Psychological Test and Assessment Modeling, Volume 59, 2017 (3), 269-295

# Detecting unmotivated individuals with a new model-selection approach for Rasch models

*Jochen Ranger*[1] *& Jörg-Tobias Kuhn*[2]

**Abstract**

In low-stakes tests some test takers do not work with high motivation but respond carelessly. This has serious consequences for item response models as careless responses impair model calibration and trait inference. In this manuscript we describe an approach to data analysis that reduces the negative implications of careless responding and allows for the identification of the poorly motivated test takers. The approach has been inspired by the Rasch model answer tree (also Rasch tree) suggested by Strobl, Kopf, and Zeileis (2015). The Rasch model answer tree subdivides the sample into several strata in a data driven way by means of significance tests and fits a distinct Rasch model to each stratum. In our new approach we build on this idea of partitioning the data into strata via a sequence of splits. Contrary to this approach we determine multi-group Rasch models by enforcing theoretically motivated configurations of hierarchical data splits and select among the configurations via information criteria. By using the response times of the test takers for partitioning, a stratum can be isolated that contains the motivated test takers and allows for unbiased model calibration. The performance of this new approach with respect to parameter recovery and the detection of unmotivated test takers are compared to alternative models for low-stakes tests in a simulation study, namely the latent class model of Meyer (2010) and a finite mixture model for the response times. The simulation study demonstrates that the new approach reduces the bias caused by low motivation under certain circumstances. An empirical application underscores the usefulness of our suggestion.

Key words: response time, rapid guessing, Rasch model answer tree

[1] *Correspondence concerning this article should be addressed to:* Jochen Ranger, PhD, Department of Psychology, Martin-Luther-University Halle-Wittenberg, Emil-Abderhalden-Str. 26-27, 06108 Halle (Saale), Germany; email: jochen.ranger@psych.uni-halle.de

[2] University of Münster

## 1 Introduction

Psychological tests in general and achievement tests in particular are regularly used for psychological assessment, the evaluation of educational institutions and psychological research. The responses of the test takers to the test items are usually analysed with an item response model. Item response models are based on the assumption that the observed responses are governed by a specific trait, which is latent and can not be observed directly. Item response models specify the relation between the observable responses and the latent trait. This relation can then be used for inferring the trait level of each test taker from his/her responses given in the test. The process of trait inference crucially depends on how well the item response model is able to represent the relation between the trait and its manifestations. Valid inference requires the choice of an adequate item response model as well as precise estimates of the model's parameters. Trait inference can seriously be wrong in case the parameter estimates are biased and deviate sharply from the true values.

With the exception of test applications in psychological or educational assessment, the test results usually do not have major personal consequences for the test takers. Such a situation is called low-stakes testing. As there is no extrinsic reward for high test scores, some test takers have little motivation to perform as well as they could, especially when there is little freedom to choose whether to take the test or not. Unmotivated test takers make little effort to regularly solve the test items, but rely on approximations and short cut strategies to avoid as much mental effort as possible. One extreme form is carelessly responding, where answers are given fast, without any serious engagement in active problem solving. It is well known that carelessly responding distorts the estimation of the item parameters (Bolt, Cohen, & Wollack, 2002; Oshima, 1994; Schnipke, 1999) and undermines score validity (Wise & DeMars, 2006; Wise & Kong, 2005). Items for example appear more difficult and the item discrimination is reduced. This can have serious consequences for psychological assessment. Hence, it would be beneficial to identify unmotivated individuals and to reduce their distorting effect on model calibration. Several methods have been suggested for this purpose so far.

A first approach to handle careless responding relies on discrete mixtures of item response models (Rost, 1990). The approach is based on the assumption that the test takers can be divided into several classes with respect to their mode of responding. In the simplest case, just two classes are assumed: A first class, which consists of the responders that respond always in a regular way and a second class, which consists of the responders that respond irregularly by using short cut strategies in at least a subset of the items. The mixture item response models account for this structure of the data by allowing for different item response models in the two subgroups of test takers; see Bolt et al. (2002) for an application to low-stakes tests. Alternatively, instead of dividing the test takers into just two classes, one can assume one class of regular responders and several subclasses of irregular responders defined by the position in the test where the subjects start to respond carelessly. Although originally these models were supposed for test speededness and the effect of running out of time (Yamamoto & Everson, 1997), they can also be used for situations where individuals lose their motivation during the test, for example

when items become too difficult (Cao & Stokes, 2008). As mixture item response models allow for different parameter values in the different latent classes, they are able to recover the item parameters in the class of the regular responders and can be used for segregating the test takers according to their mode of responding. Mixture item response models however are notoriously hard to estimate and require large sample sizes. Besides, being based on the responses solely, they ignore useful information about the response mode. As low motivated subjects are supposed to respond faster, it might be beneficial to use the response times also for separating the different classes. In this spirit, Meyer (2010) supplemented the mixture Rasch model with a mixture model for the response times.

Mixture models based on a limited number of classes differing in a stable response mode are not the only way to deal with careless responding. Instead of assuming that test takers behave consistently over the whole test, one can also assume that individuals switch from responding regularly to responding carelessly and vice versa several times during the test. As individuals can not be characterized by a stable mode of responding, there is no latent class membership to identify. Nevertheless, it would be helpful to know the item specific response mode of each test taker in each item. This knowledge, when incorporated into model calibration, could be used to reduce the effects of responding carelessly; see Wise and DeMars (2006) for further details. Several methods for the identification of the item specific response mode have been suggested so far. Most of these approaches infer the mode from the response times without using any information provided by the responses. The simplest approach consists in considering all responses faster than a lower threshold as a careless response. Ways to identify this lower threshold have been described by Kong, Wise, and Bhola (2007) and Wise and Kong (2005). Alternatively, one can infer the response mode from the response times with a mixture model. Schnipke and Scrams (1997) used such a mixture model to estimate the frequency of careless responding in each item. Both approaches, the approach of Wise and DeMars (2006) to the removal of estimation bias and the approach of Schnipke and Scrams (1997) to the identification of the response mode were combined by Yang (2007) into a single model.

Motivation is a complex phenomenon and most models for low-stakes tests are strong simplifications of the response behavior. Motivation is not a fixed binary phenomenon as motivation develops dynamically during the test. Test taking behavior is also more stable than the unsystematic switching between a regular and an irregular response mode some models imply. Recently, several alternative models have been proposed that claim to account for the complexity of motivational processes and circumvent some of the limitations of the earlier models. Rapid guessing can be modeled with the dynamic item response models proposed for change processes. Cao and Stokes (2008) for example suggested an item response model with partially decreasing ability levels that can be used when test takers reduce their effort during the test; see also Goegebeur, De Boeck, Wollack, and Cohen (2008) for a model where the tendency to respond by guessing increases during the test. Alternatively, one can try to disentangle motivation and ability with the help of a process model from cognitive psychometrics as these models distinguish between information processing and motivational aspects of the solution process; see Tuerlinckx and De Boeck (2005), van der Maas, Molenaar, Maris, Kievit, and

Boorsboom (2011) and Rouder, Province, Morey, Gomez, and Heathcote (2014). Last but not least, it is always possible to remove unmotivated test takers after they have been identified with a test of person-fit (Artner, 2016).

The manuscript makes several contributions to the existing literature. First and foremost we suggest a new approach to the problem of careless responding. The approach was inspired by the Rasch model answer tree, also known as Rasch tree (Strobl et al., 2015). Similar to the Rasch model answer tree, the new approach divides the sample into several subgroups via a hierarchical structure of splits and fits a separate Rasch model to each subgroup. Contrary to the Rasch model answer tree, where a subdivision is recursively built from the data, we enforce several configurations of splits and select among the alternative configurations via information criteria. By defining the splits with respect to the response times of the test takers, the approach provides response time thresholds that allow for a separation of careless from deliberate responses. This improves the common practice to set such thresholds a priori or after a visual inspection of the response time distribution. The new approach serves two purposes. It allows for a recovery of the item parameters one would have obtained if careless responses had been absent. It can also be used to classify the subjects according to their way of responding. Contrary to earlier approaches, the new approach avoids strong assumptions about the response process, does not require a fully specified response time model and is easy to apply. The performance of the new approach is investigated in a simulation study. In this study we compare the new approach to several alternative analysis methods suggested for low stakes tests. Surprisingly, little can be found about the relative performance of different ways to handle low motivation in the literature. And finally, we compare the results of the different methods in a real data set.

## 2 The Rasch model answer tree of Strobl et al. (2015)

Strobl et al. (2015) suggested a method to detect differential item functioning that combines two approaches, the Rasch model and classification trees. Although the authors denoted their approach as Rasch tree, this denomination is suboptimal as it is prone to misinterpretations and might evoke wrong associations. In the following, we refer to their method with the term Rasch model answer tree in order to stress the two components of the model, namely the Rasch model and answer trees. Similar to the mixture item response models, the Rasch model answer tree tries to identify subgroups of test takers for which Rasch models with different item parameters hold. Contrary to the mixture item response models, where subgroups are identified by means of the responses alone, the Rasch model answer tree uses covariates, which are supposed to be related to the differential functioning of the items. The subgroups are determined by recursive partitioning, whereby the sample is successively divided. The algorithm starts with the selection of one of the covariates, for which a cut point is determined. All subjects above the cut point are sorted into a first subgroup and all subjects below the cut point are assigned to a second. The covariate and the cut-point are selected with the intention to identify two subgroups that require different Rasch models for the test data. This step of variable selection and cut point specification is repeated for each subgroup, such that the

groups of earlier steps may be successively divided into further subgroups in later steps. As a result, the data set is partitioned into subgroups via a branch like series of splits defined by different cut points and different covariates; see Figure 1 as an illustration of the Rasch model answer tree in a test with ten items. In the first split the respondents are sorted on basis of the covariate *T*2 into two subgroups defined by $T2 \leq -0.204$ and $T2 > -0.204$. To each subgroup a distinct Rasch model is fit whose item locations are indicated by dots in Figure 1. The values for the 10 items are plotted along the *x*-axis. The plot on the left side represents the item locations of the first subgroup ($T2 \leq -0.204$) and the plot on the right side represents the item locations of the second subgroup ($T2 > -0.204$). Note that the item locations differ in the two subgroups.

Each subdivision requires the choice of a covariate and the specification of a cut point. Covariate selection is based on the generalized *M*-fluctuation test of Zeileis and Hornik (2007). This test evaluates whether a particular split results into subgroups that require

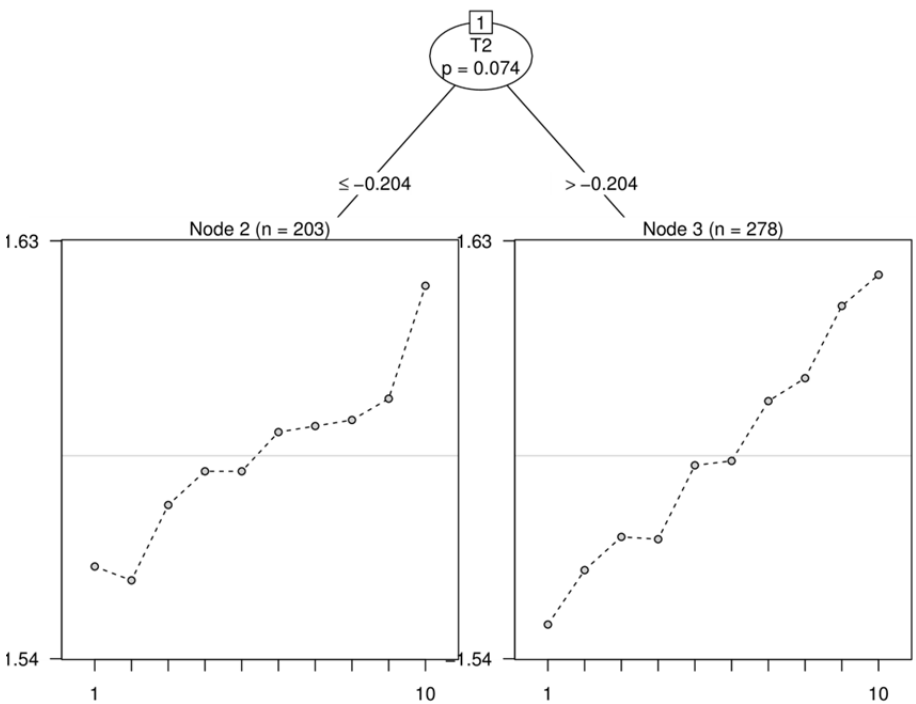


**Figure 1:**
A Rasch model answer tree for a test of 10 items with two subroups defined by the response time in the second quarter of the test (*T*2) and the cut point −0.204. The item locations of the first group with 203 subjects are represented by dots on the left side. The item locations of the second group with 278 subjects are given on the right side. Note that testing for any differences in the item locations between the two groups yields a *p*-value of 0.074 when using the *M*-fluctuation test

different Rasch models. To choose a variable all covariates are tested and the one with the smallest *p*-value is selected. Having chosen a variable, the cut point is selected by determining the level that maximizes the partitioned log-likelihood function. Finally, the subjects are sorted into the corresponding subgroups. These steps are repeated until no more subgroups can be identified that differ systematically, that is, until all *p*-values are larger than some fixed Type-I error rate α. More details concerning the algorithm can be found in Strobl et al. (2015). The Rasch model answer tree (Rasch tree) is implemented in the package `psychotree` (Zeileis, Strobl, Wickelmaier, Kopf, & Abou El-Komboz, 2014) of the statistical environment `R` (R Development Core Team, 2009).

Although originally developed for differential item functioning, the Rasch model answer tree can be used for the detection and correction of careless responding as well. This is due to the fact that responses from compliant and non-compliant subjects cannot be modeled with the same item response model. Responses given without proper information processing usually manifest in an item response model with high item difficulties and very low item discriminations. A good indicator for the renouncement of any mental effort should be the time needed to respond. Very fast responses are only possible for subjects with an irregular mode of information processing. Hence, by using the response times as the covariates, the Rasch model answer tree should be able to subdivide the subjects into a group of regular and irregular responders. Only the item parameters of the first group should be considered for psychological assessment. Response times in single test items are not very reliable. Therefore it might be better to use the average response time needed by an individual to solve the items in the first, second, third and last quarter of the test. In case individuals systematically employ different response modes and these response modes go along with substantial response time differences, there is no loss in using aggregated response time measures. It however is recommendable to standardize the response times first to remove systematic item effects. Note that the usage of response times to classify individuals into two subgroups with different item response models is similar to the analysis of Partchev and De Boeck (2012). However, these authors did a median split to form the subgroups while the Rasch model answer tree identifies the optimal cut point below which responses should be considered as invalid.

The Rasch model answer tree has several advantages. First, the model is implemented in standard software and therefore easy to use. Second, the approach includes the response times but does not require a fully specified response time model. This makes the approach less susceptible to model misspecification. Third, the Rasch model answer tree is very flexible as no assumption about the number of subgroups has to be made. Different groups of subjects can be formed, like a group which is careless from the start and a group that gets careless later etc. Fourth, it is easy to include further information about the motivation of the test takers like test scores from additional motivation questionnaires. Fifth, the approach automatically provides cut points below which responses should be regarded as invalid. Hence, the Rasch model answer tree seems to be an attractive solution for the problem of careless responding.

## 3 Multi-group Rasch models defined by hierarchical data splits

In the Rasch model answer tree of Strobl et al. (2015) the resulting tree structure is data driven as the number of splits depends on the results of the *M*-fluctuation tests. In the present context the assumed form of irregular responding however implies a specific configuration of splits. Therefore we deviated from the Rasch model answer tree and took a different, more hypothesis driven approach. Instead of having the data determine the tree structure we enforced a configuration of splits beforehand, namely a configuration with no split, with one split and with two splits. Only the variable and the specification of the cut points underlying the splits were determined via the data by using the values that maximize the partitioned log-likelihood function. Among the rivaling configurations we chose the configuration that resulted in the best approximation of the data structure via means of the Bayes information criterion (BIC; Schwarz, 1978). Using the response times for splitting as suggested above, the approach resulted in a multi-group Rasch model where the groups were defined by distinct response time patterns. As this practice deviates from the Rasch model answer tree, we will refer to our approach as multi-group Rasch model in order to stress the difference between the two approaches.

Although our approach was inspired by the Rasch model answer tree of Strobl et al. (2015), the decision to fix the configuration of splits beforehand and to replace hypothesis testing with model selection is in conflict with standard partition tree analysis that seeks flexible models for complex relations. Nevertheless, we considered our multi-group approach more adequate for several reasons. A first motivation for this decision was the intention to compare different approaches to handle careless responding and this requires that the way to select the optimal model is comparable. Model selection with information criteria is universally applicable. A second motivation for the decision against hypothesis testing was the fact that we considered the problem of irregular responding more to be a question of model selection than to be a question of hypothesis testing. The models are used to reduce the effects of irregular responding and not to draw valid conclusions about the true number of groups. Falsely detecting two subgroups is not a mistake in case the larger subgroup provides correct parameter estimates. Truly detecting two subgroups is meaningless in case the subgroup differences are irrelevant. Hence, there is only a weak correspondence between hypothesis testing and the research question. The research question at present consists in selecting the optimal model that is complex enough to yield parameters with little bias, but not too complex, such that the item parameters can still be estimated precisely. Such a problem can be addressed with information criteria like the BIC index. The usage of the BIC index in this context might not be fully justified as fitting a model with a fixed configuration of splits is different to standard maximum likelihood estimation. In theory, one should account for the selection of the covariates and the split points in the penalty term. However, the BIC index has proven promising for the selection of the number of classes in latent class analysis (Nylund, Asparouhov, & Muthen, 2007) and in mixture item response models (Li, Cohen, Kim, & Cho, 2006) and might perform well in the present context also.

## 4 Alternative approaches to the detection of careless responding

From the approaches to careless responding reviewed in the introduction we selected two as a standard of comparison for the proposed multi-group approach. The two alternatives were the latent class model of Meyer (2010) and an approach in line with Schnipke and Scrams (1997). Both approaches are sophisticated representatives of a specific strategy to handle the problem of careless responding.

The model of Meyer (2010) assumes $g = 1,...,G$ latent classes that differ with respect to the distribution of the responses and the response times. The distribution of the responses in each latent class is supposed to conform to the Rasch model with different item locations in the different latent classes. The response times are assumed to be distributed according to a log-normal distribution with class specific parameters. More specifically, the density of the response time of subject *i* in item *j* is

$$\mathrm{f}\left(t_{\mathrm{ij}} \mid g_i; \lambda_{jg_i}, \tau_{jg_i}\right) = \sqrt{\frac{\tau_{jg_i}}{2\pi}} \frac{1}{t_{\mathrm{ij}}} \exp\left[-\frac{\tau_{jg_i}}{2}\left(\log\left(t_{\mathrm{ij}}\right) - \lambda_{jg_i}\right)^2\right], \tag{1}$$

where $g_i$ denotes the latent class of subject *i*, $\lambda_{jg_i}$ is the scale parameter and $\sqrt{1/\tau_{jg_i}}$ the shape parameter of the log-normal distribution in item *j* and class $g_i$. Conditionally on the latent class membership, the response times are assumed to be independent. Note that the response time model does not include a continuous, subject specific random effect such that the latent class membership is the only systematic cause for individual differences. Contrary to our approach the model of Meyer (2010) makes strong assumptions about the response time distribution that have to be met when applying the model. This is definitely not wanted as the focus is on careless responding and not on response time modeling, which is a nontrivial topic.

The second approach was based on the mixture model of Schnipke and Scrams (1997). The model assumes a finite mixture distribution for the response times and can also be interpreted as a latent class model. The model postulates the existence of $g = 1,...,G$ latent classes that represent different modes of responding. In each latent class, the logarithmized response times are assumed to follow a multivariate normal distribution with class specific mean vector $\boldsymbol{\mu}_g$ and class specific variance covariance matrix $\boldsymbol{\Sigma}_g$. Note that there are no restrictions on the class specific covariance matrices. Hence, the log response times $\mathbf{t}_i' = \left(\log\left(t_{i1}\right),...,\log\left(t_{iJ}\right)\right)^t$ needed by subject *i* with class membership $g_i$ to respond to item *j*=1,…,*J* are distributed according to the density function of the multivariate normal distribution

$$\mathrm{f}\left(\mathbf{t}_{\mathrm{i}}' \mid g_i; \boldsymbol{\mu}_{g_i}, \boldsymbol{\Sigma}_{g_i}\right) = \frac{1}{\left(2\pi\right)^{J/2}\left|\boldsymbol{\Sigma}_{g_i}\right|^{1/2}} \exp\left(-1/2\left(\mathbf{t}_{\mathrm{i}}' - \boldsymbol{\mu}_{g_i}\right)^t \boldsymbol{\Sigma}_{g_i}^{-1}\left(\mathbf{t}_{\mathrm{i}}' - \boldsymbol{\mu}_{g_i}\right)\right). \tag{2}$$

This finite mixture model has been implemented in the package `mclust` of the statistical software `R` where several constraints concerning the class specific variance covariance matrix $\boldsymbol{\Sigma}_g$ can be imposed; see Fraley and Raftery (2007) for more details. With the finite mixture model subgroups of responders can be identified and specific item response models can be used for each subgroup by considering the predicted class membership as the true one.

## 5 Simulation study

In order to assess the capability of the three methods to correct the effects of irregular responding a simulation study was conducted. The focus of the simulation study was on the performance of the three methods to reduce the estimation error and to detect the irregular responders. In doing so we especially wanted to explore the performance of the methods under less optimal conditions, for example in case of small effects and small subgroups. This complements previous simulation studies about rapid guessing that always considered rather favorable conditions. It is hardly surprising that the methods work very well in case the response time distributions of the two groups are bimodal and have little overlap such that the two groups can already be separated visually. In our experience, this is seldom the case, as individuals do not simply guess rapidly without any information processing but use short cut strategies with reduced mental effort, that is, respond carelessly. Additionally, we wanted to explore the performance under different forms of distortions like different motivational effects or test speededness.

In the simulation study we employed two different strategies of data analysis. In the first strategy each approach was implemented as a two group version. The item parameters of the Rasch model in the larger subgroup were considered as the true item parameters and the test takers were classified as regular and irregular responders. The first strategy served mainly for an assessment of the method's capability to identify the irregular responders. Additionally we wanted to assess the costs of using the three methods mechanically as a routine tool for data analysis in case there is no irregular responding. Some of the earlier simulation studies seem to imply that using a model for rapid guessing is always beneficial, although sometimes superfluous. These findings could stimulate practitioners to use an overly complex model just to be on the safe side. Such a strategy is not optimal as using a more complex model necessarily comes at the price of increased standard errors. The size of this increase was also investigated by means of the first strategy of data analysis. In the second strategy of data analysis the number of groups was not fixed to two, but determined empirically. Therefore, each of the three models was fit to the data in a version for one group, two groups and three groups. Then the version was chosen that corresponded to the lowest BIC index. For this version the largest group was identified and the item parameters in that group were considered as the correct estimates. The second strategy of data analysis is probably more common as one usually does not assume the existence of irregular responders a priori.

In the simulation study the data in a test of 10 items was considered. The simulation study consisted of four different simulation scenarios that were characterized by a dis-

tinct form of irregular responding. In the first simulation scenario all test takers responded regularly. The first simulation scenario therefore allowed for an assessment of the costs of using one of the methods in question. The second and third scenario dealt with the effects of demotivated test takers. In the fourth scenario we analysed the effects of test speededness. In addition to the four forms of irregular responding several other factors were varied. For each simulation scenario different simulation conditions were defined by crossing three effect sizes with two mixing proportions and two sample sizes.

### 5.1 Simulation scenario I

In the first simulation scenario, all individuals responded regularly. Hence, the mixing proportion of the subjects responding carelessly was zero and no distortion had to be accounted for. The first scenario permits the assessment of the costs of using the models without need. Data sets were generated according to the hierarchical model of van der Linden (2007). This model assumes a unidimensional factor model for the log response times and a standard item response model for the responses. Both submodels depend on distinct latent traits that are distributed according to a bivariate normal distribution. Contrary to van der Linden (2007) a Rasch model was used for the responses instead of the three parameter logistic model. The data sets were generated in three steps. In the first step, the latent abilities and latent speeds of the fictitious test takers were independently drawn from the bivariate standard normal distribution. In the second step, the responses of each test taker to the ten items of an imaginary test were generated according to the Rasch model with item locations equally spread from −1 to 1. In the third step the response times were simulated. For each test taker and each item a linear combination of the latent speed and a normally distributed residual term was formed. The linear combination was then exponentiated to render it positive. The intercept term of the linear combination was 2.0, the weight of the latent speed was 0.25 and the variance of the residual term was 0.25 in all items. In this way altogether 250 simulation samples with 500 and 1000 test takers were simulated. The simulation setting implied marginal solution probabilities ranging from 0.71 to 0.28. The response times had an expectation of 7.64 and a standard deviation of 1.93. More details concerning the simulation study (the item parameters, the `R` scripts etc) can be obtained from the authors on request.

The data sets were analyzed with the three methods described above. For the multi-group Rasch model the item response times were summarized to simplify the analysis. First, the response times were logarithmized and standardized to remove item effects. Then, the test was divided into four parts consisting of the items {1,2}, {3,4}, {5,6,7} and {8,9,10}. The response times of each part were summed. This proceeding generated four variables that summarized the total time spent on each quarter of the test. A fifth variable was created for the total testing time. The five variables were used for the determination of the subgroups. Data analysis was based on the package `raschtree` (Zeileis et al., 2014) of the statistical environment `R`. Although this package fits the original Rasch model answer tree of Strobl et al. (2015), it can also be used for the present application by enforcing a specific structure of the splits in the following way. In a first analysis, the $\alpha$ level of the *M*-fluctuation test that determines the splits was set to a level that guaran-

teed at least one split. Then a Rasch model was fit to the two subgroups defined by the first split. The original Rasch model answer tree relies on conditional maximum likelihood estimation, but we decided to reestimate the item parameters by marginal maximum likelihood estimation via the package `ltm` (Rizopoulos, 2006). The common slope was thereby estimated freely in each subgroup. Conditional maximum likelihood estimation does not require distributional assumptions about the latent trait, but is less efficient than marginal maximum likelihood estimation, that is, yields larger standard error of estimation. This impairs the comparability to the other models that are based on marginal maximum likelihood estimation. We therefore changed the estimation approach to make the comparison as fair as possible. In general however, the way the item parameters were estimated was of little importance because the results were virtually identical. Having estimated the item parameters in the two subgroups, the estimates from the larger one were considered as the true values. We also classified the test takers as regular and irregular responders. In a second analysis, we enforced configurations with no split/one group, one split/two groups and two splits/three groups. This was again achieved by manipulating the α level of the *M*-fluctuation test and using just the first splits. For each version the BIC index was determined. This was accomplished by penalizing the log-likelihood function with a term depending on the number of parameters and the sample size. The model with the lowest BIC index was chosen as the best version of the model. The item parameters in the largest group of the best version were again considered as the correct estimates.

The time variables were also used for a segmentation of the sample with the finite mixture model given in Equation 2. Thereby, we analyzed the four variables measuring the time spent on each quarter of the test with the package `mclust` (Fraley, Raftery, & Scrucca, 2014). The total testing time was excluded from the analysis to avoid linear dependencies. Altogether, three different versions of the finite mixture model were fit by requesting one, two or three latent classes. The BIC index was determined for each of the fitted versions. The fitted versions were then used to classify the subjects by an assignment to the most probable class. The data set was then split into the identified classes and to each class a separate Rasch model was fit. The item parameters were estimated by marginal maximum likelihood estimation as described above. The estimates in the largest class were considered as the true values. Finally, the latent class model of Meyer (2010) was calibrated with marginal maximum likelihood estimation. The model was implemented for one, two and three latent classes. Models with more latent classes were not considered as already in the version with three latent classes the estimator had difficulties to converge. The BIC index was calculated for the three versions of the model and the best version of the model was identified. The estimates in the largest latent class were considered as the true estimates. In addition to parameter estimation the subjects were classified into the most probable class and labeled as regular or irregular responders. This was accomplished with the version for two latent classes.

Results concerning item parameter recovery for the three models and the two strategies of data analysis can be found in Table 1. The first strategy of data analysis refers to the two group version of the models while the second strategy of data analysis to the version with the lowest BIC index. Parameter $\beta_0$ denotes the item locations of the Rasch model

and parameter $\beta_1$ the slope parameter. This parameter reflects the variance of the latent ability in marginal maximum likelihood estimation. Table 1 contains the bias, average absolute bias and the relative efficiency of the parameter estimates. The average absolute bias is reported for the item locations in order to prevent that positive bias in some items compensates negative bias in other items. The relative efficiency reflects the ratio of the mean squared error of estimation achieved by one of the methods to the mean squared error of estimation that results from fitting a standard Rasch model to the whole data set. The relative efficiency allows for an evaluation of the methods' estimation performances in comparison to the performance of a standard Rasch model. A value smaller than 1 implies an error reduction by using the more sophisticated model while a value larger than or equal to 1 signifies that using the more sophisticated model does not pay off or even worsens estimation.

Table 1 corroborates that the first strategy of data analyses, that is, the automatic and inconsiderate application of one of the three models in a two group version, has high costs in case all respondents are responding regularly. Although the unnecessary usage of one of the three models does not bias the parameter estimates, it increases the mean squared error of estimation. This finding cautions against the routine application of corrective measures without assessing their empirical justification. The performance of the methods improves in case the best version of the model is determined with the BIC index. No price has to be paid when the multi-group Rasch model or the mixture model is used. Both methods were able to detect the correct number of groups in all data sets and performed as well as the standard Rasch model. The latent class model of Meyer (2010) however overestimated the number of groups and tended to prefer an overly complex model with three latent classes. This tendency was combined with rather poor parameter estimates. The poor performance was due to the fact that the latent class model was misspecified in the first simulation scenario. While the log response times were generat-

**Table 1:**
Bias, average absolute bias and relative efficiency of parameter estimates for different models and two strategies of data analysis in the two simulation conditions of the first simulation scenario.

| Strategy | N | $\beta_1$ | | | | | | $\beta_0$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Bias | | | Rel. Eff. | | | Ave. Abs. Bias | | | Rel. Eff. | | |
| | | MG | LC | MM | MG | LC | MM | MG | LC | MM | MG | LC | MM |
| 1 | 1000 | 0.01 | 0.00 | 0.00 | 1.31 | 1.92 | 1.50 | 0.01 | 0.01 | 0.00 | 1.69 | 2.21 | 1.54 |
| | 500 | 0.00 | 0.00 | 0.00 | 1.40 | 2.32 | 1.68 | 0.01 | 0.01 | 0.01 | 1.64 | 2.01 | 1.57 |
| 2 | 1000 | 0.00 | 0.00 | 0.00 | 1.00 | 2.05 | 1.00 | 0.01 | 0.00 | 0.01 | 1.00 | 2.44 | 1.00 |
| | 500 | 0.00 | 0.00 | 0.00 | 1.00 | 2.40 | 1.00 | 0.01 | 0.01 | 0.01 | 1.00 | 2.50 | 1.00 |

*Note*: Results based on 250 simulation samples. Results for parameter $\beta_0$ are averaged over different items. Strategy 1: Results concerning the two group solution, Analysis 2: Results concerning the best BIC solution. MG: Multi-group Rasch model, LC: Latent class model of Meyer (2010), MM: Finite mixture model. $\beta_0$ : Item location, $\beta_1$ : Common item discrimination

ed via a standard factor model, the model of Meyer (2010) is a latent class model that assumes a more restricted dependency structure of the response times. The only way to account for the more complex covariance matrix was to increase the number of latent classes.

### 5.2 Simulation scenario II

In the second simulation scenario, the data sets were generated according to the model of Meyer (2010). The model was implemented by assuming two latent classes. The first latent class contained the regular responders and was characterized by long response times and low item difficulties. The second latent class comprised the careless responders with fast response times and high item difficulties. Twelve simulation conditions were defined by systematically combining three effect sizes with two mixing proportions and two sample sizes. The effect size was defined as the separation of the response time distributions in regular and careless responders. In the small effect condition the mean of the log response times was 2.0 for the regular responders versus 1.9 for the careless responders in all items. In the moderate effect condition, the respective means were 2.1 and 1.8. In the large effect condition the means were 2.5 and 1.5. The standard deviation was always 0.25. The subgroup specific components of the response time distribution as well as the resulting marginal densities are plotted in Figure 2 for the three effect sizes and a mixing proportion of 0.25.

The responses of the regular responders were generated as in the first simulation scenario by using a Rasch model with item locations equally spread from −1 and 1. The responses of the careless responders were simulated according to a Rasch model whose item locations were increased by the constant amount of 0.5 and whose slope parameter was attenuated from 1.0 to 0.8. The size of the non-complying class, that is, the mixing propor-

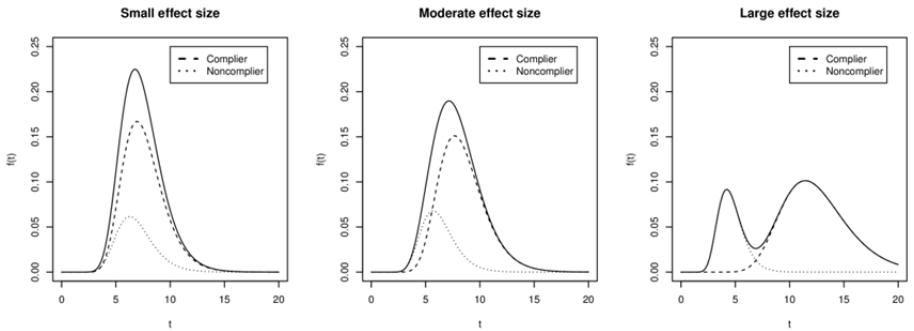


**Figure 2:**
Marginal density (solid line) as well as subgroup specific components (dotted/dashed line) of the response time distribution in the second simulation scenario for the three effect sizes. Note that the proportion of careless responders (Noncomplier) is 0.25

**Table 2:**
Bias, average absolute bias, relative efficiency of parameter estimates for the three models and the best BIC version as well as the classification performance of the models in the simulation conditions of the second simulation scenario

| Effect | Class | N | $\beta_1$ | | | | | | $\beta_0$ | | | | | | Class. Perfor. | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Bias | | | Rel. Eff. | | | Ave. Abs. Bias | | | Rel. Eff. | | | | | |
| | | | MG | LC | MM | MG | LC | MM | MG | LC | MM | MG | LC | MM | MG | LC | MM |
| Large | Large | 1000 | -0.03 | 0.00 | 0.00 | 1.00 | 0.84 | 0.84 | 0.11 | 0.00 | 0.00 | 1.00 | 0.53 | 0.53 | 0.79 | 1.00 | 1.00 |
| | | 500 | -0.03 | 0.00 | 0.00 | 1.00 | 1.02 | 1.02 | 0.11 | 0.01 | 0.01 | 1.00 | 0.73 | 0.73 | 0.76 | 1.00 | 1.00 |
| | Small | 1000 | -0.02 | 0.00 | 0.00 | 1.00 | 0.92 | 0.92 | 0.06 | 0.01 | 0.01 | 1.00 | 0.77 | 0.77 | 0.79 | 1.00 | 1.00 |
| | | 500 | -0.01 | 0.01 | 0.01 | 1.00 | 1.07 | 1.07 | 0.06 | 0.00 | 0.00 | 1.00 | 0.94 | 0.94 | 0.76 | 1.00 | 1.00 |
| Mod. | Large | 1000 | -0.03 | 0.00 | 0.00 | 1.00 | 0.88 | 0.87 | 0.11 | 0.01 | 0.01 | 1.00 | 0.52 | 0.51 | 0.76 | 0.98 | 0.97 |
| | | 500 | -0.03 | 0.00 | 0.00 | 1.00 | 0.97 | 0.95 | 0.10 | 0.01 | 0.01 | 1.00 | 0.78 | 0.76 | 0.75 | 0.97 | 0.97 |
| | Small | 1000 | -0.02 | 0.00 | 0.00 | 1.00 | 0.91 | 0.89 | 0.07 | 0.01 | 0.01 | 1.00 | 0.73 | 0.72 | 0.78 | 0.98 | 0.98 |
| | | 500 | -0.02 | 0.00 | 0.00 | 1.00 | 1.11 | 1.09 | 0.05 | 0.01 | 0.00 | 1.00 | 0.97 | 0.94 | 0.76 | 0.98 | 0.98 |
| Small | Large | 1000 | -0.03 | -0.03 | -0.03 | 1.00 | 1.00 | 1.06 | 0.11 | 0.11 | 0.09 | 1.00 | 1.00 | 0.96 | 0.67 | 0.74 | 0.70 |
| | | 500 | -0.04 | -0.04 | -0.03 | 1.00 | 1.00 | 1.03 | 0.10 | 0.10 | 0.10 | 1.00 | 1.00 | 0.99 | 0.67 | 0.73 | 0.69 |
| | Small | 1000 | -0.02 | -0.02 | -0.02 | 1.00 | 1.00 | 1.00 | 0.06 | 0.06 | 0.06 | 1.00 | 1.00 | 1.02 | 0.71 | 0.80 | 0.68 |
| | | 500 | -0.02 | -0.02 | -0.02 | 1.00 | 1.00 | 0.98 | 0.07 | 0.07 | 0.07 | 1.00 | 1.00 | 1.01 | 0.71 | 0.79 | 0.69 |

*Note*: Results based on 250 simulation samples. Results for parameter $\beta_0$ are averaged over different items. MG: Multi-group Rasch model, LC: Latent class model of Meyer (2010), MM: Finite mixture model. $\beta_0$: Item location, $\beta_1$: Slope parameter.

tion, was 0.25 or 0.15 depending on the simulation condition. Two sample sizes of 500 and 1000 subjects were considered. The second simulation scenario was supposed to mimic the effects of poor motivation combined with some partial effort to solve the items. Note that carelessly responding affected all items as it is the case when individuals have little motivation to take the test from the start. The data sets were analysed as in the first simulation scenario. The three models were fit to the data and each model was implemented in a version for two groups first. This version was used to classify the test takers as regular or careless responders. Then the best version of each model was determined according to the BIC index, and the item parameters in the largest group were considered as the true estimates. The results are reported in Table 2. Here, the bias, average absolute bias and the relative efficiency of the parameter estimates are given for the best BIC version of each model. Results concerning parameter recovery with the forced two group version (first strategy of data analysis) are not reported anymore. Table 2 additionally contains the average rate by which the test takers were correctly identified as regular or careless responders for each of the three methods. Note that the simple strategy of always predicting the larger class identifies the correct class with a rate of 0.75 in the large class condition and with a rate of 0.85 in the small class condition.

In the case that low motivation causes a slight shift of the item locations, the utility of the corrective measures depends on the size of the response time effect and the number of affected individuals. A large reduction of the bias and the standard error of estimations can be noted for the latent class model of Meyer (2010) and the mixture model for the response times in case the two groups are well separated and the proportion of careless responders is high. The multi-group Rasch model is not capable of improving parameter estimation. This is a consequence of its tendency to prefer the one group solution. Note that contrary to the model of Meyer (2010) and the mixture model the response times are used as predictors in the multi-group Rasch model. As inference is conditional on the response times and no fully specified response time model is set up, the results depend mainly on the responses for which the group differences are small. Differences in the response time distribution do not contribute to the separation of the classes directly. This explains the suboptimal performance of the multi-group Rasch model in the second simulation scenario. The results concerning the classification performance parallel the results for parameter recovery. In case the groups are well separated the latent class model of Meyer (2010) and the mixture model are capable of classifying the subjects correctly. The multi-group Rasch model is more error prone.

### 5.3 Simulation scenario III

In the third simulation scenario, the data sets were again generated according to the model of Meyer (2010) with two latent classes. The members of the first latent class responded regularly. The members of the second latent class responded irregularly by rapid guessing, choosing the correct response on chance level. Rapid guesses were generated by random draws from the binomial distribution with success probability of 0.2. As the rapid guessers responded on a random basis, the Rasch model was not valid any more in

**Table 3:**
Bias, average absolute bias, relative efficiency of parameter estimates for the three models and the best BIC version as well as the classification performance of the models in the simulation conditions of the third simulation scenario

| Effect | Class | N | $\beta_1$ | | | | | | $\beta_0$ | | | | | | Class. Perfor. | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Bias | | | Rel. Eff. | | | Ave. Abs. Bias | | | Rel. Eff. | | | | | |
| | | | MG | LC | MM | MG | LC | MM | MG | LC | MM | MG | LC | MM | MG | LC | MM |
| Large | Large | 1000 | 0.01 | 0.00 | 0.00 | 0.18 | 0.17 | 0.17 | 0.01 | 0.01 | 0.01 | 0.09 | 0.08 | 0.08 | 0.99 | 1.00 | 1.00 |
| | | 500 | 0.02 | 0.00 | 0.00 | 0.30 | 0.28 | 0.28 | 0.02 | 0.01 | 0.01 | 0.16 | 0.14 | 0.14 | 0.98 | 1.00 | 1.00 |
| | Small | 1000 | 0.00 | 0.00 | 0.00 | 0.22 | 0.21 | 0.21 | 0.01 | 0.00 | 0.00 | 0.17 | 0.17 | 0.17 | 0.99 | 1.00 | 1.00 |
| | | 500 | 0.02 | 0.01 | 0.01 | 0.43 | 0.41 | 0.41 | 0.02 | 0.01 | 0.01 | 0.29 | 0.26 | 0.26 | 0.98 | 1.00 | 1.00 |
| Mod. | Large | 1000 | 0.02 | 0.00 | 0.02 | 0.20 | 0.17 | 0.18 | 0.03 | 0.00 | 0.03 | 0.09 | 0.08 | 0.09 | 0.97 | 0.98 | 0.98 |
| | | 500 | 0.03 | 0.01 | 0.02 | 0.33 | 0.29 | 0.30 | 0.04 | 0.01 | 0.03 | 0.18 | 0.15 | 0.15 | 0.96 | 0.98 | 0.98 |
| | Small | 1000 | 0.01 | 0.00 | 0.01 | 0.29 | 0.27 | 0.28 | 0.02 | 0.01 | 0.02 | 0.18 | 0.16 | 0.16 | 0.98 | 0.99 | 0.98 |
| | | 500 | 0.02 | 0.00 | 0.01 | 0.49 | 0.34 | 0.36 | 0.05 | 0.00 | 0.02 | 0.37 | 0.28 | 0.28 | 0.95 | 0.99 | 0.98 |
| Small | Large | 1000 | 0.09 | 0.05 | 0.09 | 0.84 | 0.66 | 0.87 | 0.29 | 0.20 | 0.33 | 0.71 | 0.59 | 0.83 | 0.74 | 0.87 | 0.71 |
| | | 500 | 0.10 | 0.10 | 0.10 | 1.00 | 0.99 | 0.96 | 0.37 | 0.39 | 0.37 | 0.93 | 1.00 | 0.92 | 0.71 | 0.84 | 0.69 |
| | Small | 1000 | 0.08 | 0.08 | 0.08 | 0.99 | 1.00 | 0.98 | 0.23 | 0.24 | 0.23 | 0.97 | 1.00 | 0.97 | 0.75 | 0.89 | 0.68 |
| | | 500 | 0.08 | 0.08 | 0.08 | 1.00 | 1.00 | 0.99 | 0.24 | 0.24 | 0.23 | 1.00 | 1.00 | 0.99 | 0.72 | 0.84 | 0.68 |

*Note*: Results based on 250 simulation samples. Results for parameter $\beta_0$ are averaged over different items. MG: Multi-group Rasch model, LC: Latent class model of Meyer (2010), MM: Finite mixture model. $\beta_0$: Item location, $\beta_1$: Slope parameter.

the second latent class. This was the only change to the second simulation scenario. The response times were generated as before. Similar to the second simulation scenario twelve simulation conditions were considered. These conditions paralleled the conditions of the second scenario with respect to effect size, mixing proportion and sample size. The third scenario mimicked the (probably rare) situation that some individuals do not even make the slightest effort to solve the items. The data was analysed as before. All models were fit in a two group version to classify the subjects as regular and careless responders. Then each model's version with the lowest BIC index was determined. The item parameters in the largest group were considered as the true estimates. The results are reported in Table 3.

As long as the effect of rapid guessing on the response time distribution is at least moderate, all methods are able to detect rapid guessing and to counteract its effects. The three methods manage to reduce the bias of the parameter estimates effectively and achieve a lower standard error of estimation than the simple Rasch model. Differences between the three methods are small. The methods are also capable of identifying the rapid guessers with high probability. The rate of correct classification is near 1.00 in case the effect is moderate or larger.

### 5.4 Simulation scenario IV

The fourth simulation scenario dealt with the effects of test speededness. The generation of the responses and response times was based on the model of van der Linden (2007) as in the first simulation scenario. However, this time a fixed time limit was set for the whole test and as soon as the test takers ran out of time they responded to the remaining items rapidly, by choosing the correct response on chance level. Contrary to the second and third simulation scenario the data sets were not generated via a standard latent class model with a fixed mixing proportion. To vary the amount of irregular responding the time limit was varied. This also affected the proportion of rapid guessing in the test. In the large effect condition, a time limit of 75 was imposed. With this severe time limit about 50 % of the test takers were running out of time. The proportions of the test takers responding regularly were about 0.98, 0.92, 0.74 and 0.49 in the last four items. In the moderate effect condition the time limit was set to 85 such that the proportion of the test takers responding regularly was 0.98, 0.92 and 0.77 in the last three items. Hence, in the last item about 25 % of the subjects guessed rapidly. In the small effect condition the time limit was set to 90. The rates of rapid guessing were 0.99, 0.96 and 0.86 in the last three items. So in this condition about 15 % of the test takers were forced to guess rapidly. Note that rapid guessing affected only the last items that were also the most difficult; see the first simulation scenario for the values of the item locations. As in the last item the marginal solution probability of 0.30 was similar to the guessing probability of 0.20, the simulation setting is rather difficult. Again, two sample sizes of 500 and 1000 subjects were considered. The combination of the three time limits with the two sample sizes defined six simulation conditions, for which 250 data sets were simulated. The data sets were analysed as described above. The three methods were first estimated in a version

**Table 4:**
Bias, average absolute bias, relative efficiency of parameter estimates for the three models and the best BIC version as well as the classification performance of the models in the simulation conditions of the fourth simulation scenario.

| Effect/Class | N | $\beta_1$ | | | | | | $\beta_0$ | | | | | | Class. Perfor. | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Bias | | | Rel. Eff. | | | Ave. Abs. Bias | | | Rel. Eff. | | | | | |
| | | MG | LC | MM | MG | LC | MM | MG | LC | MM | MG | LC | MM | MG | LC | MM |
| Large | 1000 | -0.08 | 0.00 | -0.02 | 1.00 | 0.35 | 0.85 | 0.06 | 0.01 | 0.02 | 1.00 | 1.66 | 2.89 | 0.73 | 1.00 | 0.96 |
| | 500 | -0.09 | -0.01 | -0.04 | 1.00 | 0.53 | 1.03 | 0.06 | 0.01 | 0.03 | 1.00 | 1.74 | 2.73 | 0.72 | 1.00 | 0.95 |
| Mod. | 1000 | -0.03 | 0.00 | 0.00 | 1.00 | 0.82 | 1.71 | 0.02 | 0.00 | 0.01 | 1.00 | 1.17 | 2.46 | 0.77 | 1.00 | 0.97 |
| | 500 | -0.03 | 0.00 | -0.01 | 1.00 | 1.06 | 1.72 | 0.02 | 0.01 | 0.01 | 1.00 | 1.22 | 2.14 | 0.73 | 1.00 | 0.96 |
| Small | 1000 | -0.02 | 0.00 | 0.00 | 1.00 | 2.24 | 1.44 | 0.01 | 0.01 | 0.01 | 1.00 | 2.14 | 1.65 | 0.78 | 1.00 | 0.98 |
| | 500 | -0.01 | 0.01 | 0.01 | 1.00 | 1.80 | 1.37 | 0.02 | 0.01 | 0.01 | 1.00 | 2.03 | 1.54 | 0.76 | 1.00 | 0.97 |

*Note*: Results based on 250 simulation samples. Results for parameter $\beta_0$ are averaged over different items. MG: Multi-group Rasch model, LC: Latent class model of Meyer (2010), MM: Finite mixture model. $\beta_0$: Item location, $\beta_1$: Slope parameter.

for two groups, which was used to classify the test takers as regular and irregular responders. Then the best BIC version of each method was determined. The item parameters in the largest group were considered as the true estimates. Results concerning parameter recovery by the best BIC version and the rate of correct classification can be found in Table 4.

Findings are less promising. The BIC always preferred the multi-group Rasch model in its one group version. Consequently, the results were not different from fitting a standard Rasch model to the whole sample. With respect to the latent class model of Meyer (2010) and the finite mixture model for the response times, the three group version was chosen most often. This capability for detecting the presence of test takers responding differently however did not have a positive effect on parameter recovery. In some conditions the standard error of estimation almost tripled. This is somewhat surprising as the two methods are capable of identifying the rapid guessers with high probability. Although this is positive it also enforces model versions with a large number of groups, which are difficult to estimate. Estimation is additionally affected by the fact that the mixing proportion of the target group, that is, the group of subjects responding regularly in all items, becomes smaller with increasing effect size. Note that in the large effect condition only half of the test takers responded regularly in all items.

## 6 Empirical application

In addition to the simulation study the three methods were used for the analysis of a real data set. This data set consisted of the responses and the response times in a test for chess playing proficiency; for a detailed description of the test and the data see van der Maas and Wagenmakers (2005). Here, the items of the 'Choose a move'-scale were used. This scale consists of 40 items, each displaying a chess position, for which the subjects have to indicate the best move. The response format is free and responses are scored as either true or false. The test is time-constrained with a time limit of 30 s per item. The scale can be divided into three subscales, the first assessing endgame knowledge (EndMove), the second assessing positional knowledge (PosMove) and the last one assessing tactical knowledge (TacMove). Overall, 259 subjects completed the test. The data were collected during a chess tournament in Dieren. Participation was voluntary and no reward was given for participation.

Each subscale was analysed separately. First, all subjects with missing data were removed. Then, the response times in the single items were logarithmized and standardized. Each subscale was divided into four parts, for which the response times were summed. The total testing time, or to be more precise, the total log time was determined as well. The multi-group Rasch model was fit to the data, using the five generated variables as the covariates when subdividing the data set. A structure with one, two and three groups was fit and the best solution with respect to the BIC index was determined. In addition, the latent class model of Meyer (2010) was fit to the data with one, two and three latent classes. Again, the best model was chosen with respect to the BIC index.

**Table 5:**
Optimal number of classes in the three subscales of the chess test as suggested by the different models when using the BIC index as decision criterion for the optimal number of classes.

| Scale | MG | LC | MM |
|---|---|---|---|
| EndMove | 1 | 2 | 1 |
| PosMove | 2 | 3 | 4 |
| TacMove | 1 | 3 | 2 |

*Note*. MG: Multi-group Rasch model, LC: Latent class model of Meyer (2010), MM: Finite mixture model.

And finally, different versions of the finite mixture model for the response times were estimated and the best one was chosen. The suggested number of latent classes for the three subscales and the three models can be found in Table 5.

According to Table 5 there is little consensus with respect to the optimal number of classes. For the PosMove scale the multi-group Rasch model suggested two classes defined by the total log time and the cut point −0.88. For the other scales the multi-group model managed with just one class. The latent class model of Meyer (2010) pointed to two and three classes. However, as the model of Meyer (2010) tends to overestimate the number of classes in case of individual differences in work pace, these results should be considered with care. The finite mixture model, which does not assume a strict form of conditional independence between the response times but is based on the multivariate normal distribution, coincides sometimes with the multi-group Rasch model (EndMove scale), sometimes tends more to the latent class model's solution (PosMove) and sometimes lies between the two (TacMove). Again, one has to be careful when interpreting these findings as the model overestimates the number of classes in case the response time distribution is misspecified. Overall, there is little guidance which solution is best. The kernel density estimates of the total response time distribution given in Figure 3 support slightly the solution of the multi-group Rasch model. Except for the PosMove scale the distributions seem to be unimodal. The absence of strong forms of careless responding might be due to the fact that participation was voluntary. Unmotivated individuals also might prefer to abort the test than to hurry through it. Note that only subjects with complete data were selected. This might have made the motivation of the subjects more homogenous than it usually is in typical low-stakes tests.

The data from the PosMove scale was analyzed further. This time all models were restricted to a two class solution. Having calibrated the respective models the membership of the subjects was estimated. The models differed with respect to the class sizes. The multi-group Rasch model divided the subjects into two classes with sizes of 174 and 39, the latent class model of Meyer (2010) into classes of size 141 and 72 and the finite mixture model into two classes of size 134 and 79. The larger class was the class with the longer response times in all three models. Altogether, there was good agreement with respect to the class membership. The multi-group and the latent class model coincided in

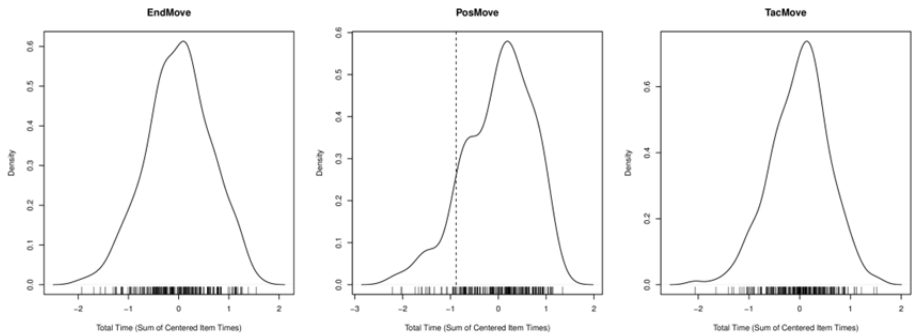


**Figure 3:**
Kernel density estimate of the distribution of the sum of the logarithmized standardized response times in the single items for the three subscales. The dotted line illustrates the cut point suggested by multi-group Rasch model for the PosMove scale

81 %, the multi-group and the finite mixture model in 81 % and the latent class model and the finite mixture model in 83 % of the cases.

Somewhat counter to intuition, the item location parameters of the Rasch model were lower in the smaller class with shorter response times – a finding that seems to contradict the claim that this class consisted of the test takers with poorer motivation on first sight. However, the marginal maximum likelihood estimates of the Rasch parameters are confounded with the scale and location of the latent traits and therefore can not be interpreted arbitrarily. The smaller class also contained the subjects with the higher chess playing proficiencies (as assessed by an external criterion) and a part of the increase in the intercepts might be due to the higher ability level in this group. Besides, the small sample size in the second class induces large standard errors of estimation such that these results should not be over-interpreted.

To assess the meaningfulness of the identified classes their moderator effect on the external validity of the test was assessed. As an external criterion of the chess playing proficiency the elo-score of each subject was used. The elo-score is an index that aggregates the past chess playing performance of a subject. It subsumes the number of wins and losses of a chess player and determines his/her position in the world ranking. It can therefore be considered as an excellent proxy for the actual chess playing proficiency of an individual. We hypothesized that in case the classes really differed with respect to their commitment to the test, the predictive power of the test should be lower in the class with the shorter response times. This hypothesis was tested as follows. The latent traits of all test takers were estimated with the Rasch model using the parameter estimates from the larger class. The estimated traits were then used to predict the elo-score of the individuals with a linear regression model. Separate linear regression models were fit to the data of the two classes. The regression coefficients as well as the coefficients of determination were regarded as a measure of the predictive validity and are reported in

**Table 6:**
Predictive validity of the latent traits from the PosMove scale in the two classes identified by the three models: Regression coefficient (b) and coefficient of determination $\left(R^2\right)$ when predicting the elo-score via a linear regression model.

| Model | Slow Class | | Fast Class | |
|---|---|---|---|---|
| | $R^2$ | $b$ | $R^2$ | $b$ |
| MG | 0.37 | 207.44 | 0.18 | 107.51 |
| LC | 0.32 | 181.48 | 0.34 | 174.42 |
| MM | 0.35 | 201.81 | 0.30 | 174.69 |

*Note*. MG: Multi-group Rasch model, LC: Latent class model of Meyer (2010), MM: Finite mixture model for response times.

Table 6. Note that the coefficient of determination depends on the range of the predictor and therefore might not be comparable in case the two classes differ with respect to the distribution of the latent traits.

The results for the larger class with longer response times (slow class) are similar for the three approaches. The coefficient of determination is moderate suggesting considerable validity of the test. There is also a strong relation between the trait and the elo-score as reflected by the regression coefficient *b*. For the small class with fast response times (fast class) the results depend on the chosen method. In the class identified by the multi-group Rasch model, the coefficient of determination drops and the regression coefficient halves. This implies that the test has less predictive validity for the subjects that were identified as slow responders. Somewhat surprisingly, the same does not occur in the classes identified by the two other methods. Only the multi-group Rasch model succeeds in identifying subjects whose trait estimates appear to be of little use.

## 7 Discussion

Low-stakes tests are tests with few personal consequences for the test takers. Consequently, as there is little to be gained from making full effort, some test takers hurry through the test without properly thinking about the items. The consequences of this careless response behavior are threefold. On the subject level, the real ability level is underestimated, as long as the real ability level is defined as the ability level under maximal motivation. On the level of the institution the subjects are from, the average ability level of all subjects attending the institution is underestimated as well. And on the level of the test, the test is miscalibrated as the item parameter estimates are biased. All these effects can have serious consequences in practice.

In the present manuscript we have compared three different methods that can be used to counteract the effects of careless responding, one of which was new. The focus of the manuscript was on the methods' capability to recover the true values of the item parameters and to identify the test takers that responded carelessly. The problem of estimation

bias due to data contamination is a general problem in statistics and has stimulated research into robust estimation. Robust estimation is still an underdeveloped field in item response theory although for generalized linear mixed models such estimators exist (Moustaki & Victoria-Feser, 2006, Sinha, 2004, Yau & Kuk, 2002). From this perspective, the three models could also be interpreted as tools for robust estimation.

The simulation study revealed that bias reduction works well with all three methods, at least as long as the effects of careless responding on the responses and the response times are not too small. Under less optimal conditions, that is, few careless responders and small differences, the methods differ. The approaches with a precise specification of the response and response time distribution such as the latent class model of Meyer (2010) perform better in case all assumptions are met. This is hardly surprising as by fully specifying the distribution, the information of the responses and response times can be used most efficiently (Altham, 1984). However, in case the response time distribution is misspecified or the model assumes the wrong form of dependency between the data, the results can even change for the worse. An example is the poor performance of the latent class model of Meyer (2010) in the first and fourth simulation scenario. This illustrates the advantage of a statistical model that makes as little assumptions as possible. The multi-group Rasch model is good in this respect as it avoids any assumption about the response time distribution. Although it does not perform as good as the two alternative models with respect to bias reduction, the model never performs worse than the standard Rasch model. Besides, the multi-group Rasch model is simple to use and provides a cut-off that distinguishes regular from careless responders.

The identification of individuals with aberrant mode of responding is treated in item response theory under the label of person fit analysis. Several approaches exist for the detection of person misfit by means of their responses; see Meijer (1996), Meijer and Sijtsma (2001) and Artner (2016). Unfortunately, these standard tests of person fit have low power in short tests (Ranger & Kuhn, 2015). This might be due to the fact that the responses alone give little insight into the solution process. Low scores can be due to careless responding or to low ability. Contrary to the classical approaches to person misfit the three methods are able to detect the irregular responders in the simulation study well, as long as the regular and careless responders differ with respect to the response time distribution moderately. This illustrates the value of the response times for drawing inferences about the response process.

Although we tried to consider the most relevant factors in the simulation study, its scope was necessarily limited as it is always the case with simulation studies. First, we used a rather short test with just ten items. We repeated parts of the study with a longer test but the results were virtually the same. Second, we did not use the single response times on the item level, but average response times in parts of the test for the multi-group Rasch model. This was motivated by our experience that in data analysis it usually is not a good strategy to feed dozens of variables into a statistical model in the hope that the model identifies the good predictors. In order to facilitate variable selection we chose to limit the analysis to a small number of aggregated response time measures that we thought to be more indicative for rapid guessing than a single response time. Nevertheless, we admit that one also could have used the single response times or alternative measures like

the intraindividual variance. Note that response times of automatic guesses should be rather uniform. Third, the conditions considered in the simulation scenarios might have favored the mixture models. Data were usually generated according to a latent class model. The log response times were normally distributed. With non-normal data the mixture models might have tended to overestimate the number of classes to approximate the true distribution; note that mixture models are also used for density estimation (Fraley et al., 2014). So the results for the model of Meyer (2010) and the finite mixture model for the response times might be too optimistic. A replication of the simulation study with different response time distributions and different forms of demotivation (e.g. a gradual change from regular to careless responding) might be interesting. Fourth, we assumed the conditional independence between the responses and the response times in the simulation study. This assumption is common practice in response time modeling and there is evidence that supports this claim (van der Linden & Glas, 2009). The multi-group Rasch model might however perform less well in case of a more complex relation between the responses and the response times. This problem could be addressed from the perspective of sample selection as the formation of subgroups might be conceived as such.

The manuscript was written from the perspective of careless responding. Careless responding was identified by means of response time data. Although response times are probably the first choice for this purpose, the multi-group Rasch model is a flexible approach that is capable of considering additional indicators of careless responding. As an anonymous reviewer pointed out one could also include scores from a motivational scale. All one had to do is to include another covariate, the motivational score, in addition to the response time measures. One could also test the hypothesis that specific test takers respond differently by defining an indicator variable, which is then entered as a regular covariate into the Rasch model answer tree. Response time was considered here as indicative of careless responding. This however is not the only application of the Rasch model answer tree or the modified multi-group Rasch model with response time covariates. The method should be able to detect different response strategies in general. Different response strategies are omnipresent in attitudinal scales, where individuals can take a rapid peripheral route of information processing or a more time intensive central one (Mayerl, 2013, Böckenholt, 2012). Responding in a social desirable way might also be associated with a distinct form of response time pattern. So there might be much more uses of the approach than just bias correction in low stakes testing. This however is a field for future research.

## Acknowledgements

We would like to thank the Editor Prof. Kubinger and two anonymous reviewers for their feedback concerning earlier versions of the paper that helped to improve it considerably. We are also grateful to Prof. van der Maas and Prof. Wagenmakers for their permission to use the chess data set.

## References

Altham, P. (1984). Improving the precision of estimation by fitting a model. *Journal of the Royal Statistical Association, B*, *46*, 118–119.

Artner, R. (2016). A simulation study of person-fit in the Rasch model. *Psychological Test and Assessment Modeling*, *58*, 531–563.

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods*, *17*, 665–678.

Bolt, D., Cohen, A., & Wollack, J. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement*, *39*, 331–348.

Cao, J., & Stokes, S. (2008). Bayesian IRT guessing models for partial guessing behaviors. *Psychometrika*, *73*, 209–230.

Fraley, C., & Raftery, A. (2007). Model-based methods of classification: Using the mclust software in chemometrics. *Journal of Statistical Software*, *18*.

Fraley, C., Raftery, A., Murphy, T., & Scrucca, L. (2014). Mclust – Normal mixture modeling for model-based clustering, classification, and density estimation [Computer Software Manual]. Department of Statistics, University of Washington. (Version: 4.3)

Goegebeur, Y., De Boeck, P., Wollack, J., & Cohen, A. (2008). A speeded item response model with gradual process change. *Psychometrika*, *73*, 65–87.

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

Li, F., Cohen A., Kim, S.-H., & Cho, S.-J. (2006). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurements*, *33*, 353–373.

Mayerl, J. (2013). Response latency measurement in surveys. Detecting strong attitudes and response effects. *Survey Methods Insights from the Field*, Retrieved from `http://surveyinsights.org/? p=1063.`

Meijer, R. (1996). Person-fit research: An introduction. *Applied Measurement in Education*, *9*, 3–8.

Meijer, R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*, 107–135.

Meyer, J. (2010). A mixture Rasch model with item response time components. *Applied Psychological Measurement*, *34*, 521–538.

Moustaki, I., & Victoria-Feser, M.-P. (2006). Bounded-influence robust estimation in generalized linear latent variable models. *Journal of the American Statistical Association*, *101*, 644–653.

Nylund, K., Asparouhov, T., & Muthen, B. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, *14*, 535–569.

Oshima, T. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement*, *31*, 200–219.

Partchev, I., & De Boeck, P. (2012). Can fast and slow intelligence be differentiated? *Intelligence*, *40*, 23–32.

R Development Core Team. (2009). R: A language and environment for statistical computing [Computer software Manual]. Vienna, Austria: Retrieved from `http://www.R-project.org` (ISBN 3-900051-07-0)

Ranger, J., & Kuhn, J. (2015). Assessing person fit with the information matrix test. *Methodology*, *11*, 3–12.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, *17*, 1–25. Retrieved from `http://www.jstatsoft.org/v17/i05/`

Rost, J. (1990). Rasch models in latent classes: An integration of two appoaches to item analysis. *Applied Psychological Measurement*, *14*, 271–282.

Rouder, J., Province, J., Morey, R., Gomez, P., & Heathcote, A. (2015). The lognormal race: A cognitive-process model of choice and latency with desirable psychometric properties. *Psychometrika*, *80,* 491-513.

Schnipke, D. (1999). *The influence of speededness on item-parameter estimation* (Tech. Rep. No. LSAC-R-96-07). Princeton: Law School Admission Council.

Schnipke, D., & Scrams, D. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, *34*, 213–232.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Sinha, S. (2004). Robust analysis of generalized linear mixed models. *Journal of the American Statistical Association*, *101*, 451–460.

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika*, *80*, 289–316.

Tuerlinckx, F., & De Boeck, P. (2005). Two interpretations of the discrimination parameter. *Psychometrika*, *70*, 629–650.

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

van der Linden, W., & Glas, C. (2009). Statistical tests of conditional independence between responses and/or response times on test items. *Psychometrika*, *75*, 120–139.

van der Maas, H., Molenaar, D., Maris, G., Kievit, R., & Boorsboom, D. (2011). Cognitive psychology meets psychometric theory: On the relation between process models for decision making and latent variable models for individual differences. *Psychological Review*, *118*, 339–356.

van der Maas, H., & Wagenmakers, E. (2005). A psychometric analysis of chess expertise. *American Journal of Psychology*, *118*, 29–60.

Wise, S., & DeMars, C. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, *43*, 19–38.

Wise, S., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*, 163–183.

Yamamoto, K., & Everson, H. (1997). Modeling the effects of test length and test time on parameter estimation using the HYBRID model. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 89–98). New York: Waxman.

Yang, X. (2007). Methods of identifying guessers from item response data. *Educational and Psychological Measurement*, *67*, 745–764.

Yau, K., & Kuk, A. (2002). Robust estimation in generalized linear mixed models. *Journal of the Royal Statistical Society, B*, *64*, 101–117.

Zeileis, A., & Hornik, K. (2007). Generalized m-fluctuation tests for parameter instability. *Statistica Neerlandica*, *61*, 488–508.

Zeileis, A., Strobl, C., Wickelmaier, F., Kopf, J., & Abou El-Komboz, A. (2014). psychotree – recursive partitioning based on psychometric models [Computer software manual]. Department of Statistics, University at Innsbruck. (Version: 0.13-0)