Educational and Psychological Measurement
Volume 69 Number 2
April 2009 232-244
© 2009 SAGE Publications
10.1177/0013164408322021
http://epm.sagepub.com
hosted at
http://online.sagepub.com

Applications of the Linear Logistic Test Model in Psychometric Research

Klaus D. Kubinger
University of Vienna

The linear logistic test model (LLTM) breaks down the item parameter of the Rasch model as a linear combination of some hypothesized elementary parameters. Although the original purpose of applying the LLTM was primarily to generate test items with specified item difficulty, there are still many other potential applications, which may be of use for psychometric research on various testing conditions. This article provides some examples of such applications. The examples include (a) position effect of item presentation (in particular, learning and fatigue effects); (b) content-specific learning effect; (c) effect of speeded item presentation; and (d) effect of item response format.

Keywords: *Rasch model; LLTM; test administration; learning effect; multiple choice response format*

As is well known, the Rasch model (1-PL model) defines the probability that a test taker v with the ability parameter ξ_v will solve item i with the difficulty parameter σ_i as follows:

$$P(+|\xi_v, \sigma_i) = \frac{\mathrm{e}^{\xi_v - \sigma_i}}{1 + e^{\xi_v - \sigma_i}}$$

What is less well known is (see, for instance, Wilson & de Boeck, 2004) that Fischer (1972) introduced a special model in which all the difficulty parameters σ_i $(i = 1, 2, \ldots k)$ of the Rasch model are postulated as a linear combination of certain hypothesized elementary parameters η_j:

$$\sigma_i = \sum_j^p q_{ij}\eta_j.$$

The number $p < k$ of elementary parameters is also hypothesized; q_{ij} are suggested as being fixed and known weights. For this purpose, the model is called the linear logistic test model (LLTM). The methodical advantage of this model is that only p parameters have to be estimated instead of k parameters, which means an optimal gain from the data's information. Given that the Rasch model holds (for customary model checks of the Rasch model, see Kubinger, 2005), it acts as the saturated model and a goodness-of-fit test is applied by using a Likelihood-Ratio test: The data's likelihood in the LLTM, L_{LLTM}, is opposed to their likelihood in the Rasch model, L_{RM}, so that $-2\ln(L_{\mathrm{RM}}/L_{\mathrm{LLTM}})$ is asymptotically χ^2-distributed with $df = k - p$ (it should be pointed out that the respective likelihoods are based on conditional maximum likelihood parameter estimations; cf. Fischer, 1972).

An illustrative example of the original application of the LLTM is demonstrated with the Viennese Matrices (Formann & Piswanger, 1979), which are often used in German-speaking countries. The difficulty of an item is hypothesized as only depending on the kind and number of elementary operations (logical rules) that are used in order to reach the solution. These elementary operations are rules like, "increase the number of elements step-wise" or "vary their number," and "apply the rule horizontally as well as vertically." The practical advantage of modeling a test like the Viennese Matrices is, of course, in case the model actually holds, that an indefinite number of items could be created with whichever item difficulties a psychologist wants (cf. Fischer & Pendl, 1980).

Kubinger (1979, 1980) gives an example of a nontraditional but potential application of the LLTM. In addition to content-based elementary operations, some formal attributes of the item presentation were hypothesized and parameterized, for instance, item position and length of the text. Apart from item position effects, group-specific learning effects were also taken into account for the first time. Based on these studies, others began to investigate item position effects by applying the LLTM (cf. Dissauer, 1979; Gittler & Wild, 1989; Hahne, 1999). Yet, a systematic reflection of how the LLTM might work for this and similar purposes is still nonexistent. One should bear in mind that if effects of the position of item presentation do exist, then adaptive testing is absolutely unwarranted. Therefore, this article illustrates how psychometric research on various aspects of testing conditions can be carried out with the help of the LLTM.

LLTM for Different Testing Conditions

Position Effect

First, the effects of the position of item presentation are of interest. To test such effects, it is necessary that different subgroups of test takers are administered different sequences of (partly) the same items—at least one single item must have different positions within a certain test. For example, in the most extreme case, the sequence of item presentation would be completely reversed.

The LLTM should then be conceptualized as follows: From now on, we prefer to call an item with content h the "item root" h ($h = 1, 2, \ldots r$), however, we differentiate between such an item root (h) administered at position i, which therefore has the item difficulty parameter σ_i, and the same item root (h) administered at position l, which has the item difficulty parameter σ_l. That is to say, σ_i quantifies the difficulty of the "virtual" item i, and σ_l quantifies the difficulty of the virtual item l. We now define the difficulty of item root h as the item root difficulty and use σ_h^* to symbolize this. Bear in mind that this difficulty is the presumed difficulty of item root h in a standardized position "*" within the test. Hence, the first r LLTM elementary parameters η_h are redefined so that the special case results as $\eta_h = \sigma_h^*$. However, certain additional elementary parameter(s) are hypothesized according to the position in which the item root under consideration is administered. The interpretation of such a position parameter is that the difficulty increases or decreases the probability of solving an item root depending solely on a given position. So, $\sigma_i = \Sigma_j^p \; q_{ij}\eta_j$ is simplified to $\sigma_i = \sigma_h^* + \eta_{r+x}$, if x is the respective position. We can formalize our situation in detail, given, for instance, that $r = 4$ and there are two different sequences of presentation, the second one being completely reversed:

$$
\begin{aligned}
\sigma_1 &= \sigma_1^* + \eta_1 \\
\sigma_2 &= \sigma_2^* + \eta_2 \\
\sigma_3 &= \sigma_3^* + \eta_3 \\
\sigma_4 &= \sigma_4^* + \eta_4 \\
\sigma_5 &= \sigma_1^* + \eta_4 \\
\sigma_6 &= \sigma_2^* + \eta_3 \\
\sigma_7 &= \sigma_3^* + \eta_2 \\
\sigma_8 &= \sigma_4^* + \eta_1
\end{aligned}
$$

The structure of the LLTM's linear combination under consideration can now be better represented by only presenting the matrix of weights $((q_{ij}))$. In our case, there are $r = 4$ item roots, as a consequence of which there are $2 \times 4 = k = 8$ virtual items, and $4 = p - r$ position parameters (see Figure 1).

As p equals k, there is no need for the LLTM at all. However, we can, of course, think of more than two different sequences of presentation. For instance, there are $k = 16$ virtual items, if again, $r = 4$ item roots are used and $4 = p - r$ position parameters are supposed, and every item root is presented at every position – as a consequence of which $p = 8$ is much less than k.

Of principle importance is that certain hypotheses are tested. Given that the Rasch model holds for the k virtual items, the null hypothesis is as follows:

$$H_0\colon \eta_{r+x} = 0, \text{ for every } x = 1, 2, \ldots p - r; \text{this is equivalent to}$$

$$H_0\colon \sigma_i = \sigma_h^*, \text{ for every } i = h + r(x - 1); (i = 1, 2, \ldots, k), (h = 1, 2, \ldots p - r).$$

Obviously, $H_1\colon \eta_{r+x} \neq 0$. If H_0 is rejected, then position effects are given. Of course, any specific hypotheses, H_1^s, are possible, that is to say that a certain few position parameters $\eta_{r+x} \neq 0$.

Figure 1
The Linear Logistic Test Model's Matrix of Weights ((q_{ij})) for the Illustration

elementary operation j	1	2	3	4	5	6	7	8
virtual item i	item root A	item root B	item root C	item root D	position 1 within test	position 2 within test	position 3 within test	position 4 within test
1	1				1			
2		1				1		
3			1				1	
4				1				1
5	1							1
6		1					1	
7			1			1		
8				1	1			

Note: There are four item roots and two different sequences of presentation, the second one being completely reversed.

Furthermore, there are many other alternative hypotheses H_1^x in addition to H_1 and H_1^s: To start with, one could hypothesize a linear position effect. That is to say, a constant gradual increase or decrease of difficulty is assumed for the item presentation. In this case, the number of parameters of the LLTM is reduced to $p = 5$, and the weights q_{i5} are all either 1, 2, 3, or 4, depending on the virtual item i's position within the test. This is, therefore, a very strong hypothesis. A nonlinear function of position and difficulty seems a more likely alternative and feasible hypothesis; for instance, the weights q_{i5} could be fixed according to a logistic function: Instead of 1, 2, 3, and 4, the weights would then be 0.73, 0.88, 0.95, and 0.98 or likewise. Bear in mind that the latter has never been applied so far.

An empirical example. An adaptive test called the AID 2 (Adaptive Intelligence Diagnosticum–Version 2.1; Kubinger & Wurst, 2000) uses a branched test design. For instance, there are three interesting subsets of five items each within the subtest "Applied Computing." Eight- to 9-year-old test takers start with the arbitrarily labeled subset 4, and 10- to 11-year-old test takers start with subset 5. In the case that an 8- or 9-year-old test taker solves at least four items from subset 4, then subset 5 is next administered to him or her. And, if the test taker solves two or three items in this subset, then he or she finally gets subset 12. If, on the other hand, a 10- or 11-year-old test taker solves only one item at the most from subset 5, then

Figure 2
Data Design for Analyzing Item Position Effects Within the Adaptive Intelligence Diagnosticum–Version 2.1 (AID 2) Subtest "Applied Computing"

4-1	4-2	4-3	4-4	4-5	5-1	5-2	5-3	5-4	5-5	4-1	4-2	4-3	4-4	4-5	5-1	5-2	5-3	5-4	5-5	12-1	12-2	12-3	12-4	12-5	Item root
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	virtual item
1	2	3	4	5	6	7	8	9	10																Group A
										6	7	8	9	10	1	2	3	4	5						Group B
1	2	3	4	5																11	12	13	14	15	Group C
															1	2	3	4	5	11	12	13	14	15	Group D
1	2	3	4	5	6	7	8	9	10											11	12	13	14	15	Group E
										6	7	8	9	10	1	2	3	4	5	11	12	13	14	15	Group F

Note: There are 15 item roots that evolve into 25 virtual items. Six different groups of test takers have been administered 10 or 15 item roots in different sequences. The 15 item roots are grouped into three subsets ("4", "5", and "12") times five items; for instance, 12-5 represents the fifth item within subset 12. The gray shadowed boxes indicate that the respective item has been administered to the group in question. The numbers within the gray shadowed boxes refer to the position that the respective item root was administered.

subset 4 is next administered to him or her. And, if the test taker solves at least four items in this subset, then he or she finally gets subset 12. To go into more detail, the following six groups of test takers are of interest: Group A includes 8- or 9-year-old test takers, who have been tested with subset 4 first and then with subset 5. Group B includes 10- or 11-year-old test takers, who have been tested with subset 5 first and then with subset 4. Group C includes 8- or 9-year-old test takers, who have been tested with subset 4 first and then with subset 11, which is of no interest in this case, and finally with subset 12. Group D includes 10- or 11-year-old test takers, who have been tested with subset 5 first and then with subset 6, which is of no interest in this case, and finally with subset 12. Group E includes 8- or 9-year-old test takers, who have been tested with subset 4 first, then with subset 5, and finally with subset 12. And, Group F includes 10- or 11-year-old test takers, who have been tested with subset 5 first, then with subset 4, and finally with subset 12. The respective design is given in Figure 2. The structure of the LLTM's linear combination is then hypothesized as follows: There are $k = 25$ virtual items and $r = 15$ item root parameters with an additional linear effect position parameter.

Altogether, there were 176 test takers from the standardization sample. The 25 virtual items stood the test with regard to the Rasch model's fit of the data—the software LPCM-Win (Fischer & Ponocny-Seliger, 1998) and eRm (Mair & Hatzinger, 2006; cf. also Poinstingl, Mair, & Hatzinger, 2007) were used. That is not at all surprising, as the subtest as a whole was calibrated according to the Rasch model. Although the application of the LLTM disclosed a significant Likelihood-Ratio test, $\chi^2 = 23.07$, $df = 9$ ($\chi^2_{.01} = 21.67$), the goodness of fit is, descriptively seen, impressive: The Pearson correlation coefficient of the 25 virtual item parameter Rasch model estimations and those based on the hypothesized item root parameters and the position parameter amounts to .9659. In other words, the very restrictive hypothesized model to explain the virtual item difficulties does indeed work for practical purposes. Therefore, it is of interest as to whether the position effect significantly differs from zero. For this, an additional LLTM analysis had to be done. Now, only the item root parameters are hypothesized, not the position parameter. When the data's likelihood according to that structure of LLTM's linear combination, L_{LLTM^*}, was then opposed to their likelihood in the Rasch model again, a significant Likelihood-Ratio test close to the critical value resulted: $\chi^2 = 26.41$, $df = 10$ ($\chi^2_{.01} = 23.21$). The Pearson correlation coefficient equals .9655. Of more importance, however, is the comparison of the likelihoods L_{LLTM} and L_{LLTM^*}, which leads to a nonsignificant $\chi^2 = 3.34$, $df = 1$ ($\chi^2_{.01} = 6.64$). That is to say, there is no need to take a position effect on the item root difficulties into account. This means that adaptive testing is justified; otherwise, the test achievements of different test takers who were tested with the same items in different sequences would not be comparable in a fair manner.

Learning and Fatigue Effects

Position effects can obviously be interpreted as either learning effects or effects of fatigue. However, if changes in the probability of item solutions are due to psychological aspects of the test taker, then there is no sense in calculating any change of the item difficulty parameter. From a formal point of view, only the ability parameter ξ_v changes if test taker v works on items administered later on. That is to say, this parameter ξ_v is, in fact, modeled as a linear combination $\xi_v = \xi_v^* + \eta_i$. ξ_v^* is now the "ability root" of test taker v at the very beginning of the test administration. The elementary parameter η_i is then a learning (or fatigue) parameter, which depends on the fact that test taker v has worked on $i - 1$ items beforehand. Keep in mind that as η_i does not have a suffix v, we have hypothesized effects here that are independent of the test taker. Thus, the exponent of the numerator in the Rasch model formula, $\xi_v - \sigma_i$, can instead be broken down into $(\xi_v^* + \eta_i) - \sigma_i$ and also be broken into $\xi_v^* - (\sigma_i - \eta_i)$. In this way, the model actually becomes the LLTM. Of course, if we took such individually different effects η_{vi} into account, the LLTM would not work. A dynamic test model from Kempf (1977) exists, which takes individual learning effects into consideration, depending on the number of previously solved items. As this model aims to estimate the ability parameter within psychological assessment but is not suitable for fundamental research, it will not be considered further in this article.

Position effects considered until now have not been distinguished by any specific content but are, rather, either learning or fatigue effects. Although it has never been done before, a specific fatigue effect H_1^x could be tested: For instance, it is hypothesized that there is no fatigue effect up to a certain number of administered items; however, after this point, a fatigue effect occurs. In other words, up to the position $l = k_1$, all the weights amount to $q_{lj} = 0$, but from $l = k_1 + 1$ onward, it is $q_{lj} \neq 0$ (e.g., $q_{lj} = 1$). Once again, it is, of course, possible to hypothesize a linear or even a nonlinear function starting from position k_1.

Effect of Speeded Item Presentation

There are further testing conditions that can be analyzed with the LLTM; these are actually very specific position effects. On one hand, we have the warming-up effect, and on the other hand, we have the effect of speeded item presentation. For instance, the latter occurs within a group testing situation when there is a time limit to work on the given k items. As a result, some test takers only manage to finish k_1 items, so that the last $k - k_1$ items are not finished. Given, again, that different groups of test takers were tested with different sequences of item presentation, such an effect can also be analyzed by using the LLTM.

An empirical example. The Family Relations Reasoning Test (unpublished) has $r = 38$ item roots. It is assumed that a relevant number of test takers work on a maximum of 24 items. For this, a study was designed with four groups of test takers, A through D, which had a maximum of 16 items to work on (e.g., A: 1, 3, 6, 8, 10, 11, 13, 14, 16, 18, 20, 23, 26, 31, 36, 37), as well as an additional Group E, which had all 38 item roots. Hence, for the LLTM analysis, a respective matrix of weights resulted with $k = 52$ virtual items and a speed effect hypothesized for the last 14 items for Group E only.

Altogether, there were 264 test takers. The 52 virtual items stood the test with regard to the Rasch model's fit of the data. The application of the LLTM disclosed a significant Likelihood-Ratio test, $\chi^2 = 102.43$, $df = 13$ ($\chi^2_{.01} = 27.72$). That is to say, the item difficulties of the virtual items are not only explained by certain item root parameters and the hypothesized speed effect parameter. On the other hand, calculation—only for descriptive reasons—of the data's likelihood of an even more restrictive structure of LLTM's linear combination without a speed effect parameter, L_{LLTM^*}, led to a comparable Likelihood-Ratio test with $\chi^2 = 261.30$, $df = 14$. This means that LLTM's model fit became even worse. Hence, there is empirical support for assuming a certain speed effect. Therefore, another speed effect was hypothesized and another LLTM analysis was carried out. In this analysis, it was assumed that a progressive additional difficulty, instead of a constant one, arises due to speed from the 25th administered item onward. Instead of a 1 as the speed-based weight for each of the last 14 items of Group E, the weights 1 through 14 were used. That is to say, a linear increasing speed effect was hypothesized instead of a constant effect. As a matter of fact, the respective Likelihood-Ratio test was not significant, $\chi^2 = 16.95$, $df = 13$ ($\chi^2_{.01} = 27.72$). The Pearson correlation coefficient of the 52 virtual item parameter Rasch model estimations and those based on the hypothesized item root parameters and this speed effect parameter amounts to .9919. Although the results above do not make a comparison necessary, the comparison of the new likelihood L_{LLTM} and the former likelihood L_{LLTM^*} without any speed parameter leads to $\chi^2 = 244.34$, $df = 1$ ($\chi^2_{.01} = 6.64$). That is to say, serious linear speed effects, which increase depending on the number of items administered, have been established.

Effect of Item Response Format

There is hardly any evidence concerning the extent to which the difficulty of an item (root) depends on the chosen item response format. For instance, a free response format versus a multiple choice format would be of interest. In this case, the same item roots presented in different response formats create the virtual items.

An empirical example. The subtest “Knowledge of Test Inventory” from the Psychological Assessment Education Test, given in a textbook by Kubinger (2006), has $r = 14$ item roots. To establish the effect of the multiple choice response format on item difficulty, 5 of the 14 item roots were administered with two different response formats. Three different response formats were actually applied: a free response format (“F”), a multiple choice response format with one correct response option and five distracters (“1 from 6”), and a multiple choice response format with five response options, of which either none or one, two, three, four, or even all five may be correct—in this case, the test takers are not told how many response options are correct for each item, and an item was scored as solved only if all correct response options and none of the distracters were chosen by the test taker (“x from 5”). Because of the complexity of this example, the detailed matrix of weights is given in Figure 3. There are 14 item root parameters and three response format effect parameters for the $k = 19$ virtual items. Of course, any variation of doubled item roots with two different response formats would have been possible, and even every item root could easily have been used twice; furthermore, it would have been possible to design not only pairs but also triples of virtual items based on the same item root.

Figure 3
The Linear Logistic Test Model's Matrix of Weights ((q_{ij})) as an Example of Taking the Response Format Effects Into Account

→ *j*

virtual item *i*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	F	"x from 5"	"1 from 6"
1	1														1		
2		1														1	
3			1														1
4				1											1		
5					1											1	
6						1											1
7							1								1		
8								1								1	
9									1						1		
10										1					1		
11											1					1	
12												1			1		
13													1			1	
14														1	1		
15										1						1	
16											1						1
17												1					1
18													1				1
19														1		1	

Note: There are 14 item roots and two different subgroups of virtual items administered to two different groups of test takers. Three different response format effects are hypothesized.

As the same item root cannot be administered twice to the same test taker, even if another item response format is used, the 173 test takers were randomly allocated to two different groups. The first group was tested with the virtual items 1 through 14, and the second group with the virtual items 1 through 9 and 15 through 19.

The 19 virtual items stood the test with regard to the Rasch model's fit of the data. The application of the LLTM disclosed a nonsignificant Likelihood-Ratio test, $\chi^2 = 3.85$, $df = 3$ ($\chi^2_{.01} = 11.34$). The Pearson correlation coefficient of the 19 virtual item parameter Rasch model estimations and those based on the hypothesized item root and response format parameters amounts to .9932. Hence, there is no need to suppose a specific item difficulty of any virtual item; the difficulty actually only depends on the item root difficulty and the response format effect. An additional LLTM analysis was carried out to establish whether these effects significantly differ from zero. Here, only the item root parameters were hypothesized and the last three columns in Figure 3 were deleted. When the data's likelihood, according to the specific structure of the LLTM's linear combination, L_{LLTM^*}, was then opposed to their likelihood in the Rasch model, a significant Likelihood-Ratio test resulted, $\chi^2 = 156.97$, $df = 5$ ($\chi^2_{.01} = 15.09$). Alternatively, one could compare the likelihoods L_{LLTM} and L_{LLTM^*}, leading to $\chi^2 = 153.12$, $df = 2$ ($\chi^2_{.01} = 9.21$). That is to say, serious item response format effects were established.

Comparison of the resulting parameter estimations produces the following results (take into account that parameter estimations have been standardized, with the effect of the free response format ["F"] as zero): Response format "1 from 6" reduces the item difficulty by 0.6568 (the respective parameter estimation amounts to −0.6568), whereas response format "x from 5" increases it by 1.9515; bear in mind that the easiest item root has an estimated item difficulty parameter of −2.6930 and the most difficult one a parameter of 2.3013. To summarize, it can be said that, as expected, the multiple choice response format "1 from 6" reduces item difficulty in comparison to the free response format; however, the multiple choice response format "x from 5" increases the difficulty in comparison to the free response format, which is probably not intended.

Technical Considerations for LLTM Applications

To estimate the LLTM's parameters, each given matrix of weights $((q_{ij}))$ has to be standardized to a certain anchor, routinely $\sigma_1^* = 0$ and, if $x > 1$, $\eta_{r+x} = 0$; otherwise, the matrix would have full rank and the estimations become unequivocal. Hence, in the last example (in Figure 3), the first column, as well as the third to the last column "F," has to be removed. Then, the effects of response formats "x from 5" and "1 from 6" are simply interpreted in relation to the free response format, and the difficulty of any item root is simply interpreted in relation to item root 1.

As indicated, each suggestion in this article for an application of the LLTM for fundamental research is based on the existence of at least two different groups of test takers. As a consequence, none of the test takers is administered every virtual item; rather, every test taker has missing data with respect to a large proportion of the virtual items. That is to say, a data structure exists like the one in Figure 2. As shown in this example, it is absolutely necessary that the virtual items are linked to one another. In other words, statistically spoken (cf. Rasch & Kubinger, 2006), a connected, yet incomplete, balanced block design of virtual items and groups of test takers must be given. There must always be a path crossing the groups, in the sense that an item is administered starting from each certain virtual item *i* and ending at all the other virtual items $l \neq i$ (cf. Figure 2: For instance, a path starts from the virtual items 1 through 10 in Group E and bridges over to the virtual items 21 through 25 in the same group, whereas the virtual items 11 through 20 are linked via the same virtual items 21 through 25 in Group F. In other words, there would be an overlap of Group E's and Group F's virtual items if 1 through 10 were arranged on the tail instead of on the head of the virtual items 11 through 25, and obviously all 25 items would then be linked).

As traditional software for item response theory (IRT) models, in most cases, only deals with tests where the items are administered to every test taker, obtaining the estimations needed is now a matter of having relevant software at a researcher's disposal. However, the software LPCM-WIN and eRm, referred to above, actually deal with different groups of test takers, who are tested with different subsets of items.

Discussion

The question arises as to whether or not the LLTM is, in fact, necessary for the investigation of all aspects of formal conditions of psychological testing that have been discussed. Of course, each matrix of weights conforms to incomplete analysis of variance (ANOVA) designs, so that conventional statistical approaches would also seem to serve the purpose.

However, the LLTM in fact uses interval scaled (ability) parameters for testing each alternative hypothesis, whereas an ANOVA is only based on the ordinal scaled numbers of solved items; keep in mind that one of the most important features of the Rasch model is that the ordinal scaled numbers of solved items are transformed into interval scaled parameters (cf. Fischer, 1995). If the design is cross-classified and incomplete, then there are neither commonly known elaborations for two- or multiple-way ANOVAs for ranks, nor is there pertinent software available for researchers.

On the other hand, using Rasch model item parameter estimations for conventional statistical approaches instead of the LLTM would suffice. For instance, the example of the item response format effect would simply lead to an analysis using a paired *t*-test: All that has to be done is to insert the difficulty parameter estimations of the paired virtual items, which have the same item root and a different response format, in the respective formula. It is unfortunate, however, that the paired *t*-test would have to be applied three times ("F" vs. "1 from 6," "F" vs. "x from 5," and "1 from 6" vs. "x from 5"), which would mean a comparatively high Type I risk. Position effects could be analyzed in a similar manner (ANOVA for dependent samples), given that there are not too many positions and each item has been administered at every position—if the latter does not hold, then an ANOVA for independent samples would work.

Nevertheless, the LLTM does have an advantage: It serves to consecutively test a system of (alternative) hypotheses (cf. the example of speed effect). This system refers to a hierarchy of alternative hypotheses that result from the degree to which the model in question (the matrix of weights in question) comes close to the saturated model. This hierarchy may, for instance, concern a particular sequence as follows: (a) specific position effects versus (b) a linear position effect versus (c) a logistic position effect. In a similar way, although only exploratively and not inference statistically, one could look for the point on the trace line where a speed effect first occurs. With regard to the effects of different response formats, one could, for instance, test whether the effect of using 3 or 4 or 5 or 7 distracters on a multiple choice format proceeds in a linear manner or not. In addition, the hypothesis as to whether the effect of such a number of distracters disappears for any number larger than 7 but finally equals the effect of the free response format can be tested. There are, moreover, many other hypotheses that may be tested in this way.

References

Dissauer, G. (1979). *Lern-bzw. Übungseffekte innerhalb von Testreihen* [Learning and training effects within psychological tests]. Unpublished doctoral dissertation, University of Vienna, Austria.

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

Fischer, G. H. (1995). Derivations of the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models* (pp. 15-38). New York: Springer.

Fischer, G. H., & Pendl, P. (1980). Individualized testing on the basis of the dichotomous Rasch model. In L.J.D. van der Kamp, W. F. Langerak, & D.N.M. de Gruijter (Eds.), *Psychometrics for educational debates* (pp. 171-188). New York: John Wiley.

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

Formann, A. K., & Piswanger, K. (1979). *Wiener Matrizen-Test (WMT)* [Viennese Matrices]. Weinheim: Beltz.

Gittler, G., & Wild, B. (1989). Der Einsatz des LLTM bei der Konstruktion eines Itempools für das adaptive 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). Munich: PVU.

Hahne, J. (1999). *Lerneffekte innerhalb von Leistungstests* [Learning effects on achievement tests]. Unpublished master's thesis, University of Vienna, Austria.

Kempf, W. (1977). Dynamic models for the measurement of traits in social behavior. In W. Kempf & B. H. Repp (Eds.), *Mathematical models for social psychology* (pp. 14-58). Berne: Huber.

Kubinger, K. D. (1979). Das Problemlöseverhalten bei der statistischen Auswertung psychologischer Experimente. Ein Beispiel hochschuldidaktischer Forschung [Problem-solving behavior in the case of statistical analyses of psychological experiments. An example of research on university didactics]. *Zeitschrift für Experimentelle und Angewandte Psychologie*, *26*, 467-495.

Kubinger, K. D. (1980). Die Bestimmung der Effektivität universitärer Lehre unter Verwendung des Linearen Logistischen Testmodells von Fischer. Neue Ergebnisse [The evaluation of effectiveness of university lecturing with the help of the linear logistic test model by Fischer. New results]. *Archiv für Psychologie*, *133*, 69-79.

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. (2006). *Psychologische Diagnostik—Theorie und Praxis psychologischen Diagnostizierens* [Psychological assessment—Theory and application]. Göttingen: Hogrefe.

Kubinger, K. D., & Wurst, E. (2000). *Adaptives Intelligenz Diagnostikum–Version 2.1 (AID 2)* [Adaptive Intelligence Diagnosticum]. Göttingen: Beltz.

Mair, P., & Hatzinger, R. (2006). eRm: Extended Rasch modeling. R package version 0.9.5 [Computer software]. Retrieved from http://r-forge.r-project.org/

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., & Kubinger, K. D. (2006). *Statistik für das Psychologiestudium—Mit Softwareunterstützung zur Planung und Auswertung von Untersuchungen sowie zu sequentiellen Verfahren* [Statistics for the study of psychology—Software support for the planning of studies and sequential procedures]. Munich: Spektrum.

Wilson, M., & de Boeck, P. (2004). Descriptive and explanatory item response models. In P. de Boeck & M. Wilson (Eds.), *Explanatory item response models* (pp. 43-74). New York: Springer.