{p} q_{il}\eta_l \tag{1}$$

whereas q_{il} are the element of the weight matrix **Q** for the elementary parameter η_l in item i. The q_{il} elements are known and defined. The parameters of the above model can be estimated using the Conditional Maximum Likelihood (CML) method (cf. Fischer, 1995), as long as the number of basic parameters does not exceed the number of item parameters. If the Rasch model holds true for a certain set of items, hypotheses about the elementary parameters can be tested using Andersen's Likelihood Ratio test (Andersen, 1973; see also Fischer, 1995, Kubinger, 2005):

$$-2\ln\left(\frac{L_{RM}}{L_{LLTM}}\right)\overset{as}{\sim}\chi^2 \tag{2}$$

where L_{RM} is the likelihood of the data estimated by the Rasch model and L_{LLTM} is the likelihood of the data estimated by the LLTM, with $df=k-p$.

Hence, LLTM analyses do not only provide useful information with respect to item difficulty. Furthermore, it is theoretically possible to generate any number of new items follow-

ing item construction rules and to predict item difficulty. Examples of item construction with the Linear Logistic Test Model are the Viennese Matrices (WMT, Formann & Piswanger, 1979), the Three-dimensional Cube test (3DW, Gittler, 1990), and the Adaptive Matrices (AMT, Hornke, Rettig & Etzel, 1999). Further extensive research was conducted by Embretson (1998, 1999, 2002), who developed a matrix item generator for Raven-like Matrices, reporting good item fit with the 2-PL model.

Cognitive operations can be described by the LLTM, and it is also possible to factor non-cognitive components into the item parameters. These may be position effects, learning effects or, contrastingly, fatigue effects (cf. Kubinger, 2008). Gittler (1990, see also Gittler & Wild, 1989) tested items of the Three-dimensional Cube test (3DW) for global transfer effects independent of the cognitive operations involved. The results showed good fit for the LLTM when position effects for the first eight items were taken into account. Hornke and Rettig (1989) evaluated matrix test items for learning effects, but found none.

LLTM for experimental designs

In addition to its use with cognitive processes, the LLTM is also used to analyze experimental designs. For this purpose, each item I_i with the corresponding item parameter σ_i is transferred into two (or more) virtual items I_j^* and I_l^* with the corresponding item parameters σ_j^* and σ_l^*. For an estimation of the change that a real item undergoes in different conditions, effect parameters are introduced. The virtual items' difficulty is some weighted sum of the actual item problem's difficulty under different conditions. Rather than representing cognitive processes, the elementary parameters show the change of item difficulty under different experimental conditions. The examinees, obviously, are given the same item problem just under different circumstances (i.e. test condition). An item problem under two different test conditions is thus formally treated as two virtual items. Because one person generally does not undergo each experimental condition, the data is incomplete by design (structural incompleteness).

The significance of the effect parameters can again be tested using the Likelihood Ratio test:

$$-2\ln\left(\frac{L_0}{L_1}\right)\overset{as}{\sim}\chi^2 \tag{3}$$

L_1 is the likelihood of the data under the alternative hypothesis, meaning that the likelihood is estimated including all parameters. L_0 is the likelihood of the data without the effect parameters in question (null-hypotheses), *df=k-p*.

Materials and methods

The aim of the study was to determine if there were any position effects on examinees solving reasoning items. This question was addressed using the LLTM for structurally incomplete data. The Viennese Matrices (Formann & Piswanger, 1979) was selected as the

testing material. The test contains 24 items. The item difficulties can be explained by means of three cognitive operations:

a) *Nature of elements*, referring to the shape (e.g. circle), pattern (e.g. striated), number and spatial arrangement (left-right)
b) *Type of rule*, referring to the basic operation applied to the elements. The operations are continuation, variation and superimposition
c) *Direction of the rule*, referring to vertical, horizontal, or both.

To solve the items, the examinee sees various elements arranged in a 3 × 3 matrix, must correctly identify similarities between adjoining elements and must then find and apply the logical principles behind the matrix elements to determine the correct solution to be selected from a set of eight answer options. The following is an example:

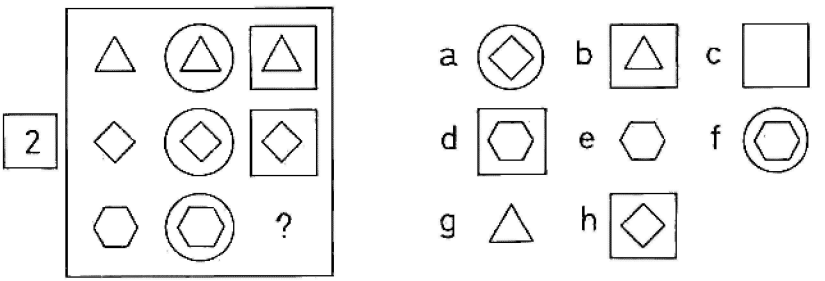

Figure 1:
Item 2 from the Viennese Matrices. The correct answer is d.

The test was chosen for two reasons: First, the items of this test are known to fit the Rasch model, which obviously is a prerequisite when working with the LLTM. Second, there is some evidence that learning may occur in Matrices tests (Raven, Raven & Court, 1998). The study sample consisted of high school students. The data collection took place in group sessions in two different schools. For all examinees, informed consent was obtained from the parents. The Viennese Matrices are usually administered in a power test setting. This point was extremely important for the present study: The students were instructed to work as long as they needed and, in agreement with school administration, no time limit was set for the testing procedure. Six different types of test booklets were created, each type with different item order (see table 1).

Whenever an examinee did not know the answer to an item, he/she had the possibility of checking the category labeled *I don't know*. This deviated from the usual test protocol and was introduced to diminish guessing effects.

The study sample consisted of 405 high school students, 251 (62%) boys and 154 (38%) girls. The mean age was 15.51 years (s = 1.18). The examinees were tested in group sessions; the mean duration of the test was 27.13 minutes (s = 7.15). Complete data sets were obtained for all examinees. The 405 students were randomly assigned to one of the six test booklets. The items for each booklet were the same, but their position in the test was changed. One group was tested with the original test form that contained items in ascending

order of item difficulty. Another group solved the items in descending order, beginning with the most difficult item. The other four groups received test booklets with randomly selected item orders. Each session started with an introductory example to familiarize the examinees with the testing material and three exercise items (see Table 1). As a warming up item, the first test item was excluded from the analyses. The LPCM-win program (Fischer & Ponocny-Seliger, 1998) was used for all analyses.

Table 1:
Test protocol: G1 (2,3,4,5,6) refers to Group 1 (2,3,4,5,6). *I* refers to the introductory example, e1-3 are the exercise examples 1 to 3. As indicated by the perforated line, only the first and the last items are shown

G1	*I*	e1-3	*I*	2	3	4	5	----	20	21	22	23	24
G2			*I*	10	16	12	8	----	11	17	7	13	9
G3			*I*	9	18	2	14	----	11	17	20	3	8
G4			*I*	24	23	22	21	----	7	6	5	4	3
G5			*I*	16	10	14	18	----	15	9	19	13	17
G6			*I*	18	9	7	13	----	16	10	6	24	19

Rasch Model analyses

LLTM can only be applied as long as the data shows fit with the Rasch model. This was tested by means of Andersen's Likelihood Ratio test. When using this test the sample of examinees is divided into subsamples in line with content-related considerations. The pooled likelihood of the data of each subsample is compared to the likelihood of the data of the entire sample. The formula to be used is again (3). The degrees of freedom are according to the difference of the estimated parameters.

For the present study the criteria *score* separated in examinees with a high and examinees with a low score. Age, sex and test duration were used as well as a criterion. The sample was divided for all criteria except sex by means of median splits. Table 2 shows cut off values and distribution for the partition criteria.

Table 2:
Cut off values for test duration (in minutes), score and age (in years) and distribution of sex ('f' female, 'm' male).

group	**test duration (min.)**				**score**				**age (yrs.)**				**sex**		***n***
		n		*n*		*n*		*n*		*n*		*n*	*n*(m)	*n*(f)	
1	<27	32	≥27	35	<16	32	≥16	35	<15	37	≥16	30	38	29	67
2	<27	31	≥27	37	<16	34	≥16	34	<16	36	≥16	32	45	23	68
3	<26	34	≥26	34	<18	35	≥18	32	<16	38	≥16	29	38	29	68
4	<28	34	≥28	33	<17	36	≥17	33	<16	35	≥16	33	41	27	67
5	<28	35	≥28	33	<16	33	≥16	35	<16	35	≥16	33	42	22	68
6	<28	33	≥28	34	<17	35	≥17	32	<16	33	≥16	34	47	20	67

The results of the Likilihood Ratio tests are shown in Table 3. Because the first item of each booklet was used as a warming up item, 23 item parameters per group were estimated. The critical value is $\chi^2(1\%) = 40.92$, $df = 22$. As can be seen, the items showed good fit with the Rasch model.

Table 3:
Results of the Rasch model analyses

group	criterion	χ^2
1	score	28,04290
	age	16,29413
	sex	16,17168
	time	10,45272
2	score	20,54809
	age	18,89184
	sex	22,12476
	time	33,34823
3	score	31,00931
	age	22,46003
	sex	23,04322
	time	25,31805
4	score	24,54910
	age	26,27065
	sex	26,90836
	time	18,07122
5	score	23,59339
	age	16,69636
	sex	24,61706
	time	23,82070
6	score	36,48543
	age	30,67309
	sex	33,73163
	time	19,58271

LLTM analyses

Three models were postulated for the LLTM analyses. Model 1 estimates item parameters regardless of their position in the test, i.e. $\sigma_{i1}^* = \sigma_{i2}^* = \sigma_{i3}^* = \sigma_{i4}^* = \sigma_{i5}^* = \sigma_{i6}^*$ plus one position or learning parameter λ. It is assumed that the latter continuously increases with each processed item, regardless of whether the item was solved or not and independently of the cognitive operations involved. Hence, 22 item difficulty parameters are estimated and the parameter λ representing the position within the respective test booklet. Figure 2 shows the design matrix.

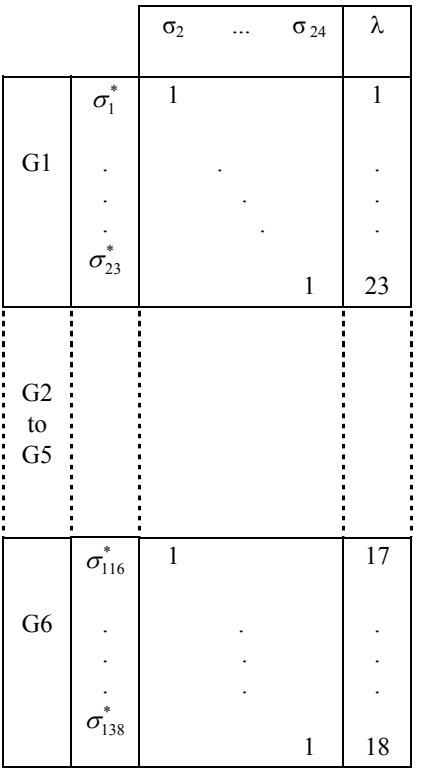

Figure 2:
Weight matrix **Q** of model 1 for the groups 1 and 6. The first item was excluded from the analysis

σ_1^* to σ_{138}^* are the item parameters for the virtual items of the six test booklets, σ_2 to σ_{24} are the parameters for the real item problems (Item 1 is the warming up item and therefore excluded from the analysis). There are $p = (k - 1) \times h = 23 \times 6 = 138$ virtual items (item problems $i = 2, \ldots, k$; $k = 24$ – groups $g = 1, \ldots h$; $h = 6$). The null-hypothesis is H_0: $\sigma_{ih} = \sigma_{ih}$ and the alternative hypothesis is H_1: $\sigma_{ih} = \sigma_i + q_{(ig)}\lambda$; $q_{(ig)} = l = 0, 1, 2, \ldots k - 1$, dependent on item problem i's position.

Model 2 ignores the different testing conditions, i.e. item positions, and the item parameters for each item are estimated independently of their position. That is, it is hypothesized that $\sigma_{i1}^* = \sigma_{i2}^* = \sigma_{i3}^* = \sigma_{i4}^* = \sigma_{i5}^* = \sigma_{i6}^*$. For this model, the null-hypothesis is again H_0: $\sigma_{ih} = \sigma_{ih}$ the alternative hypothesis is H_1: $\sigma_{ih} = \sigma_i$. That means, the 138 virtual items can be reduced to 23 items.

Model 3 supposes as many position parameters as item positions are given, ignoring which item problem is given at this position. That is, in this model the item difficulty only refers to its position within the test booklet. Again, H_0: $\sigma_{ih} = \sigma_{ih}$, but now H_1: $\sigma_{ih} = \lambda_{(ig)}$.

Table 4 provides a formal description of the models:

Table 4:
Definition of the models

model 0	G1: G2: G3: G4: G5: G6:	$\sigma^*_{i1} = \sigma_{i1}$ $\sigma^*_{i2} = \sigma_{i2}$ $\sigma^*_{i3} = \sigma_{i3}$ $\sigma^*_{i4} = \sigma_{i4}$ $\sigma^*_{i5} = \sigma_{i5}$ $\sigma^*_{i6} = \sigma_{i6}$	$u = 6 \times 22 = 132$
model 1	G1, G2, G3, G4, G5, G6:	$\sigma^*_{i1} = \sigma^*_{i2} = \sigma^*_{i3} = \sigma^*_{i4} = \sigma^*_{i5} = \sigma^*_{i6} = \sigma_i + q_{(ig)}\lambda$	$u = 23$
model 2	G1, G2, G3, G4, G5, G6:	$\sigma^*_{i1} = \sigma^*_{i2} = \sigma^*_{i3} = \sigma^*_{i4} = \sigma^*_{i5} = \sigma^*_{i6} = \sigma_i$	$u = 22$
model 3	G1, G2, G3, G4, G5, G6:	$\sigma_{ih} = \sum_{h,l} q_{ihl}\eta_l + \lambda_{(ih)}$	$u = 22$

G1 (2,3,4,5,6) refers to groups 1 (2,3,4,5,6). σ^*_{ih} are the item parameters of the Rasch model analysis. σ_i refers to the item problem parameters ($k = 23$). λ is a global position or learning parameter, u is the number of parameters to be estimated within the 4 models. One parameter per model is not to estimate because of standardizing conditions.

The LLTM can be applied to an experimental design with hierarchical testing, in which independent parameters are estimated for each item under each experimental condition in a so-called *saturated model*. This constitutes the core condition of hierarchical testing. The saturated model can then be opposed to any model containing a restricted number of parameters using Likelihood Ratio tests. To estimate a saturated model for the present study, each person would have to answer to all test forms, which is impossible because of massive transfer effects. The conditions of independent parameters per item and experimental condition can also be met using a quasi-saturated model, where subgroups of persons work with different test forms. One parameter is estimated for each item under each condition, resulting in 6 × 23=138 virtual item parameters for the quasi-saturated model (model 0).

Results

The estimated log-likelihoods for the models can be seen in Table 5:

Table 5:
Degrees of freedom and log-Likelihood of the models

model		*df*	–ln L
model 0	quasi saturated model	132	3467,639947
model 1	23 item parameters, 1 position parameter	23	3524,062746
model 2	23 item parameters	22	3525,280193
model 3	23 position parameters	22	4565,347449

Andersen's Likelihood Ratio tests can determine whether models 1 through 3 describe the data as well as the quasi-saturated model 0. Hence, model 0 always represents the null-hypothesis, to be tested against models 1 through 3, which represent the respective alternative hypotheses. Table 6 gives a summary.

Table 6:
Results of the LRT

	df	χ^2	$\chi^2_{\alpha=.05}$
model 0 vs. model 1	109	112.8456	136.59
model 0 vs. model 2	110	115.2805	135.48
model 0 vs. model 3	110	297.5784	135.48

As shown, models 1 and 2 fit well, indicating that the data are not described any worse by a reduction from 138 (quasi-saturated model) to 23 (model 1) and 24 (model 2) parameters. Model 3, in which the likelihood of the data is explained solely through item position, must be rejected.

To test whether the position parameter contributes significantly to the model, one can easily stay within the framework of hierarchical testing: The Likelihood Ratio test opposes the probability of the data in model 1 to that of the data in model 2:

$$-2\left(3524.062746\text{-}3525.280193\right)=2.434894\overset{as}{\sim}\chi^2$$

As the critical χ^2 for 5% (df=1) is 3.841, the data are not explained any worse with the fewer number of parameters of model 2, the position parameter does not seem to be of value. Two graphs demonstrate the results:

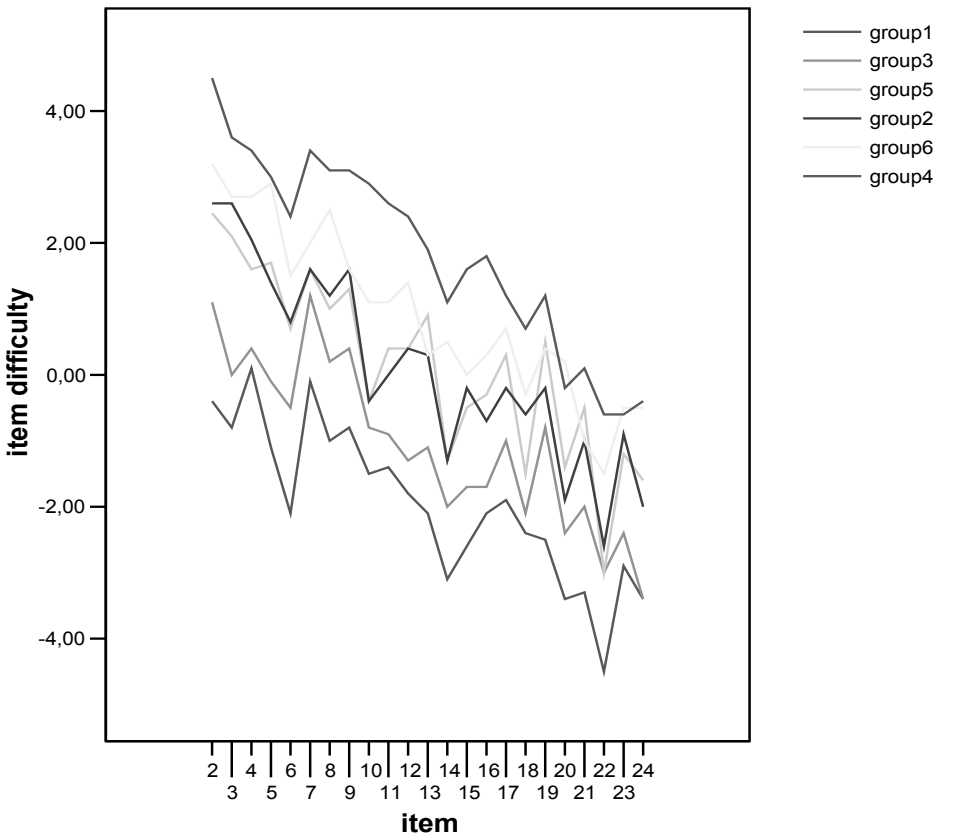

Figure 3:
Item difficulty parameters for all test groups.

In Figure 3, item difficulties of all test groups are ordered according to test Group 1, i.e. to the original item order of the Viennese Matrices. Please note that the item parameters are in fact item easiness parameters. The descending lines indicate that the items for all groups show nearly the same difficulties relative to the items preceding or following.

By contrast, Figure 4 shows the item difficulty parameters for the study groups ordered for item position. Each position, 1 through 23, in fact shows six different item difficulties. As position does not contribute to item difficulty, no distinct pattern can be seen.

Conclusion

The results indicate no evidence for position effects within the Viennese Matrices. In the context of this study, position effects were assumed as linear and constant for each position within the test series. In this case, the data are not explained any worse by means of only 23

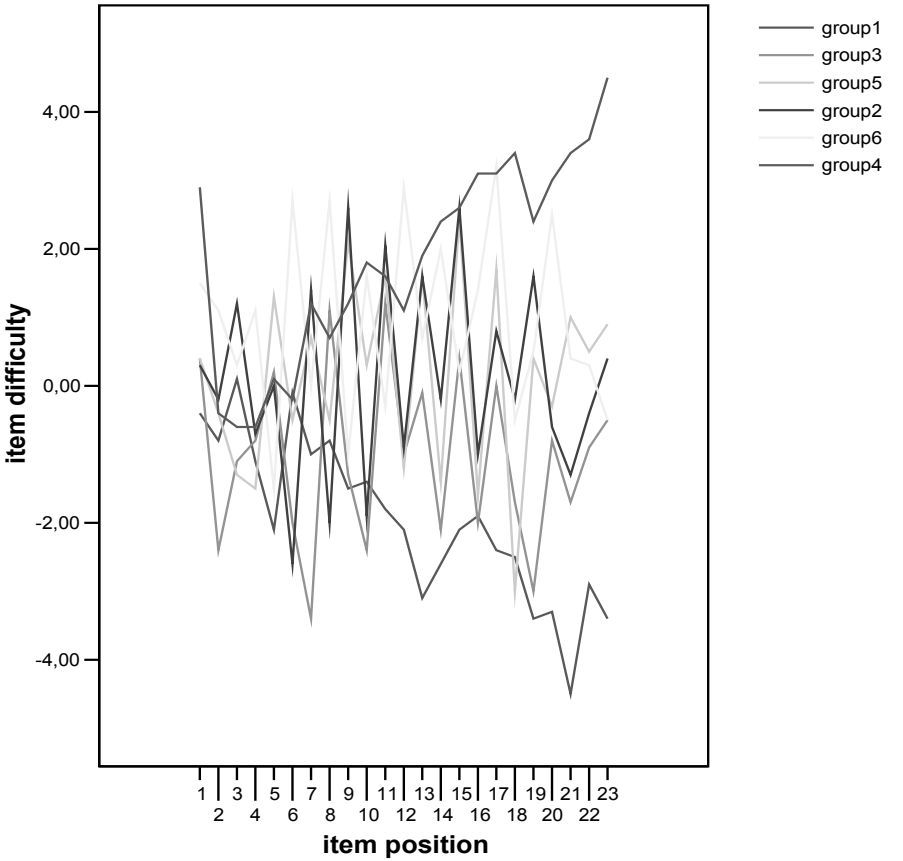

Figure 4:
Item difficulty parameters arranged according to item position

actual item problem parameters as opposed to 138 (virtual) item parameters in the quasi-saturated model. The LLTM makes it possible to deal with the problem of position effects rather economically. The testing occasion, however, may have contributed to this result insofar, as 5 items were presented before the test session started: One introductory example, three exercise items and one warming up item. The exercise examples include all three elementary operations that are responsible for item difficulty, even though in a very easy version.

References

Andersen, E. B. (1973). *Conditional interference and models for measuring.* Kopenhagen: Mentalhygienjnisk Forskingsinstitut.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359-374.

Fischer, G. H. (1995). Linear Logistic test model. In G. H. Fischer & I. W. Molenaar (eds.), *Rasch models. Foundations, Recent Developments, and Applications*. New York: Springer.

Fischer, G. H., & Ponocny-Seliger, E. (1998). *Structural Rasch modeling*. Handbook of Usage of LPCM-WIN 1.0.Groningen: ProGAMMA.

Formann, A. K., & Piswanger, K. (1979). *Viennese Matrizen test. Ein Rasch-skalierter sprachfreier Intelligenztest.* Weinheim: Beltz test.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3,* 380-396.

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrica, 64,* 407-433.

Embretson, S. E. (2002). Generating abstract reasoning items with cognitive theory. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development*. Mahwah, NJ: Lawrence Erlbaum.

Gittler, G. (1990). *Dreidimensionaler Würfeltest (Three-dimensional Cube test).* Weinheim: Beltz.

Gittler, G., & Wild, B. (1989). Der Einsatz des LLTM bei der Konstruktion eines Itempools für das adaptive Testen. [Using the LLTM for adaptive testing] In K. D. Kubinger (ed.), *Moderne Testtheorie. [Modern psychometrics]*. Weinheim: Beltz.

Hornke, L. F., & Rettig, K. (1989). Regelgeleitete Itemkonstruktion unter Zuhilfenahme kognitionspsychologischer Überlegungen. [Rule based item construction by means of cognition theory] In K. D. Kubinger (ed.), *Moderne Testtheorie. [Modern psychometrics]*. Weinheim: Beltz.

Hornke, L. F., Rettig, K., & Etzel, S. (1999). *Adaptive Matrices.* Mödling: Schuhfried.

Kubinger, K. D. (2005). Psychological test calibration using the Rasch model - Some critical suggestions on traditional approaches. *International Journal of testing, 5*, 377-394.

Kubinger, K. D. (2008). On the revival of the Rasch model-based LLTM: From composing tests by item generating rules to measuring item administration effects. *Psychology Science Quarterly, 50,* 311-327.

Raven, J. C., Raven J. & Court, J. H. (1998). *Raven's Progressive Matrices and Vocabulary Scales.* Frankfurt: Swets test Services.