Psychology Science Quarterly, Volume 51, 2009 (4), pp. 370-384

On designing data-sampling for Rasch model calibrating an achievement test

KLAUS D. KUBINGER[1], DIETER RASCH[2] & TAKUYA YANAGIDA[3]

Abstract

In correspondence with pertinent statistical tests, it is of practical importance to design data-sampling when the Rasch model is used for calibrating an achievement test. That is, determining the sample size according to a given type-I- and type-II-risk, and according to a certain effect of model misfit which is of practical relevance is of interest. However, pertinent Rasch model tests use chi-squared distributed test-statistics, whose degrees of freedom do not depend on the sample size or the number of testees, but only on the number of estimated parameters. We therefore suggest a new approach using an *F*-distributed statistic as applied within analysis of variance, where the sample size directly affects the degrees of freedom. The Rasch model's quality of specific objective measurement is in accordance with no interaction effect in a specific analysis of variance design. In analogy to Andersen's approach in his Likelihood-Ratio test, the testees must be divided into at least two groups according to some criterion suspected of causing differential item functioning (DIF). Then a three-way analysis of variance design $(A \succ B) \times C$ with mixed classification is the result: There is a (fixed) group factor *A*, a (random) factor *B* of testees within *A*, and a (fixed) factor *C* of items cross-classified with $A \succ B$; obviously the factor *B* is nested within *A*. Yet the data are dichotomous (a testee either solves an item or fails to solve it) and only one observation per cell exists. The latter is not assumed to do harm, though the design is a mixed classification. But the former suggests the need to perform a simulation study in order to test whether the type-I-risk holds for the $A \times C$ interaction *F*-test – this interaction effect corresponds to Rasch model's specific objectivity. If so, the critical number of testees is of interest for fulfilling the pertinent precision parameters. The simulation study (100 000 runs for each of several special cases) proved that the nominal type-I-risk holds as long as there is no significant group effect. Analysing a certain DIF, this *F*-test has fair power, consistently higher than Andersen's test. Hence, we advise researchers to apply our approach as long as there is no significant group effect, and only to use other Rasch model tests if it is significant. Keep in mind that this is true only for some special cases and needs to be generalized in further research. Then a formula needs to be provided which will allow explicit calculation of the number of testees, given a type-I-, a type-II-risk, and a relevant effect as concerns Rasch model misfit.

Key words: Rasch model; sample size; type-I- and type-II-risk; analysis of variance; mixed model

---

[1] *Correspondence concerning this article should be addressed to:* Prof. Klaus D. Kubinger, Ph.D., Head of the Division for Psychological Assessment and Applied Psychometrics, Faculty of Psychology, University of Vienna, Liebiggasse 5, A-1010 Vienna, Austria; email: klaus.kubinger@univie.ac.at

[2] Institute of Applied Statistics and Computing, University of Natural Resources and Applied Life Sciences, Vienna

[3] University of Vienna, Faculty of Psychology, Division of Psychological Assessment and Applied Psychometrics

Introduction

There is no doubt that the well-known Rasch model, nowadays often called 1-PL model, is applied world-wide for test calibration. Though originally intended to measure different psychological aptitudes (cf. Rasch, 1960/1980; see also Fischer, 1974), in the last decades it has captured the market via large scale assessments within educational frameworks. This is primarily because of the advantage of using different test-booklets with partially different items to nevertheless achieve fair comparisons of the testees' test scores – an advantage shared by most Item Response Theory (IRT) models, though the Rasch model (as well as its generalizations) is the only one that provides "specific objective" measuring and therefore fulfils the basic requirements of measurement theory (cf. e. g. Scheiblechner, 2009). As a matter of fact, some Rasch model generalizations applied in large scale assessments, above all the well-known PISA study (OECD, 2007), have finally led to the Rasch model's popularity. Googling "Rasch model" now leads to about 647.000 initial hits.

Since the early 70's, several statistical approaches have been taken for testing the Rasch model. The most established test is Andersen's Likelihood-Ratio test (LRT; Andersen, 1973). Furthermore, see Glas and Verhelst (1995) for a current review of Rasch model tests[4]. However, if a researcher calibrates an achievement test using such tests, there is always the problem of designing data sampling, i.e. determining the sample size.

In the first instance, a researcher aims for a sample size as large as economically acceptable. This is because parameter estimation's accuracy depends enormously on a proper sample size; as concerns the Rasch model, usually the data from no less than 200 testees are sampled; however, sample sizes up to 1000 testees and even more are common (cf. for instance Kubinger, 2009). In the second instance, however, he/she is aware of the fact that a too large sample size would most likely lead to a significant result when testing the model, even if this result is based only on a minor effect: the null-hypothesis is rejected though model contradiction is hardly of practical relevance. This possibility may mislead the researcher to adopt the strategy of almost ignoring the result of the statistical significance test.

What is needed is the approach usually applied in designing an experiment, particularly if Student's *t*-test is planned to be used for analysis: Given H_0 and H_1, a certain type-I-risk α and a certain type-II-risk β – that is the probability of rejecting the null-hypothesis though it is correct on the one side and the probability of accepting the null-hypothesis though it is wrong on the other side – is determined at the very beginning of planning, as well as a certain effect δ referring to the deviation of the parameter in question from H_0 to H_1 which is supposed to be of practical importance. Using such "precision" requirements, the sample size must be calculated so that such an effect or even larger ones lead to a type-II-error with at most the probability of the fixed type-II-risk. That is, the sample size is determined in such a way that for given α and δ, the type-II-risk is equal to the given β. If the actual effect size exceeds δ, the type-II-risk is smaller than β and therefore we are on the safe side insofar as we detect each effect equal to or larger than δ with at least the probability 1- β, the power of the statistical test.

[4] We strictly distinguish between (Rasch-) "model tests" which test some model implications or are, so to speak, performed according to specific objective measurement and "goodness-of-fit tests" which only measure the model's appropriateness.

Several attempts have been made to establish at least some statistical effect size parameter as concerns Andersen's Likelihood-Ratio test (cf. Müller-Philipp & Tarnai, 1989; Goethals, 1994; Alexandrowicz, 2002); apart from this, Goethals (1994) provided a rule of thumb: Any difference of parameter estimations not greater than a tenth of the range of the parameters is hardly of practical relevance (cf. Kubinger, 2005).

However, currently the problem in designing the data-sampling for Rasch model calibrating an achievement test is that the pertinent test-statistic is chi-squared distributed – and this statistic's degrees of freedom do not at all depend on the sample size, but only on the number of estimated parameters. Consequently, this statistic cannot be used for designing the data-sampling, that is it does not offer a means for sample size calculation given any precision requirements.

Hence, we try a new approach in this paper. We aim for an F-distributed statistic as applied within analysis of variance, because then the sample size directly affects the degrees of freedom. Therefore it becomes possible to calculate the sample size according to this distribution, given a certain type-I- and type-II-risk and some specified alternative hypothesis *via* δ.

Method

Of course, nowadays the Rasch model can be interpreted as a special case of generalized linear models (McCullagh & Nelder, 1989); within traditional Rasch model research, Kelderman (1984) was the first who used this fact deliberately for a class of model tests. So see for instance De Boeck and Wilson (2004) or Raudenbush and Bryk (2002) for details on how the Rasch model can be formulated as a generalized linear model for binary data with one observation per cell and a logit link function.

Among all the assumptions and properties of the Rasch model, the one most frequently referred to is that item difficulty parameters are statistically independent of the person ability parameters, or in other words that specific objectivity is given if the Rasch model holds. As a consequence – used in particular by Andersen's LRT – item parameter estimations do not, for instance, depend on which sub-sample of a given population of testees is taken into account.

First attempt

Now, thinking in terms of analysis of variance, if the different items of an item pool which shall be calibrated according to the Rasch model are considered as the different levels of a first (fixed) factor and the testees as the different levels of a second (random) factor, then specific objectivity means: there is no interaction effect between the factors – irrespective of the probably strong main effects – because the testees will differ with respect to their test performance just as the items may differ with respect to their frequencies of being solved within the sample. The first factor is a fixed one, because we are interested in just these given items; but the second factor is a random one, as we have an almost randomly chosen sample of testees who are part of a certain intended population.

However, the sketched design of analysis of variance suffers from at least two problems: Firstly, this design (see Figure 1) establishes just a single observation within each cell (n = 1); and hence there is no test-statistic or corresponding distribution function, if, as given, we have to deal with a mixed model (i.e. one factor being fixed, the other being random). Secondly, this design is applied to dichotomous, not interval-scaled – and not remotely normally distributed – data.

	A	Items					
B		1	2	…	*i*	…	a
	1	y_{11}	y_{21}		y_{i1}		y_{a1}
	2	y_{12}	y_{22}		y_{i2}		y_{a2}
Testees	…						
	j	y_{1j}	y_{2j}		y_{ij}		y_{aj}
	…						
	b	y_{1b}	y_{2b}		y_{ib}		y_{ab}

Figure 1:
Rasch model data-design interpreted as a two-way layout (mixed model). The items as the levels of a fixed factor A, the testees as the levels of a random factor B. y_{ij} is either 1 or 0, depending on whether Testee j has solved Item i or not

There are several test-statistics at a researcher's disposal in order to test the hypothesis of no interaction effect if both the factors are fixed, even when n = 1. The most well-known of such additivity tests is based on Tukey (1949). Rasch, Rusch, Šimečkova, Kubinger, Moder, and Šimeček (2009) furthermore proved *via* simulation studies that some modification of the latter actually keeps the type-I-risk even for the mixed factor design, given interval-scaled data; the power function has been established there as well. Thus, for this case, sample size might be calculated with reference to pertinent precision requirements. Unfortunately, the same is not at all true if dichotomous data are used (cf. Rasch et al., 2009). There are cases where the actual type-I-risk far exceeded .25, instead of the nominal .05; therefore, we must state that the two-way analysis of variance approach does not solve our problem.

A new attempt

In analogy to Andersen's approach in his LRT, we now consider grouping the testees. We therefore establish a third factor in the analysis of variance design, that is the group factor A with a levels, i.e. the groups. These groups need to be defined in advance, as a consequence of which this factor is a fixed one. As above, the two other factors are the testees (random factor B) and the items (fixed factor C with c levels). Obviously the factor B is nested within A, that is A is a partition of the total set of testees (for instance according to a testee's sex). This leads to a mixed classification $(A \succ B) \times C$, where C is crossed with $A \succ B$. For simplification, we select $a \cdot b$ testees in such a way that each of the a groups has

equal size *b* (see the design in Figure 2). Again we have the special case *n* = 1, and the data are dichotomous. The model equation of this model is then:

$$\boldsymbol{y}_{ijk} = \mu + a_i + \boldsymbol{b}_{ij} + c_k + (ac)_{ik} + \boldsymbol{e}_{ijk}\ ^{5} \qquad (1)$$

The table of analysis of variance can be found in "*Uebersicht* 1" in procedure 1-61/3300 in Rasch, Herrendörfer, Bock, Victor, and Guiard (2007) and the expected mean squares in the second column in "*Uebersicht* 3" of the same procedure[6]. The consequence of this new approach is that now specific objectivity means: there is no interaction effect between groups and items, that is between the two fixed factors *A* and *C*. From "*Uebersicht* 3" of the procedure 1-61/3300, we find that the statistic for testing our null-hypothesis is $F = \frac{MS_{AC}}{MS_{BCwithinA}}$, which is *F*-distributed under the null-hypothesis with (*a*-1)(*c*-1) and *a*(*b*-1)(*c*-1) degrees of freedom.

A		*C*						
			Items					
Groups	*B*		1	2	…	*k*	…	*c*
		1	y_{111}	y_{112}		y_{11k}		y_{11c}
1		2	y_{121}	y_{122}		y_{12k}		y_{12c}
	Testees	…						
		j	y_{1j1}	y_{1j2}		y_{1jk}		y_{1jc}
		...						
		b	y_{1b1}	y_{1b2}		y_{1bk}		y_{1bc}
2		*b+1*	$y_{2(b+1)1}$	$y_{2(b+1)2}$		$y_{2(b+1)k}$		$y_{2(b+1)c}$
		…						
...		…						
i		*j*	y_{ij1}	y_{ij2}		y_{ijk}		y_{ijc}
...		…						
a		*a·b=b'*	$y_{ab'1}$	$y_{ab'2}$		$y_{ab'k}$		$y_{ab'c}$

Figure 2:
Rasch model data-design interpreted as a three-way analysis of variance design with mixed classification $(A \succ B) \times C$. The items are levels of a fixed factor *C* and the testees are levels of a random factor *B*, nested within a fixed factor *A* of different groups. y_{ijk} is either 1 or 0, depending on whether Testee *j* from Group *i* has solved Item *k* or not

[5] Random variables are printed in bold.

[6] In this column there are two printing errors: In the row of *A* levels as well as in the row of *B* levels within *A* levels the term with $\sigma^2_{bc(a)}$ must be deleted.

In order to assess the type-I-risk, a simulation study was performed: The question was whether this test, applied in our case, keeps the nominal type-I-risk, and – given that it does – what its power is? For this study, the number of levels for the fixed factor *C* (items) was established as $c = 6$ and 20; the number of levels of the random factor *B* (testees) for each level of *A* was chosen as $b = 25$, 50, and 100, and the number of levels of the fixed factor *A* (groups) was restricted to $a = 2$ for the present. The *c* levels with parameters c_k (matches σ_k within Rasch model terminology – see Formula (2)) of the fixed factor *C* were set as equally spaced within the interval [-2.5, 2.5] for $c = 6$ and [-3, 3] for $c = 20$, which basically corresponds to the whole spectrum of item difficulties that arise in practice. The levels of the random factor $\mathbf{b}_{ij}$ (matches ξ_j within Rasch model terminology – see Formula (2)) were drawn randomly from a $N(0, 1.5)$, again corresponding to the values of person parameters that are likely to occur in practice. In each step of the simulation – the random number generator of R was used as implemented in the program package *eRm* (*extended Rasch modelling*; Mair & Hatzinger, 2006; cf. also Poinstingl, Mair, & Hatzinger, 2007). A data set was generated by calculating the probability *P* that testee *j* solves (+) item *i* according to the pertinent Rasch model formula:

$$P\left(+\middle|\xi_{\mathrm{v}},\sigma_{\mathrm{i}}\right)=\frac{\mathrm{e}^{\xi_{\mathrm{v}}-\sigma_{\mathrm{i}}}}{1+\mathrm{e}^{\xi_{\mathrm{v}}-\sigma_{\mathrm{i}}}} \tag{2}$$

Then a Bernoulli trial was carried out with the probability *P*, which led to a matrix of data based on the Rasch model. 100 000 simulation replications were performed, i.e. 100 000 data matrices were generated for each combination of $j(i)$ and *k*. A significance level of $\alpha = .05$ was applied. The main question of interest was whether the *F*-test for the interaction effect $A \times C$ holds this nominal type-I-risk.

If so, then a type-II-risk investigation, i.e. power analyses, should be made. Violations of the Rasch model could have been taken into account in a way similar to the approach of Suarez-Falcon and Glas (2003), but were intentionally restricted here to the case of DIF (differential item functioning) as concerns specific item pairs.

Results

We used the program package R (R Development Core Team, 2008) after their problem-specific routine had been tested by typical applications analysed with SAS and SPSS.

In preparation, we tested whether the analysis of variance in question actually works for $n = 1$ when normally distributed data are given. Data simulation was based on the null-hypotheses that there are no main or interactions effects. Using 100 000 simulation replications each, the largest difference of nominal and actual type-I-risk – that is the relative frequency of wrong rejections of the null-hypothesis – amounted to 0,00178 ($\alpha = .05$; $a = 2$, now $c = 3$, $18 \le b \le 32$); and the power with respect to the interaction $A \times C$ resulting for a given effect of an upper bound $[(ac)_{max} - (ac)_{min}] = 0.5 \cdot \sigma(\mathbf{e})$ and $[(ac)_{max} - (ac)_{min}] = 0.67 \cdot \sigma(\mathbf{e})$ was ~ .70 in both cases, with $b = 32$ in the former case and $b = 18$ in the latter case ($\alpha = .05$; $a = 2$, $c = 3$). That is, the non-standard application for $n = 1$ does no harm.

Coming then to the simulation study of data based on the Rasch model, the first scenario was no main effect as concerns the group factor A; as described above, there are severe main effects as concerns the testee factor $B(A)$ and the item factor C. Table 1 gives all the F-tests' results, though only the one concerning the interaction effect $A \times C$ is of focussed interest. As it turns out, actual type-I-risk of the F-test of the interaction effect $A \times C$ is near to the nominal one.

Table 1:

The F-tests in a three-way analysis of variance design $(A \succ B) \times C$ with mixed classification. A is a fixed factor with the $a = 2$ levels (groups from the same population), B is a random factor nested within A with the levels $b = 25$, 50, and 100 (testees) for each of the $a = 2$ levels, and C is a fixed factor with $c = 6$ levels (items). Given are the actual type-I-risks of the F-test for the interaction effect $A \times C$ and of the F-test of the main effect of A, as well as the power of the F-tests of the main effects of $B(A)$ and C – estimated using 100 000 simulation replications of Rasch model based data. The nominal type-I-risk is 5%

		p (F-test)			
b	effect	A	$B(A)$	C	$A \times C$
		groups	testees	items	
25		.05024	.99753	1.00000	.05371
50		.05104	1.00000	1.00000	.05276
100		.04938	1.00000	1.00000	.05208

The second scenario again involved Rasch model based data, but now an additional main effect, A, was taken into account: While the first group exhibited ξ_j, drawn randomly from $N(-0.5, 1.5)$, the second groups exhibited ξ_j, drawn randomly from $N(0.5, 1.5)$ – this corresponds in terms of the model equation (1) with $a_1 + E(b_1) = -0.5$ and $a_1 + E(b_1) = 0.5$. A main effect A is likely in practice, as such a group factor due to some critical testees' attitudes is explicitly looked for within Rasch model analyses in order to test the model (cf. in particular Andersen's LRT). Table 2 gives all the F-tests' results. As a matter of fact, type-I-risk of the F-test for the interaction effect $A \times C$ is artificially high and comes up to more than 16% in the case of $b = 100$: The actual type-I-risk increases monotonously by an increasing b.[7]

[7] We also tried Andersen's original approach of a-posteriori partition of the sample of testees according to their score. When doing so, $a = 5$ as $c = 6$ (testees with a score of 0 or 6 were deleted); as a result, even for $b = 25$, Rasch model based simulated data led to an artificial type-I-risk of .7920.

Table 2:

The *F*-tests in a three-way analysis of variance design $(A \succ B) \times C$ with mixed classification. *A* is a fixed factor with $a = 2$ levels (groups with different means), B is a random factor nested within *A* with the levels $b = 25$, 50, and 100 (testees) for each of the $a = 2$ levels, and *C* is a fixed factor with $c = 6$ levels (items). Given are the actual type-I-risks of the *F*-test of the interaction effect *A*×*C* and the power of the *F*-tests of the main effects of *A*, *B*(*A*), and *C* – estimated using 100 000 simulation replications of Rasch model based data. The nominal type-I-risk is 5%

		p (*F*-test)			
b	effect	*A*	*B*(*A*)	*C*	*A*×*C*
		groups	testees	items	
25		.45695	.99707	1.00000	.07777
50		.75257	1.00000	1.00000	.10440
100		.96232	1.00000	1.00000	.16515

For this, our new attempt needs to be restricted on a grouping factor which does not plausibly disclose an effect, for the present. This means a random partition into two groups, at worst. Nevertheless, the power of this approach is of interest.

We restricted the simulation study of the first scenario (i.e. there is no main effect *A*) to two cases. The first case refers to parameters c_k (matches σ_i) as [-2.5, -1.5, **-0.5**, **0.5**, 1.5, 2.5] for group $i = 1$ and as [-2.5, -1.5, **0.5**, **-0.5**, 1.5, 2.5] for $i = 2$ – this corresponds in terms of the model equation (1) with c_k is [-2.5, -1.5, 0.0, 0.0, 1.5, 2.5], $(ac)_{1k}$ is [0.0, 0.0, -0.5, 0.5, 0.0, 0.0], and $(ac)_{2k}$ is [0.0, 0.0, 0.5, -0.5, 0.0, 0.0]. That is, there actually is, apart from main effects of *B*(*A*) and *C*, only an interaction effect *A*×*C* due to Items 3 and 4. This means a DIF of both these items with respect to the two groups of testees. The results of the actual type-I-risk are given in Table 3a. The second case refers to parameters c_k (matches σ_i) as [-2.5; **-1.5;** -0.5; 0.5; **1.5**; 2.5] for group $i = 1$ and as [-2.5; **-0.5**; -0.5; 0.5; **0.5**; 2.5] for $i = 2$. The difference between these two cases is that in the latter case not only a two-item DIF, but also a difference in the variation of the parameters σ_i applies – that variation corresponds in terms of the model equation (1) with the nominator of the non-centrality parameter of the $(ac)_{ik}$. The respective results are given in Table 3b. It follows that in the case of the same variation of parameters σ_i, the *F*-test has more power, though the magnitude of DIF is smaller.

As a given main effect in *A* disclosed an artificial type-I-risk for the interaction effect *A*×*C* (i.e. an unacceptable risk of rejecting the null-hypothesis that the data conform to the Rasch model) we analogously analysed a second scenario in order to test the influence of such a main effect on the power of the *A*×*C* interaction *F*-test. The same group effect was applied as above and the case of two-item DIF with differences in the variation of parameters σ_i. The results of this analysis are given in Table 4. Now the artificial effect is almost constant and amounts (cf. Table 3b) "only" to 3 to 5 percent.

Table 3a:

The *F*-tests in a three-way analysis of variance design $(A \succ B) \times C$ with mixed classification. *A* is a fixed factor with $a = 2$ levels (groups with the same mean), *B* is a random factor nested within *A* with the levels $b = 25$, 50, and 100 (testees) for each of the $a = 2$ levels, and *C* is a fixed factor with $c = 6$ levels (items), having the same mean and the same variation of parameters σ_i at both levels of *A*. Given are the actual type-I-risks of the *F*-test of the interaction effect *A*×*C* and of the *F*-test of the main effect of *A*, as well as the power of the *F*-tests of the main effects of *B*(*A*) and *C* – estimated using 100 000 simulation replications of DIF based data: Within the first group, Rasch model based data were used with a two-item DIF as compared to the second group's Rasch model based data. The nominal type-I-risk is 5%

		p (*F*-test)			
b	effect	*A*	*B*(*A*)	*C*	*A*×*C*
		groups	testees	items	
25		.05010	.99771	1.00000	.37736
50		.04926	1.00000	1.00000	.68144
100		.05083	1.00000	1.00000	.94896

Table 3b:

The *F*-tests in a three-way analysis of variance design $(A \succ B) \times C$ with mixed classification. *A* is a fixed factor with $a = 2$ levels (groups with the same mean), *B* is a random factor nested within *A* with the levels $b = 25$, 50, and 100 (testees) for each of the $a = 2$ levels, and *C* is a fixed factor with $c = 6$ levels (items), having the same mean but a different variation of parameters σ_i at both levels of *A*. Given are the actual type-I-risks of the *F*-test of the interaction effect *A*×*C* and of the *F*-test of the main effect of *A*, as well as the power of the *F*-tests of the main effects of *B*(*A*) and *C* – estimated using 100 000 simulation replications of DIF based data: Within the first group, Rasch model based data were used with a two-item DIF as compared to the second group's Rasch model based data. The nominal type-I-risk is 5%

		p (*F*-test)			
b	effect	*A*	*B*(*A*)	*C*	*A*×*C*
		groups	testees	items	
25		.04866	.99838	1.00000	.30158
50		.04971	1.00000	1.00000	.57028
100		.05007	1.00000	1.00000	.89294

Table 4:

The *F*-tests in a three-way analysis of variance design $(A \succ B) \times C$ with mixed classification. *A* is a fixed factor with $a = 2$ levels (groups with different means), *B* is a random factor nested within *A* with the levels $b = 25$, 50, and 100 (testees) for each of the $a = 2$ levels, and *C* is a fixed factor with $c = 6$ levels (items), having the same mean but a different variation of parameters σ_i at both levels of *A*. Given is the power of the *F*-tests of all effects, that is of *A*×*C*, *A*, *B*(*A*), and *C* – estimated using 100 000 simulation replications of DIF-based data: Within the first group, Rasch model based data were used with a two-item DIF as compared to the second group's Rasch model based data. The nominal type-I-risk is 5%

		p (*F*-test)			
b	effect	*A*	*B*(*A*)	*C*	*A*×*C*
		groups	testees	items	
25		.45878	.99827	1.00000	.33456
50		.75764	1.00000	1.00000	.62696
100		.96459	1.00000	1.00000	.92727

Because $c = 6$ is rather an unusual case, we also investigated the case of $c = 20$. Table 5a gives the results specifically as concerns the type-I-risk of the *A*×*C* interaction *F*-test, and Table 5b as concerns its power. Both times no main effect *A* is assumed. Referring to the analysis of the actual type-I-risk, that is when the null-hypothesis is true, the parameters c_k (= σ_i) were [-3, -2.5, -2, -1.75, -1.5, -1.25, -1, -0.75, -0.5, -0.25, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2, 2.5, 3]; referring to the analysis of the power, that is when a specific alternative hypothesis is true, the parameters c_k (= σ_i) were [-3, -2.5, -2, -1.75, -1.5, -1.25, -1, -0.75, **-0.5**, -0.25, 0.25, **0.5**, 0.75, 1, 1.25, 1.5, 1.75, 2, 2.5, 3] for group $i = 1$ and -3, -2.5, -2, -1.75, -1.5, -1.25, -1, -0.75, **0.5**, -0.25, 0.25, **-0.5**, 0.75, 1, 1.25, 1.5, 1.75, 2, 2.5, 3] for $i = 2$. As a result, type-I-risk holds as well but the power of the interaction *F*-test is reduced in comparison to the case of $c = 6$ (cf. Table 3a).

Table 5a:

The *F*-tests in a three-way analysis of variance design $(A \succ B) \times C$ with mixed classification. *A* is a fixed factor with $a = 2$ levels (groups from the same population), *B* is a random factor nested within *A* with the levels $b = 25$, 50, and 100 (testees) for each of the $a = 2$ levels, and *C* is a fixed factor with $c = 20$ levels (items). Given are the actual type-I-risks of the *F*-test of the interaction effect *A*×*C* and of the *F*-test of the main effect *A*, as well as the power of the *F*-tests of the main effects of *B*(*A*) and *C* – estimated using 100 000 simulation replications of Rasch model based data. The nominal type-I-risk is 5%

		p (*F*-test)			
b	effect	*A*	*B*(*A*)	*C*	*A*×*C*
		groups	testees	items	
25		.04904	1,00000	1.00000	.05514
50		.05116	1.00000	1.00000	.05463
100		.05044	1.00000	1.00000	.05318

Table 5b:

The *F*-tests in a three-way analysis of variance design $(A \succ B) \times C$ with mixed classification. *A* is a fixed factor with $a = 2$ levels (groups with the same mean), *B* is a random factor nested within *A* with the levels $b = 25$, 50, and 100 (testees) for each of the $a = 2$ levels, and *C* is a fixed factor with $c = 20$ levels (items), having the same mean and the same variation of parameters σ_i at both levels of *A*. Given are the power of the *F*-test of the interaction effect $A \times C$ and of the *F*-tests of the main effects of $B(A)$ and *C*, as well as the actual type-I-risk of the *F*-test of the main effect of *A* – estimated using 100 000 simulation replications of DIF based data: Within the first group, Rasch model based data were used with a two-item DIF as compared to the second group's Rasch model based data. The nominal type-I-risk is 5%

		p (*F*-test)			
b	effect	*A*	*B*(*A*)	*C*	*A*×*C*
		groups	testees	items	
25		.05031	1.00000	1.00000	.21493
50		.04939	1.00000	1.00000	.42929
100		.04940	1.00000	1.00000	.79093

Discussion

Apart from some side issues as concerns possible group effects, our problem has generally been solved. The interaction effect testing *F*-test for the given analysis of variance design holds its type-I-risk α (bear in mind that within statistics a test is defined as 20%-robust if even in the case of a violation of its distribution assumptions the actual type-I-risk does not differ from the nominal value by more than 20 percent; cf. for instance Rasch & Guiard, 2004). And this *F*-test's power 1 - β and type-II-risk β, respectively, depend (monotonously) on the number of testees (*b*). For instance, as concerns the alternative hypothesis of a simple two item DIF with respect to two specific groups dealt with here, we can now design the data sampling, i.e. determine the sample size for Rasch model calibrating an achievement test: If there are $c = 20$ items with parameters likely to be equally spaced on the interval [-3, 3], if there is a nominal significance level of $\alpha = .05$, an aimed-for power of 1 - $\beta = .80$ (as is usually the case in designed studies), and a defined relevant effect of parameter difference with respect to two interesting groups of at least two times one (-0.5 minus 0.5 and 0.5 minus -0.5), if there are good reasons that this relevant DIF does not change the variation of parameters σ_i, and finally if there are good reasons that no main effect exists between the two interesting groups, then, according to Table 5b, a sample size of 100 testees in each of both groups is needed. – We additionally carried out a simulation study in order to achieve an exact power of .80 instead of .79093 as given in Table 5b; actually, a $b = 101$ is needed.

Obviously, there are many issues to consider in depth. Most of them involve the need for additional simulation studies, since we have dealt only with a handful of cases. However, the main issue is the artificial results of the type-I-risk of the $A \times C$ interaction *F*-test, given that a main effect of *A* exists.

Primarily, the question is the cause for this artefact since having no explanation could cast doubt on our entire simulation procedure. There is, however, an explanation. If there is a

main effect in A – that is $a_1 \neq a_2$ within our designed analysis of variance model's equation –, this is due to the nested $\mathbf{b}_{ij}$ which match the Rasch model's ξ_j, the latter stemming from two populations with different means. And this mean difference does not affect $\mathbf{y}_{ijk}$ additively with respect to k, the level of C: Though the items will be solved ($y_{ijk} = 1$) more often in the one group than in the other, this does not apply to every item with the same rate. This would only be the case if there were no main effect C, i.e. if every item had the same parameter. Because the first group has ξ_j drawn randomly from $N(-0.5, 1.5)$ while the second groups has ξ_j drawn randomly from $N(0.5, 1.5)$, the second group's testees will, on average, achieve more solutions to moderate items than the first group's testees, but not to easy or difficult items, because both group members will tend to solve the easy ones and tend to fail to solve the difficult ones. As a consequence, only/primarily in dependence on k, the y_{ijk} differ in average between the groups $i = 1$ and $i = 2$. And that means a significant $A \times C$ interaction effect, though the Rasch model holds.

In the first instance, one could conclude that our new approach is of no use, because we do not know when to reasonably assume that there is no difference in mean between both the groups for which a DIF is suspected. Besides the objection that such cases might yet happen – take for instance into account that there is hardly empirical evidence to support the assumption of different intelligence tests scores between different countries but a DIF seems plausible –, we may argue as follows: The new approach works as long as no significant main effect of A occurs. Then a significant interaction effect leads to the rejection of the Rasch model, a non-significant interaction effect to its acceptance. Thereby we can base our decision on precision requirements stated in advance, that is a (minimal) relevant magnitude of DIF has either to be claimed (risking an error with a percentage of α) or has to be disclaimed (risking an error with a percentage of β).

We must state that there was no intention of establishing a new statistic to test the Rasch model – though, under the given restrictions, our approach would actually work. The intention was only to allow designing data-sampling for Rasch model calibrating an achievement test. And, as illustrated for a typical case, this goal was achieved. It is irrelevant whether there actually is no main effect A, given that we might use our new approach only for determining the sample size needed to fit our precision requirements. This is necessary, above all, to avoid sampling too large samples which results in rejecting the Rasch model even when model contradiction is hardly of practical relevance. That is, after determining the sample size, we could simply apply Andersen's LRT. Of course, this advice is good as long as the correlation of our $A \times C$ interaction F-test approach and the LRT are not known. For this see Table 6 which juxtaposes the power of both tests for the same data. As a surprising fact, the $A \times C$ interaction F-test approach proves throughout to be more powerful. Hence, the given advice is rather poor.

Although there was no intention to evoke objections against Andersen's LRT besides the fact that it does not allow the determination of the sample size according to certain given precision requirements, our results show some faults of the LRT. With respect to the very restricted case of a two-item DIF between two given groups as analysed here, our approach is ultimately to be preferred[8]. That is, the LRT – or some other pertinent test for the Rasch model – should only be applied if the main effect A results in significance.

[8] Interested colleagues who prefer to apply R instead of SAS and SPSS might ask the authors for the syntax.

Table 6:
The power of the $A \times C$ interaction F-tests in a three-way analysis of variance design $(A \succ B) \times C$ with mixed classification as opposed to Andersen's LRT – F-test's were estimated using 100 000 simulation replications, LRT's using 10 000. There are DIF based data: Within the first group, Rasch model based data were used with a two-item DIF as compared to the second group's Rasch model based data. The nominal type-I-risk is 5%

		p (F-test)	p (LRT)
	b	$A \times C$	
c = 6 levels (items) having the same mean and the same variation of parameters σ_i at both levels of A.	25	.37736	.3063
	50	.68144	.6207
	100	.94896	.9196
c = 6 levels (items) having the same mean but a different variation of parameters σ_i at both levels of A.	25	.30158	.2453
	50	.57028	.5172
	100	.89294	.8562
c = 20 levels (items) having the same mean and the same variation of parameters σ_i at both levels of A.	25	.21493	.1874
	50	.42929	.3773
	100	.82285	.7177

Given, however, the serious problem that in the presence of a main effect of A, artificial results of the type-I-risk of the $A \times C$ interaction F-test are observed, further research is needed. Maybe applying our approach directly to the simulated probabilities P (that Testee j solves Item i) instead of using a Bernoulli trial in order to get y_{ijk} = 0/1 data would help. Presumably the artefact phenomenon discussed would be of less consequence in this case.

Conclusion

For the present, we can advise researchers calibrating a 20 items achievement test according to the Rasch model to use 100 testees in each of two groups for which a DIF of 1 or even more is suspected with respect to at least a pair of items. If our approach of a three-factorial analysis of variance with an $(A \succ B) \times C$ design with mixed classification is then applied and no significant group effect discloses the $A \times C$ interaction, the F-test rejects the Rasch model with a type-I-risk of .05 and accepts the Rasch model with a type-II-risk of approximately .20 – given a DIF of the discussed magnitude; that is, the F-test then has a power of .80[9]. In the case of a significant group effect, a Rasch model test like Andersen's LRT has to be applied; but be aware that the power of this LRT is lower.

[9] We also have analysed a similar case as given in Table 3a, in order to test Goethals' (1994) rule of thumb (any difference of parameter estimations not greater than a tenth of the range of the parameters is hardly of practical relevance): For b = 100 the power of the test amounts just to .37611.

Several challenges remain:

1. investigating many other cases of b and c
2. investigating many other cases of α and β, for instance $\alpha = .01$ and $\beta = .10$ or .05
3. tabulating the needed b for given c, α, β, and a given magnitude of relevant DIF
4. investigating DIF for a single item only, and for more than a pair of items
5. investigating other cases of groups' different variations of item parameters
6. investigating different variances of ξ_j between the $a = 2$ groups[10]
7. investigating other violations of the Rasch model than DIF, for instance according to Suarez-Falcon and Glas (2003)
8. providing the formula for calculating b explicitly for given precision requirements and developing respective software
9. considering the case of using different test-booklets with partially different subsets of items.

References

Alexandrowicz, R. (2002). *Die Teststärke des Likelihood-Quotienten-Tests nach Andersen bei der Überprüfung der Modellgüeltigkeit des dichotomen logistischen Modells nach Rasch* [Andersen-Likelihood-Ratio test's power as a model check of the Rasch model]. Unpubl. doctoral thesis, University of Vienna, Vienna.

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

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach.* New York: Springer.

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

Goethals, R. (1994). *Die praktische Erprobung von Alternativen zur multiple-choice-Vorgabe bei Computertests* [Experiences with alternatives of traditional multiple choice response format at computerized testing]. Unpubl. doctoral thesis, University of Vienna, Vienna.

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests* [Introduction in psychological test theory]. Berne: Huber.

Kelderman, H. (1984). Loglinear Rasch model tests. *Psychometrika, 49*, 223-245.

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. (2009). *Adaptives Intelligenz Diagnostikum – Version 2.2 (AID 2) samt AID 2-Türkisch* [Adaptive Intelligence Diagnosticum, AID 2-Turkey included]. Göttingen: Beltz.

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

McCullagh, P., & Nelder, J. M. (1989). *Generalized linear models.* London: Chapman and Hall.

Müller-Philipp, S., & Tarnai, C. (1989). Signifikanz und Relevanz von Modellabweichungen beim Rasch-Modell [Significance and relevance of model misfits in the Rasch model]. In K. D. Kubinger (ed.), *Moderne Testtheorie – Ein Abriß samt neuesten Beiträgen* [Modern psychometrics – A brief survey with recent contributions]. (2nd ed.) (pp. 239-258). Munich: PVU.

[10] This is due to a reviewer's suggestion.

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.

R Development Core Team (2008). *R: A language and environment for statistical computing.* Wien: R Foundation for Statistical Computing (URL http://www.R-project.org).

Rasch, G. (1980, reprint). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press.

Rasch, D. & Guiard, V. (2004). The robustness of parametric statistical methods. *Psychology Science, 46*, 175-208.

Rasch, D., Herrendörfer, G., Bock, J., Victor, N., & Guiard, V. (2007) (Eds.). *Verfahrensbibliothek* [Encyclopedia of statistical procedures]. (2nd ed.). Munich: Oldenbourg.

Rasch, D., Rusch, T., Šimečkova, M., Kubinger, K. D., Moder, K. & Šimeček, P. (2009). Tests of additivity in mixed and fixed effect two-way ANOVA models with single sub-class numbers. *Statistical Papers, 50*, 905-916.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods.* Thousand Oaks, CA: Sage.

Scheiblechner, H. H. (2009). Rasch and pseudo-Rasch models: suitableness for practical test applications. *Psychology Science Quarterly, 51*, 181-194.

Suarez-Falcon, J., & Glas, C. A. W. (2003). Evaluation of global testing procedures for item fit to the Rasch model. *British Journal of Mathematical and Statistical Psychology, 53*, 127-143.

OECD (2007). *PISA – The OECD Programme for International Student Assessment.* Paris: Organization for Economic Co-Operation and Development.

Tukey, J. W. (1949). One degree of freedom for non-additivity. *Biometrics, 5*, 232-242.