Educational Research and Evaluation
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/nere20

A new approach for testing the Rasch model

Klaus D. Kubinger [a] , Dieter Rasch [b] & Takuya Yanagida [a]
[a] Division of Psychological Assessment and Applied Psychometrics, Faculty of Psychology, University of Vienna, Vienna, Austria
[b] Institute of Applied Statistics and Computing, University of Natural Resources and Applied Life Sciences, Vienna, Austria

Available online: 30 Nov 2011

To cite this article: Klaus D. Kubinger, Dieter Rasch & Takuya Yanagida (2011): A new approach for testing the Rasch model, Educational Research and Evaluation, 17:5, 321-333

To link to this article: http://dx.doi.org/10.1080/13803611.2011.630529

Educational Research and Evaluation
Vol. 17, No. 5, October 2011, 321–333

A new approach for testing the Rasch model

Klaus D. Kubinger[a]*, Dieter Rasch[b] and Takuya Yanagida[a]

[a]*Division of Psychological Assessment and Applied Psychometrics, Faculty of Psychology, University of Vienna, Vienna, Austria;* [b]*Institute of Applied Statistics and Computing, University of Natural Resources and Applied Life Sciences, Vienna, Austria*

Though calibration of an achievement test within psychological and educational context is very often carried out by the Rasch model, data sampling is hardly designed according to statistical foundations. However, Kubinger, Rasch, and Yanagida (2009) recently suggested an approach for the determination of sample size according to a given Type I and Type II risk, and a certain effect of model misfit when testing the Rasch model is supported by some new results. The approach uses a three-way analysis of variance design $(A \succ \boldsymbol{B}) \times C$ with mixed classification. There is a (fixed) group factor A, a (random) factor $\boldsymbol{B}$ of testees within A, and a (fixed) factor C of items cross-classified with $(A \succ \boldsymbol{B})$. The simulation study in this article deals with further item parameter ranges and ability parameter distributions, and with larger sample sizes and item numbers than the original paper. The results are: The approach works given several restrictions, and its main aim, the determination of the sample size, is attained.

Keywords: Rasch model; sample size; Type I and Type II risk; analysis of variance; mixed model

Introduction

Recently, Kubinger, Rasch, and Yanagida (2009) suggested a very simple approach for testing the Rasch model (Rasch, 1980), nowadays also known as the 1-PL model. Most importantly, the authors strictly distinguish between (Rasch-) ''model tests'', on the one hand, which test model implications or are, so to speak, performed according to specific objective measurement, and ''goodness-of-fit tests'', on the other hand, which only measure the model's appropriateness. Their approach deals with model tests. And although there are several statistical approaches for testing the Rasch model, among which Andersen's Likelihood-Ratio test is best established (LRT; Andersen, 1973) – see Glas and Verhelst (1995), for a current review of Rasch model tests –, the authors emphasize that all these tests offer no clear procedure for planning a study for item calibration. That is, if a researcher tries to calibrate within psychological or educational context an achievement test according to the Rasch model, data sampling is not designed based on statistical foundations concerning the determination of the sample size for fulfilling certain ''precision'' requirements. So

*Corresponding author. Email: klaus.kubinger@univie.ac.at

ISSN 1380-3611 print/ISSN 1744-4187 online

http://dx.doi.org/10.1080/13803611.2011.630529
http://www.tandfonline.com

they point out that it is necessary to calculate the sample size given a certain Type I risk α and a certain Type II risk β – that is, the probability of rejecting the Rasch model though it is correct, on the one side, and the probability of accepting the Rasch model though it is wrong, on the other side –, and of course given a certain effect δ. This effect refers to the degree of item parameters' misfit to the Rasch model, which is supposed to be of practical importance. That is, 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. To achieve this, the authors aimed to use an *F*-distributed statistic where the sample size directly affects the degrees of freedom – bear in mind that Andersen's LRT is chi-squared distributed and the statistics' degrees of freedom do not at all depend on the sample size, but only on the number of estimated parameters. In contrast, the proposed *F*-distributed statistic enables the researcher 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.

Though Kubinger et al. (2009) focussed just on a proper approach to sample size calculation or rather to planning a study for Rasch model calibrating an achievement test, they suggested by the way a new approach to testing the Rasch model. They disclosed that this new approach is – despite some severe shortcomings – seriously competitive to Andersen's LRT. But as they restricted their simulation study to a very small sample of scenarios, we now try in this article to give some support to their results and to sound out the approach's practicability in greater detail. We therefore conduct a simulation study with a broader range of scenarios.

Method

Like Andersen's LRT, the approach of Kubinger et al. (2009) shares the most frequently referenced assumption of the Rasch model – that is that the item difficulty parameters are statistically independent of the person ability parameters. The approach is now a three-way analysis of variance.

Analysis of variance[1] is a special case of the so-called general linear model. That is, some character y has to be modelled as a random variable $\boldsymbol{y}$ due to a linear function of certain model parameters. Thus, models of analysis of variance differ with respect to such parameters. The main aim of analysis of variance is to test the null hypothesis that some of these parameters are zero. The pertinent test is an *F*-test which presupposes a random variable $\boldsymbol{y}$ which is normally distributed. Given different conditions or treatments by which this variable is sampled, the *F*-test presupposes equal variances as well. These conditions or treatments establish the different levels of so-called factors and might move the random variable's $\boldsymbol{y}$ mean. In educational research, such factors could be sex and age of the pupils, social status of their parents, localisation of the school, and many other variables. Dependent on the number p of such factors there, we have a p-way layout of analysis of variance. At least two different data situations and models must be distinguished. Case 1: All or certain levels a of any factor are of interest and included in the analysis. That is, the levels are fixed and not (randomly) sampled; hence, the factor is a *fixed factor*. From the factors mentioned above, sex and age of pupils as well as social status of the parents are fixed factors. We call this situation Model I. Case 2: There are many levels of the interesting factor, the number of which has theoretically to be considered as infinite. Those a levels included into the study have been randomly

selected by drawing a random sample from the population of all levels. Thus, the factor has to be modelled as a random variable, too. Hence, the factor is a *random factor*. From the factors mentioned above, the school would be a random factor if we consider not all schools from a country but only take a random sample of schools. We call this situation Model II. Besides these two models for a single factor or even for p factors, there are several mixed models when $p - x$ factors are fixed and x factors are random ($p = 2, 3, \ldots ; x = 1, 2, \ldots p - 1$). A further characterization of analysis of variance refers to a cross-classification of the $p > 1$ factors versus a hierarchical classification or to say "nested factors". Let us consider the case of $p = 3$ like in the following, dealing with the factors A, B, and C having a, b, and c levels, respectively. If every level of any factor is combined with every level of the other factors, then the design is a cross-classification. This occurs for instance in the case we would have girls and boys (levels of factor A = sex) combined completely with the ages 10, 11, and 12 years (levels of factor B = age) and all five social groups of parents (levels of factor C = social status). That is, we would have a three-way cross-classification (Model I) with $a = 2$, $b = 3$, and $c = 5$ levels. However, if $p = 2$ and the levels of, for example, the factor B were realized only in a certain level of another factor, say A, then B is nested within A; we would have a nested factor. In case of $p > 2$, we of course could gain some mixed classification. In the following, we need a mixed three-way classification: $(A \succ \boldsymbol{B}) \times C$. That is, A and C are fixed factors, and $\boldsymbol{B}$ (in bold) is a random factor, and $\boldsymbol{B}$ is nested within A; the combination $\boldsymbol{B}$ nested in A $(A \succ \boldsymbol{B})$ is cross-classified with C.

Now, the different items of an item pool to be calibrated according to the Rasch model are considered as the c different levels of a first (fixed) factor C, and the testees as the b different levels of a second (random) factor $\boldsymbol{B}$ – factor C is a fixed one, because just these given items are of interest, and factor $\boldsymbol{B}$ is a random one, as there is an almost randomly chosen sample of testees who are part of a certain intended population. Finally, the second (fixed) factor A is due to grouping of the testees in a different levels. These groups need to be defined in advance and therefore establish a fixed factor; they might be, for instance, male versus female testees. Obviously, then, the factor B is nested within A, that is, A is a partition of the total set of testees. This leads to a mixed classification $(A \succ \boldsymbol{B}) \times C$, where C is crossed with $(A \succ \boldsymbol{B})$. Now, Rasch model fitting means that there is no interaction effect between the fixed factors – irrespective of the presumably strong main effect C due to different items being solved more or less frequently within the sample. For simplification, $a \cdot b$ testees will be selected in such a way that each of the a groups has equal size b (see the design in Figure 1). Then the equation for this model is:

$$\mathbf{y}_{ijk} = \mu + a_i + \boldsymbol{b}_{ij} + c_k + (ac)_{ik} + \boldsymbol{e}_{ijk}{}^2 \tag{1}$$

For instance, according to Rasch, Herrendörfer, Bock, Victor, and Guiard (2008), the statistic for testing the null hypothesis H_0: $(ac)_{ik} = 0$ for every i and k is $F = \frac{MS_{AC}}{MS_{BCwithinA}}$, which is F-distributed with $(a-1)(c-1)$ and $a(b-1)(c-1)$ degrees of freedom – MS the mean squares.

The presented design of analysis of variance suffers from two problems: Firstly, the design establishes just a single observation within each cell ($n = 1$) of a mixed model; secondly, this design is applied to dichotomous, not interval-scaled – and not remotely normally distributed – data. A simulation study needs to assess the actual

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 1. Rasch model data design interpreted as a three-way analysis of variance design with mixed classification $(A \succ \boldsymbol{B}) \times C$.
Note: The items are levels of a fixed factor C, and the testees are levels of a random factor $\boldsymbol{B}$, 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.

Type I and Type II risk as these violations of the analysis of variance's test assumptions could destroy the test statistic's distribution.

Kubinger et al. (2009) ran several scenarios: $c = 6$ and 20; $b = 25$, 50, and 100; $a = 2$. The c levels with parameters c_k (matches item difficulty parameters σ_k within Rasch model terminology – see Formula (2)) were equally spaced within the interval [–2.5, 2.5] for $c = 6$ and [–3, 3] for $c = 20$; the levels of the random factor $\boldsymbol{b}_{ij}$ (matches person ability parameters ξ_j within Rasch model terminology – see Formula (2)) were randomly drawn from a $N(0, 1.5)$. One hundred thousand 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. However, similar scenarios were used for power analyses. Violations of the Rasch model were restricted to the case of differential item functioning (DIF) as concerns specific item pairs. As concerns $c = 6$, 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 $i = 1$ in A 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 as

[–2.5, –1.5, 0.0, 0.0, 1.5, 2.5], $(ac)_{1k}$ as
[0.0, 0.0, –0.5, 0.5, 0.0, 0.0], and $(ac)_{2k}$ as
[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 $A \times C$ due to items 3 and 4. This means there is a DIF of both these items with respect to the two groups of testees. 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 $i = 1$ in A 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 $(ac)_{ik}$. As concerns $c = 20$, for the analysis of the actual Type I risk (that is when the null hypothesis is true), the parameters c_k $(=\sigma_i)$ 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];

and referring to the analysis of the power (that is when a specific alternative hypothesis is true), the parameters c_k $(=\sigma_i)$ 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 $i = 1$ in A 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 $i = 2$.

The main results were:

(1) If a main effect of A exists, an artificially high Type I risk of the $A \times C$ interaction F-test results – that is, the new approach works as long as no significant main effect of A occurs.
(2) Given no main effect of A and the null hypothesis is correct, a significant interaction effect $A \times C$ occurs with a probability very near to the actual Type I risk.
(3) Given no main effect of A and the null hypothesis is wrong, a significant interaction effect $A \times C$ occurs with an acceptable probability for a defined two-item DIF, depending on the number of testees (and the size of the item pool).
(4) Throughout the investigated scenarios, the $A \times C$ interaction F-test approach proves to be more powerful than Andersen's LRT.

We now perform an additional simulation study in order to test several other scenarios. Above all, a broader range of item parameter than [–3, 3] is of interest, as well as peaked rather than equally spaced item parameters within that interval. Other variances of person ability parameters are also of interest. Finally, we wondered whether there is much difference in the power of the approach under discussion when there is only a single-item DIF or, on the other side, two times a two-item DIF.

Hence, the following scenarios (in combination) were under consideration:

(1) number of items: $c = 6$; 25; 30; 40; 60; 100; 500;
(2) number of testees (within each of $a = 2$ groups): $b = 25$; 50; 100; 150; 500;
(3) standard deviations of person ability parameters: normally distributed with $N(0, 1)$; $N(0, 1.5)$; $N(0, 2)$; $N(0, 2.5)$, and uniquely distributed in [–3, 3]; [−4, 4];
(4) interval of item parameter: [–3, 3]; [–4, 4];
(5) item parameters' concentration: equally spaced versus a peaked dispersion;
(6) number of DIF: a single-item DIF; a two-item DIF; two times a two-item DIF; three times a two-item DIF;
(7) magnitude of DIF: 0.6.

Based on the already given results, we restricted the simulation study to cases where there is no main effect A.

In each step of the simulation, the random number generator of R (R Development Core Team, 2008) was used as implemented in the program package *extended Rasch modeling* (*eRm*; Mair, Hatzinger, & Maier, 2010; cf. also Poinstingl, Mair, & Hatzinger, 2007). A data set was generated by calculating the probability P that testee j with person ability parameter ξ_v solve (+) item i with item difficulty parameter σ_i according to the pertinent Rasch model formula:

$$P(+|\xi_v,\sigma_i) = \frac{e^{\xi_v-\sigma_i}}{1+e^{\xi_v-\sigma_i}} \tag{2}$$

Then a Bernoulli trial was carried out with the probability P, which led to a matrix of data based on the Rasch model. In contrast to Kubinger et al. (2009), who performed 100,000 simulation replications for each scenario, we only used for the moment 10,000 replications if $b<150$ and only 1,000 replications if $b \geq 150$ and $c > 6$. In the case of $c = 500$, we performed 1,000 replications for all conditions. That is, 10,000 or 1,000 data matrices were generated for each combination of $j(i)$ and k. A significance level of $\alpha = .05$ was applied. Of course, simulation studies in statistics are always based on 100,000 replications, because otherwise chance effects might be established leading to unacceptable impreciseness. Nevertheless, we chose smaller replication numbers in order to be able to take more relevant scenarios into account. In the case of incongruent results, we had a larger number of replications up our sleeve.

Results

We used the program package R for the calculation of all F-tests (main effects A, B, C, and the interaction effect $A \times C$).

The first question of interest was whether the F-test for the interaction effect $A \times C$ holds the nominal Type I risk of 5%. Table 1 gives the results of Kubinger et al. (2009). Table 2 gives our results for the case of equally spaced item parameters within the interval [–3, 3] and the ability parameters randomly distributed according to $N(0, 1.5)$. Table 3 does the same; however, now the interval is [–4, 4]. As a matter of fact, the corresponding cases in Table 1 and Table 2 (see the shadowed cells) led to

Table 1. The actual Type I risk of the F-test for the interaction effect $A \times C$ in a three-way analysis of variance design $(A \succ \boldsymbol{B}) \times C$ with mixed classification (the nominal Type I risk is 5%).

	p (F-test) $A \times C$		
	b		
c	25	50	100
6	.05371	.05276	.05208
20	.05514	.05463	.05318

Note: A is a fixed factor with $a = 2$ levels (groups from the same population), $\boldsymbol{B}$ 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 $c = 6$, 20 levels (items). Estimations are based on 100,000 simulation replications of Rasch model-based data; the item parameters c_k $(=\sigma_i)$ are equally spaced within the interval [–2.5, 2.5] in the case of $c = 6$ and within the interval [–3, 3] in the case of $c = 20$; the ability parameters are randomly distributed according to $N(0, 1.5)$.

Table 2. The actual Type I risk of the *F*-test for the interaction effect $A \times C$ in a three-way analysis of variance design $(A \succ \boldsymbol{B}) \times C$ with mixed classification (the nominal Type I risk is 5%).

	p (*F*-test) $A \times C$				
	b				
c	25	50	100	150	500
6	.0564	.0533	.0532	.0546	.0514
25	.0573	.0562	.0575	.046	.057
30	.0581	.0583	.0532	.055	.056
40	.0576	.0597	.0560	.057	.060
60	.0664	.0612	.0572	.056	.066
100	.0635	.0665	.0617	.065	.072
500	.098	.095	.087	.085	.102

Note: A is a fixed factor with $a = 2$ levels (groups from the same population), $\boldsymbol{B}$ is a random factor nested within A with the levels $b = 25$, 50, 100, 150, and 500 (testees) for each of the $a = 2$ levels, and C is a fixed factor with $c = 6$, 25, 30, 40, 60, 100, and 500 levels (items). Estimations are based on 1,000 or 10,000 simulation replications of Rasch model-based data; the item parameters c_k $(=\sigma_i)$ are equally spaced within the interval [–3, 3], the ability parameters randomly distributed according to $N(0, 1.5)$.

Table 3. The actual Type I risk of the *F*-test for the interaction effect $A \times C$ in a three-way analysis of variance design $(A \succ \boldsymbol{B}) \times C$ with mixed classification (the nominal Type I risk is 5%).

	p (*F*-test) $A \times C$				
	b				
c	25	50	100	150	500
6	.0623	.0631	.0680	.0573	.0635
25	.0705	.0680	.0682	.066	.064
30	.0682	.0675	.0661	.068	.068
40	.0682	.0692	.0715	.077	.069
60	.0756	.0701	.0769	.075	.074
100	.0852	.0808	.0844	.075	.100
500	.134	.119	.143	.140	.153

Note: A is a fixed factor with $a = 2$ levels (groups from the same population), $\boldsymbol{B}$ is a random factor nested within A with the levels $b = 25$, 50, 100, 150, and 500 (testees) for each of the $a = 2$ levels, and C is a fixed factor with $c = 6$, 25, 30, 40, 60, 100, and 500 levels (items). Estimations are based on 1,000 or 10,000 simulation replications of Rasch model-based data; the item parameters $c_k(=\sigma_i)$ are equally spaced within the interval [–4, 4], the ability parameters randomly distributed according to $N(0, 1.5)$.

almost the same values. But, specifically Table 3 discloses a trend that the actual Type I risk increases when b and c increase; and if the item parameters have a greater range, then the actual Type I risk is almost always beyond the nominal Type I risk plus 20%. For this reason, we analysed in greater detail the behavior of the actual Type I risk in dependence on (a) the interval of the item parameters, (b) the standard deviation of the ability parameters, and (c) the item parameters' concentration. Although we did this for every combination of b and c, we summarize the results in Table 4 only for a representative combination of them, that is $b = 100$ and $c = 40$.

Hence, for the moment the approach suggested by Kubinger et al. (2009) for testing the Rasch model has to be brought back down to earth somewhat. In

Table 4. The actual Type I risk of the *F*-test for the interaction effect $A \times C$ in a three-way analysis of variance design $(A \succ \boldsymbol{B}) \times C$ with mixed classification in dependence of (a) the interval of the item parameters, (b) the standard deviation of the ability parameters, and (c) the item parameters' concentration (the nominal Type I risk is 5%).

		testees' ability					
items' difficulty		$N(0, 1)$	$N(0, 1.5)$	$N(0, 2)$	$N(0, 2.5)$	uniquely distributed in [–3, 3]	uniquely distributed in [–4, 4]
equally spaced	[−3, 3]	.0591	.0560	.0604	.0671	.0581	
	[−4, 4]	.0710	.0715	.0736	.0807		.0716
peaked	[−3, 3]	.0559	.0538	.0529	.0552	.0477	
	[−4, 4]	.0600	.0628	.0577	.0673		.0655

Note: For simplicity, only the results for a representative combination of b (=100) and c (=40) are given. Estimations are based on 10,000 simulation replications of Rasch model-based data.

addition to the already mentioned restriction that there must not be a significant main effect in A (a partition of the total set of testees), we now also have to restrict this approach to applications of an item parameter interval [–3, 3] (at most) with a peaked dispersion (at best) and a (normal) distribution of the testees' ability parameters with a standard deviation not larger than 1.5 – and even then the number of items should not be larger than 100 (better not larger than 40), and the number of testees no larger than two times 150. Of course, there are two ways out or attempts to rescue the approach. On the one hand, Kubinger, Rasch, and Yanagida (2009) did not really aim for a new Rasch model test but only tried to use a three-way analysis of variance design $(A \succ \boldsymbol{B}) \times C$ with mixed classification as a means for determining the sample size according to a given Type I and Type II risk, and according to a certain effect of model misfit when testing the Rasch model. And for this purpose, the approach will work, given the mentioned restrictions. On the other hand, we could of course apply some correction of the nominal Type I risk in order to gain an actual Type I risk that fits. However, we can not detect any mathematical function for doing so in a well-founded manner.

So the second question of interest, the power of the *F*-test for the interaction effect $A \times C$, is to be restricted to cases where the nominal Type I risk has proven to hold tolerably – that is, an item parameter interval [–3, 3] (equally spaced) and a normal distribution of the testees' ability parameter with a standard deviation of 1.5. And the case of $c = 500$ is not considered any more. As indicated, we investigated a single-item DIF, a two-item DIF, and a two times two-item DIF. The magnitude of the DIF has been fixed to 0.6 as this is, in case of an item parameter interval [–3, 3], a 10th of the item parameters' range, which is, according to some rules of thumb, a relevant effect size (cf. Kubinger, 2005). The single-item DIF is, for instance, for the case of $c = 25$ and DIF's location at the 15th item as follows: the parameters c_k $(=\sigma_i)$ were

[–3.00, –2.75, –2.50, –2.25, –2.00, –1.75, –1.50, –1.25, –1.00, –0.75, –0.50, –0.25, 0.00, 0.25, **0.50**, 0.75, 1.00, 1.25, 1.50, 1.75, 2.00, 2.25, 2.50, 2.75, 3.00] for group $i = 1$ in A and
[–3.02, –2.77, –2.52, –2.27, –2.02, –1.77, –1.52, –1.27, –1.02, –0.78, –0.52, –0.28, –0.03, 0.22, **1.10**, 0.72, 0.98, 1.23, 1.48, 1.73, 1.98, 2.23, 2.48, 2.73, 2.98] for $i = 2$.

The two-item DIF is, for instance, for the case of $c = 25$ and DIF's location at the 12th and 14th items as follows: the parameters c_k $(=\sigma_i)$ were

[–3.00, –2.75, –2.50, –2.25, –2.00, –1.75, –1.40, –1.25, –1.00, –0.75, –0.50, **–0.30**, 0.00, **0.30**, 0.50, 0.75, 1.00, 1.25, 1.40, 1.75, 2.00, 2.25, 2.50, 2.75, 3.00] for group $i = 1$ in A and
[–3.00, –2.75, –2.50, –2.25, –2.00, –1.75, –1.40, –1.25, –1.00, –0.75, –0.50, **0.30**, 0.00, **–0.30**, 0.50, 0.75, 1.00, 1.25, 1.40, 1.75, 2.00, 2.25, 2.50, 2.75, 3.00] for $i = 2$.

The two times a two-item DIF is, for instance, for the case of $c = 25$ and DIF's location at the 5th, 7th, 19th, and 21st items as follows: the parameters c_k $(=\sigma_i)$ were

[–3.00, –2.75, –2.50, –2.25, **–2.00**, –1.75, –**1.40**, –1.25, –1.00, –0.75, –0.50, –0.30, 0.00, 0.30, 0.50, 0.75, 1.00, 1.25, **1.40**, 1.75, **2.00**, 2.25, 2.50, 2.75, 3.00] for group $i = 1$ in A and
[–3.00, –2.75, –2.50, –2.25, –**1.40**, –1.75, **–2.00**, –1.25, –1.00, –0.75, –0.50, –0.30, 0.00, 0.30, 0.50, 0.75, 1.00, 1.25, **2.00**, 1.75, **1.40**, 2.25, 2.50, 2.75, 3.00] for $i = 2$.

And the three times a two-item DIF is, for instance, for the case of $c = 25$ and DIF's location at the 5th, 7th, 12th, 14th, 19th, and 21st items as follows: the parameters c_k $(=\sigma_i)$ were

[–3.00, –2.75, –2.50, –2.25, **–2.00**, –1.75, **–1.40**, –1.25, –1.00, –0.75, –0.50, **–0.30**, 0.00, **0.30**, 0.50, 0.75, 1.00, 1.25, **1.40**, 1.75, **2.00**, 2.25, 2.50, 2.75, 3.00] for group $i = 1$ in A and
[–3.00, –2.75, –2.50, –2.25, **–1.40**, –1.75, **–2.00**, –1.25, –1.00, –0.75, –0.50, **0.30**, 0.00, **–0.30**, 0.50, 0.75, 1.00, 1.25, **2.00**, 1.75, **1.40**, 2.25, 2.50, 2.75, 3.00] for $i = 2$.

That is, in accordance with the results of Kubinger et al. (2009), we avoided the case that the relevant DIF changes the variation of the item parameters – and, remember, no main effect between the two interesting groups (factor A) has been assumed.

Of course, the effect of DIF depends on the absolute value of the item parameter. Hence we analysed different localisations of the DIF, too. We always took into account four different locations on the item parameter continuum in the case of a single-item DIF and always three different couples or quadruples of locations in the cases of a two-item and two times two-item DIF; only in the case of three times a two-item DIF did we just study a single form of localisations. Table 5 summarizes the results of the single-item DIF and the two-item DIF, Table 6 for two and three times a two-item DIF.

The main result of Tables 5 and 6 is as follows: Obviously, and not surprisingly (cf. the item standard error of estimation's dependency on the parameter itself), the power of the F-test is considerably greater if the same DIF happens to occur for items with moderate item difficulty. Then, there is of course a big difference in the power between a sample size of 2×150 or 2×500; but, as Table 5 (a single-item DIF and a two-item DIF) shows, since a DIF of 0.6 leads in general to very low power, differences between sample sizes of 2×25 (50, 100) and 2×150 are almost negligible. And, also expectedly, the power decreases if the number of DIF-involved items becomes relatively small. For instance, a two-item DIF in a medium position results in power of about .435 when there are 25 items (with 2×150 testees), however only about .365 when there are 40 items. All in all, Table 5 discloses that the F-test's

Table 5. The power of the *F*-test for the interaction effect $A \times C$ in a three-way analysis of variance design $(A \succ \boldsymbol{B}) \times C$ with mixed classification (the nominal Type 1 risk is 5%).

			p (*F*-test) $A \times C$				
			b				
number of DIF	localisation of DIF	c	25	50	100	150	500
1	1	6	.0693	.0821	.1162	.1612	.5117
	3		.1159	.1826	.3156	.4668	.9501
	4		.1063	.1598	.2838	.4196	.9270
	6		.0628	.0750	.0923	.1165	.3260
1	5	25	.0667	.0745	.1019	.143	.312
	11		.0805	.1018	.1555	.216	.704
	15		.0731	.0969	.1395	.174	.655
	21		.0671	.0702	.0794	.124	.266
2	5/7	25	.0814	.1019	.1588	.256	.786
	12/14		.0985	.1587	.2974	.435	.971
	19/21		.0776	.1049	.1557	.226	.776
1	6	30	.0655	.0792	.0979	.128	.394
	13		.0716	.0949	.1456	.193	.674
	18		.0696	.0887	.1324	.183	.629
	25		.0641	.0714	.0845	.113	.239
2	6/9	30	.0778	.0955	.1528	.245	.750
	14/17		.0953	.1448	.2688	.414	.957
	22/25		.0724	.1016	.1507	.230	.753
1	7	40	.0658	.0740	.0891	.101	.310
	16		.0704	.0920	.1340	.185	.570
	25		.0732	.0828	.1165	.159	.489
	34		.0669	.0694	.0697	.089	.172
2	7/11	40	.0702	.0902	.1281	.164	.631
	18/23		.0863	.1271	.2348	.365	.943
	30/34		.0758	.0907	.1261	.175	.608
1	10	60	.0680	.0706	.0844	.091	.252
	24		.0739	.0910	.1154	.137	.495
	37		.0708	.0805	.1032	.127	.403
	51		.0654	.0682	.0792	.084	.149
2	10/16	60	.0738	.0808	.1131	.157	.511
	27/34		.0887	.1159	.1931	.268	.887
	45/51		.0749	.0889	.1155	.139	.514
1	18	100	.0718	.0731	.0897	.093	.203
	41		.0747	.0828	.1050	.135	.371
	60		.0715	.0763	.0998	.113	.301
	83		.0678	.0697	.0768	.083	.146
2	18/28	100	.0771	.0802	.1045	.137	.452
	45/56		.0849	.1052	.1513	.232	.771
	73/83		.0758	.0834	.1016	.137	.411

Note: A is a fixed factor with $a = 2$ levels (groups from the same population), $\boldsymbol{B}$ is a random factor nested within A with the levels $b = 25$, 50, 100, 150, and 500 (testees) for each of the $a = 2$ levels, and C is a fixed factor with $c = 6$, 25, 30, 40, 60, and 100 levels (items). Estimations are based on 1,000 or 10,000 simulation replications of DIF-based data: Within the first group, Rasch model-based data were used with either a single- or a two-item DIF as compared to the second group's Rasch model-based data. There is no difference in the variation of the item parameters in both the groups. The item parameters are equally spaced within the interval [–3, 3], the ability parameters randomly distributed according to $N(0, 1.5)$.

power for the given effect size of a DIF of 0.6 is acceptable if there are 25 to 40 items, 2×500 testees, and a two-item DIF. According to Table 6, the situation is better – though no case of sample size 2×100 or smaller satisfies. On the other hand, a

Table 6. The power of the *F*-test for the interaction effect $A \times C$ in a three-way analysis of variance design $(A \succ \boldsymbol{B}) \times C$ with mixed classification (the nominal Type I risk is 5%).

number of DIF	localisation of DIF	c	p (*F*-test) $A \times C$				
			b				
			25	50	100	150	500
4	5/7/12/14	25	.1246	.2190	.4580	.670	1.000
	5/7/19/21		.0987	.1568	.3178	.478	.992
	12/14/19/21		.1282	.2222	.4607	.675	1.000
6	5/7/12/14/19/21	25	.1513	.3049	.6264	.850	1.000
4	6/9/14/17	30	.1220	.2041	.4187	.619	1.000
	6/9/22/25		.0991	.1459	.2855	.468	.989
	14/17/22/25		.1170	.2108	.4357	.661	.998
6	6/9/14/17/22/25	30	.1527	.2888	.5901	.796	1.000
4	7/11/18/23	40	.1086	.1794	.3637	.576	.998
	7/11/30/34		.0987	.1286	.2362	.375	.974
	18/23/30/34		.1124	.1699	.3504	.527	.999
6		40	.1264	.2358	.4827	.737	1.000
4	10/16/27/34	60	.1039	.1552	.2881	.418	.992
	10/16/45/51		.0815	.1117	.1903	.283	.944
	27/34/45/51		.0989	.1549	.2892	.445	.990
6	10/16/27/34/45/51	60	.1118	.1951	.3964	.638	1.000
4	18/28/45/56	100	.0934	.1292	.2276	.335	.963
	18/28/73/83		.0835	.1087	.1713	.240	.838
	45/56/73/83		.0968	.1241	.2316	.363	.972
6	18/28/45/56/73/83	100	.1034	.1664	.3234	.473	.997

Note: A is a fixed factor with $a = 2$ levels (groups from the same population), $\boldsymbol{B}$ is a random factor nested within A with the levels $b = 25$, 50, 100, 150, and 500 (testees) for each of the $a = 2$ levels, and C is a fixed factor with $c = 6$, 25, 30, 40, 60, and 100 levels (items). Estimations are based on 1,000 or 10,000 simulation replications of DIF-based data: Within the first group, Rasch model-based data were used with either two times or three times a two-item DIF as compared to the second group's Rasch model-based data. There is no difference in the variation of the item parameters in both the groups. The item parameters are equally spaced within the interval [–3, 3], the ability parameters randomly distributed according to $N(0, 1.5)$.

sample size of 2×500 is generally too large; there is no realistic Type II risk then (apart probably from a single case). Hence, interpolating the numerical results, a sample size of 2×200 or 2×250 promises to fit best – however, be aware that in order to hold the Type I risk, the sample size should not be larger than 2×150. As concerns the sample size 2×150, a three times two-item DIF of magnitude 0.6 will be discovered with an acceptable power of about .80, if the number of items is not larger than 40.

Discussion and conclusion

As already indicated, with our results we have to constrain the approach suggested by Kubinger et al. (2009) for testing the Rasch model. The approach is not only restricted to the case where no significant main effect in A (a partition of the total set of testees) results, but also to an item parameter interval of [–3, 3] and a (normal) distribution of the testees' ability parameter with a standard deviation not larger than 1.5. Then, the number of items should not be larger than 40, and, from the actual Type I error's point of view, the number of testees not larger than two times

150, from the power's point of view, not smaller than 150. However, the very last restriction is based on a single effect size of an item DIF of 0.6, and we did not investigate more than a three times two-item DIF, which would most likely increase the power of the approach.

While a larger DIF as well as a DIF involving more items could probably meet a researcher's aspirations for the power of this approach (hence further research is needed), the problem of the extremely raised actual Type I risk stated here is hardly understandable. Again, further research may disclose which combination of item parameter interval and ability parameter's standard deviation leads to analyses where the nominal Type I risk holds. At any rate, a peaked dispersion leads somehow to better results, which is, by the way, most likely the common case.

Maybe the problem is based on the fact that we have analysed only cases where the range of the ability parameters is larger than the range of the item parameters – as is well-known, for instance, for $N(0, 1)$, the ability parameter ranges from –3 to 3. Thereby, a lot of simulated testees will not master even the easiest items, and a lot of simulated testees will master every item; hence, any specific unlikely observation (an unlikely solution or an unlikely failure) might happen for one of the observed items in the one group, for another item in the other group, or to say: Chance effects with generally a very low probability become likely to polarize at least two items in both the groups. However, our selective results for [–3, 3] and N (0, 0.5) do not support this argumentation.

Nevertheless, the main aim of the approach, that is, to use a three-way analysis of variance design $(A \succ \boldsymbol{B}) \times C$ with mixed classification as a means for determining the sample size – according to a given Type I and Type II risk, and according to a certain effect of model misfit when testing the Rasch model – is to some extent attained. We now know in much more detail which sample size fits. Given that a DIF of 0.6 (a 10th of the item parameters' range) which occurs at least with respect to three couple of items is of relevance, for instance, a sample size of two times 150 will detect such an effect in the case of 30 items at a nominal Type I risk of 1% with a power of almost 80%. Of course, a table of various combinations of the number of items, the number of testees, the relevant DIF, the Type I risk, and the power within the stated frame of restrictions would be of further interest.

Notes

1. Because of a reviewer's suggestion, we give here a short introduction into analysis of variance (see for a deliberate presentation, e.g., Kubinger, Rasch, & Yanagida, 2011; Rasch, 1995; or Rasch, Kubinger, & Yanagida, 2011).
2. Random variables are printed in bold.

References

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

Glas, C.A.W., & Verhelst, N.D. (1995). Testing the Rasch model. In G.H. Fischer & I.W. Molenaar (Eds.) *Rasch models: Foundations, recent developments, and applications* (pp. 69–95). New York, NY: Springer.

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., Rasch, D., & Yanagida, T. (2009). On designing data-sampling for Rasch model calibrating an achievement test. *Psychology Science Quarterly*, *51*, 370–384.

Kubinger, K.D., Rasch, D., & Yanagida, T. (2011). *Statistik in der Psychologie – vom Einführungskurs bis zur Dissertation* [Statistics in Psychology – Introduction course up to doctoral thesis]. Göttingen, Germany: Hogrefe.

Mair, P., Hatzinger, R., & Maier, M. (2010). eRm: extended Rasch modeling. R package version 0.13-0 [Computer software]. Retrieved from http://cran.r-project.org/web/packages/eRm/

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, Germany: Pabst.

Rasch, D. (1995). *Mathematische Statistik*. Berlin, Germany: Wiley.

Rasch, D., Herrendörfer, G., Bock, J., Victor, N., & Guiard, V. (2008). *Verfahrensbibliothek Versuchsplanung und -auswertung. Elektronisches Buch* [Collection of procedures in design and analysis of experiments. Electronic book]. München, Germany: Oldenbourg.

Rasch, D., Kubinger, K.D., & Yanagida, T. (2011). *Statistics in psychology – Using R and SPSS*. Chichester, UK: Wiley.

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

R Development Core Team. (2008). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org.