Psychology Science Quarterly, Volume 50, 2008 (3), pp. 391-402

Examining item-position effects in large-scale assessment using the Linear Logistic Test Model

CHRISTINE HOHENSINN[1], KLAUS D. KUBINGER[2], MANUEL REIF[3], STEFANA HOLOCHER-ERTL[3], LALE KHORRAMDEL[3] & MARTINA FREBORT[3]

Abstract

When administering large-scale assessments, item-position effects are of particular importance because the applied test designs very often contain several test booklets with the same items presented at different test positions. Establishing such position effects would be most critical; it would mean that the estimated item parameters do not depend exclusively on the items' difficulties due to content but also on their presentation positions. As a consequence, item calibration would be biased. By means of the linear logistic test model (LLTM), item-position effects can be tested. In this paper, the results of a simulation study demonstrating how LLTM is indeed able to detect certain position effects in the framework of a large-scale assessment are presented first. Second, empirical item-position effects of a specific large-scale competence assessment in mathematics (4th grade students) are analyzed using the LLTM. The results indicate that a small fatigue effect seems to take place. The most important consequence of the given paper is that it is advisable to try pertinent simulation studies before an analysis of empirical data takes place; the reason is, that for the given example, the suggested Likelihood-Ratio test neither holds the nominal type-I-risk, nor qualifies as "robust", and furthermore occasionally shows very low power.

Key words: Rasch model, LLTM, large-scale assessment, item-position effects, item calibration

[1] Christine Hohensinn, M.Sc., Center of Testing and Consulting, Division of Psychological Assessment and Applied Psychometrics, Faculty of Psychology, University of Vienna, Liebiggasse 5, A-1010 Vienna, Austria, Europe; email: christine.hohensinn@univie.ac.at

[2] Klaus D. Kubinger, Ph. D., Chief of the Division of Psychological Assessment and Applied Psychometrics, Center of Testing and Consulting, Faculty of Psychology, University of Vienna, Liebiggasse 5, A-1010 Vienna, Austria, Europe; email: klaus.kubinger@univie.ac.at

[3] Manuel Reif, email: manuel.reif@univie.ac.at; Stefana Holocher-Ertl, M.Sc., email: stefana.holocher-ertl@univie.ac.at; Lale Khorramdel, M.Sc., email: lale.khorramdel@univie.ac.at; Martina Frebort, M.Sc., email: martina.frebort@univie.ac.at. All: Center of Testing and Consulting, Division of Psychological Assessment and Applied Psychometrics, Faculty of Psychology, University of Vienna, Liebiggasse 5, A-1010 Vienna, Austria, Europe.

1. Introduction

Of course, every test author has to be aware that there may be some item-position effects. There are two possible kinds of item-position effects. First, examinees generally become familiar with the test material and the task, respectively, or even detect problem solving strategies for the items in question. In other words, practice or learning effects could take place, so that an item at the beginning of a test is more difficult than the same item administered at the end of the test. Second, some fatigue effects could occur, so that an item at the beginning of a test is less difficult than the same item administered at the end of the test. Now, while such position effects do not raise any problems if they happen in general, that is with no notable individual differences, there are two cases of test application within psychological assessment where they cause considerable troubles.

Administration of a test in a conventional way, where every examinee is administered every item in the same sequence, means that any item-position effect as described above is absorbed into the item difficulty parameter. For instance, in terms of the Rasch model (1-PL model) item parameter, the calibrated σ_i represents not only item i's difficulty, say σ^*_i, but also some position effect λ, so that $\sigma_i = \sigma^*_i + \lambda$. This circumstance is hardly of any consequence, and the Rasch model might actually hold (nevertheless, for a review of unintended item-position effects on test scores, see Leary & Dorans, 1985; Zwick, 1991). However, if there are different sequences of item presentation, then the composition $\sigma^*_i + \lambda$ is of great importance, since it means that test performances of different examinees presented with different sequences of items are most likely not compared in a fair manner. This is due to the fact that one examinee might have an advantage working on a certain item at a certain position, while the other might be handicapped. Such different sequences of item presentation occur systematically within large-scale assessments where various test booklets with partly different item subsets are used. Of course, they also occur within adaptive testing, where (roughly speaking) every examinee is administered different items in different item sequences, because at any step of test administration an item is needed and sought whose difficulty best matches the current estimated ability parameter of the examinee in question. However, this paper focuses on item-position effects in the context of large-scale assessment.

Superficially there is no need for evaluation of the bias of test scores due to item-position effects in large-scale assessments, as these assessments regularly aim only for ability parameter estimations averaged within the given sample of the interesting population, not individually valid ability parameter estimations (cf. for instance the well-known PISA study, OECD, 2005). And then, of course, any item position can be balanced over all the test booklets, as a consequence of which the averaged ability parameter estimation becomes unbiased. However, some large-scale assessments provide additional feedback to every individual examinee, as well as sometimes to the individual class or school; in this case, when several test booklets have been used for instance to limit cheating, any item-position effect would invalidate the individuals' test results.

The Rasch model-based linear logistic test model (LLTM; see Fischer, 1972) is a proper means for analysing item-position effects (cf. Kubinger, 2008) and will therefore be applied in this study. The LLTM is particularly well suited, as several concurrent hypotheses of occurring item-position effects can be hypothesized and modeled, respectively; as a consequence of the results of hypotheses testing, the effect becomes specified and its size most accurately quantified.

While the Rasch model hypothesizes the probability of an item solution, given examinee v with ability parameter ξ_v and item i with item difficulty parameter σ_i, as:

$$P\left(+\middle|\xi_{\mathrm{v}},\sigma_{\mathrm{i}}\right)=\frac{\mathrm{e}^{\xi_{\mathrm{v}}-\sigma_{\mathrm{i}}}}{1+\mathrm{e}^{\xi_{\mathrm{v}}-\sigma_{\mathrm{i}}}} \tag{1}$$

the LLTM applies the linear combination:

$$\sigma_{\mathrm{i}}=\sum_{j}^{p} q_{ij}\eta_j \;\; ; p < k, k \text{ the number of items.} \tag{2}$$

That is, the item difficulty parameter σ_i is assumed to be a linear combination of the basic parameters η_j and some fixed weights q_{ij}. For parameter estimation, conditional maximum likelihood method (CML) may be used, which implicates advantageous mathematical properties in comparison to other parameter estimation methods (cf. Fischer, 1974).

The present paper now demonstrates how to apply the LLTM for testing item-position effects in the particular context of large-scale assessments (see, for instance, Gittler & Wild, 1989, as well as Hahne, 2008, for LLTM analyses of item-position effects of specific psychological tests). In order to do this, we first test in a simulation study whether at all the LLTM detects certain relevant item-position effects. Next, real data from a large-scale assessment are used in order to illustrate the indicated LLTM approach.

2. Method

In order to test item-position effects, a researcher must have access to a data structure encompassing different sequences of item presentation. Then, according to our considerations above, we must distinguish between items as content-specific tasks and items as some combination of content-specific tasks and concrete item-position (cf. Kubinger, 2008). Let us call the latter "virtual items". Then the difficulty of any virtual item is postulated as a linear combination of the given content-specific task h and the effect of the given position of that task within the test. That is, σ^*_h represents the difficulty of the content specific task h, which we call the "item root". Thus, if item root h is administered in different test booklets, for instance at l different presentation positions, then for the LLTM, l different virtual items result. As a consequence, the number of virtual items k amounts in sum to the number of item roots r multiplied with the number of different positions l at which the item roots are administered. Therefore, $\sigma_i = \sum_{j}^{p} q_{ij}\eta_j \;\rightarrow\; \sigma_i = \sum_{h}^{r} q_{ih}\sigma^*_h + q_{i(r+1)}\lambda$. However, $q_{i(r+1)}\lambda$, representing the item-position effect, can now be modeled in various ways. For example, a linear effect can be assumed, that is $q_{i(r+1)}$ is something like $t_i \in 1, 2, \ldots m$, according to virtual item's i position within the test – m the number of different item-positions. Alternatively, a non-linear effect can be assumed, for example $q_{i(r+1)} = 0$ or $q_{i(r+1)} = 1$, depending on whether the virtual item i is administered at item-position $t_i \leq t^*$ or at item-position $t_i > t^*$.

Application of the LLTM entails testing whether the model holds at all. For this, first the Rasch model has to hold for the entire virtual item pool (pertinent model checks are given, for instance, by Kubinger, 2005; see also Glas & Verhelst, 1995). If the Rasch model (RM) holds, it acts somehow like a saturated model. Then the question is whether the virtual items' σ_i are explainable by the LLTM's hypothesized linear combination $\sum_j^r q_{ij}\sigma_h^* + q_{i(r+1)}\lambda$. This question can be tested using a Likelihood-Ratio test suggested by Fischer (1974):

$$\chi^2 = -2\ln\left\{\frac{L_{LLTM}}{L_{RM}}\right\} \text{ asymptotically } \chi^2\text{-distributed with } df = k - (r+1) \tag{3}$$

– one degree of freedom will be gained if no item-position effect is hypothesized. Furthermore, it is possible to specifically test the hypothesis $\lambda = 0$ again according to Fischer (1974) using another Likelihood-Ratio test:

$$\chi^2 = -2\ln\left\{\frac{L_{LLTM(\lambda = 0)}}{L_{LLTM(\lambda \neq 0)}}\right\} \text{ asymptotically } \chi^2\text{-distributed with } df = 1 \tag{4}$$

3. Results of the simulation study

The data were simulated and analysed using the *R*-package *eRm* (Mair & Hatzinger, 2007a, b; see also Poinstingl, Hatzinger & Mair, 2007). Data simulation according to the Rasch model was done by analogy to some given empirical data, which actually should be analysed in the sequence.

The empirical data stem from a large-scale assessment using several test booklets, each of which consisted of 35 items (item roots); however, they belonged to four different subtests assessing four different mathematical competencies, so that from each subtest only about 9 items (item roots) were administered to an examinee. Bear in mind that the subtests' items were mixed in each test booklet; any given subtest's items were not administered successively but each item was followed by an item from another subtest. The subtest under consideration consists of 9 items (item roots), all of them administered in three different test booklets. 2 of these 9 operated as linking items. This means that they were presented at the very same position within all three test booklets and therefore allow a conjoint (virtual) item parameter calibration – in all the three test booklets they were presented at positions 34 and 35. The remaining 7 item roots were, however, presented as follows: Compared to Booklet 1, they were fully reversed in Booklet 2, and in Booklet 3 they were arranged randomly (cf. Fig. 1) – as a matter of fact, this resulted, by chance, in item root 6 being placed at the same position, 16, as in the Booklet 1 and 2 (nevertheless, only item roots 1 and 2 were treated as linking items).

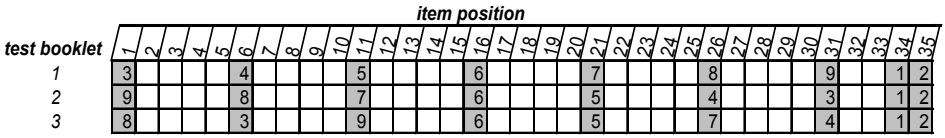

Figure 1:
Test design for the simulated data: There are three different test booklets, each containing 35 items (item roots), 9 of which are of interest and numbered from 1 to 9. Apart from item roots 1 and 2, the item roots were arranged in three different sequences.

Data was simulated for three different scenarios: (a) there is a linear learning effect, (b) there is a non-linear fatigue effect, (c) there is no position effect of item presentation.

The matrices $Q = ((q_{ij}))$ were defined according to these hypotheses. Table 1 outlines the scheme of Q for the hypothesis of a linear learning effect. That is, item root number 3 is used to create three virtual items; hence the difficulty of item root number 3 is added in Booklet 2 to the linear learning effect $q_{i(r+1)} = 31$ times λ.

For each hypotheses (a) and (b), the size of the simulated effect λ was varied. Furthermore, two different sample sizes, both seeming realistic for a large-scale assessment, were used. In the one case, an examinee sample size of $n = 300$ was suggested for each test booklet and in the other case, $n = 500$, so that for the planned number of three test booklets, a total sample size of $n = 900$ and $n = 1500$ resulted. Data were simulated 1000 times for each given condition. Table 2 gives an overview of the design of the simulation study.

Table 1:
A scheme of the matrix Q of LLTM analysis for the hypothesis of a linear learning effect; missing entries means $q_{ij} = 0$.

test booklet	Item root number	σ^*_1	σ^*_2	σ^*_3	λ
1	3	1			1
1	4		1		6
1	5			1	11
2	3	1			31
2	4		1		26
2	5			1	21
3	3	1			6
3	4		1		31
3	5			1	21
…	…	…	…	…	…

Table 2:
Design of the simulation study: There are three hypotheses, two different sample sizes of simulated examinees, and two effect sizes

	linear learning effect hypothesis		non-linear fatigue effect hypothesis		null-hypothesis
	λ= -0.01 $q_{i(r+1)}$ = 1, 2, …, 35 according to the position of item presentation	λ= -0.1 $q_{i(r+1)}$ = 1, 2, …, 35 according to the position of item presentation	λ= 0.2 $q_{i(r+1)}$ = 0 for $t_i \leq$ 24 and $q_{i(r+1)}$ = 1 for t_i > 24.	λ= 0.5 $q_{i(r+1)}$ = 0 for $t_i \leq$ 24 and $q_{i(r+1)}$ = 1 for t_i > 24.	λ= 0
n = 1500	1000 simulations	1000 simulations	1000 simulations	1000 simulations	1000 simulations
n = 900	1000 simulations	1000 simulations	1000 simulations	1000 simulations	1000 simulations

Table 3 presents the results of the power analyses of the given Likelihood-Ratio tests (LRT) for several hypotheses (i.e. the relative frequency of significant results discovered at the 1000 simulations each). Remember that data simulation for the virtual items was applied according to the given hypotheses; this means that, except for a type-I-error, the virtual item pool fits the Rasch model. The question of interest is, however, whether or not the Rasch model-based virtual item pool fits the hypothesized linear combination of LLTM-parameters as well: In the case of testing just that hypothesis according to which the data were simulated, the relative frequency of significant results indicates to what extent the nominal type-I-error actually holds. In the case of testing the hypothesis $\lambda = 0$ given that the data have been simulated according to $\lambda \neq 0$, the relative frequency of significant results indicates the power of the statistical test (i.e. 1 minus the type-II-risk).

First, Table 3 (Column II) discloses that the LRT does not at all hold the nominal type-I-risk. That is, given an item-position effect, parameterizing this within the LLTM (see rows A to D with $\lambda \neq 0$, but also row E with $\lambda = 0$) and comparing the LLTM and the Rasch model for all the virtual item parameters, the LRT must not be significant apart from a relative frequency corresponding to the used significance level of α = .01 due to chance effects (type-I-risk). However, instead of .01, the respective relative frequency is as high as .031. In no case but one does this frequency achieve 20-percent robustness (.008 ≤ actual estimated α ≤ .012), which for instance Rasch and Guiard (2004) have established to be acceptable for statistical tests. Of course, using 1000 simulations is not a state-of-the-art method within mathematical statistics (there, rather, 100 000 has proven to be proper). However, in the case of the given 1000 simulations, the analyses still lasted almost 7 hours. Table 4 also shows the accuracy of item parameter estimation for the case of the smaller fatigue effect for n = 900, where the means of the estimated item parameters are compared to the item parameters the simulation was based on: the initial item parameters were re-estimated most accurately with a quite low standard deviation. For the moment, we must conclude that the suggested asymptotic distribution of the LRT in Formula (3) does not apply for the given conditions. Perhaps it would be enough if the degrees of freedom were smaller, that is, if the number of virtual item parameters were not restrained by such a small number of LLTM parameters. Further research is needed.

Table 3:

Power analysis of the Likelihood-Ratio tests (LRT) for several hypotheses by simulation study. Columns II, III and IV give the relative frequency of significant results. Within every cell, the upper value represents the result for $n = 1500$ simulated examinees, the value below for $n = 900$. All analyses are based on the significance level of $\alpha = .01$. Comparisons with the Rasch model mean that all the virtual item parameters have their own and specific item parameter, whereas for the LLTM analyses the number of parameters is restricted according to the hypothesized linear combination. Column I gives the parameter estimations $\hat{\lambda}$ of λ.

Simulation scenario (1000 data sets each)		I		II	III	IV
		Estimation of $\hat{\lambda}$ by LLTM		LRT LLTM ($\lambda \neq 0$) vs. Rasch model $df = 12$ (cf. formula (3))	LRT LLTM ($\lambda = 0$) vs. Rasch model $df = 13$ (cf. formula (3))	LRT LLTM ($\lambda = 0$) vs. LLTM($\lambda \neq 0$) $df = 1$ (cf. formula (4))
		Mean of $\hat{\lambda}$	Standard deviation of $\hat{\lambda}$	Relative frequency $\chi^2 > \chi^2_{.01}$	Relative frequency $\chi^2 > \chi^2_{.01}$	Relative frequency $\chi^2 > \chi^2_{.01}$
A	learning effect: $\lambda = -0.1$ $q_{i(r+1)}=1,2,\ldots.35)$	-0.100 -0.100	0.005 0.006	.024 .022	1.000 1.000	1.000 1.000
B	learning effect: $\lambda = -0.01$ $q_{i(r+1)}=1,2,\ldots.35)$	-0.010 -0.010	0.004 0.005	.016 .017	.156 .088	.489 .281
C	fatigue: effect $\lambda = 0.5$ $q_{i(r+1)} = 0$ for $t_i \leq 24$ and $q_{i(r+1)} = 1$ for $t_i > 24$	0.497 0.494	0.082 0.105	.013 .011	.966 .722	1.000 .980
D	fatigue: effect $\lambda = 0.2$ $q_{i(r+1)} = 0$ for $t_i \leq 24$ and $q_{i(r+1)} = 1$ for $t_i > 24$	0.199 0.199	0.084 0.112	.016 .031	.128 .087	.417 .235
E	none position effect: $\lambda = 0.0$	0.000 0.000	0.004 0.005	.020 .013	.018 .014	.009 .003

Second, Table 3 (column III) discloses a very high power of the LRT if the item-position effect is large (row A): In this case, when the respective effect reaches 31 × -0.1 = -3.1 (which comes close to the typical difficulty of a very easy item), the power equals 1 – this applies even for the stronger fatigue effect (row C) if the sample size is large enough. Beside the former case, however, a smaller sample size of simulated examinees diminishes the power, even though the (fatigue) effect seems to be large (row C). Furthermore, in all cases of only moderate effects, the power of the LRT is very disappointing. As a matter of fact, all these conclusions hold for the LRT of formula (4) as well (cf. column IV) – though type-I-risk (row E) and power (rows B to D) come closer to the ideal.

As a consequence, we must be aware that we will not discover small item-position effects and that a slightly increased type-I-risk is given; for both handicaps of the desired statistical test, we finally have to apply rather Formula (4) as the basis of our interpretation.

Table 4:

Re-estimation of the item parameters. The simulation was based on a small fatigue effect (λ = 0.2) and sample size n = 900: means and standard deviations

number item root	initial item parameter	Rasch model: item parameter estimation		LLTM (λ = 0.2): item paramater estimation		LLTM (λ = 0): item paramater estimation	
		mean	standard deviation	mean	standard deviation	mean	standard deviation
1	0.70	0.70	0.08	0.70	0.08	0.70	0.08
2	-0.80	-0.81	0.09	-0.81	0.09	-0.81	0.09
3	2.95	2.98	0.25	2.96	0.13	3.02	0.13
3	3.15	3.19	0.27	3.16	0.15	3.02	0.13
3	2.95	2.97	0.25	2.96	0.13	3.02	0.13
4	2.20	2.22	0.21	2.21	0.12	2.33	0.10
4	2.40	2.41	0.21	2.40	0.11	2.33	0.10
4	2.40	2.42	0.21	2.40	0.11	2.33	0.10
5	1.45	1.46	0.17	1.45	0.09	1.45	0.09
5	1.45	1.45	0.17	1.45	0.09	1.45	0.09
5	1.45	1.45	0.17	1.45	0.09	1.45	0.09
6	-0.05	-0.05	0.15	-0.05	0.07	-0.05	0.07
6	-0.05	-0.05	0.15	-0.05	0.07	-0.05	0.07
6	-0.05	-0.05	0.15	-0.05	0.07	-0.05	0.07
7	-1.55	-1.57	0.18	-1.56	0.10	-1.49	0.09
7	-1.55	-1.57	0.18	-1.56	0.10	-1.49	0.09
7	-1.35	-1.36	0.17	-1.36	0.12	-1.49	0.09
8	-2.10	-2.12	0.20	-2.11	0.13	-2.24	0.10
8	-2.30	-2.31	0.21	-2.31	0.11	-2.24	0.10
8	-2.30	-2.31	0.21	-2.31	0.11	-2.24	0.10
9	-2.85	-2.88	0.24	-2.86	0.14	-2.99	0.12
9	-3.05	-3.07	0.26	-3.06	0.13	-2.99	0.12
9	-3.05	-3.08	0.26	-3.06	0.13	-2.99	0.12

4. Results for empirical large-scale assessment data

The mathematics competence test for 4th grade students of the Austrian educational standards project (cf. Kubinger et al., 2006; Kubinger et al., 2007) was designed to be used for large-scale assessment. In order to test many items for building a large item pool, 18 different test booklets with linked items were used. The test consists of four subtests testing four different abilities. In the following, we refer only to the first subtest "*Modelling,*" which includes 48 items (item roots).

The test was administered to 1792 students in all federal states of Austria. Each test booklet contained 35 items (item roots), including about 9 items from each of the four subtests. Test design used an item-wise allocation of items to test booklets. The sequence of item presentation of the subtest in question was chosen to vary arbitrarily, which is more or less randomly. Of course, the test being additionally restricted to a very limited number of test booklets made a completely balanced design of all the items over all positions impossible. Nearly every student completed the test within the given time limit of one class period.

Both a linear item-position effect and, alternatively, a non-linear item-position effect (in each case a learning/practice effect or a fatigue effect) were hypothesized – both effects were considered as applying to all items of a test booklet: Though there are four different subtests, the measured abilities are very likely to correlate; this means that both general learning effects and general fatigue effects could occur. That is, no subtest-specific effects were hypothesized. From a psychological point of view a linear increasing practice or fatigue effect is plausible and also a non-linear effect of the type, that working on the first items could produce a practice effect or – in contrast – solving the last items could be made more difficult through a fatigue effect.

Rasch model and LLTM analyses were again conducted using *eRm*. First, the data were analysed according to the Rasch model using state-of-the-art model checks, specifically Andersen's Likelihood-Ratio test and an additional graphical model check (cf. Kubinger, 2005). Item-fit-statistics were also calculated. Four partition criteria were chosen (nominal type-I-risk $\alpha = .01$): low *vs.* high score, male *vs.* female students, German mother tongue *vs.* mother tongue other than German, and Western *vs.* Eastern geographical region in Austria.

Before starting analyses, the overall sample was randomly split using a 70:30 ratio into a calibration sample and a validation sample ($n_c = 1187$, $n_v = 605$). Items with poor model fit in the calibration sample were deleted from the item pool. The result was an a-posteriori model-fitting item pool. In order to test whether that a-posteriori model fit was not just artificial, some kind of cross validation was necessary. Thus, the model fit of the reduced item pool was tested in the validation sample. In the calibration sample, only 2 of 48 items needed to be deleted. The remaining 46 items indicated good model fit in the validation sample (see the results in Table 5).

For further analyses, only the calibration sample was used. Because of the random allocation used, each item root was split into several virtual items according to its presentation position within the given test booklets as follows: Only when the item-position of the same item root differed between two test booklets at least by the number 5 then that item root was split into two different virtual items. If this had not been done, the test design would not offer the linking (virtual) items which are necessary for an analysis according to the Rasch model. The data matrix resulted in 103 virtual items. Testing whether the Rasch model holds for the virtual items also resulted in a proper model fit (see the results in Table 6).

Table 5:
Results of Andersen's Likelihood-Ratio tests with respect to the calibration and validation sample. The results refer to the analysis after the deletion of 2 non-fitting items.

	calibration sample after deletion of 2 items			validation sample		
sample's partition	χ^2 (LRT)	df	$\chi^2(\alpha = .01)$	χ^2 (LRT)	df	$\chi^2(\alpha = .01)$
Score	64.544	44	68.710	62.786	45	69.957
Gender	62.916	45	69.957	36.117	45	69.957
Mother tongue	62.771	45	69.957	70.535	45	69.957
Geographical Region	54.675	45	69.957	44.288	45	69.957

Table 6:
Results of Andersen's Likelihood-Ratio tests for all the $k = 103$ virtual items

sample' partition	χ^2 (LRT)	*df*	$\chi^2(\alpha = .01)$
Score	120.51	96	131.14
Gender	113.89	102	138.13
Mother tongue	112.10	97	132.31
Geographical Region	116.48	102	138.13

First, the hypothesis of a linear item-position effect was tested. Matrix Q was defined in a way analogous to the matrix in the simulation study, thus resulting in 46 item root difficulty parameters and the learning parameter λ. The weights q_{i47} were fixed to be equal to the item presentation number. LRT (Formula (3)) proved that if λ is set to zero, the LLTM does not fit ($\chi^2_{LRT} = 102.43$, $df = 57$, $\chi^2_{.01} = 84.73$). However, using LRT (Formula (3)) to test the given matrix with a learning parameter $\lambda \neq 0$ also results in significance ($\chi^2_{LRT} = 91.72$, $df = 56$, $\chi^2_{.01} = 83.51$). Nevertheless, the LRT of Formula (4) established that taking an item-position-effect parameter into account results in much better data fit ($\chi^2_{LRT} = 10.71$, $df=1$, $\chi^2_{.01} = 6.635$). The estimated learning parameter $\hat{\lambda}$ disclosed a small effect of 0.01, with a confidence interval at the 95%-level of: [0.04; 0.016]. But be aware, that for higher presentation numbers – given, for example, the presentation number 30 – the effect increases to 0.3. As a matter of fact, as the effect is positive, it is rather a continuous fatigue effect than a practice effect.

Second, the hypothesis of a non-linear item-position effect was tested. The last 10 items of a test booklet were assumed to establish this effect, so that item-position weights $q_{i(r+1)}$ were set to 0 for item roots presented at positions 1 to 24, but otherwise (for positions 25 to 35) were set to 1. The results are very similar to the first analyses: Even for $\lambda \neq 0$, LRT (Formula (3)) shows significance ($\chi^2_{LRT} = 91.82$, $df = 56$, $\chi^2_{.01} = 83.51$), but Formula (4) discloses a better fit with a hypothesized item-position parameter than without: ($\chi^2_{LRT} = 10.61$, $df=1$, $\chi^2_{.01} = 6.635$). The estimated parameter $\hat{\lambda} = 0.240$ (with the confidence interval

of [0.095; 0.384]) indicates a fatigue effect, though a small one in comparison to the effect in our simulation study.

5. Discussion

As pointed out in the introduction, the occurrence of item-position effects must be tested when the same items (item roots) are administered in different test booklets at different positions (or the test is administered adaptively at all); this is because ignoring given position effects means that different examinees would be compared in an unfair manner. At the very least, if only fatigue effects take place, then test administration should be changed by either shortening the number of administered test items or by implementing a resting period.

To summarize, LLTM has proven to be a proper means for testing item-position effects. We have focussed in this paper on the application of LLTM in the context of typical large-scale assessments.

As a byproduct, however, we established that before any analysis of empirical data is tried, simulation studies should be performed in accordance with the given empirical conditions. For instance, in our case, we disclosed that the nominal significance level does not hold; the type-I-risk comes up to three times of the given α.

Hence, the significant result in our empirical study is not at all unequivocal.

Using simulation studies, we have also realized that the power of the LRT is poor, in particular as concerns Formula (3).

Given that in our case Formula (4) nevertheless leads to significance, we conclude that an item-position effect indeed exists with respect to both types of hypotheses, a linear as well as a non-linear effect. As a matter of fact, both significant results are to be interpreted in the same way: especially at the very end of the test, the item solutions suffer from a fatigue effect.

Of course, besides general, examinee-independent learning or fatigue effects as dealt within this paper, there also might be examinee-specific effects. In most cases, these may be subsumed in the examinee's ability parameter estimation but are not to be taken further into account, unless such effects are again to some extent general. For instance, if an effect depends on the number of previous solved items, the dynamic test model by Kempf (1974), which is also Rasch model-based, may come into question.

References

Fischer, G. H. (1972). Conditional maximum-likelihood estimations of item parameters for a linear logistic test model. *Research Bulletin, 9,* Psychological Institute University of Vienna, Vienna.

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests. Grundlagen und Anwendungen* [Introduction into the theory of psychological tests - basics and applications]. Bern: Huber.

Gittler, G., & Wild, B. (1989). Der Einsatz des LLTM bei der Konstruktion eines Itempools für das adptive Testen [Using LLTM for adaptive test construction]. In K. D. Kubinger (Ed.), *Moderne Testtheorie - Ein Abriß samt neuesten Beiträgen* [Modern psychometrics – A brief survey with recent contributions] (pp. 115-139). Weinheim: Beltz.

Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch Model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models* (pp. 69-95). New York: Springer.

Hahne, J. (2008). Analyzing position effects within reasoning items using the LLTM for structurally incomplete data. *Psychology Science Quarterly, 50,* 379-390.

Kempf, W. F. (1974). Dynamische Modelle zur Messung sozialer Verhaltensdispositionen. [Dynamic models for measuring social behaviour dispositions]. In W. F. Kempf (Ed.), *Probabilistische Modelle in der Sozialpsychologie* [Probabilistic models in social psychology] (pp. 13-55). Bern: Huber.

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.

Kubinger, K. D., Frebort, M., Holocher-Ertl, S., Khorramdel, L., Sonnleitner, P., Weitensfelder, L., Hohensinn, C., & Reif, M. (2007). Large-Scale Assessment zu den Bildungsstandards in Österreich: Testkonzept, Testdurchführung und Ergebnisverwertung [Large-scale assessment of the Austrian education standards: conceptualization of the tests, test administration, and utilization of the test results]. *Erziehung und Unterricht, 157,* 588-599.

Kubinger, K. D., Frebort, M., Holocher-Ertl, S., & Pletschko, T. (2006). Standard-Tests zu den Bildungsstandards in Österreich – Wissenschaftlicher Hintergrund und Hinweise zur Interpretation der Ergebnisse der Standard-Tests [Tests of the Austrian education standards – psychometric background and interpretation rules of the test scores]. *Broschüre, Das Zukunftsministerium*: bmbwk, Wien.

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.

Mair, P., & Hatzinger, R. (2007a). eRm: Extended Rasch modeling. R package Vs 0.9.7. http://cran.r-project.org/

Mair, P., & Hatzinger, R. (2007b). CML based estimation of extended Rasch models with the eRm package in R. *Psychology Science* [laterly: *Psychology Science Quarterly*]*, 49,* 26-43.

OECD. (2005). *PISA 2003 Technical Report.* Paris: OECD.

Poinstingl, H., Mair, P. & Hatzinger, R. (2007). *Manual zum Softwarepackage eRm (extended Rasch modeling). Anwendung des Rasch-Modells (1-PL Modell) – Deutsche Version* [Manual of eRm. To apply the Rasch model – German Version]. Lengerich: Pabst.

Rasch, D., & Guiard, V. (2004). The Robustness of Parametric Statistical Methods. *Psychology Science* [laterly: *Psychology Science Quarterly*]*, 46,* 175-208.

Sonnleitner, P. (2008). Using the LLTM to evaluate an item generating system for reading comprehension. *Psychology Science Quarterly, 50*, 345-362.

Zwick, R. (1991). Effects of item order and context on estimation of NAEP reading proficiency. *Educational Measurement:Issues and Practice, 10*, 10-16.