Psychology Science Quarterly, Volume 50, 2008 (3), pp. 311-327

On the revival of the Rasch model-based LLTM: From constructing tests using item generating rules to measuring item administration effects

KLAUS D. KUBINGER[1]

Abstract

The "linear logistic test model" (LLTM) breaks down the item parameter of the Rasch model into a linear combination of certain hypothesized elementary parameters. Apart from the originally intended primary application of generating an indefinite number of items composed of whichever item difficulties the examiner chooses, there are many other potential applications. They all deal with measuring certain item administration effects. This paper illustrates several of these approaches as well as how to design data sampling using the respective LLTM's structure matrix. These approaches deal with: a) Rasch model item calibration using data sampled consecutively in time but partly from the same examinees; b) measuring position effects of item presentation, in particular, learning and fatigue effects – specific for each position, as well as linear or non-linear; c) measuring content-specific learning effects; d) measuring warming-up effects; e) measuring effects of speeded item presentation; f) measuring effects of different item response formats. It is pointed out that the given LLTM approaches have the advantage of "elegance," as a hierarchical system of concurrent (alternative) hypotheses can be tested.

Key words: Rasch model, LLTM, item generating rules, position effects, multiple choice format

[1] Klaus D. Kubinger, PhD, University of Vienna, Faculty of Psychology, Head of the Division of Psychological Assessment and Applied Psychometrics, Liebiggasse 5, A-1010 Vienna, Austria, Europe; email: klaus.kubinger@univie.ac.at

1. Introduction

Originally, Scheiblechner (1972) – not yet calling his approach "linear logistic test model" – suggested predicting the Rasch model item parameter estimations of a psychological test by means of a (multiple) regression analysis and some cognitive operations which were hypothesized to be responsible for solving items. Fischer (1972, 1973) picked up this suggestion, but used parameter decomposition directly within the Rasch model formula instead of the multiple linear regression model. That is, he used a linear combination of some hypothesized elementary operation parameters η_j (j = 1, 2, … $p < k$) in order to explain every Rasch model item parameter σ_i, i = 1, 2, … k: $\sigma_i = \sum_j^p q_{ij}\eta_j$ – q_{ij} are postulated as being fixed and known weights, whose actual values are part of the hypothesis as well.

In detail: Whereas the well-known Rasch model (1-PL model) defines the probability of an examinee v with the ability parameter ξ_v of solving item i with the difficulty parameter σ_i as follows:

$$P\left(+\middle|\xi_v,\sigma_i\right)=\frac{e^{\xi_v-\sigma_i}}{1+e^{\xi_v-\sigma_i}} \tag{1}$$

the "linear logistic test model" (LLTM) specializes

$$P\left(+\middle|\xi_v;\sigma_i=\sum_j^p q_{ij}\eta_j\right)=\frac{e^{\xi_v-\sum_j^p q_{ij}\eta_j}}{1+e^{\xi_v-\sum_j^p q_{ij}\eta_j}} \tag{2}$$

Just as a conditional maximum likelihood estimation of the parameters σ_i in the Rasch model is possible, the same is true for the parameters η_j in the LLTM (cf. Fischer, 1995a). As a consequence, model checks according to Andersen's Likelihood-Ratio test are available as well. This means that in the first instance, the Rasch model must hold for the data under question. Given that the Rasch model holds, in particular according to Andersen's Likelihood-Ratio test (for state-of-the-art model checks of the Rasch model, see Kubinger, 2005), then a goodness-of-fit test is applied by using another Likelihood-Ratio test: The data's likelihood in the LLTM, L_{LLTM}, is opposed to its likelihood in the Rasch model, L_{RM}, so that $-2\ln(L_{\text{RM}} / L_{\text{LLTM}})$ is asymptotically χ^2-distributed with $df = k - p$ (or to say more precisely: the degrees of freedom are the number of estimated parameters in the Rasch model minus the number of estimated parameters in the LLTM).

However, while the Rasch model has spread particularly in the last two decades because of pertinent not "specifically objective" parameter estimations (i.e. some non-conditional-maximum-likelihood-based estimation approaches) and different generalizations (i.e. the mixed Rasch model by Rost, 1990, and the multidimensional Rasch model by Adams, Wilson & Wang, 1997), the LLTM has almost vanished from scientific research work (a most important exception is the book of De Boeck & Wilson, 2004, which contains even some generalizations of the LLTM, and of course the generalization of Embretson, 1997). For this

reason, the following paper serves to remind researchers of the LLTM's original feasibility within psychological test construction and furthermore suggest that the LLTM may serve as a proper means for fundamental research concerning different item administration effects within psychological testing.

2. Constructing tests using item generating rules

In constructing a psychological test, a researcher is always faced with the problem of validity. However, within psychological test theory there is also the concept of content validity, or the idea that a test is valid by definition. This is true, for instance, when solving the items requires exactly that competence by definition which a researcher aims to measure. If some cognitive rules must be applied in order to solve an item, and the solution is obtainable in no other way, then a high test score may be validly interpreted as a high level of competence in mastering these rules. This means that psychological test constructors may look for such cognitive rules with which to construct test items, and which are of practical relevance for diagnosing an examinee (given a certain psychological problem).

The LLTM can be used to this purpose. The first historic application of the LLTM gives an illustrative example. Based on a doctoral thesis, the *Viennese Matrices* (Formann & Piswanger, 1979) reasoning test contains items with difficulties that were hypothesized at test construction and have been proven to depend only on the kind and number of elementary operations (logical rules) necessary to reach the solution. For the sample item in Figure 1, these elementary operations are the rules: "increase the number step-wise" (with regard to the dots in the left bar), "vary the number" (with regard to the bars), and „apply a rule horizontally as well as vertically" (with regard to both the previous rules). Hence, what the test actually measures is known: the ability to master those well-defined rules. Admittedly, this disregards the question of whether that competence is of any ecological validity, but we are confident that it is.

Though not actually implemented within the *Viennese Matrices* or in another test being used for public counseling, the elementary operations parameters, once established, enable the test constructor principally to compose an almost infinite number of new items with any given item difficulty – an extremely useful possibility. This means that there are not only

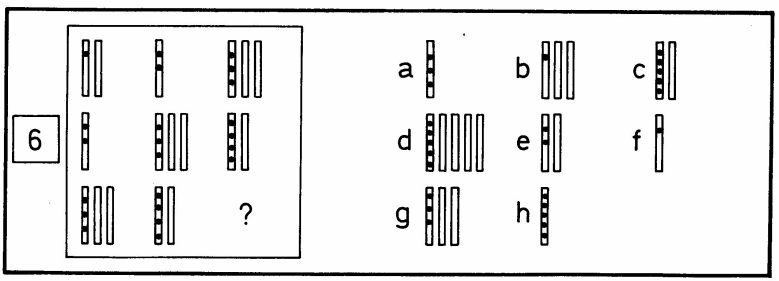

Figure 1:
Item 6 from the *Viennese Matrices* – the solution is "h".

item generating rules based on elementary (cognitive) operations but the difficulties of those operations are also known. Fischer and Pendl (1980) illustrated in detail using a non-empirical but arbitrarily numerical example, how item construction could work, and a few researchers followed – in particular very recently – with practical applications (cf. Gorin & Embretson, 2006; Newstead et al., 2006). This is of great importance, in particular for adaptive testing. As this requires having exactly that item at the examiner's (or computer's) disposal which best fits the currently estimated ability parameter for every examinee (cf. Kubinger, 2003), the optimal item could immediately be constructed by a respective combination of elementary operations.

In order to illustrate these first considerations, the so-called structure matrix $((q_{ij}))$ of the item difficulties' linear combination of elementary operation parameters is sketched for the *Viennese Matrices* in Figure 2.

Although composing tests using item generating rules has not yet been realized very often, LLTM's original feasibility led to its occasional application as a means of testing a psychological theory or of establishing psychological effects on rational decision-making models.

First of all, Kubinger (1979, 1980) investigated the problem-solving behavior of students taking a course in statistics. Of course, in this topic there is strict reglimentation, for instance as concerns testing the null-hypothesis that all means of several interesting populations are equal. That is, given the data of a certain study, the theory of statistics demands specific decisions following a (more or less) specific sequence in order to find the correct and best statistical test. Like the logical rules of the *Viennese Matrices,* these decisions can be inter-

elementary operation *j* / item *i*	1 increase the number step-wise	2 vary the number	3 apply a rule horizontally as well as vertically	4 …	5 …	6 …	…	p …
1				1	1			
2					1			1
3			1			1	1	
4		1	1					1
5	1							
6	1	1	2					
…						2		
…							2	1
…	1				1	2		
k		2	1	1				

Figure 2:

The structure matrix $((q_{ij}))$ of the item difficulties' linear combination of elementary operation parameters, abstractly reconstructed for the *Viennese Matrices* – only the weights q_{6j} for Item 6 are original; all the other items, elementary operations, and weights are fictitious (an empty cell means $q_{ij} = 0$).

preted as those elementary operations necessary for the solution of the items of a university statistics exam. For instance, the elementary operations and decisions are, respectively: "the variable in question is neither nominally scaled nor ordinally scaled but interval (or ratio) scaled", "there is the need to test or to rationally explain whether the variable is normally distributed", "the given samples delivers matched (paired) data", "there is the need to test the populations' homogeneity of variances", etc. So the structure matrix $((q_{ij}))$ is similar to that of Figure 2. The aim of the studies was to discover which of the decisions are more difficult for the students to make in a correct way and which are less difficult – and to accordingly revise the didactic measures. However, analyses proved that the examination items' difficulties were not explained by the elementary operations' (decisions') difficulties; the goodness-of-fit test – the data's likelihood in the LLTM opposed to its likelihood in the Rasch model – resulted in significance. As a consequence, it was concluded that students' problem-solving behavior does not fit the requirements of statistical theory. There must be characteristics of the items which determine their difficulties beyond the theory. As a matter of fact, taking some additional formal condition (elementary operation) parameters into account resulted in a non-significant goodness-of-fit test. Such formal condition parameters were in particular the position of the item within the examination booklet and the length of the item text (short *vs.* long). In other words, specific psychological effects on rational decision-making models have been established. Nevertheless, the analyses served their intended purpose, as the difficulties of the theory-based decisions could be quantified separately from the formal condition difficulties.

Secondly, Sonnleitner (2008) used the LLTM for testing specific cognitive models of reading comprehension. Again, some formal condition (elementary operation) parameters proved to be of great importance when testing reading comprehension, in particular the number of offered answer options of a multiple choice response format. And thirdly, Poinstingl (2008) analyzed a certain (lexical) reasoning test in order to determine to which extent the logical rules (elementary operations) necessary for a solution indeed exhaustively determined the items' difficulties. Once more, certain formal parameters established dramatic effects, due above all to the amount of irrelevant information given in the item. Finally, Wilson and De Boeck (2004) even tried to explain verbal aggression behavior through certain typical theory-based components (elementary operations); though they did not achieve a non-significant goodness-of-fit test, their results lead to a slightly better understanding of the theoretical background of the phenomenon of aggression.

3. Measuring item administration effects

As reviewed, Kubinger (1979, 1980) already used the LLTM for analyzing position effects of item presentation, and Sonnleitner (2008) established psychological effects of the design of a multiple choice response format using LLTM. That is to say, LLTM offers an appropriate method for fundamental research on different item administration effects within psychological testing. We will deal with various such approaches in the following sections.

3.1. Taking learning effects into account because of testing at different sessions

First of all, it was again Kubinger (1979, 1980) who was also compelled to take certain learning effects into account, because in order to have a sufficient sample size of examinees as well as of items at his disposal, he had to use the data from different examination sessions. The fact that several examinees took the examination (with different items) up to four times because of failing the test, raised the following problems: a) not every examinee is administered all the items of a given item pool but different groups of examinees are administered different subsets of items, b) one and the same examinee who is tested at two or more sessions in time is identical as concerns the actual personal identity but may differ with respect to the degree of the (intended) measured ability – that is, it is likely that the ability of an individual has changed between exams because of some learning effects. As concerns a), the consequence was to apply a software for Rasch model and LLTM parameter estimation that masters such data; nowadays we speak of having "missing values" (instead of having different subgroups of examinees administered different subsets of items) and preferably use the program package *eRm* (Mair & Hatzinger, 2006; cf. also Poinstingl, Mair & Hatzinger, 2007). Of course, parameter estimation is possible only if the item subsets of different examinee groups overlap in some manner through so-called linking items – we will deal with this topic in detail later.

As concerns b), Kubinger (1979) initially generalized the linear combination of the hypothesized elementary operation parameters in the expression $\xi_v - \sigma_i = \xi_v - \sum_j^p q_{ij}\eta_j$ as $\xi_v - \sigma_i = (\xi_v + \lambda_v) - \sum_j^p q_{ij}\eta_j$; $\lambda_v \neq 0$ if examinee v is tested the second, third, or forth time. Of course, this model is no longer a LLTM because now the ability parameter is also decomposed so that Fischer's entire algorithm of parameter estimation does not apply. However, if it is assumed that the learning (or repetition or re-administration) effect λ_v is constant for all examinees with the same combination of drawn sessions x and y ($y = x+1, x+2, \ldots x+4 \leq 4$; $x = 0, 1, \ldots, 4$), that is $\lambda_v = \lambda^{xy}$, then Fischer's linear combination only needs to be redefined. Given x and y, then $\xi_v - \sigma_i = \xi_v + (\lambda^{xy} - \sum_j^p q_{ij}\eta_j)$. Now σ_i is also a function of x and y; that is, an item's difficulty also depends on the combination of sessions at which examinee v has been tested, so σ_i must be specified as σ_i^{xy}. For this the labeling "virtual item" was found: one and the same problem i is represented in at least two different virtual items, one of them presented to any examinee tested the first time and another presented to any examinee tested the second (or third or forth) time. So not the actual problem i changes but its difficulty does: initially its difficulty is σ_i, then it is $\sigma_i^{xy} = \sigma_i + \lambda^{xy}$. In the most simple case – if there is no linear combination of elementary operation parameters to explain σ_i – then the LLTM reduces to: $\xi_v - \sigma_i^{xy} = \xi_v - (\sigma_i - q^{xy}\lambda^{xy})$, $q^{xy} = 1$ if examinee v is administered item i at session y after having already been tested at session x; otherwise, $q^{xy} = 0$. Of course, $(\sigma_i - q^{xy}\lambda^{xy})$ is a linear combination as well, so this is just a special LLTM. To illustrate: think of $k = 15$ problems i, $i = 1, 2, \ldots 5$ administered at session $x = 1$, $i = 6, 7, \ldots 10$ administered at session $x = y = 2$, and $i = 11, 12, \ldots k = 15$ administered at session $y = 3$. There are five subgroups of

examinees, subgroup 1/0/0 with examinees v, $v = 1, 2, \ldots n_{1/0/0}$, tested only at session $x = 1$, subgroup 1/1/0 with examinees v, $v = n_{1/0/0}+1, n_{1/0/0}+2, \ldots n_{1/0/0}+n_{1/1/0}$ also tested at session $x = 1$ and additionally at session $y = 2$, subgroup 1/0/1 with examinees v, $v = n_{1/0/0}+n_{1/1/0}+1, n_{1/0/0}+n_{1/1/0}+2, \ldots n_{1/0/0}+n_{1/1/0}+n_{1/0/1}$ tested at session $x = 1$ and $y = 3$, subgroup 0/1/0 with examinees v, $v = n_{1/0/0}+n_{1/1/0}+n_{1/0/1}+1, n_{1/0/0}+n_{1/1/0}+n_{1/0/1}+2, \ldots n_{1/0/0}+n_{1/1/0}+n_{1/0/1}+n_{0/1/0}$ tested only at session $x = 2$, and finally subgroup 0/1/1 with examinees v, $v = n_{1/0/0}+n_{1/1/0}+n_{1/0/1}+n_{0/1/0}+1, n_{1/0/0}+n_{1/1/0}+n_{1/0/1}+n_{0/1/0}+2, \ldots n_{1/0/0}+n_{1/1/0}+n_{1/0/1}+n_{0/1/0}+n_{0/1/1}$ tested at session $x = 2$ and $y = 3$. Then we have 15 problems but 20 virtual items, as the problems 6 to 10 are administered both to examinees without any respective examination routine and to examinees with a certain examination routine or general learning effect. Figure 3 demonstrates the interrelation of sessions, problems, virtual items, subgroups of examinees, and item difficulties. Obviously, not every examinee is administered all the virtual items of a given pool of virtual items. But Rasch model (virtual) item parameter estimation is possible (for 20 virtual items), as is LLTM parameter estimation (for 15 problems and the learning effect).

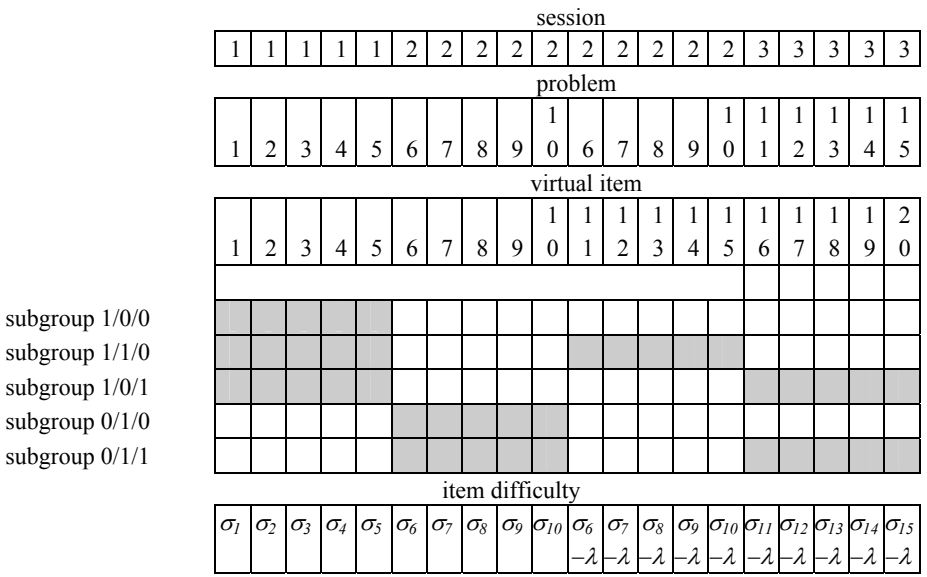

Figure 3:

An example of the interrelation of sessions, problems, virtual items, subgroups of examinees, and item difficulties of a data design to be analyzed by a special LLTM (gray shaded boxes symbolize that the respective virtual item is administered to the respective subgroup). In this case, different subgroups of examinees are tested at different terms and by different items. Hence, some learning parameter λ has to be taken into account, as a consequence of which not the original items (problems), but virtual items are analyzed using the Rasch model. If the LLTM holds, the learning parameter can indeed be separated and the problems can be calibrated according to the Rasch model.

Of course, this assumes that the learning effect is the same for every examinee – this presupposition can be tested by splitting the sample of examinees as Andersen's Likelihood-Ratio test does.

To summarize: the described method makes it possible to calibrate an item pool according to the Rasch model even if subgroups of examinees are tested repeatedly at different times using different items. This may be particularly relevant for large scale assessments where data sampling for the item calibration occurs consecutively in time but partially using the same examinees.

3.2. Analyzing position effects of item presentation

The fundamental considerations described above can be used to deal with the problem of position effects of item presentation. Again, applying the LLTM requires some transformation of the original items (problems) into virtual items and a data design with different subgroups of examinees. Examinees must be administered different sequences of (partly) the same items – at least one item must have different positions within two different versions of a given psychological test. For example, the most extreme case would be if the sequence of item presentation is completely reversed in two test versions.

The LLTM may then be conceptualized as follows.

From now on, we will call the content of an item (problem) h the "item root" h (h = 1, 2, … r) and distinguish between such an item root (h) administered at position i, which has the item difficulty parameter $\sigma_{hi} = \sigma_i$, and the same item root (h) administered at position l, which has the item difficulty parameter $\sigma_{hl} = \sigma_l$. That is to say, σ_i quantifies the difficulty of the virtual item i, σ_l quantifies the difficulty of the virtual item l. We now define the difficulty of the item root h as σ^*_h (irrespective of any position, or assuming that it is at some abstract standardized position "*" within the test). Hence, the first r LLTM elementary parameters η_h are redefined so that $\eta_h = \sigma^*_h$. In addition, a specific elementary parameter is hypothesized according to the position at which the given item root is administered. Like the learning parameter λ^{xy} in the section above, the interpretation of such a position parameter η_{r+x} (x is the respective position) is that its magnitude increases or decreases the probability of solving an item root, depending solely on the given position. So the expression $\xi_v - \sigma_i = \xi_v - \sum_j^p q_{ij}\eta_j$ is simplified to $\xi_v - \sigma_i = \xi_v - (\sigma^*_h + \eta_{r+x})$. As an illustration, think of r = 4 item roots and four different sequences of presentation. While the first two sequences might use a completely reversed item presentation, the third and the forth differ from both rather arbitrarily. Then we have 4×4 = k = 16 virtual items, and we will hypothesize 4 = p – r position parameters. The resulting structure matrix $((q_{ij}))$ is shown in Figure 4.

It is of principle importance to test certain hypotheses. Given that the Rasch model holds for the k virtual items, the null-hypothesis is

H_0: $\eta_{r+x} = 0$, for every x = 1, 2, … $p - r$ – this is equivalent to
H_0: $\sigma_i = \sigma^*_h$, for every $i = h + r(x - 1)$; (i = 1, 2, … k), (h = 1, 2, … r), (x = 1, 2, … $p - r$).

elementary operation *j* / virtual item *i*	1 item root A	2 item root B	3 item root C	4 item root D	5 position 1 within test	6 position 2 within test	7 position 3 within test	8 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			
9	1					1		
10		1						1
11			1		1			
12				1			1	
13	1						1	
14		1			1			
15			1					1
16				1		1		

Figure 4:
The LLTM's matrix of weights ((q_{ij})) for the case: 4 item roots and four different sequences of presentation – each position within item presentation is hypothesized as having its own special effect.

Obviously, H_1: $\eta_{r+x} \neq 0$. If H_0 is rejected, then position effects are corroborated. Of course, any special hypotheses (H_1^s) are possible, where a certain few position parameters are set $\eta_{r+x} \neq 0$. Furthermore, a number of alternative hypotheses H_1^x in addition to H_1 and H_1^s exist: 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 order of item presentation. This case would result in the structure matrix ((q_{ij})) shown in Figure 5. Here 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. This is therefore a very strong hypothesis. A non-linear 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.

elementary operation j / virtual item i	1 item root A	2 item root B	3 item root C	4 item root D	5 position within test
1	1				1
2		1			2
3			1		3
4				1	4
5	1				4
6		1			3
7			1		2
8				1	1
9	1				2
10		1			4
11			1		1
12				1	3
13	1				3
14		1			1
15			1		4
16				1	1

Figure 5:
The LLTM's matrix of weights $((q_{ij}))$ for the case: 4 item roots and four different sequences of presentation – a linear position effect is hypothesized.

Also bear in mind that if effects of the position of item presentation do exist, then adaptive testing is absolutely unwarranted. This is because there are always at least two examinees being administered the same item but at different positions. This means that position effects alter the difficulty of the (original) item and therefore one examinee is advantaged and another handicapped, depending on when they were administered the item – this is true as long as that position effect parameter is not taken into account.

In fact, using the LLTM in that way, Gittler and Wild (1989) have established for the first time significant position effects; they investigated a test of spatial ability. More recently, Hahne (2008) similarly analyzed the *Viennese Matrices*, but no significant effects were found. But Hohensinn et al. (2008) disclosed such an effect for a mathematics test for large scale assessment by the application of that LLTM's approach.

The position effects considered up until now are not characterized by any specific content, but are rather general, test-immanent learning or fatigue effects. To be more concrete, however, a specific fatigue effect H_1^x could be tested – though this has not yet been attempted. For instance, we could hypothesize that there is no fatigue effect up to a certain number of administered items, but after this point a fatigue effect occurs. In other words, up to the position $l = k_1$, all the weights of any position parameter amount to $q_{lj} = 0$, but from $l = k_1 + 1$, it is $q_{lj} \neq 0$, e.g. $q_{lj} = 1$ (compare the structure matrix $((q_{ij}))$ in Fig. 5 and substitute for

instance all $q_{i5} < 4$ with $q_{i5} = 0$, and all $q_{i5} = 4$ with $q_{i5} = 1$). Once again, it is of course possible to hypothesize a linear or even a non-linear function starting from position k_1.

Another specific effect might be a certain content-specific learning effect. For instance, Kubinger, Formann and Farkas (1991) disclosed that for the well-known *Standard Progressive Matrices,* pairs of items with the same content in terms of applied logical rules but incorporating different material components differ in difficulty depending on which item of the pair was administered first. For an illustration, see the structure matrix $((q_{ij}))$ in Figure 6; item 1 and item 2 are such a pair, item 3 and item 4 another one – all the items, 1 to 4, are administered in the same sequence for every examinee.

elementary operation *j* / item *i*	1 item root 1	2 item root 2	3 learning effect item typ 1	4 learning effect item typ 3
1	1			
2	1		1	
3		1		
4		1		1

Figure 6:
The LLTM's matrix of weights $((q_{ij}))$ as an example of content-specific learning effects: 2 pairs of items; the first item of each pair produces an effect on the second.

Given that the Rasch model holds for the k items (now there is no need for virtual items), the null-hypothesis is H_0: $\eta_{r+x} = 0$, for every $x = 1, 2, \ldots r$ (r being the number of pairs). Obviously, H_1: $\eta_{r+x} \neq 0$. If H_0 is to be accepted, then the pairs of items not only have the same logical rationale, but their difficulties have been empirically proven equal.

There are further item administration effects which can be analyzed by the LLTM: the warming-up effect and the effect of speeded item presentation. The warming-up effect occurs when an examinee is not familiar with the test's instructions or principle, but becomes acquainted with them after completing a few items. For this reason, so-called training items are usually given. Yet sometimes warming-up effects occur even when carefully constructed training items are administered. For instance, such effects are likely on the first and second administered items. Again, analyzing different subgroups of examinees who were tested with different sequences of items allows those effects to be tested. Figure 7 shows a structure matrix $((q_{ij}))$ for the case of $r = 4$ item roots, administered in two different sequences.

Given that the Rasch model holds for the k virtual items, then the null-hypothesis is H_0: $\eta_{r+x} = 0$ for every $x = 1, 2, \ldots p - r$ (r being the number of item roots, $p - r$ the number of warming-up effect parameters). Again, H_1: $\eta_{r+x} \neq 0$. If H_0 is rejected, then warming-up effects take place.

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

Figure 7:
The LLTM's matrix of weights ((q_{ij})) for the case: 4 item roots and two different presentation sequences – a warming-up effect for the first and second administered item is hypothesized.

The effect of speeded item presentation occurs within a group testing situation when there is a maximum time limit to work on the given k items. As a consequence, some examinees only manage to finish k_1 items, so that the last $k - k_1$ items are not finished. Given again that different subgroups of examinees are tested with different sequences of item presentation, then such an effect can also be analyzed using the LLTM. Take for instance $k = 9$ items and let us hypothesize that an important number of examinees work on a maximum of $2k/3$. Figure 8 shows the corresponding structure matrix ((q_{ij})), given exactly two different sequences of item presentation. Within the first sequence, the item roots G, H, and I are administered last so that a speed effect is assumed for them; this holds for the item roots D, E, and F within the second sequence.

Given that the Rasch model holds for the k virtual items, the null-hypothesis is H_0: $\eta_{r+1} = 0$, and H_1: $\eta_{r+1} \neq 0$. If H_0 is rejected, then a speed effect takes place. A numerical example of such an application of LLTM is given by Kubinger (in print).

3.3. Analyzing effects of the item response format

In particular because of guessing effects, which happen to occur if a multiple choice response format is used, it might be of fundamental research interest to analyze the psychometric difference of different kinds of multiple choice designs. Bear in mind that the well-known 3-PL model (and specifically the "difficulty plus guessing PL model", Kubinger & Draxler, 2006) takes an item-specific guessing parameter into account in order to estimate an examinee's ability parameter in a fair manner, but it does not show the general extent of guessing effects in dependence on a given kind of multiple choice response format. For instance, free response format *vs.* multiple choice response format would be of interest.

elementary operation *j*	1	2	3	4	5	6	7	8	9	10
	item root									
virtual item *i*	A	B	C	D	E	F	G	H	I	speed effect
1	1									
2		1								
3			1							
4				1						
5					1					
6						1				
7							1			1
8								1		1
9									1	1
10				1						1
11					1					1
12						1				1

Figure 8:
The LLTM's matrix of weights ((q_{ij})) for the case: $r = 9$ item roots and two different sequences of presentation, yielding $k = 12$ virtual items, because an effect of speeded item presentation is hypothesized for the last three administered items (these being item roots G, H, and I in the first sequence, D, E, and F in the second).

The LLTM applies if the same item root has different response formats. Again different subgroups must be tested using the same item roots but different response formats so that none of the examinees has to work on the same item root twice. For example, suppose that there are 5 item roots, all of which are administered with three different response formats: a free response format ("F"), a multiple choice response format with 4 distractors ("4"), and a multiple choice response format with 3 distractors ("3"). The respective structure matrix ((q_{ij})) is given in Figure 9. Of course, any variation is possible; one could, particularly, not use triples of virtual items based on the same item root, but pairs.

Given that the Rasch model holds for the k virtual items, the null-hypothesis is H_0: $\eta_{r+x} = 0$, for every $x = 1, 2, \ldots p - r$ ($p - r$ is the number of response format effect parameters). Evidently, H_1: $\eta_{r+x} \neq 0$. If H_0 is rejected, then the response formats differ with respect to their difficulty. This is most plausible to interprete, that in one case more lucky guessing takes place than in another case; the ratios of the respective parameter estimations indicate how many times guessing is (on average) more lucky using the one response format than using the other. A numerical example of such an application of LLTM is given by Kubinger (in print).

elementary operation j	1	2	3	4	5	6	7	8
	item root					response format effect		
virtual item i	Z	Y	X	W	V	"F"	"4"	"3"
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

Figure 9:
The LLTM's matrix of weights $((q_{ij}))$ for the case: $r = 5$ item roots yielding $k = 15$ virtual items administered to at least three different subgroups of examinees – three different response format effects are hypothesized.

4. Technical presuppositions

In order to estimate the LLTM's parameters, each given structure matrix $((q_{ij}))$ has to be standardized to some "anchor", this is routinely $\sigma^*_1 = 0$ and (given $p - r > 1$) $\eta_{r+1} = 0$; otherwise the matrix would have full rank and the estimations would become not unequivocal. Hence, in the last example (Fig. 9) the first column has to be removed as well as the one before both the last columns. Then the effects of response formats "4" and "3" 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 already indicated in Chapter 2, some of the suggestions for an application of the LLTM given in this paper are based on the existence of at least two different subgroups of examinees. In this case, none of the examinees is administered every virtual item, but rather every examinee has missing data with respect to a large proportion of the virtual items. That is to say, there is a data structure like the one in Figure 3. There the gray shaded boxes symbolize which of the virtual items has been administered to which subgroup of examinees. As

shown in this example, it is absolutely necessary that the virtual items are linked to each other. In other words, described statistically (cf. Rasch & Kubinger, 2006), a connected, though incomplete, balanced block design of virtual items and subgroups of examinees must be given: There always has to be a path crossing the subgroups, in the sense that an item is administered starting from each specific virtual item i and ending at all the other virtual items $l \neq i$ (in Fig. 3, this happens if the columns of the virtual items as well as the rows of the subgroups are properly interchanged; see Fig. 10).

As also already indicated in Chapter 2, this requires software which can handle different subgroups of examinees who are administered different groups of items. Though it has not long been at a researcher's disposal, *eRm* stood the test even in this respect. Above all, it is open source software and thus free of charge.

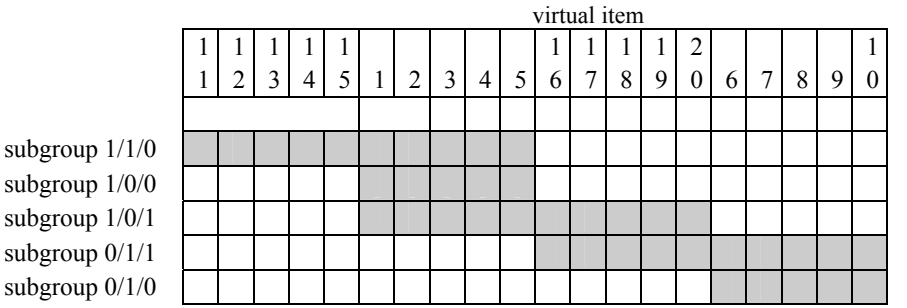

Figure 10:
Rearrangement of columns and rows of the data design given in Figure 3 (gray shaded boxes symbolize that the respective virtual item is administered to the respective subgroup).

5. Discussion

The question arises as to whether or not the LLTM is in actual fact necessary for the investigation of the aspects that have been discussed. Of course, some structure matrices ((q_{ij})) conform to incomplete analysis of variance designs so that conventional statistical approaches would also seem to serve this purpose: Given that the Rasch model holds for the (virtual) items, the respective interval-scaled (person or item) parameter might be used for testing some null-hypotheses – bear in mind that one of the most important characteristics of the Rasch model is that the ordinal-scaled numbers of solved items are transformed into interval-scaled parameters (cf. Fischer, 1995b).

Nevertheless, the given LLTM approaches have the advantage of "elegance." They consecutively test a system of hypotheses. This system refers to a hierarchy of alternative hypotheses which result from the degree to which the structure matrices ((q_{ij})) in question come close to the saturated model. This hierarchy may, for instance, concern a particular sequence as follows: 1) specific position effects *vs.* 2) a linear position effect *vs.* 3) 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 distractors on a multiple choice response format proceeds in a linear manner or

not; additionally, the hypothesis could be tested as to whether the effect of such a number of distractors disappears for any number larger than 7 but ultimately equals the effect of the free response format. Moreover, many other hypotheses may be tested in this way.

References

Adams, R. J., Wilson, M. R., & Wang, W. (1997). The multidimensional random coefficients multinomal logit. *Applied psychological Measurement, 21*, 1-23.

De Boeck, P., & Wilson, M. (Eds.) (2004). *Explanatory item response models*. New York: Springer.

Embretson, S. E. (1997). Multicomponent latent trait models. In W. van der Linden & R. Hambleton, *Handbook of modern item response theory* (pp. 305-322). New York: Springer.

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

Fischer, G. H. (1995a). The linear logistic test model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models* (pp. 131-155). New York: Springer.

Fischer, G. H. (1995b). 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: Wiley.

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 Konstruction 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.

Gorin, J., & Embretson, S. E. (2006). Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement, 30,* 394-411.

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

Hohensinn, C., Kubinger, K. D., Reif, M., Holocher-Ertl, S., Khorramdel, L., & Frebort, M. (2008). Examining Item-Position Effects in Large-Scale Assessment Using the Linear Logistic Test Model. *Psychology Science Quarterly, 50,* 391-402.

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 universities 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. (2003). Adaptives Testen [Adaptive testing]. In K. D. Kubinger & R. S. Jäger (Eds.), *Schlüsselbegriffe der Psychologischen Diagnostik* [Key-words of Psycho-diagnostics] (pp. 1-9). Weinheim: PVU.

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. (in print). Formal Conditions of Psychological Testing: Fundamental Research by the Linear Logistic Test-Model (LLTM). *Educational and Psychological Measurement.*

Kubinger, K. D., & Draxler, C. (2006). A comparison of the Rasch model and constrained item response theory models for pertinent psychological test data. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and Mixture Distribution Rasch Models - Extensions and Applications* (pp. 295-312). New York: Springer.

Kubinger, K. D., Formann, A. K., & Farkas, M. G. (1991). Psychometric shortcomings of Raven´s Standard Progressive Matrices (SPM) in particular for computerized testing. *European Review of Applied Psychology, 41*, 295-300.

Mair, P., & Hatzinger, R. (2006). eRm: extended Rasch models. R package version 0.9.5: http://r-forge.r-project.org/.

Newstead, S. E., Bradon, P., Handley, S. J., Dennis, I., & Evans, J. S. B. T. (2006). Predicting the difficulty of complex logical reasoning problems. *Thinking & Reasoning, 12*, 62-90.

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.

Poinstingl, H. (2008, in print). The LLTM as the basis for item generating rules of a new reasoning test: Family Relations Test. *Psychology Science Quarterly, 50.*

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.

Rost, J. (1990). Rasch model in latent classes: An integration of two approaches to item analysis. *Applied psychological Measurement, 14,* 271-282.

Scheiblechner, H. (1972). Das Lernen und Lösen komplexer Denkaufgaben [Learning and solving complex cognitive problems]. *Zeitschrift für Experimentelle und Angewandte Psychologie, 19,* 476-505.

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

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.