Psychology Science Quarterly, Volume 51, 2009 (2), pp. 123 - 134

The Linear Logistic Test Model (LLTM) as the methodological foundation of item generating rules for a new verbal reasoning test

HERBERT POINSTINGL[1]

Abstract

Based on the demand for new verbal reasoning tests to enrich psychological test inventory, a pilot version of a new test was analysed: the 'Family Relation Reasoning Test' (FRRT; Poinstingl, Kubinger, Skoda & Schechtner, forthcoming), in which several basic cognitive operations (logical rules) have been embedded/implemented. Given family relationships of varying complexity embedded in short stories, testees had to logically conclude the correct relationship between two individuals within a family. Using empirical data, the linear logistic test model (LLTM; Fischer, 1972), a special case of the Rasch model, was used to test the construct validity of the test: The hypothetically assumed basic cognitive operations had to explain the Rasch model's item difficulty parameters. After being shaped in LLTM's matrices of weights ((q_{ij})), none of these operations were corroborated by means of the Andersen's Likelihood Ratio Test.

Key words: Rasch model; LLTM; item generating rules; reasoning test

[1] Herbert Poinstingl, PhD, University of Vienna, Faculty of Psychology, Department of Developmental Psychology and Psychological Assessment, Liebiggasse 5, A-1010 Vienna, Austria;
Email: herbert.poinstingl@univie.ac.at

Introduction

A new kind of lexical reasoning test which was conceptualized more intuitively than on a theoretical basis is employed to analyse and finally establish item generating processes. For this reason, the Linear Logistic Test Model (LLTM; Fischer 1972) was used. For instance, Kubinger (2008) aimed at a revival of this model and revealed some applications of the LLTM, including test construction using item generating rules. In this study, the Linear Logistic Test Model is the method chosen to detect artefacts in the item construction process and to check if the FRRT is an appropriate procedure for measuring 'verbal reasoning'.

This assumes that the item parameters of the Rasch model (RM; Rasch, 1960) can be decomposed into a weighted sum of additive 'basic parameters' η_i plus a normalization constant c. The weights must not be random variables und have to be determined before parameter estimation. The purpose of the LLTM was to analyze change under different conditions and to describe item difficulty in terms of rules and basic cognitive operations of the item material.

$$P(+|\xi_v;\sigma_i) = \sum_j^p q_{ij}\,\eta_j = \frac{\exp(\xi_v - \sum_j^p q_{ij}\,\eta_j)}{1+\exp(\xi_v - \sum_j^p q_{ij}\,\eta_j)} \quad (1)$$

Conditional maximum likelihood estimation procedures are available for both the parameters σ_i in the Rasch model and the parameters η_j in the LLTM. Similarly, model checks like Andersen's Likelihood Ratio Test can also be applied to the LLTM. Based on the basic cognitive components that may be detected, a theoretically infinite number of items may be constructed. With our aim of disclosing or rather hypothesizing item generating rules, the rows of the LLTM's matrix of weights ((q_{ij})) thus consisted of basic cognitive operations and the columns consisted of the generated items.

The construction of the FRRT

The design of the FRRT is rather simple. The examinee/testee has to read a short story consisting of relationships between several family members and to find the correct relationship between two distinct members of the family; the solution is hidden in a large number of choices.

For example: "Bill is the father of Mary und Susan. Cathy is the daughter of Susan. What is the relationship between Bill and Cathy?" In this simple example the solution is 'grandfather'.

The items of the FRRT were composed using only two construction rules. The first rule is called 'complexity of family relationships' and distinguishes between four groups. Group 1 is called 'nuclear family'. Examples are 'father', 'mother', 'brother', 'sister', and so on. Group 2 concerns a 'relation in the second degree'. Representatives of this kind are 'uncle', 'cousin', 'aunt', and so an. In Group 3, family members were extended by in-laws, such as 'brother-in-law' and 'sister-in-law'. Group 4 is called 'patchwork family'. Representatives of this group

are 'stepfather' and 'stepbrother'. As mentioned above, the item difficulty is most likely determined to some extent by the group membership of the solution. It is assumed that relationships from Group 2 are more difficult to name than relationships from Group 1. Solutions from Group 3 are more demanding than solutions from Group 2. And the most sophisticated relationships can be found in Group 4, the group of the in-laws.

The second rule to denote the item difficulty is called 'total number of relations used in the item'. Theoretically, items can be constructed by the application of these item generating rules. A number of items were created for different levels of difficulty. All in all, roughly 100 items were generated.

Examples of the FRRT
Easy item: Kurt's son, Tobias, has a son. What is the relationship between this son and Kurt? (Answer: grandson)
Difficult item: Angela has only one cousin called Rafaela. Rafaela is the daughter of Edith and Engelbert. Edith is the aunt of Angela und the sister of her father, Helmut. Edwin is Edith's husband. What is the relationship between Edwin and Helmut's niece? (Answer: stepfather)

Figure 1:
Examples for different levels of difficulty

Method

Since a fitting Rasch model is necessary in order to conduct LLTM analyses, the first step of analysis consisted of applying dichotomous Rasch model analyses. The basic equation of the Rasch model (cf. Kubinger, 2009) defines the probability that a test taker with ability parameter ξ_v solves item i with the difficulty parameter σ_i. There are several feasible ways to test the fit of the Rasch model. In this article, the Likelihood Ratio Test (LRT, Andersen, 1973), the graphical model test (Rasch, 1960), and a Wald-type test (Glas & Verhelst, 1995) were the preferred methods for examining the psychometric qualities of the FRRT. All tests are supported by the software package 'extended Rasch modeling' (eRm, Mair & Hatzinger, 2007; see Poinstingl, Hatzinger & Mair, 2008). The LRT (cf. Kubinger, 1989) provides only an examination of the total data set; the Wald type-test was used for itemwise examination of the data (cf. Kubinger, 2005).

Testing the psychometric qualities of the FRRT

In a first sample (Skoda, 2005), (secondary school) students (n=264) were tested with the FRRT while no time limit was set. The four different test booklets consisting of overlapping link items were administered to four subgroups of students. Specifically, the fourth group was used to link the first three groups in order to assure comparability between the results of the subgroups. The age of the students varied from 14 years to 18 years. Missing values by design are handled by the software eRm (Mair & Hatzinger, 2007) selected for the Rasch model estimations in this study, and the few missing values caused by test takers were re-

corded as not solved. A second sample (Placek, 2005) provided a fifth group (n=134), again with overlapping or linked items. This time, however, the early conclusion of testing due to an administrative time limit caused a high percentage of missing values for the items administered at the end of each booklet. Consequently, only the first 25 administered items could be used for research work, in order to ensure that there was no impact on items administered in the test situation with a time limit and that the two samples are comparable. The age of the students varied from 12 to 14 years.

As the validity of the Rasch model is a necessary condition for a fitting LLTM, the fit of the Rasch model was first tested. Applications of the Rasch model entailed the deletion of several items. Two split criteria (score, age) were used for Andersen's LRT (cf. Table 1). All in all, 49 (of the 100) items did not fit the Rasch model and had to be deleted. An examination of the content of these items did not reveal any obvious reason for their misfit. Another 4 items had to be eliminated because of erroneous item construction. All in all, 47 items remained in the item pool.

Table 1:

H_0: 'The Rasch model is valid' holds in a data set with 47 items

Likelihood Ratio Test				
Split criterion	χ^2	*df*	$\chi^2_{krit\ (\alpha = .05)}$	*p*
Score	53.921	46	62.8296	0.106
Sample	17.219	19	30.14353	0.575
Sex	36.738	45	61.65623	0.805

All in all, a data set consisting of 398 testees and 47 items was used for further psychometric investigation.

Construct validity of the collected data set

The basic cognitive operations needed for testing the hypothesis about item generation and for checking the construct validity were determined using theoretical a priori assumptions and thorough item analyses. By shaping the assumed item generating rules in the LLTM's matrix of weights ((q_{ij})), hypotheses are specified in this data matrix as well; by examining the data with the LLTM, these hypotheses are tested. If the hypotheses concerning the item construction process result in a valid LLTM, then construct validity is assumed and the components determining the item difficulty are identified.

In this study, a close investigation of the items suggested a considerable number of basic item components. These components are listed in Table 2.

Table 2:
Basic item components modelled in LLTM's matrices of weights ((q_{ij}))

Nr	Basic item component	Description
1	Complexity of family relations	All items of the FRRT were generated in an intuitive process by applying the rules 'total number of relations in the item" and 'degree of complexity of the family relations'.
2	Total number of relations	
3	Position effects	The position of an item in the test (beginning, middle, end) could lead to position effects like fatigue and learning.
4	Number of names used in the item	It is assumed that the occurrence of many names indicates a large number of relations or a high complexity of the relations in the items. This is assumed to influence reasoning ability and working memory load.
5	Number of words used in the item	A long story text increases the item difficulty by increasing memory load.
6	Number of characters in the item	A high number of characters indicates a long story text and a high memory load.
7	Number of relations needed to solve the item	The total number of relations equals the number of relations needed to solve the item plus the number of unnecessarily mentioned relations. This means that the number of needed relations is the number of unnecessary relations subtracted from the total number of relations. Unnecessary relations are not needed in the item solving process and only distract the test person. The difference between the number of needed relations and the total number of relations is quite large in some items.
8	Difference 'total number of relations in the item' - 'number of items needed for solving'	This component is dependent on the number of needed relations (7) and the total number of relations (2) in the item.

Since it was assumed that all items consist of the basic components denoted above, the postulated basic components were used to check the construct validity of the FRRT-items. Once again, the conditional Likelihood Ratio Test was used to test for fit

$$-2\left(\ln L_{LLTM} - \ln L_{RM}\right) \overset{\approx}{\sim} \chi^2 \tag{3}$$

with *df* as the number of linear independent columns in LLTM's matrices of weights ((q_{ij})) (see Kubinger, 2008). The second method used to check the fit of the LLTM was the graphical goodness-of-fit test (cf. Kubinger, 2005).

Checking the construct validity with LLTM's matrix of weights ((q_{ij}))

Two approaches are presented in order to demonstrate the examination of the construct validity. In Approach 1, the original item generating rules were used to design the LLTM's matrix of weights ((q_{ij})). In Approach 2, a promising combination of item generating rules was used to check the construct validity.

Figure 2 shows the structure of the tested LLTM's matrices of weights ((q_{ij})). The rows denote the structure matrices and the columns denote the basic components of the structure matrices. All in all, a number of LLTM's matrices of weights ((q_{ij})) were designed in order to examine the construct validity. Some of the LLTM's matrices of weights ((q_{ij})) caused difficulties by becoming singular. Since singularity is a sign of dependence in a matrix, a simple solution is impossible and the singular matrices had to be discarded from further computations. The singularity of the structure matrices was facilitated by their dichotomy. In general, LLTM's matrices of weights ((q_{ij})) can be constructed using all possible numbers, including fractions, but in this case only dichotomous values were assigned to the structure matrices. The dichotomy in the matrix of weights ((q_{ij})) (Table 3) can be exemplarily described by the basic component "number of names". In this example a left column consisting of "1" denotes that only few names occur in the items (e.g. Item 1) and a "1" in the right column denotes that many names are included in the item (e.g. Item 47).

In Figure 2 q_{1ij} (Approach 1), q_{8ij} (Approach 2) are the names of the LLTM's matrix of weights ((q_{ij})) in the rows; the columns represent the basic components (black denotes 'rule is applied', white denotes 'rule is not applied').

structure matrix	Item position in the test (3)	Complexity of family relation (1)	Number of names (4)	Number of words (5)	Number of characters (6)	Number of total relations (2)	Number of needed rel. (7)	Rel. total - Rel. needed (8)
q_{1ij}		■				■		
q_{8ij}	■	■	■	■	■			

Figure 2:
Structure of the dichotomous LLTM's matrix of weights ((q_{ij})) used for examining construct validity of the FRRT

Approach 1: an examination of the construct validity using LLTM's matrix of weights ((q_{1ij}))

Since the basic operations 'number of relations' and 'complexity of family relations' were used in the FRRT for item construction, these rules were modelled in the LLTM's matrix of weights ((q_{1ij})) in order to estimate a LLTM. Afterwards, the estimated basis parameters were used to predict the item parameters in order to obtain the model fit for the data sets. A first feasibility check for the data set is the graphical model check, where the rescaled item parameters of the LLTM are plotted against the item parameter estimates of the Rasch model. In Figure 3, the rather poor result of the graphical model check is presented. A second feasibility check is the Likelihood Ratio Test, where the validity of the linear decompo-

sition of the LLTM is tested. The results (Table 4) denote a significant χ^2 -value. In conclusion, the hypothesis that the items were constructed by the exertion of the original two rules had to be discarded. The results showed that the two originally applied item generating rules were not reflected in the data and that construct validity was clearly not given under the assumption that the original item construction rules were used in the FRRT.

Table 3:
The original item generating rules "Complexity of family relations" and "Number of relations" modelled by LLTM's matrix of weights ((q_{1ij}))

item	Complexity of family relations			Number of relations		item	Complexity of family relations			Number of relations	
1	1	0	0	1	0	**25**	0	1	0	0	1
2	1	0	0	1	0	**26**	0	0	1	0	1
3	1	0	0	1	0	**27**	0	0	1	0	1
4	1	0	0	1	0	**28**	0	0	1	0	0
5	1	0	0	1	0	**29**	0	0	1	0	0
6	1	0	0	1	0	**30**	0	0	1	0	0
7	1	0	0	1	0	**31**	0	0	1	0	0
8	1	0	0	1	0	**32**	0	0	1	0	0
9	1	0	0	0	1	**33**	0	0	1	0	1
10	0	1	0	1	0	**34**	0	0	1	0	1
11	0	1	0	0	1	**35**	0	0	1	0	0
12	0	1	0	0	1	**36**	0	0	1	0	1
13	0	1	0	0	1	**37**	0	0	1	0	0
14	0	1	0	0	1	**38**	0	0	0	0	0
15	0	1	0	0	1	**39**	0	0	0	0	0
16	0	1	0	1	0	**40**	0	0	0	0	0
17	0	1	0	0	1	**41**	0	0	0	0	0
18	0	1	0	0	1	**42**	0	0	0	0	0
19	0	1	0	0	1	**43**	0	0	0	0	0
20	0	1	0	1	0	**44**	0	0	0	0	0
21	0	1	0	0	1	**45**	0	0	0	0	0
22	0	1	0	0	1	**46**	0	0	0	0	1
23	0	1	0	1	0	**47**	0	0	0	0	0
24	0	1	0	0	1						

Table 4:
Results using LLTM's matrix of weights ((q_{1ij}))

Results of the LLTM
LRT (Andersen): $\chi^2 = 524.878$; $df = 41$; $\chi^2_{krit\ (\alpha = .05)} = 56.94239$; $p = 0$

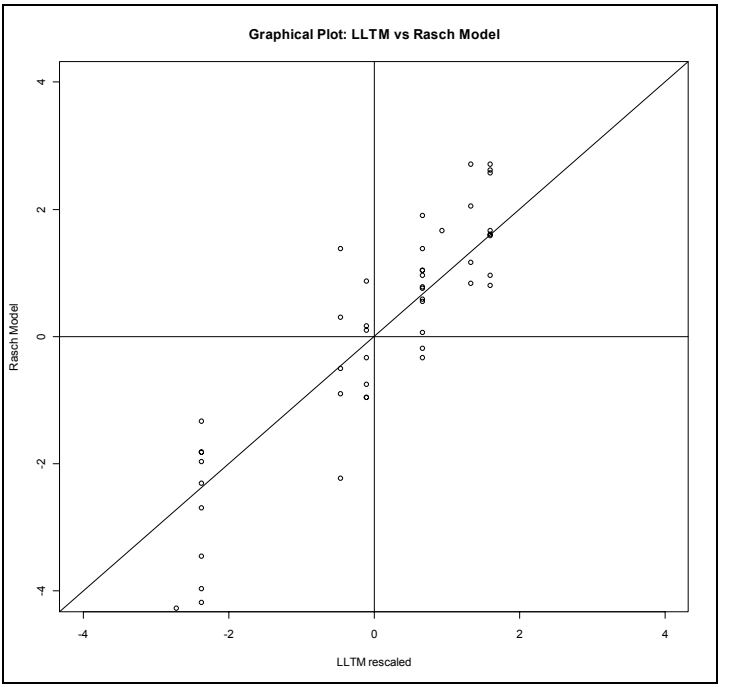

Figure 3:
Graphical model check using LLTM's matrix of weights ((q_{1ij}))

Approach 2: an examination of the construct validity using LLTM's matrix of weights ((q_{8ij}))

After the disappointing results of the first approach, further structure matrices were applied to test hypotheses about the basic cognitive operations of FRRT-items in order to find adequate item construction rules. To find hypothesized basic components by means of the LLTM, a stepwise item component elimination process was used. First the LLTM's matrix of weights ((q_{ij})) with the highest complexity was analysed, then the matrix with the second highest complexity was used and so on. In this second approach, a LLTM's matrix of weights ((q_{ij})) called q_{8ij} (Table 5) was finally found.

Table 5:
LLTM's matrix of weights ($(q8_{ij})$) consisting of logical rules for each of the 5 subgroups

Item	Position of the item in the test				Complexity of family relation			Number of names	Number of words		Number of characters	
1	1	0	0	0	1	0	0	1	1	0	1	0
2	1	0	0	0	1	0	0	1	1	0	1	0
3	1	0	0	0	1	0	0	1	1	0	1	0
11	0	1	0	0	0	1	0	1	1	0	1	0
12	0	1	0	0	0	1	0	1	1	0	0	1
13	0	1	0	0	0	1	0	1	0	1	0	1
28	0	0	1	0	0	0	1	1	0	1	0	0
29	0	0	1	0	0	0	1	1	0	1	0	1
30	0	0	1	0	0	0	1	1	0	1	0	1
31	0	0	0	1	0	0	1	1	0	1	0	1
32	0	0	0	1	0	0	1	1	0	1	0	1
33	0	0	0	1	0	0	1	1	0	1	0	0
45	0	0	0	0	0	0	0	0	0	1	0	1
46	0	0	0	0	0	0	0	1	0	1	0	0
47	0	0	0	0	0	0	0	1	0	1	0	0

The complexity of the LLTM's matrix of weights ((q_{ij})) q_{8ij} has grown and the matrix consists of the following basic components: complexity of family relations, item position effects, number of names used in the item, number of words used in the item, number of characters in the item. The 'number of relations', one of the original rules, was not included in this approach. In Figure 4, the LLTM-predicted (rescaled) item parameters were again

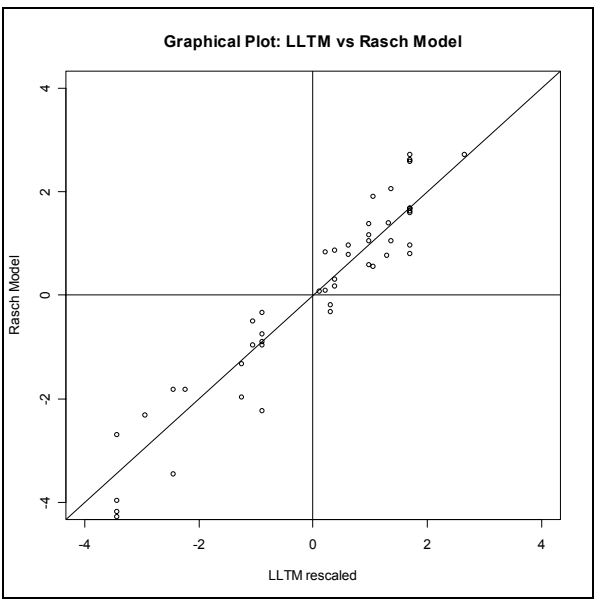

Figure 4:
Graphical model check using LLTM's matrix of weights ((q_{8ij}))

plotted against the estimated item parameters of the RM. Most data points (LLTM rescaled, RM) do not hit the 45° line and some points are far away from this line, but all in all the results look far more attractive than the graphical model test in Approach 1.

Checking the construct validity using the LRT (Table 6) again led to a significant result and the hypothesis of a valid LLTM had to be discarded; the LRT value of Approach 2 is considerably lower, but not significantly lower than the value in Approach 1. According to Kubinger (1979), a lack of fit can be explained by a very high complexity of the applied logical item generating rules or by the fact that the LRT is too sensitive. The results are on the one hand disappointing, but on the other hand the results are much better than those of the LRT in Approach 1. All in all, the LLTM's matrix of weights $((q_{ij}))$ q_{8ij} does not fit the Rasch model. But the results in Approach 2 are quite promising and provide information for the construction of a LLTM's matrix of weights $((q_{ij}))$ consisting of valid item generating rules.

Table 6:
Results of the LLTM using LLTM's matrix of weights $((q_{8ij}))$

Results of the LLTM
Likelihood Ratio Test: $\chi^2 = 265.2779$; $df = 34$; $\chi^2_{krit\,(\alpha\,=\,.05)} = 48.60237$; $p = 0$

Discussion

The first attempt to construct this new kind of test was rather explorative and the construction of the items rather informal and intuitive. Nevertheless, it was shown that the Rasch model holds under certain conditions. Despite the fact that the LLTM gives some information about applied logical rules, many additional assumptions arose. There is no clear conjecture as to which combination of hypothetically assumed basic cognitive operations was implicitly used in the item construction process. There were several reasons why the Rasch model might hold only after excluding so many items. The multiple-choice format with a high number of choices may have caused confusion among the test persons. Grammar, style and vocabulary of the short stories can affect item difficulty, but these are difficult to detect. Other reasons for the lack of validity of the LLTM may be found in the item construction process. Some items assumed to be difficult can be solved by reading only the last sentence in the story. If an examinee is clever enough to discover this concept, then these items can be solved easily.

As mentioned before, the Likelihood Ratio Test is a sensitive test statistic especially if it is used for LLTM analyses. The results in the second approach are quite promising and can give us information about which logical rules were applied in the item construction process. Similarly, the Graphical Goodness-of-fit Test is quite promising (the items show a more or less LLTM-fitting behaviour). Although none of these rules could be corroborated by means of the Likelihood Ratio Test after being modelled in LLTM's matrices of weights $((q_{ij}))$, the presented approach may nevertheless be valuable, because it was possible to start a psychometric investigation of the components influencing the item difficulties. Additionally, the investigation of hypothesized basic operations indicates which rules might have an impact on

the item difficulties. Comparing two significant results of LLTM analyses is not a justified statistical method for corroborating hypotheses, but it showed utility in generating hypotheses about the operations used in the solution process and determining the item difficulties. In conclusion, the examination of construct validity by means of the LLTM gives useful information about the consistency of the LLTM's matrix of weights ((q_{ij})) and therefore about the rules used in the item construction process. One can also take into consideration that the applied methods are valuable in indicating improvements of the item construction process for a future version of the FRRT.

Outlook

After first analyses of the FRRT and the demanding examination concerning the psychometric qualities of the test, a second version of the FRRT was created. In a first step, simple item generating rules were constructed. Example: "A is the son of B and C. D is the sister of A. D is the _____ of B." (Solution: D is the daughter of B). Through systematic variation of the item generating rules, an item universe containing all possible item combinations was created. In a further step, the items were adapted for presentation.

Example: Peter is the son of Cathy and Erik. Angie is the sister of Peter. Angie is the ____ of Cathy.

Furthermore, the item difficulties were additionally varied by adding redundant information to the items. Through this procedure, the feasibility of creating parallel tests by drawing different samples from the item universe is given while the content validity of the parallel tests is ensured.

Most recently, an internet version of the test has been made available where different item response formats are administered. This means that the psychometric qualities can be investigated again with new data and with help of the valuable experiences gained in this study.

References

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

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.

Glas, C.A.W, & Verhelst, N. (1995). Tests of Fit for Polytomous Rasch Models. In G.H. Fischer & I.W. Molenaar (Eds.), *Rasch Models: Foundations, Recent Developments, and Applications* (pp. 325-352). New York: Springer.

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. (1989). Aktueller Stand und kritische Würdigung der Probabilistischen Testtheorie [Critical evaluation of latent trait theory]. In K. D. Kubinger (Ed.), *Moderne Testtheorie –*

Ein Abriß samt neuesten Beiträgen [Modern psychometrics – A brief survey with recent contributions] (pp.19–83). Munich, Germany: Psychologie Verlags Union.

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. (2009). Applications of the Linear Logistic Test Model in Psychometric Research. *Educational and Psychological Measurement, 69*, 232-244.

Mair, P., & Hatzinger, R. (2007). Extended Rasch Modeling: The eRm Package for the Application of IRT Models in R. *Journal of Statistical Software*, *20*(9). URL http://www.jstatsoft.org/v20/i09/.

Placek, K. (2006). *Dimensionalitätserweiterung des AID 2 um eine Reasoning Komponente* [Extending the dimensionality of the AID 2 by a reasoning component]. Unpublished master thesis, Vienna.

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, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Danish Institute for Educational Research: Copenhagen.

Schechtner, C. (2009). *Entwicklung eines rationalen Itemkonstruktionsprinzips als Basis eines sprachlichen Reasoning-Tests* [Development of item construction rationals as foundation of a verbal reasoning test]. Unpublished master thesis, Vienna.

Skoda, S. (2005). *Der Verwandtschaften-Reasoning-Test* [The Family Relations Test]. Unpublished master thesis, Vienna.