AUSTRIAN JOURNAL OF STATISTICS
Volume 38 (2009), Number 4, 221–230

A Note on the Sampling Distribution of the Likelihood Ratio Test in the Context of the Linear Logistic Test Model

Rainer W. Alexandrowicz
Ludwig-Maximilians-Universität München

Abstract: One important tool for assessing whether a data set can be described equally well with a Rasch Model (RM) or a Linear Logistic Test Model (LLTM) is the Likelihood Ratio Test (LRT). In practical applications this test seems to overly reject the null hypothesis, even when the null hypothesis is true. Aside from obvious reasons like inadequate restrictiveness of linear restrictions formulated in the LLTM or the RM not being true, doubts have arisen whether the test holds the nominal type-I error risk, that is whether its theoretically derived sampling distribution applies. Therefore, the present contribution explores the sampling distribution of the likelihood ratio test comparing a Rasch model with a Linear Logistic Test Model. Particular attention is put on the issue of similar columns in the weight matrix $\mathbf{W}$ of the LLTM: Although full column rank of this matrix is a technical requirement, columns can differ in only a few entries, what in turn might have an impact on the sampling distribution of the test statistic. Therefore, a system of how to generate weight matrices with similar columns has been established and tested in a simulation study. The results were twofold: In general, the matrices considered in the study showed LRT results where the empirical alpha showed only spurious deviations from the nominal alpha. Hence the theoretically chosen alpha seems maintained up to random variation. Yet, one specific matrix clearly indicated a highly increased type-I error risk: The empirical alpha was at least twice the nominal alpha when using this weight matrix. This shows that we have to indeed consider the internal structure of the weight matrix when applying the LRT for testing the LLTM. Best practice would be to perform a simulation or bootstrap/re-sampling study for the weight matrix under consideration in order to rule out a misleadingly significant result due to reasons other than true model misfit.

Keywords: Rasch Model, Linear Logistic Test Model, Conditional Likelihood Ratio Test, Sampling Distribution.

1 Introduction

G. Rasch (1960) introduced a statistical model that allows for describing the probability of a positive response to a dichotomous item by means of two real-valued parameters, β_i ($i = 1, \ldots, k$) covering the difficulty (or easiness) of the item i and θ_v ($v = 1, \ldots, n$) characterizing person v in terms of the ability to solve this item or proneness to endorse a statement:

$$p(+|\theta_v, \beta_i) = \frac{e^{\theta_v - \beta_i}}{1 + e^{\theta_v - \beta_i}} \,. \tag{1}$$

The Linear Logistic Test Model (LLTM; Fischer, 1972, 1973, 1983, 1995; Scheiblechner, 1971, 1972) is an extension of the Rasch Model (RM). It decomposes each item difficulty parameter β_i into a weighted sum of basic parameters η_j ($j = 1, \ldots, p$), representing for example item components or cognitive operations required for solving the item. A weight matrix $\mathbf{W} = \{w_{ij}\}$ of size $k \times p$ stipulates which basic parameter contributes to each item:

$$\beta_i = \sum_{j=1}^{p} w_{ij}\eta_j + c\,. \tag{2}$$

The weight matrix $\mathbf{W}$ must have full column rank p and is defined according to theoretical considerations. Many applications work with entries of $\mathbf{W}$ equalling 0 or 1, but any other real-valued entry could be chosen as well—as long as they are justifiable from a substantive point of view and established prior to parameter estimation. The normalizing constant c is required to compensate for an (admissible) shift of the β_i (for instance for norming purposes) without affecting the η_j (Fischer, 1995). In fact, it can be eliminated by setting $c = -1/k \sum_i \sum_j w_{ij}\eta_j$, hence $w^*_{ij} = w_{ij} - 1/k \sum_i w_{ij}$ (Fischer, 1983). Applying the LLTM when the RM does not hold would be of little interest, as it would boil down to the decomposition of an item parameter that innately has not adequately described the data.

Both the RM and the LLTM allow for conditional maximum likelihood (CML) estimation (Andersen, 1970, 1972). Beside the advantageous fact that this method provides for handling of the incidental parameter problem (Neyman and Scott, 1948; revisited by Lancaster, 2000) by conditioning on the sufficient statistics, it also allows for the application of Conditional Likelihood Ratio Tests (LRT; Andersen, 1973) to assess model fit. Fischer (1995, 1997) offers two different assessments determining the interrelation of the two models: from a theoretical point of view, the LLTM can be regarded a more general model than the RM, because the RM is a special case of the LLTM if $\mathbf{W}$ is an identity matrix. From an empirical perspective, the LLTM can be considered a restriction of the RM, because it imposes a (linear) structure on the item parameters β_i. Adopting the empirical point of view allows for applying the LRT to test the adequacy of the decomposition (2). Let $\mathbf{r}$ be the vector of rawscores of a data set $\mathbf{X}$. We will call L_{RM} the maximum of the conditional likelihood function of the Rasch Model, $L_C(\mathbf{X}|\mathbf{r}, \hat{\boldsymbol{\beta}})$, and L_{LLTM} the maximum of the conditional likelihood function of the LLTM, $L_C(\mathbf{X}|\mathbf{r}, \hat{\boldsymbol{\beta}}, \mathbf{W})$. Then, the test statistic $\lambda = -2\log(L_{LLTM}/L_{RM})$ is asymptotically χ^2 distributed with $k-p-1$ degrees of freedom. If for a given data set the test does not reject the null hypothesis of whether the data can be described equally well with the LLTM and the RM (given a type-I error risk α), an estimate of the item difficulty parameter $\hat{\beta}_i$ can be determined from the vector of basic parameter estimates $\hat{\boldsymbol{\eta}}$ and the structure implied by $\mathbf{W}$.

1.1 Problem

The LRT seems to be problematic as in most applications the LLTM is rejected, as has already been shown by Fischer (1995, 1997). Two reasons seem obvious for this increased rejection rate: Either the theoretical assumptions concerning the decomposition of item parameters into basic parameters expressed in $\mathbf{W}$ are too restrictive for most empirical

data or the requirement of the RM to hold for the data does not apply. Both reasons depend on the respective set(s) available, therefore a general analysis is difficult. Another reason why the LRT may appear likely to reject the null hypothesis of model equivalence could relate to the sampling distribution of the test statistic as well: it could be possible that the theoretically derived distribution does not apply so that the test does not hold the nominal alpha risk. Some evidence for a systematic increase of the type-I error risk has been provided by Hohensinn et al. (2008).

The present study examines whether certain configurations of $\mathbf{W}$ principally result in a higher (than the nominal α) rejection rate of the null-hypothesis when applying the LRT. The specific issue that shall be dealt with in the present study is linear independence of columns: The matrix $\mathbf{W}$ must have full column rank, otherwise the decomposition of item parameters into the basic parameters is not unique. Although this rule is unambiguous from a mathematical point of view, a perturbing variation can be thought of: full column rank does not prevent columns from being similar (yet linearly independent in the algebraical sense). Similarity in this context means that the item parameter decomposition (2) leads to all but one (or a few) items sharing a certain component (described by its basic parameter). Then, the entries in the respective columns of $\mathbf{W}$ will be nearly identical except for a few cell entries. We want to investigate, whether such similarities in the columns of $\mathbf{W}$ may cause the LRT to violate the nominal alpha. This article will not consider a misspecification of the design matrix $\mathbf{W}$, that issue has been dealt with by Baker (1993) or Klein (2003), for example. Therefore, only design matrices correctly describing the true decomposition of item parameters into basic parameters shall be taken into account. We want to confine ourselves to binary entries of $\mathbf{W}$, which are most common in practical applications. Of course, other entries could be used as well, but a substantive justification of non-integer w_{ij} seems difficult: why should, for example, a basic parameter be weighted with 0.6 and not 0.7 or 0.5 for determining the difficulty of an item? Of course, if a basic parameter describes a cognitive operation required to solve a task, and the operation has to be applied twice (or 3 times, ...) when working on a certain task, then an entry w_{ij} of 2 (or 3, ...) would be appropriate. But such applications seem rare, hence non-binary weighting shall not be considered in this study.

1.2 Method

In order to assess the empirical type-I error risk alpha of the LRT with respect to variations in the weight matrix $\mathbf{W}$, a simulation study has been performed. The empirical rejection rate of the LRT was assessed for a large number of data sets complying with both the RM and a given matrix $\mathbf{W}$. The simulations were conducted according to the following scheme: Based on a set of basic parameters η_j and a weight matrix $\mathbf{W}$, the corresponding item difficulty parameters β_i were determined. Along with a vector of ability parameters $\boldsymbol{\theta}$, the item-response matrix $\mathbf{X}$ was generated according to the usual principles for generating data pursuant to the RM (van den Wollenberg, 1982). Then both the item parameters of the RM and the basic parameters according to the LLTM were estimated, resulting in a likelihood for each of the two models. The LR test statistic, the respective degrees of freedom, and the p-value were stored. This procedure was repeated $m = 1000$ times for each weight matrix $\mathbf{W}$ (see below), and the number of significant results s was

computed. The relative frequency of s gives an estimate for the empirical alpha (α_{emp}). The LRT is considered correct if α_{emp} does not differ from the nominal α (evaluated at 0.01, 0.05, and 0.10) but for random variation. This means that the empirical α does not exceed an interval ϵ of $\pm 10\%$ and 20% (the 10% interval covers values between 0.045 and 0.055 and the 20% interval covers values between 0.04 and 0.06).

1.3 Designs

In order to establish a system of matrices exhibiting increasingly similar columns, the following approach was chosen: We started with a matrix $\mathbf{W}_0$ which we would a priori consider perfect in terms of linear independence, as it consists of all 2^p patterns that can be formed with p basic parameters. Columns exhibit maximum dissimilarity in such a matrix. We then systematically added columns similar to the existing ones, thus obtaining an extended matrix $\mathbf{W}^*$. These additional columns $(p+1), \ldots, 2p$ are copies of the columns 1 to p of $\mathbf{W}_0$, in which one element per column has been altered. Each single column of $\mathbf{W}_0$ can result in up to p new columns, allowing for a maximum of $p(2^p+1)$ columns. This, of course, would in some cases violate the necessity of $p < k$, therefore the admissibility of each matrix $\mathbf{W}$ has to be checked. In the present study, we copied each of the p columns of $\mathbf{W}_0$ either one or two times and altered one entry in each copy. This process will be labelled by a three digit notation throughout the text: the first digit will denote the number of original columns, p. The second digit, c_1 will give the number of columns of $\mathbf{W}_0$ copied (and altered) for the first time and the third digit, c_2 will give the number of columns copied (and altered) a second time (note that the implementation started with the rightmost column of $\mathbf{W}_0$, which has no implications at all). Table 1 contains the design 4–4–4 as an example:

Table 1: Example of $\mathbf{W}^*$ with the description 4–4–4

	1	2	3	4	4*	3*	2*	1*	4**	3**	2**	1**
1	0	0	0	0	⋆							
2	0	0	0	1		⋆						
3	0	0	1	0			⋆					
4	0	0	1	1				⋆				
...	...	...	...	...	...	...	...	...	...	...	...	...
13	1	1	0	0								⋆
14	1	1	0	1							⋆	
15	1	1	1	0						⋆		
16	1	1	1	1					⋆			

Note: Column headers denote the number of the respective column in $\mathbf{W}_0$, and asterisks indicate first and second copies of the original columns. The cells marked with an ⋆ indicate reversed entries. Ellipses indicate that no changes compared to the corresponding cells of columns 1 to p have been made.

The simulation covered the following cases: For each p the number of copies c_1 ranged from zero to p while c_2 remained zero. Then c_2 took all values from zero to p while c_1 remained zero. And finally, c_1 and c_2 took all values from 1 to p at the same time. This

system was applied to values of p ranging from 3 to 5. Sample sizes were 250 and 750 for all designs. Basic parameters η_j ($j = 1, \ldots, p^*; p^* = p + c_1 + c_2$) were chosen equidistantly from the interval $[-1, 1]$. Moreover, some specific designs were analysed as well, but due to exorbitant runtime no systematic assessment was possible. The analysis of the specific designs will not be presented in detail, however none of them showed results contradicting those presented in this article.

All computations were performed with R (R Development Core Team, 2008), using the library eRm (Mair and Hatzinger, 2007a, 2007b).

2 Results

The empirical p-values for all designs and evaluated at a nominal alpha of 0.01, 0.05, and 0.1 are given in Table 3 along with an indication whether the values either exceed the 10% epsilon interval or the 20% epsilon interval (see note beneath Table 3 for the coding scheme). The designs 3–2–2 and 3–3–3 had to be skipped, as the corresponding design matrices $\mathbf{W}$ do not have full column rank. The number of designs indicating violations of the nominal α according to the criteria mentioned above is given in Table 2 (broken down by α_{nom} and n).

Table 2: Empirical α broken down by nominal α and sample size n

α_{nom}	0.01		0.05		0.10	
n	250/750	Total	250/750	Total	250/750	Total
α_{emp} above α_{nom}	13/12	25	11/10	21	7/6	13
exceeding $+2\epsilon$	7/7	14	4/5	9	2/2	4
exceeding $+1\epsilon$	6/5	11	7/5	12	5/4	9
correct	8/8	16	17/19	38	24/22	46
α_{emp} beyond α_{nom}	16/17	33	9/8	17	6/9	15
exceeding -1ϵ	3/9	12	4/2	6	1/0	1
exceeding -2ϵ	13/8	21	5/6	11	5/9	14
Total	37/37	74	37/37	74	37/37	74

The number of results to be considered robust increases with the nominal α: Roughly, compared to $\alpha_{nom} = 0.01$, the number of correct results for $\alpha_{nom} = 0.05$ is twice as high, and for $\alpha_{nom} = 0.1$ it is about three times as high, independent of sample size. Accordingly, the number of deviating results decreases with increasing α_{nom} but not in a certain direction: the number of designs showing a progressive characteristic ($\alpha_{emp} > \alpha_{nom}$) is approximately equal to the number of designs exhibiting a conservative tendency ($\alpha_{emp} < \alpha_{nom}$), again independent of sample size. The extent of transgression follows the same overall tendency, it decreases when α_{nom} increases.

In general, no global pattern of aberration was discernible in a designwise analysis (cf. Table 3), neither in terms of direction (α_{emp} below or above α_{nom}) nor extent (exceeding $\epsilon = 0.1$ or $\epsilon = 0.2$). However, one specific design was identified that revealed a strikingly deviant behaviour: The design 4–4–4 exceeded the nominal alpha to a remarkable extent (note that the weight matrix $\mathbf{W}$ of this design has been chosen for the exemplification

Table 3: Empirical α and robustness broken down by design

	$n = 250$						$n = 750$					
Mod	$\widehat{\alpha_{.01}}$	$\delta_{.01}$	$\widehat{\alpha_{.05}}$	$\delta_{.05}$	$\widehat{\alpha_{.10}}$	$\delta_{.10}$	$\widehat{\alpha_{.01}}$	$\delta_{.01}$	$\widehat{\alpha_{.05}}$	$\delta_{.05}$	$\widehat{\alpha_{.10}}$	$\delta_{.10}$
300	.011	0	.056	1	.102	0	.009	−1	.044	−1	.092	0
301	.015	2	.055	0	.110	0	.009	−1	.055	0	.098	0
302	.012	1	.051	0	.105	0	.010	0	.059	1	.118	1
303	.010	0	.042	−1	.101	0	.014	2	.060	2	.118	1
310	.008	−1	.057	1	.114	1	.006	−2	.041	−1	.104	0
320	.009	−1	.050	0	.111	1	.008	−1	.040	−1	.084	−1
330	.012	2	.064	2	.100	0	.008	−1	.050	0	.095	0
311	.012	1	.044	−1	.089	−1	.013	2	.052	0	.100	0
400	.010	0	.054	0	.104	0	.007	−2	.038	−2	.085	−1
401	.008	−1	.039	−2	.086	−1	.007	−2	.051	0	.110	0
402	.012	1	.051	0	.108	0	.014	2	.061	2	.104	0
403	.012	1	.057	1	.110	0	.009	−1	.042	−1	.088	−1
404	.009	−1	.054	0	.108	0	.006	−2	.051	0	.097	0
410	.019	2	.054	0	.100	0	.007	−2	.052	0	.110	0
420	.009	−1	.041	−1	.091	0	.015	2	.047	0	.101	0
430	.009	−1	.053	0	.097	0	.007	−2	.051	0	.107	0
440	.011	0	.057	1	.111	1	.011	0	.050	0	.083	−1
411	.004	−2	.039	−2	.086	−1	.010	0	.055	0	.097	0
422	.008	−1	.052	0	.098	0	.011	0	.060	1	.107	0
433	.018	2	.064	2	.127	2	.006	−2	.046	0	.095	0
444	**.024**	**2**	**.090**	**2**	**.185**	**2**	**.031**	**2**	**.102**	**2**	**.179**	**2**
400	.010	0	.054	0	.104	0	.007	−2	.038	−2	.085	−1
501	.010	0	.036	−2	.074	−2	.021	2	.068	2	.131	2
502	.009	−1	.050	0	.098	0	.011	0	.047	0	.089	−1
503	.009	−1	.054	0	.106	0	.012	1	.045	0	.092	0
504	.004	−2	.051	0	.119	1	.008	−1	.048	0	.107	0
505	.013	2	.055	0	.115	1	.009	−1	.052	0	.089	−1
510	.012	1	.056	1	.103	0	.011	1	.055	1	.104	0
520	.009	−1	.053	0	.095	0	.011	0	.051	0	.099	0
530	.008	−1	.043	−1	.100	0	.012	1	.056	1	.099	0
540	.008	−1	.043	−1	.085	−1	.009	−1	.043	−1	.085	−1
550	.011	0	.064	2	.108	0	.007	−2	.044	−1	.087	−1
511	.013	2	.060	1	.109	0	.012	1	.062	2	.108	0
522	.012	1	.052	0	.106	0	.013	2	.058	1	.118	1
533	.010	0	.055	0	.106	0	.010	0	.054	0	.093	0
544	.005	−2	.031	−2	.080	−1	.011	0	.052	0	.111	1
555	.008	−1	.056	1	.108	0	.012	1	.051	0	.102	0

Note: Mod = Three-digit model specification (p—c_1—c_2); α = empirical p-value for α_{nom}; δ = deviation indicator: -2 = empirical p-value beyond $\epsilon = 0.2$-interval; -1 = empirical p-value beyond $\epsilon = 0.1$-interval; 0 = empirical p-value within the $\epsilon = 0.1$-interval; ; $+1$ = empirical p-value exceeds $\epsilon = 0.1$-interval; $+2$ = empirical p-value exceeds $\epsilon = 0.2$-interval; $\widehat{\alpha} = \alpha_{emp}$

given in Table 1). This design did not only exhibit deviations from the nominal alpha for all levels and combinations of α_{nom}, ϵ, and n, but the deviations observed were enormous too: the observed significances α_{emp} amounted to roughly twice (or even three times) the respective α_{nom} in all cases considered. In order to check this result for both error and chance, the simulation of this specific design was repeated with $m = 300000$ ($n = 100$

only). The results were virtually identical, with empirical alpha values of 0.024, 0.103, and 0.185 for α_{nom} = 0.01, 0.05, and 0.1, respectively.

3 Discussion

The main focus of this study was, whether linearly independent yet very similar columns in $\mathbf{W}$ have an impact upon the test's actual type-I error risk given the H_0 holds. For that purpose, a system was established which allows for the generation of an indefinite number of weight matrices with similar columns, a selection of which has been analysed. The empirical alpha was evaluated using two tolerance levels ϵ, 10% and 20%, the more stricter of which (10%) will be used for our decision concerning the correctness of the LRT (cf. D. Rasch, Kubinger, Schmidtke, and Häusler, 2004).

Several designs with empirical alpha levels outside the 10% ϵ interval were found. This might, at first sight, be taken as an indicator for the sampling distribution of the test statistic to be different from the theoretically derived χ^2 distribution with degrees of freedom equalling the difference in the number of parameters estimated for the RM and the LLTM. Nevertheless, a closer inspection unambiguously revealed that the observed rejection rate was sometimes below and sometimes above the nominal alpha. But the pattern of empirical alpha values outside the ϵ interval was unpredictable, considering the respective structure of the weight matrix $\mathbf{W}$, the chosen nominal alpha, or the sample size. Therefore, the results do not allow for determining a general pattern of deviation from the nominal alpha. Rather, the observed deviations of α_{emp} from α_{nom} have to be attributed to the simulation design: Obviously, the chosen number of replications ($m = 1000$) does not suffice to obtain sufficiently stable results. This conjecture could be substantiated by means of additional simulations, which were carried out with a larger m: results were considerably more stable and the 10% interval limits were violated to a lesser extent. Therefore, we conclude that the observed deviations of α_{emp} to α_{nom} have to be considered spurious and the central χ^2 distribution (with degrees of freedom equalling the difference of the number of parameters estimated in the RM and the LLTM) seems to generally apply for the test statistic under the null hypothesis. The reason for the present study to only use 1000 replications was that it focused on facilitating numerous (37) different design matrices $\mathbf{W}$ to be analysed in finite time (computer runtime is still a limiting factor, as it took weeks to obtain the results presented here).

Nevertheless, one matrix (4–4–4) proved highly progressive according to all levels of nominal alpha (1%, 5%, and 10%) and sample sizes (250 and 750). The empirical alpha was at least twice the nominal alpha in all cases. In order to rule out that an insufficient number of replications (1000) has been chosen, the simulation of the 4–4–4 matrix has been repeated using 300000 samples of size 100. This simulation revealed identical results, therefore the highly increased empirical alpha of this matrix $\mathbf{W}$ cannot be considered spurious. Rather, the decomposition of item difficulty parameters into basic parameters according to this specific weight matrix in fact seems to lead to an increased type-I error risk.

Because most of the design matrices considered showed correct results, both the postulated chi-square distribution of the test statistic under the null hypothesis and the pur-

ported degrees of freedom seem appropriate. The fact that one specific weight matrix deviated heavily (and reproducibly) from this general tendency leads us to the conclusion that the inner structure of the weight matrix $\mathbf{W}$—always being non-singular—plays a decisive role for the factual type-I error risk of the LRT. Therefore, for practical applications, no general rule could have been established so far. Hence, the respective matrix $\mathbf{W}$ has to be examined anew. For that purpose, one could either perform an a priori simulation study comparable to the present one or a post-hoc bootstrap analysis. Concerning the latter, a parametric bootstrap (cf. Davison and Hinkley, 1997; Efron and Tibshirani, 1993; Shao and Tu, 1995) seems appropriate, as we know the model and have parameter estimates at our disposal. Such an additional analysis would allow for the distinction, whether a significant test result indicates a true model deviation or has to be considered a method artifact due to the structure of the respective weight matrix. As only one single weight matrix (or maybe a few) will be considered, a much larger number of replications than 1000 is feasible (for instance, 50000), hence results will be stable.

Altogether, the conclusions of this study are twofold: on the one hand, the LRT in general seems to keep the nominal alpha, so the theoretically assumed sampling distribution of the test statistic appears to apply. Hence, the LRT seems to work across a wide range of designs, significant test results are more likely to reflect true model violations than method artifacts. On the other hand, certain weight matrices may lead to false rejections, as has been shown in the present study. Therefore the structure of the weight matrix has to be taken into consideration. As a next step, of course, we have to find a criterion indicating which weight matrices are likely to lead to an increased type-I error risk.

Some general remarks seem necessary in this context: The results presented do not supersede the crucial requirement of the test that might easily be violated in practical application: the denominator model has, of course, to be true. Now, in our case, this is the Rasch Model, the appropriateness of which can be tested—but again we rely on tests. This means that our decision can be correct with probability $1 - \alpha$ (in the non significant case) or $1 - \beta$ (in the significant case). Concerning the first (non-significant) case, we have to bear in mind that failing to reject the null hypothesis will not prove that it is true. Only, if we succeed in not rejecting the null hypothesis of model fit for many times, then its corroboration—according to Popper—increases. However, regarding both cases, we have to keep an eye on sample size: Too small a sample will render relevant model violations undetectable and too large a sample will cause irrelevant discrepancies to become significant. Most of all, the first case will cause the erroneous acceptance of the RM and its unjustified use for testing the LLTM. Of course, one could argue that in such a case, the LRT will as well be underpowered—but actually we cannot be sure about that. Kubinger and Draxler (2007) have already pointed out that there is still a gap in the investigation of the power of the LRT. Hence, the problem of the LRT is retained, namely the lack of knowledge concerning the power of the test along with the availability of an adequate effect size measure (an interesting approach has been proposed by Draxler, 2007). Lastly, in the present context, the missing knowledge concerning test power and effect size measure reappears again: We do not know which differences between the RM and the LLTM should be considered relevant and which sample size would be required to evidence those differences with a given type-I and type-II error risk. These issues have to be tackled in further studies.

This leads us to another—and by far more elementary—problem of assessing whether premises of a test can be considered given: the application of the test of interest (the LRT for testing the restrictions expressed in $\mathbf{W}$) depends on whether or not a previously applied test (in our case the LRT for assessing the fit of the RM) has been significant or not. The same phenomenon occurs, for instance, when we decide to use either the t-test or the Welch test depending on a preliminarily applied F- or Levene-test. As these two tests (in both cases) are not independent of each other, we cannot ascertain the type-I-error-risk of the whole procedure (cf. Zimmerman, 2004).

But, finally, to the extent we succeed in constructing theoretically sound and structurally justified items, such problems diminish. Again, sophisticated methodological considerations cannot substitute a sound theoretical framework. The Rasch family of models represents this in its purest form—maybe one reason why some refrain from applying them.

References

Andersen, E. B. (1970). Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society, Series B*, *32*, 283-301.

Andersen, E. B. (1972). The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistikal Society, Series B*, *34*, 42-54.

Andersen, E. B. (1973). A goodness of fit test for the rasch model. *Psychometrika*, *38*, 123-140.

Baker, F. B. (1993). Sensitivity of the linear logistic test model to misspecification of the weight matrix. *Applied Psychological Measurement*, *17*, 201-210.

Davison, A. C., and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.

Draxler, C. (2007). *Sequentielle Tests für das Rasch Modell*. Berlin: Logos.

Efron, B., and Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Boca Raton: Chapman & Hall.

Fischer, G. H. (1972). A measurement model for the effect of mass-media. *Acta Psychologica*, *36*, 207-220.

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

Fischer, G. H. (1983). Logistic latent trait models with linear constraints. *Psychometrika*, *48*, 3-26.

Fischer, G. H. (1995). The linear logistic test model. In G. H. Fischer and I. Molenaar (Eds.), *Rasch Models: Foundations, Recent Developments, and Applications* (p. 131-156). New York: Springer.

Fischer, G. H. (1997). Unidimensional linear logistic Rasch models. In W. J. Van der Linden and R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (p. 225-243). New York: Springer.

Hohensinn, C., Kubinger, K. D., Reif, M., Holocher-Ertl, S., Khorramdel, L., and Frebort, M. (2008). Examining item-position effects in large-scale assessment using the linear logistic test model. *Psychology Science Quarterly*, *50*, 391-402.

Klein, S. (2003). *Neue Methoden zur Entdeckung von Fehlspezifikation bei Latent-Trait-Modellen der Veränderungsmessung.* Retrieved 23.01.2010, from `http://edoc.hu-berlin.de/dissertationen/klein-stefan-2003-05-09/PDF/Klein.pdf`

Kubinger, K. D., and Draxler, C. (2007). Probleme bei der Testkonstruktion nach dem Rasch-Modell. *Diagnostica*, *53*, 131-143.

Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics*, *95*, 391-413.

Mair, P., and Hatzinger, R. (2007a). CML based estimation of extended Rasch models with the eRm package in R. *Psychology Science*, *49*, 26-43.

Mair, P., and Hatzinger, R. (2007b). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software*, *20*(9), 1-20.

Neyman, J., and Scott, E. L. (1948). Consistent estimation from partially consistent observations. *Econometrica*, *16*, 1-32.

R Development Core Team. (2008). R: A Language and Environment for Statistical Computing [Computer software manual]. Vienna, Austria. Available from `http://www.R-project.org`

Rasch, D., Kubinger, K. D., Schmidtke, J., and Häusler, J. (2004). The misuse of asterisks in hypothesis testing. *Psychology Science*, *46*, 227-242.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danmarks Pædagogiske Institut.

Scheiblechner, H. (1971). *CML-parameter-estimation in a generalized multifactorial version of Rasch's probabilistic measurement model with two categories of answers* (Research Bulletin No. 4). Vienna: Department of Psychology, University of Vienna.

Scheiblechner, H. (1972). Das Lernen und Lösen komplexer Denkaufgaben. *Zeitschrift für Experimentelle und Angewandte Psychologie*, *3*, 456-506.

Shao, J., and Tu, D. (1995). *The Jackknife and Bootstrap*. New York: Springer.

van den Wollenberg, A. L. (1982). A simple and effective method to test the dimensionality axiom of the Rasch model. *Applied Psychological Measurement*, *6*, 83-91.

Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology*, *57*, 173-181.

Author's address:

Rainer W. Alexandrowicz
Ludwig-Maximilians-Universität München
Department Psychologie – Fakultät für Psychologie und Pädagogik
Lehrstuhl Methodenlehre und Psychologische Diagnostik
Leopoldstraße 13
D-80802 München