Psychological Test and Assessment Modeling, Volume 59, 2017 (1), 93-108

Using Rasch model generalizations for taking testees' speed, in addition to their power, into account

Christine Hohensinn[1] *& Klaus D. Kubinger*[2]

Abstract

It is common practice in several achievement and intelligence tests to credit quick solutions with bonus points in order to gain more information about a testee's ability. However, using models of item response theory (IRT) for respective approaches is rather rare. Within IRT, the main question is whether speed and power do actually measure unidimensionally, that is, the same ability. In this paper, analyses were carried out in a sample of 9210 7th grade students, participants of an optional assessment, *Informal K[/C]ompetence Measurement* (IKM), within the programme of the Austrian Educational Standards. The following models were used: Rasch's multi-dimensional polytomous model as well as his unidimensional polytomous model (Rasch, 1961) (see also Fischer, 1974, and Kubinger, 1989); and Fischer's speed-and-power two-steps model (Fischer, 1973; see again also Kubinger, 1989), which has never been applied since its introduction. The first one modelizes speed and power in a joint measurement approach, meaning another ability/dimension is postulated for several combinations of power performance and speed performance. The unidimensional model additionally hypothesizes that the respective combinations, in other words "response categories", all refer to the same ability and differ only in a graded manner. Fischer's model considers speed and power as two completely independent abilities, for each of which the dichotomous Rasch model applies. Apart from model tests, information criteria are applied in order to reveal which model meets the best validness.

Keywords: Missing values, Rasch model, model fit, multicategorical IRT models, speed and power

[1] *Correspondence concerning this article should be addressed to:* Christine Hohensinn, Ph.D., Division of Psychological Assessment and Applied Psychometrics, Faculty of Psychology, University of Vienna, Liebiggasse 5, 1010 Vienna, Austria; email: christine.hohensinn@univie.ac.at

[2] Faculty of Psychology, University of Vienna, Austria

Introduction

In psychological assessment, the speed-power-issue has existed almost since the beginnings of psychological testing. Most intelligence tests have a time-limit, which is often only due to organizational reasons (to make test administration possible for a group of testees, for instance). Apart from this, some intelligence and achievement tests involve speed by scoring. For example, the commonly used Wechsler tests (e.g. Wechsler Adult Intelligence Scale – Fourth Edition, WAIS IV; Wechsler, 2008) include subtests that credit quick solutions with bonus points. The desirable advantage of such a scoring procedure is an attainment of information about a testee's ability. The scores of the testees are more differentiated and thus, measurement would take place in a more precise way. Of course, the advantage of such information is only valid if the assumptions underlying the scoring procedure are correct. Particularly, using bonus points assumes that "power" and "speed" are confounded but not separated traits. This means the ability to solve an item and speededness of a testee in finding the solution are assumed to be a manifestation of the same latent trait and reflect only gradual differences in the measured trait. This assumption is to be scrutinized, as there are empirical results that show speed and power are actually separated (Carroll, 1993; Partchev, De Boeck, & Steyer, 2011). Partchev, De Boeck, and Steyer (2011) even remark that the current focus should be on avoiding a mixture between speed and power. However, in practice, scoring procedures combining speed and power do still exist, which makes the application of methods which test the validness of such a scoring procedure important.

Nowadays, it is easy to record response times for each item and there are increasing attempts to use these response times not only in psychological but also in educational assessment. Large-scale tests that have been applied for years, are currently going through a transition from paper-based administration to a computer-based one. In the first part of the special topic "Current Methodological Issues in Educational Large-Scale Assessments" by Stadler, Greiff, and Krolak-Schwerdt (2016) in this journal Bürger, Krohne, and Goldhammer (2016) give a short overview which of the broad-based international large-scale assessments have already changed their administration mode or are planning to change it in the near future. This transition to a computer-based administration provides the opportunity to record not only the response of the examinee but many other variables including the item-specific response time. Response times provide an additional information about how the testee did work on the test. So far, there are many attempts to use response times to increase the measurement accuracy and to minimize measurement errors in psychological and educational assessment. For example, a variety of studies dealt with the detection of guessing in multiple-choice items by using the response times (DeMars, 2007, 2010; Kong, Wise, & Bhola, 2007; Schnipke, & Scrams, 2007, Wise, Pastor, & Kong, 2009). Weeks, von Davier, and Yamamoto (2016) are using response times to distinguish between missing responses which were skipped and those the testee had tried on which but did not give a response.

Another option to use the information of response times in large-scale assessment is to incorporate the response times into scoring; that is analogous to intelligence tests which use some credit points for quick solutions. This approach is thought of as a means of increasing measurement accuracy which is of need especially in large-scale assessments where the number of items is limited due to organizational restrictions.

A variety of approaches were introduced for incorporating response times in assessments. Van der Linden (2011) gives a fine overview of actual IRT methods modeling response times. He distinguishes between models that include the distributions of response times without any reference to the quality of the item response, and models that integrate item responses and response times (e.g. Verhelst, Verstralen, and Jansen (1997) and also Roskam (1997) introduce IRT models for time-limited tests). Van der Linden (2007) derived a general model that is composed by a first submodel for taking the responses into account and a second one that takes the response time distribution into account.

All proposed models have in common that they primarily aim to calibrate a test. In contrast, for the present study, we search for methods to evaluate the validness of given scoring procedures and, additionally, to find out more about the relationship between speed and power in a given test. As described above, many sorts of relationships between these two aspects measured in psychological tests are possible: speed and power as separate latent traits, speed and power that are confounded in such a way that quick responses are to be qualitatively distinguished from slow item responses, and speed and power as joint constructs that only reflect graduations of one and the same latent trait.

For the evaluation whether a scoring procedure with bonus points is appropriate, we applied two approaches that are based on IRT methods. Both of them are described in the following.

Multi-dimensional polytomous Rasch model / Unidimensional polytomous Rasch model

The multi-dimensional polytomous Rasch model as a simple generalization of the well-known, that is dichotomous Rasch model (Rasch, 1960/1980; see also Fischer, 1974, and Kubinger, 2005) was introduced by Rasch (1961). It has a completely different model formulation than the multi-dimensional IRT (MIRT) models that are applied very often nowadays (a detailed overview of MIRT models is given in Reckase, 2009). The multi-dimensional polytomous Rasch model handles k items, each of them having the same m > 2 response categories. It assumes a different latent trait ξ_h for each category h. That is, a person parameter vector $\boldsymbol{\xi}_v = (\xi_{v1,} \xi_{v2,} \ldots, \xi_{vm})$ results for each person v. Needless to say, the number of latent dimensions is equal to the number of response categories. With regard to the item parameters, a separate parameter is postulated for each item and each category, resulting again in a vector of length m for each item i: $\boldsymbol{\sigma}_i = (\sigma_{i1}, \sigma_{i2,} \ldots, \sigma_{im})$. Hereinafter, σ_{ih} is referred to as "item category parameter". That is, the multi-dimensional polytomous Rasch model defines the probability that a person v responds at item i with category h as a logistic function of the item category parameter σ_{ih} and the person parameter ξ_{vh} (Fischer, 1974):

$$P(X_{vi} = h|\xi_{vh}, \sigma_{ih}) = \frac{\exp\left(\xi_{vh} + \sigma_{ih}\right)}{\sum_{j=1}^{m} \exp\left(\xi_{vj} + \sigma_{ij}\right)} \tag{1}$$

The following constraints have to be set (Fischer, 1974):

$$\sum_{h=1}^{m} \xi_{vh} = 0$$

$$\sigma_{im} = 0, i = 1, \ldots, k$$

$$\sum_{i=1}^{k} \sigma_{ih} = 0 \tag{2}$$

It can easily be seen that in the case of $m = 2$, the multi-dimensional polytomous Rasch model reduces to the dichotomous Rasch model. At any rate, it retains some important properties well-known from that Rasch model: there are sufficient statistics for each type of parameter. As a consequence, a conditional likelihood of the data can be derived and this can be used for parameter estimation. Furthermore, Andersen's Likelihood Ratio Test (Andersen, 1973) can be applied. This model test refers to the model concept of specific objective comparisons, which entails that parameter estimations have to be statistically the same, independent of the selected sub-sample of testees (or items). As a consequence, the likelihood of the total sample is to be compared with the combined likelihood of any sub-samples. If the difference between these likelihoods is negligibly small, then the model holds. Apart from this test of model validness there obvioulsy are some means of evaluating the goodness of model fit; we only refer to Rasch's graphical model check which corresponds to the concept of specific objective comparisons as well. It opposes the item parameter estimations of two arbitrarily chosen sub-samples of testees in a graphical manner.

Moreover, Fischer (1974) deals with a statistical test as to whether the multi-dimensional polytomous Rasch model can be reduced to a unidimensional polytomous model. In this case, the item category parameters are linearly dependent: $\sigma_{ih} = \varphi_h \cdot \sigma_i$. The response categories are no longer manifestations of different latent traits, but rather graduations of the same latent trait ξ. For each item i, only a single item parameter σ_i is estimated and for each response category h, an item-independent scoring parameter φ_h is estimated, which reflects the graduations between the categories. Due to normalization constraints, the scoring parameter of the lowest category is set to 0 and that of the highest category to 1 (Fischer, 1974).

This reduction condition offers the opportunity to determine whether an assumed scoring function is appropriate for a psychological test or not. If, in fact, quick solutions only mean a quantitatively higher but not a qualitatively different ability, then the reduction condition would hold for a psychological test. Furthermore, if an extra point for quick solutions is due to the factual graduation of a higher ability, then the category parameters would result to 0 for no solution, 0.5 for a slow solution and 1 for a quick solution.

Speed-power-two-steps model

Fischer (1973) proposes a composite Rasch model: first, the ability ("power") is scaled by a dichotomous Rasch model and secondly, a Rasch model measuring the speed of a testee, given that the item was solved, is applied. This results in the following model equations (cf. Kubinger, 1989):

$$P(X_{vi}=2|\xi_v,\pi_v;\sigma_i,\kappa_i)=\frac{\exp(\xi_v-\sigma_i)}{1+\exp(\xi_v-\sigma_i)}\bullet\frac{\exp(\pi_v-\kappa_i)}{1+\exp(\pi_v-\kappa_i)}$$

$$P(X_{vi}=1|\xi_v,\pi_v;\sigma_i,\kappa_i)=\frac{\exp(\xi_v-\sigma_i)}{1+\exp(\xi_v-\sigma_i)}\bullet\frac{1}{1+\exp(\pi_v-\kappa_i)}$$

$$P(X_{vi}=0|\xi_v,\pi_v;\sigma_i,\kappa_i)=\frac{1}{1+\exp(\xi_v-\sigma_i)} \tag{3}$$

where the score $X_{vi}=2$ represents a correct and fast response, $X_{vi}=1$ a correct but slow answer and $X_{vi}=0$ an incorrect item response. The "speed" part of the formula contains a second person parameter π_v that reflects the "speededness“ of person *v* and a second item parameter κ_i that represents the time requirement of item *i*.

Whereas the speed-power-two-steps model was derived specifically for this purpose, the multi-dimensional / unidimensional polytomous Rasch model is more general but can be applied as well. Both approaches have very specific fundamental concepts: the polytomous Rasch model confounds power and speed, but nevertheless separates the responses into quick and slow correct responses (as well as incorrect ones). In contrast, the speed-power-two-steps model has the goal to yield separated measures of ability and speededness of a person. We try to compare both approaches in the following.

Method

Sample and material

For the present study, data from the *Informal K[/C]ompetence Measurement (IKM)* were used. IKM is a large-scale test that can be administered by school teachers in all Austrian schools on a voluntary basis. The tests were administered online in the computer laboratories of the schools. Due to organizational reasons, the test was time-limited.

The sample consisted of *n* = 9210 7th grade students from all districts of Austria, who worked on the mathematical test of IKM in 2010. 31.4 % of the sample were male, 32.9 % female and the remaining 35.7 % did not provide information regarding their sex. 53.8 % of the sample reported German as their mother tongue, 10.5 % stated a non-German mother tongue and 35.7 % omitted this question. Because IKM was constructed as a self-evaluating tool for teachers, the only socio-demographic variables which were gathered were students' sex and their mother-tongue. However, as a matter of fact the number of missing values for both of them is rather high. Response patterns and response times reveal that students who denied to deliver the asked sociodemographic information worked nevertheless with sufficient test motivation (that is they operated the items with a reasonable long time and gave, at least partly, correct responses). Therefore it can be assumed, that these students had just tried to save their anonymity: It was easy to omit the respective questions. Coming to the point: there is no evidence at all that the achievement motivation was peculiarly low.

The mathematical test consists of four subtests: (1) Modeling (7 items), (2) Calculating and Operating (8 Items), (3) Interpreting (7 Items), (4) Explaining (8 items). The items either have a multiple-choice or open-response format. In total, the test consists of 30 items which were administered in two versions – differing only in the sequence of the items to avoid the testees copying answers. Computer-based administration allowed the response time (in seconds) for each item to be recorded.

Analyses and results

According to the data's check of meaningfulness, students with less than 10 item responses were excluded. The remaining n = 9066 students were taken for further analyses. Exploration of the response times of the missing values indicated that the reason for not answering an item was either due to a student's omission or due to the fact, that he/she did not reach the items at the end.

Taking the students' response times for solving an item into account, the approaches described in the introduction were compared. First, for both approaches the categorically given item response times had to be polarized into quick and slow solutions. Therefore, a cut-off point had to be determined. As the distribution of response times resulted as item-specific in particular due to the amount of text which has to be read by the testees, the cut-off points were based empirically, that is for each item individually according to the median of the response times of correct given responses.

According to this polarization, the data was recoded in the following ways.

1) multi-dimensional polytomous Rasch model: For the estimation of the category specific person parameters, ξ_{vh}, another data matrix $\mathbf{X}_m$ was constituted: quickly solved (x_{vi} = 0), slowly solved (x_{vi} = 1) and not solved (x_{vi} = 2). Labeling the three response categories in a different way would have been possible of course, for instance the other way round.

2) speed-power-two-steps-model: For the estimation of the first person parameters, power ξ_v, a data matrix $\mathbf{X}$ was constituted which discloses whether student v has solved item i (x_{vi} = 1) or not (x_{vi} = 0). For the estimation of the second person parameters $\pi_{\mathrm{v},}$ which describe the students' speed (that is, whether item solutions are given rather quickly or rather slowly), a second data matrix $\mathbf{X}_\mathrm{T}$ was constituted; this one discloses whether a student v has solved item i quickly (x_{vi} = 1) or slowly (x_{vi} = 0).

The speed-power-two-steps model was analyzed sequentially: the item parameters for the dichotomous Rasch model were estimated for each of the two data matrices, $\mathbf{X}$ and

$\mathbf{X}_\mathrm{T}$. To get the likelihood of the model as a whole, the likelihoods of the two models were multiplied. The Rasch model parameters and the data's likelihood were estimated with the eRm package (Mair, Hatzinger, & Maier, 2015). For the multi-dimensional polytomous Rasch models, the pcIRT package (Hohensinn, 2015) was used; both software packages are included in R (R Core Team, 2015; for the analyses R version 3.1.3 was used).

As described above, data matrix $\mathbf{X}_\mathrm{T}$ only refers to the correct responses; the incorrect ones are treated as missing values. Students with only one or no correct response had to be excluded. The sample sizes after respective deletion were as follows: for subtest 1 n = 8165, for subtest 2 n = 8678, for subtest 3 n = 7601, and for subtest 4 n = 7363. These reduced data sets were used for all further analyses (for the multi-dimensional and the unidimensional polytomous Rasch model, too).

The first research question deals with the appropriateness of the unidimensional polytomous Rasch model. However, it is to be tested in advance whether the multi-dimensional polytomous model holds at all. Finally, even if the unidimensional model suffices, the question is whether the category parameters, φ_1, φ_2, φ_3, for a quick solution (0), a slow solution (1), and no solution (2) are actually equidistantly scaled: Obviously, an extra point is only justified if this is true, but of course the three parameters could result completely differently, that is non-equidistantly.

First, the multi-dimensional polytomous IRT model was applied. As pointed out above, Andersen's Likelihood Ratio test can be used for testing this model. However, for the interpretation of the numerical result, the sample size has to be taken into account. Recently, the discussion arose that the application of pertinent model tests within IRT traditionally suffers from controlling for the type-II-risk. Planning a study statistically always means determining the sample size according to a given type-I- and type-II-risk, and according to a certain effect which is of practical relevance (cf. e.g. Rasch, Kubinger, & Yanagida, 2011). However, this approach hardly applies within IRT analyses (but see Kubinger, Rasch, & Yanagida, 2009, 2011, and Yanagida, Kubinger, & Rasch, 2015, as well as Draxler, 2010, Draxler & Alexandrowicz, 2015; and finally Draxler & Kubinger, 2017, in print): as a consequence of determining the sample size arbitrarily, particularly IRT applications regularly result in significance, although the effects are rather minimal, or of an irrelevant extent – and this also implies a rather senseless type-II-risk of almost zero. For this reason, our applied model tests based on more than 7000 students were expected to result in significance anyway. Therefore, rather descriptive measures are of interest, most notably, Rasch's graphical model check. This is true concerning the multi-dimensional polytomous Rasch model as well as the other applied models.

The graphical model check of the multi-dimensional polytomous Rasch model opposes the item category parameters estimated for two sub-samples, that is, students with a score greater than the median and students with a score lower or equal to the median of the raw score. The sum of solved items of the original data matrix was used as a raw score – that is, no scoring with bonus points was applied at this point. Andersen's Likelihood-Ratio test resulted as follows: Subtest 1: χ^2 = 222.18, df = 12, p = .000; Subtest 2: χ^2 = 205.66, df = 14, p = .000; Subtest 3: χ^2 = 243. 80, df = 12, p = .000; Subtest 4: χ^2 = 153.34, df =

14, p = .000. As the graphical model tests look relatively good – ideally, every dot would lie on the 45-degree line – we decided for a model fit (see the results in Figures 1a to 1d).

Secondly, for testing the unidimensional Rasch model, we also applied a graphical model check. This time the multi-dimensional category parameter estimations were opposed to the restricted but unidimensional category parameter estimations. The results of the respective Likelihood-Ratio test (cf. Andersen, 1980) are as follows: Subtest 1: χ^2 = 409.44, df = 5, p = .000; Subtest 2: χ^2 = 65.42, df = 6, p = .000; Subtest 3: χ^2 = 260.04, df = 5, p = .000; Subtest 4: χ^2 = 188.46, df = 6, p = .000. In addition, Table 1 presents the

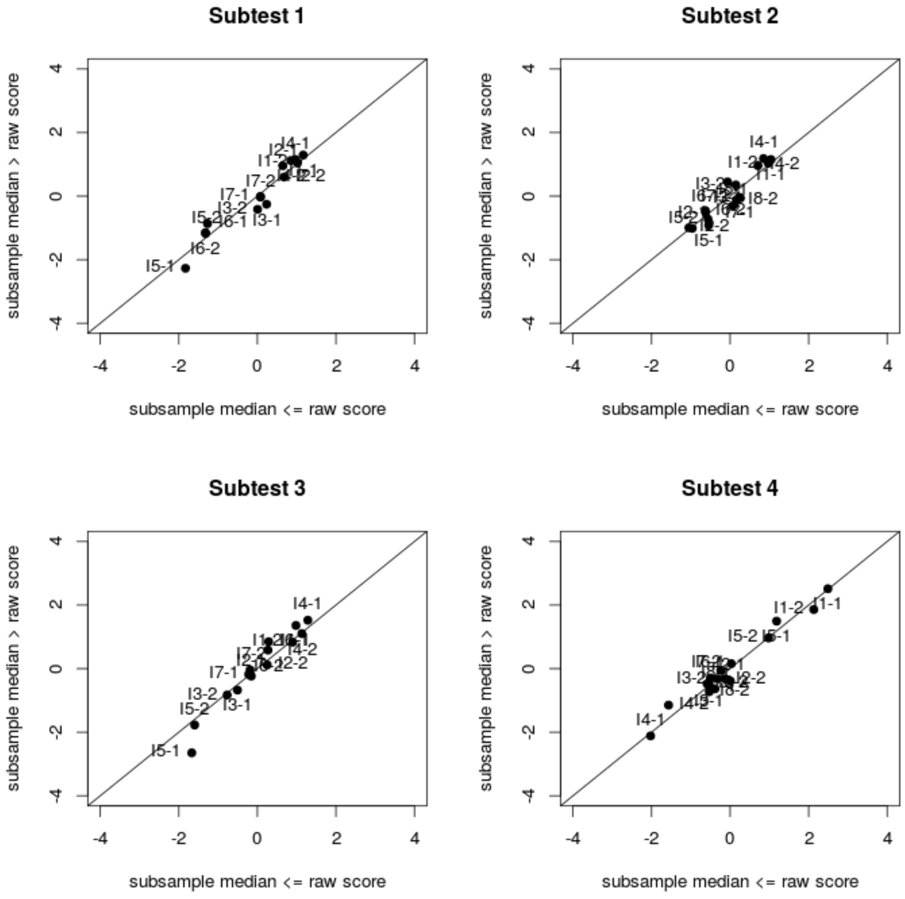

Figure 1 (a to d):
Graphical model checks for the multi-dimensional polytomous Rasch model for each subtest. The first number of the labels indicate the item number, the second number the category (1 for quick solution, 2 for slow solution). For example „I5-2“ means the item category parameter for item number 5 and category 2 (slow solution).

Akaike information criterion (AIC) and the Bayesian information criterion (BIC) for the multi-dimensional and the unidimensional model. Both information criteria are indices that provide a relative comparison of goodness-of-fit of different models. From the results, it is shown that except for Subtest 2, the unidimensional model has a lower fit than the multi-dimensional one. However, the graphical model checks (Figure 2 a to d) offer a good concordance of the unidimensional item category parameter estimates to those of the multi-dimensional model. As described in the introduction, the item category parameters of the unidimensional model are estimated by $\sigma_{ih} = \varphi_{\mathrm{h}} \cdot \sigma_i$. For each subtest, one scoring parameter φ_1 was estimated ($\varphi_0 = 0$ and $\varphi_2 = 1$ due to normalization constraints). The estimated scoring parameters for the four subtests are as follows: $\hat{\varphi}^{(1)} = 0.81$, $\hat{\varphi}^{(2)} = 0.89$, $\hat{\varphi}^{(3)} = 0.77$, $\hat{\varphi}^{(4)} = 0.79$.

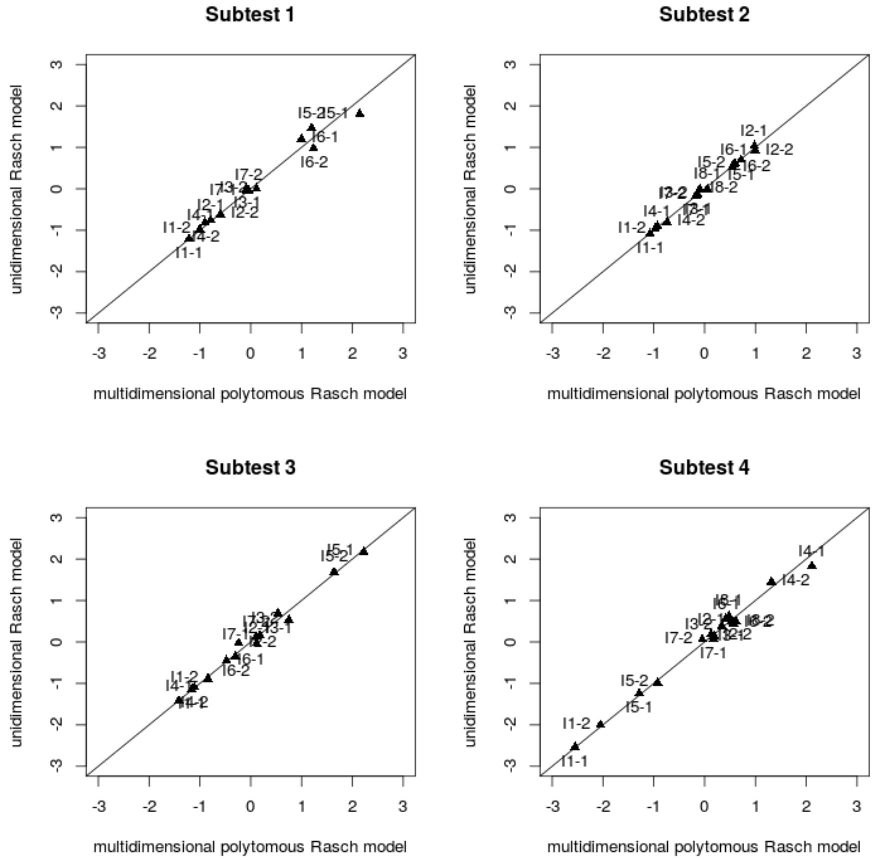

Figure 2 (a to d):
Estimated item category parameters according to the multi-dimensional and the unidimensional Rasch model. The labels are the same as in Figure 1.

Thirdly, another graphical model check was used to test the hypothesis that a quick solution actually credits an additional point, that is, instead of 1 point, 2 points. As already indicated, if this scoring reflects the empirical difference between a quick and a slow solution, the item category parameters for a quick solution must have a ratio of 2:1 to the item category parameters for a slow solution. In this case the unidimensional item-category parameters for a slow solution opposed to those for a quick solution should lie on a line with a slope of 1:2, i.e. 0.5. See the results in Figures 3a to 3d.

Apparently the item category parameters of the four subtests are quite compatible with each other. With only a few exceptions (item 5 in Subtest 1 or item 4 in Subtest 4), the

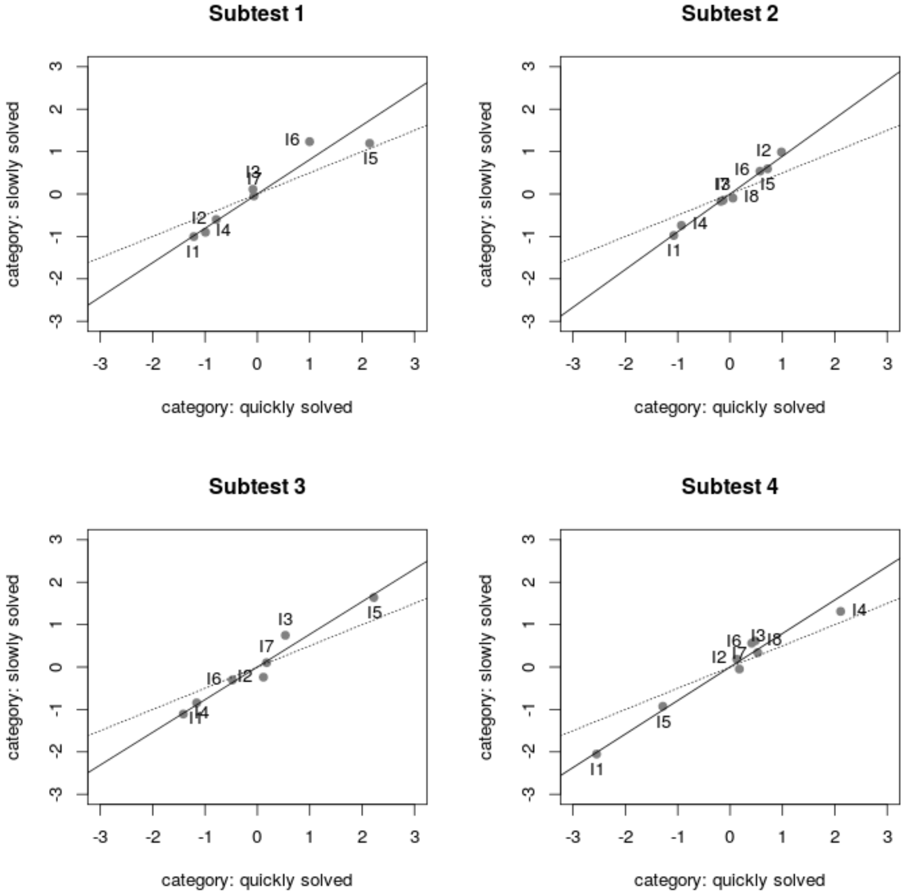

Figure 3 (a to d):
Estimated parameters of categories „quickly solved“ and „slowly solved“ for the items of each subtest. In each plot, the dotted line represents the implicated slope of 0.5, the dashed line discloses the empirically resulted scoring parameter.

relation between the item category parameters of the categories "quickly" and "slowly solved" is the same for all items. As a matter of fact, the estimated scoring parameters come close to 1 which means that both categories, "slow" and "fast solution", are almost equally difficult. As a consequence, a differentiation between quick and slow solutions seems to not be very important.

As described above, the speed-and-power-two-steps model is composed of two Rasch models, one for power and the other for speed, given the power (see Equation 3). The speed-and-power-two-steps model was tested in the same way as the multi-dimensional polytomous Rasch model, that is, by applying Andersen's Likelihood Ratio test using the

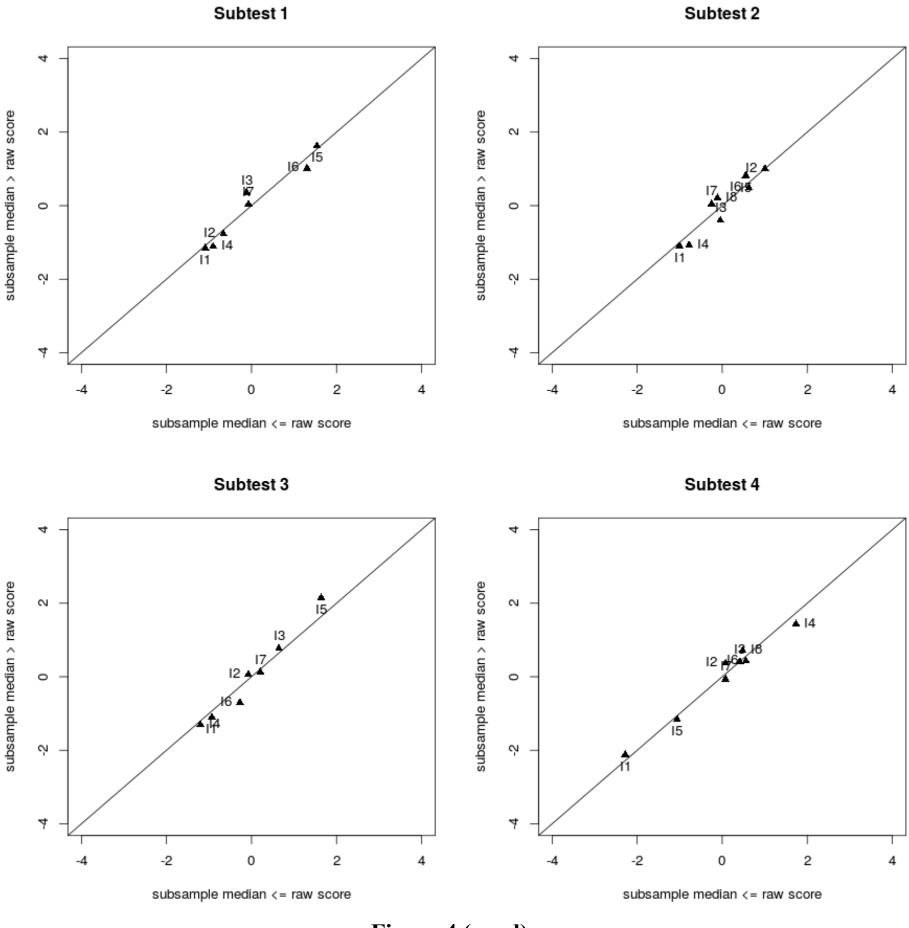

Figure 4 (a – d):
Graphical model checks for the power component of the speed-and-power-two-steps model for each subtest.

common partition criterion “score”; in addition, the graphical model check was applied. For Andersen’s Likelihood Ratio test, the likelihoods of the two Rasch models (power and speed) were combined. The results are as follows: $\chi^2 = 224.71$, $df = 10$, $p = .000$; $\chi^2 = 231.77$, $df = 12$, $p = .000$; $\chi^2 = 237.94$, $df = 10$, $p = .000$; $\chi^2 = 131.90$, $df = 12$, $p = .000$. For the graphical model checks, the item parameters for the subsamples high scorers and low scorers were opposed in a plot. Because the item parameters were estimated separately for the speed and power, the graphical model checks are displayed in two different figures as well (see Figures 4a to 4d and 5a to 5d). The plots offer good model fits. For

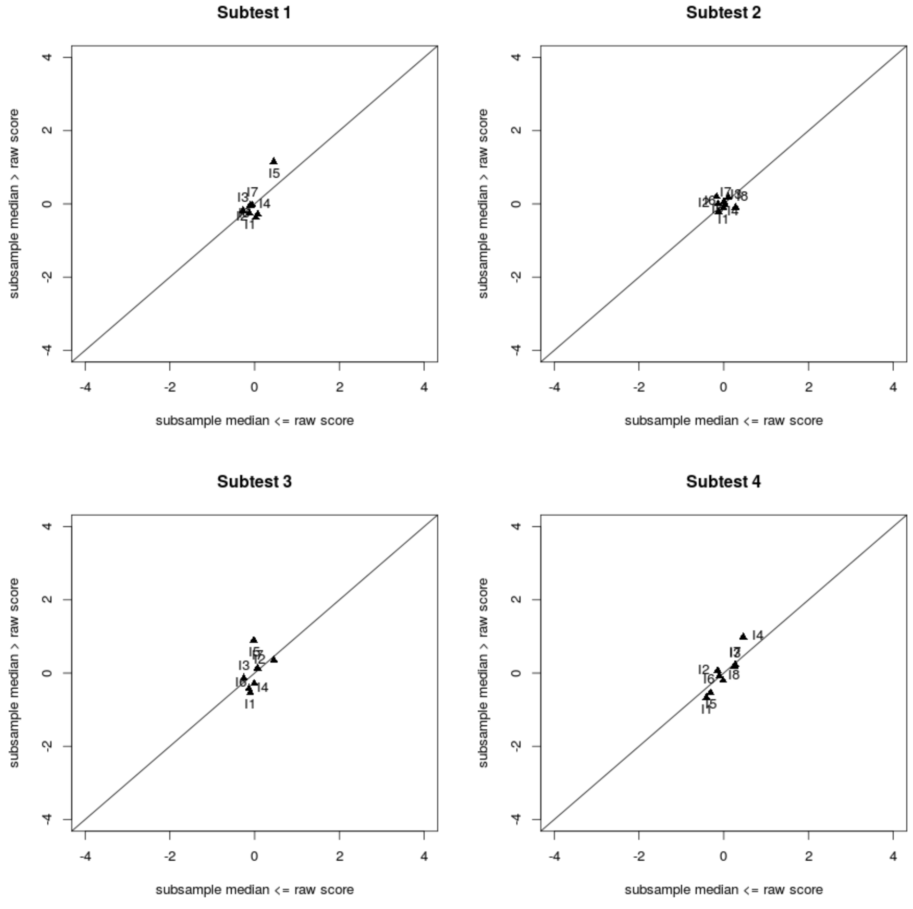

Figure 5 (a - d):
Graphical model checks for the speed component of the speed-and-power-two-steps model for each subtest.

speed, the item parameter estimations evidently only show a small variance. This indicates that there is almost no difference between a fast and a slow solution for every item.

The second research question aims to compare the two approaches chosen here when speed of solving an item is incorporated into scoring, that is, the polytomous Rasch models on the one side and the speed-and-power-two-steps model on the other side. As described above, the data sets for these models are different which makes a model comparison very difficult. Neither a likelihood ratio test nor the information criteria AIC and BIC can be applied in this context.

However, to get at least some descriptive impression of the fit of these two approaches, the probabilities of the given item responses were calculated – one time given the multi-dimensional polytomous Rasch model, the other time given the speed-and-power-two-steps model. The product of the item response probabilities, which is the likelihood of the given response vector, was calculated for each person *v*. To get a rough comparison of the two models, we compared the likelihoods of the two response vectors for each person *v*. The response vector likelihood in Subtest 1, given the multi-dimensional poly-

Table 1:
-2·Loglikelihood, the number of estimated parameters (n_p), AIC and BIC subtest-wise for the two models.

		-2· LogLikelihood	n_p	AIC	BIC
Subtest 1	multi-dimensional Rasch model	59545.7	12	59569.7	59653.8
	uni-dimensional Rasch model	59955.2	7	59969.2	60018.2
Subtest 2	multi-dimensional Rasch model	81709.3	14	81737.3	81836.3
	uni-dimensional Rasch model	81774.7	8	81790.7	81847.3
Subtest 3	multi-dimensional Rasch model	54917.9	12	54941.9	55025.1
	uni-dimensional Rasch model	55177.9	7	55191.9	55240.5
Subtest 4	multi-dimensional Rasch model	57981.7	14	58009.7	58106.3
	uni-dimensional Rasch model	58170.1	8	58186.1	58241.4

tomous Rasch model, was in 93.42 percent of the times greater than that given the speed-power-steps model. For Subtest 2, in 88.23 percent of the observations it was greater; for Subtest 3, 91.33 percent; and for Subtest 4, 94.06 percent.

Discussion

In the present study two IRT-based approaches, which incorporate the item solutions' speed into scoring, were compared. Before our research, there were some empirical results that suggest to take power and speed as separate traits. The speed-power-two-steps model is based on this assumption and models speed as a power independent, separate trait. Nevertheless, our results show that the speed-power-two-steps model seems to have a lower model fit compared to the multi-dimensional polytomous Rasch model. Of course, a direct comparison of the two approaches is difficult and only a descriptive impression using a comparison of the likelihoods of the item response vectors could be used. However, estimation of the item category parameters shows that it is approximately equally difficult to solve an item quickly as it is to solve it slowly. Moreover, a reduction of the multi-dimensional polytomous Rasch model to a unidimensional one is (according to the information criteria) not appropriate, though we decided at first (due to the graphical model check) that even this model holds: the multi-dimensional category parameter estimations of "quickly" vs. "slowly" solved differed too much in the end with respect to the items that they could not be explained by a linear function over all items. Incidentally, the scoring parameters are not at all 0.5 as indicated by the apodictically chosen scoring rule. This shows, at least, that the scoring of 2 points for quick correct responses and 1 point for slow correct responses does not at all reflect actual performances.

Of course, our results only refer to the particular given tests. There might be other tests for which our results do not hold. However, this does not necessarily mean that any scoring rule which provides bonus points for speed is not to be justified according to our analyses.

There are further limitations of this study: first, the categorization into quick and slow solutions was set by the median of the response times. We repeated the analysis by setting the cut-off point between quick and slow solutions in the first quartile of the response times. This means that a quick solution discloses a more "excellent" achievement than before. The results only differ marginally, the overall conclusions of the study remain the same. This indicates that the results hardly depend on the cut-off point of speed categorization. Another limitation is that the majority of the items in the mathematical tests use a multiple-choice format. For this response format, the occurrence of rapid lucky guessing is well-known: students give a very quick item response without really "solving" the item. Of course, these effects could have taken place as well and could have led to a distortion of the distribution of the response times. Excluding students that probably solved items with lucky guessing could have led to a shift in the cut-off point of the response time. Again, the results of our second analysis with the cut-off point set at the first quartile should be taken with great caution as an indicator that a shift in the cut-off point should not have led to dramatically different results.

References

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

Bürger, S., Kröhne, U., & Goldhammer, F. (2016). The Transition to Computer-Based Testing in Large-Scale Assessments: Investigating (Partial) Measurement Invariance between Modes. *Psychological Test and Assessment Modeling, 58,* 587-606.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytical studies.* New York, NY: Cambridge University Press.

Fischer, G.H. (1974). *Einführung in die Theorie psychologischer Tests.* [Introduction into theory of psychological tests]. Bern: Huber.

Fischer, G. H. (1973). Ein Beitrag zum Speed-Power-Problem [A contribution to the speed and power problem]. In G. Reinert (Ed.), *Bericht über den 27. Kongreß der Deutschen Gesellschaft für Psychologie, Kiel 1970* [Proceedings of the 27th congress of the German association of psychology, Kiel 1970] (pp. 389-404). Göttingen: Hogrefe.

DeMars, C. E. (2007). Changes in rapid-guessing behavior over a series of assessments. *Educational Assessment, 12,* 23-45.

DeMars, C. E., & Wise, S. L. (2010). Can differential rapid-guessing behavior lead to differential item functioning? *International Journal of Testing, 10,* 207-229.

Draxler C. (2010). Sample size determination for Rasch model tests. *Psychometrika*, 75, 708–724.

Draxler, C. & Alexandrowicz, R. W. (2015). Sample size determination within the scope of conditional maximum likelihood estimation with special focus on testing the Rasch model. *Psychometrika*, 80, 897–919.

Draxler, C. & Kubinger, K.D. (2016, in print). Power and sample size considerations in psychometrics. D. Rasch et al. (eds.), *Proceeding of the 8th International Workshop on Simulation, Vienna 2015.* Springer.

Hohensinn, C. (2015). pcIRT: IRT models for polytomous and continuous item responses. R package version 0.2.1.

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement, 67,* 606-619.

Kubinger, K.D. (1989). Aktueller Stand und kritische Würdigung der Probabilistischen Testtheorie [Critical evaluation of latent trait theory]. In K.D. Kubinger (Ed.), *Moderne Testtheorie - Ein Abriß samt neuesten Beiträgen* [Modern psychometrics – A brief survey with recent contributions] (pp. 19-83). Munich: PVU.

Kubinger, K.D. (2005). Psychological Test Calibration using the Rasch Model - Some Critical Suggestions on Traditional Approaches. *International Journal of Testing, 5,* 377-394.

Kubinger, K.D., 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). A new approach for testing the Rasch model. *Educational Research and Evaluation, 17,* 321-333.

Mair, P., Hatzinger, R., & Maier M. J. (2015). eRm: Extended Rasch Modeling. 0.15-5.

R Core Team (2015). *A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

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

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press.

Rasch, G. (1961). On general laws and the Meaning of Measurement in Psychology. In: *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability.,* 4, pp. 321-333.

Reckase, M. D. (2009). *Multi-dimensional item response theory.* New York, NY: Springer.

Roskam, E. E. (1997). Models for speed and time-limit tests. In W. J. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 187-2008). New York, NY: Springer.

Stadler, M., Greiff, S., & Krolak-Schwerdt, S. (2016). Current Methodological Issues in Educational Large-Scale Assessments – Guest Editorial. *Psychological Test and Assessment Modeling, 58,* 593-595.

Van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72,* 287-308.

Van der Linden, W. J. (2011). Modeling response times with latent variables: Principles and applications. *Psychological Test and Assessment Modeling, 53,* 334-358.

Verhelst, N., Verstralen, H., & Jansen, M. (1997). Models for time-limit tests. In W. J. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 169-185). New York, NY: Springer.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale – Fourth Edition*. San Antonio: Pearson.

Weeks, J.P., von Davier, M., & Yamamoto, K. (2016). Using Response Time Data to Inform the Coding of Omitted Responses. *Psychological Test and Assessment Modeling, 58,* 661-689.

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: implications for test development and measurement practice. *Applied Measurement in Education, 22,* 185-205.

Yanagida, T., Kubinger, K. D., & Rasch, D. (2015). Planning a study for testing the Rasch model given missing values due to the use of test-booklets. *Journal of Applied Measurement*, 16, 432–444.