Psychological Test and Assessment Modeling, Volume 58, 2016 (3), 531-563

A simulation study of person-fit in the Rasch model

Richard Artner[1]

Abstract

The validation of individual test scores in the Rasch model (1-PL model) is of primary importance, but the decision which person-fit index to choose is still not sufficiently answered. In this work, a simulation study was conducted in order to compare five well known person-fit indices in terms of specificity and sensitivity, under different testing conditions. Furthermore, this study analyzed the decrease in specificity of Andersen´s Likelihood-Ratio test in case of person-misfit, using the median of the raw score as an internal criterion, as well as the positive effect of removing suspicious respondents with the index C*. The three non-parametric indices Ht, C* and U3 performed slightly better than the parametric indices OUTFIT and INFIT. All indices performed better with a higher number of respondents and a higher number of items. Ht, OUTFIT, and INFIT showed huge deviations between nominal and actual specificity levels. The simulation revealed that person-misfit has a huge negative impact on the specificity of Andersen´s Likelihood-Ratio test. However, the removal of suspicious respondents with C* worked quite well and the nominal specificity can be almost respected if the specificity level of C* is set to 0.95.

Keywords: Rasch model, type-I- and type-II-risk, simulation study, person-fit, Andersen's Likelihood-Ratio test

[1] *Correspondence concerning this article should be addressed to:* Richard Artner, PhD, University of Vienna, Austria; email: Richardartner@gmx.at

Introduction

Which index is best to analyze potential *person-misfit* (e.g. cheating, guessing, careless responding, distorting behavior, fatigue) in the item response theory? The answer is still not entirely clear despite the rich body of literature in this area. The two main issues regarding past research are the method of comparison between different indices and the way person-misfit was operationalized. This simulation study has the purpose to shed (some more) light on the detection skills of certain indices for person-misfit in the dichotomous Rasch model. It takes a close look at the influence of certain parameters (e.g. number of items[2]) and the differences between nominal and actual type-I-risks.

How important is the detection of person-misfit if the Rasch homogeneity of certain items is tested? How strong is the support a person-fit index can offer in that case? Up to this point no study can answer this issue.

The Rasch model

The Rasch model (also known as *one parametric logistic model*) is the most prominent model in the item response theory (IRT). The basic assumption of IRT is that it is not possible to directly observe traits of interest. Therefore these traits are called *latent*. However, it is possible to infer from discrete responses of a person, particularly the answers given on a test, the individual characteristic of the latent ability trait of interest. This work analyses person-fit in the context of the dichotomous Rasch model, where for each item a certain response (e.g. "The correct one") is quantified with 1 and all other possible responses are quantified with 0[3]. In this dichotomous Rasch model the response each person (respondent) gives to each individual item is a Bernoulli random variable (1 = correct response, 0 = incorrect response). Equation (1) shows the well-known probability function of this model which depends on the latent ability trait and the item difficulty parameter.

$$P(X_{ni} = x_{ni} \mid \xi_v, \sigma_i) = \frac{e^{x_{ni}(\xi_n - \sigma_i)}}{1 + e^{(\xi_n - \sigma_i)}} \tag{1}$$

X_{ni} is the response (hereinafter also referred to as "answers") that respondent n gives to item i, ξ_n the latent ability trait of respondent n, and σ_I the item difficulty parameter of item i.

---

[2] In this work real life concepts like objects and phenomena (e.g. respondent, test, item, cheating) are used as a placeholder for the underlying statistical and mathematical operationalization of the certain real life concept which ultimately is just a certain sequence of binary code. The context should always make it clear if a word is used in the common sense or in the specific meaning it has in this simulation study.

[3] In real life it does occur that no response is given to a certain item and therefore a third possible value beside zero and one, indicating a missing response, is of need (e.g. -99). However, missing values cannot occur in this simulation study.

The probability that respondent n answers item i correctly can be computed by setting x_{ni} to 1 in (1) and is equal to 1 minus the probability that respondent n answers item i incorrectly, which in turn can be computed by setting x_{ni} to 0. Note that the probability is smaller than 0.5 if the exponent of the exponential function is smaller than 0 and that centralizing (i.e. adding or subtracting a constant such that the expected value of the parameter becomes 0.) both parameters facilitates their interpretation.

The Rasch model has important assumptions/properties. If all items measure the manifestation of the same latent trait (=*person parameter*) and if the probability of a correct response only depends on the latent trait beside the item difficulty (=*item parameter*), we know that the Rasch model holds and vice versa. That means that two respondents with the same latent trait have the same probability for a correct response on each item. This property is often called *unidimensionality*. Furthermore, we have the so-called *local stochastic independence* for items and respondents. The former is an independence of the answers a respondent gives on two items if conditioned on the person parameter and the item parameters. The latter is an independence of the answers of two respondents on a certain item if conditioned on the item parameter as well as on the two person parameters. It is well known that in case of local stochastic independence the raw score r_n (the number of correct responses of respondent n) is a sufficient statistic for the person parameter, and that p_i (The number of respondents with correct responses on item i) is a sufficient statistic for the item difficulty parameter[4]. Therefore, two respondents have the same person parameter estimate if they both answered k out of I items correctly, even though they answered different items correctly.

A very important property of a test where the Rasch model holds is the so-called *specific objectivity*. This property says that the ranking order of the items is the same for each possible population subgroup when ranked according to their difficulty (e.g. item j is harder than item k for every respondent even though both are easy for someone with a high person parameter and both are hard for someone with a low person parameter.) Moreover, the relative difference in difficulty between two items is the same in each possible population subgroup. If specific objectivity does not hold for an item, we have a so-called *differential item functioning* (DIF), that is to say, a different relative difficulty of the item in two subgroups. If no DIF exists for any of the items, a comparison of the person parameter estimations of two respondents is valid, even though they solved different items or even worked on different items. They must not even answer the same number of items. Furthermore, the difficulty parameters of different items can be compared, even though they were estimated in a different sample, hence, with different respondents answering them. The estimation of difficulty and person parameters is therefore *sample-independent*. A more detailed analysis and proofs of these properties can be found in Fischer´s work (Fischer, 1974).

The Rasch model is suited for a test with I items if equation (1) holds for each and every item, which can only be the case if the respondents try their best to solve each item. The

[4] A statistic is sufficient in respect to a certain parameter if it contains all important information for a parameter estimation.

goodness-of-fit for individual items is generally called *item fit*. To test whether the items of a test conform the Rasch model or not, Andersen´s Likelihood-Ratio (LR) test for two or more groups is often used (Andersen, 1973). Andersen´s LR test works in the following way: The sample is split into groups according to an external criterion (e.g. gender) or an internal criterion (e.g. the raw score of each person), and it is investigated how much the Likelihood of the data can be enlarged if different item difficulty parameters in the subgroups are allowed. If the assumption of specific objectivity is violated, this improvement will be higher as in the case of no DIF. A more technical and detailed description of Andersen´s LR test is omitted in this work and can, for example, be found in Andersen´s paper (Andersen, 1973).

Person-fit

If the behavior of a respondent violates the assumption, that only the latent variable of interest systematically influences the probability of correct responses to the items, we speak of *person-misfit*. Types of person-misfit, are among others, cheating, distorting, inattentive behavior and careless behavior. In order to quantify, specify and measure the type and magnitude of person-misfit, the Guttman scale, named after the Israeli mathematician Louis Guttman, is of great help (Mokken, 1971).

Instead of a continuous probabilistic model, let us now consider a simple deterministic approach: A respondent answers an item correctly if his person parameter is greater or equal to a certain value, and answers it incorrectly if it is smaller. If this deterministic model holds in a test with I items, it is certain that a respondent with a raw score of k answered the k easiest items correct and the I-k hardest items wrong. Let us now - without any loss of generality - rank the items according to their difficulty parameters, with the first item being the easiest. Given a raw score of k, the perfect Guttman scale then corresponds to a correct response to the items 1 to k and an incorrect response to the items $k+1$ to I. If a pool of I items follows the perfect Guttman scale and we know that a respondent answered item k wrong, we can therefore conclude that he or she also answered all items $k+i$ wrong with i being an integer between 1 and I-k.

For a four item test we have 2^4=16 possible response vectors:

0 0 0 0*, 0 0 0 1°, 0 0 1 0, 0 0 1 1°, 0 1 0 0, 0 1 0 1, 0 1 1 0, 0 1 1 1°,
1 0 0 0*, 1 0 0 1, 1 0 1 0, 1 0 1 1, 1 1 0 0*, 1 1 0 1, 1 1 1 0*, 1 1 1 1*

Five of these 16 vectors (labeled with *) follow the perfect Guttman scale. Another three (labeled with °) follow the reversed Guttman scale.

In the case of the Rasch model, we obviously cannot expect the response vectors a respondent to always correspond with the perfect Guttman scale. That being said, given a raw score of k, the perfect Guttman scale is the most likely response vector. Let us, for example, take a Rasch conform test consisting of ten items, all being different in difficulty, and a respondent with a raw score of 6.

Here are six possible response vectors:

V1 = 1 1 1 1 1 1 0 0 0 0* V2 = 1 1 1 1 1 0 1 0 0 0 V3 = 1 1 1 1 0 0 1 1 0 0
V4 = 1 1 1 0 1 0 1 0 1 0 V5 = 1 0 0 1 1 1 0 0 1 1 V6 = 0 0 0 0 1 1 1 1 1 1°

With the help of equation (1) and the local stochastic independence property, it is easy to verify that: P(V1) > P(V2) > P(V3) > P(V4) > P(V5) > P(V6), with P standing for probability.

Without any loss of generality it, therefore, can be concluded that in the Rasch model, "strong" deviations from the perfect Guttman scale are unlikely. If we want to test if the Rasch model holds, we can therefore look at the response vectors of the respondents and compare them to the perfect Guttman scale. The greater the difference, the unliklier the results. Strong deviations from the perfect Guttman scale are labeled as a model *underfit*, very small deviations are labeled as a model *overfit*. If the term person-misfit appears in some scientific work, the researcher(s) most likely address a model underfit. However, in this work overfit and underfit are both seen as a potential model misfit and the context should always make it clear if the former or the latter is addressed.

C* and U3 (non-parametric)

C* and U3 are Guttman error based non-parametric person-fit indices. Non-parametric means that no model parameters are estimated (in our case: item difficulty and person parameter). The modified caution index (C*) was developed by Harnisch and Linn (Harnisch and Linn, 1981) by modifying Sato´s caution index (Sato, 1975) in order to limit the possible index values to real numbers on the interval [0, 1]. U3 was developed by Van der Flier (Flier, 1982). C* and U3 belong to the family of *group-based person-fit statistics* (Meijer & Sijtsma, 2001). All indices in this family satisfy the general equation (2), for some non-negative weight w_i. To measure the magnitude of deviation from the perfect Guttman scale the items must be sorted ascending according to their difficulty. In the case of C* and U3 the items are ordered according to their proportion-correct score (p_i/N) with p_i being the number of respondents which answered item i correctly and N being the number of respondents. It is easy to see that the numerator cannot be bigger than the denominator in (2) if the weights (w_i) are a monotonic increasing function of p_i/N. C* uses p_i/N as the weight (equation (3)). U3 uses a more complicated weight which includes the natural logarithm (equation (4)).

$$General_n = \frac{\sum_{i=1}^{r_n} w_i - \sum_{i=1}^{I} X_{n,i} * w_i}{\sum_{i=1}^{r_n} w_i - \sum_{i=I-r_n+1}^{I} w_i} \tag{2}$$

$$C_n^* = \frac{\sum_{i=1}^{r_n} \frac{p_i}{N} - \sum_{i=1}^{I} X_{n,i} * \frac{p_i}{N}}{\sum_{i=1}^{r_n} \frac{p_i}{N} - \sum_{i=I-r_n+1}^{I} \frac{p_i}{N}} \tag{3}$$

$$U3_n = \frac{\sum_{i=1}^{r_n} \ln\left(\frac{\frac{p_i}{N}}{1-\frac{p_i}{N}}\right) - \sum_{i=1}^{I} X_{n,i} * \ln\left(\frac{\frac{p_i}{N}}{1-\frac{p_i}{N}}\right)}{\sum_{i=1}^{r_n} \ln\left(\frac{\frac{p_i}{N}}{1-\frac{p_i}{N}}\right) - \sum_{i=I-r_n+1}^{I} \ln\left(\frac{\frac{p_i}{N}}{1-\frac{p_i}{N}}\right)} \tag{4}$$

C* and U3 are sensitive to Guttman errors because the numerator in (3) as well as in (4) gets smaller each time an easier item *i* is answered wrong instead of a more difficult item *j* ($p_i > p_j$, hence, $w_i > w_j$). Both indices take values between 0 and 1. 0, in case of the perfect Guttman scale, and 1, in case of the reversed Guttman scale. The higher the value of C* and U3, the stronger the model underfit. The proportion-correct score for item *i* (*pi/N*) acts as an estimator for the probability that a random person from the population (for which the Rasch model holds) gives a correct response on item *i*. If person-misfit in the sample of *N* respondents systematically distort the estimate *pi/N* (e.g. all *N* respondents cheat on a difficult item k and the estimate of p_k/N is one) detecting person-misfit gets harder or even becomes impossible. One way to address this problem, particularly in cases where *N* is small compared to the population size, is to replace the proportion-correct scores with other estimators (e.g. the proportion-correct scores from another (bigger) sample from the past). In particular, the proportion-correct score is not adequate if some respondents did not answer some of the items as, for instance, in adaptive test settings.

Ht (non-parametric)

Ht, another non-parametric person-fit index, was proposed by Sijtsma (1986). Just like C* and U3 this index is sensitive to Guttman errors. Let us once again rank respondents (increasingly) according to their raw score r_n. Equation (5) then gives the index value for person *n*, with $r_{n,m}$ being the sum of all items where person *n* and person *m* both answered correctly.

$$H_n^T = \frac{\sum_{n \neq \mathrm{m}} \left(\frac{r_{n,m}}{I} - \frac{r_n * r_m}{I^2}\right)}{\sum_{n>m} \left(\frac{r_m}{I} - \frac{r_n * r_m}{I^2}\right) + \sum_{n<m} \left(\frac{r_n}{I} - \frac{r_n * r_m}{I^2}\right)} \tag{5}$$

Ht can take values between minus infinity and 1, although negative values can only be obtained by an absurdly high level of person underfit since the denominator is always non-negative and the numerator is the sum of the sample covariances of the responses person *n* (fixed) and person *m* ($m \neq n$) give on a random item of the test. The higher the value of Ht, the stronger the model overfit. The value 1 is only reached if person *n* answered all items correctly that at least one person with a lower raw score has answered

correctly, and all items incorrectly that at least one person with a higher raw score answered incorrectly. Since Ht evaluates person-fit by comparing the response vectors of the respondents, a high proportion of person-misfit in the sample is expected to be problematic and in contrast to C* and U3 it is not possible to modify this index to become more "sample-independent".

OUTFIT and INFIT (parametric)

OUTFIT and INFIT are parametric indices, since they involve an estimation of the item difficulty parameters and the person ability parameters. Both indices are based on the differences between the observed and the expected responses, the so-called residuals. Equation (6) shows how the standardized residual for respondent n and item i is computed. Based on these residuals Wright and Masters proposed the OUTFIT mean squared error (equation (7)) as well as the INFIT mean squared error (equation (8)) (Wright & Masters, 1990). The former is the average of the sum of the squared residuals (i. e. unweighted), while the latter weights the sum of the squared residuals by the variance of the response.

$$z_{ni} = \frac{\left(X_{ni} - \mathrm{E}\left(X_{ni}\right)\right)}{\sqrt{Var\left(X_{ni}\right)}} \tag{6}$$

$$OUTFIT_n = \frac{\sum_{i=1}^{I} {z_{ni}}^2}{N} \tag{7}$$

$$INFIT_n = \frac{\sum_{i=1}^{I} Var\left(X_{ni}\right) * {z_{ni}}^2}{\sum_{i=1}^{I} Var\left(X_{ni}\right)} \tag{8}$$

The index values from the equations (7) and (8) can be standardized with the Wilson-Hilferty transformation. After transformation, OUTFIT and INFIT are asymptotically Student t-distributed with infinite degrees of freedom (i.e. standard normal distributed), if the Rasch model holds. A detailed description of this transformation as well as the computation of the expected values and the variances are given in Wright and Masters work (Wright & Masters, 1990).

High (positive) values of OUTFIT and INFIT correspond to a model underfit, low (negative) values correspond to a model overfit.

Method

Simulation design

For the simulation, R was chosen as the programming language (R Core Team, 2014). Four non-basic R packages were used in this simulation: *PerFit* (Tendeiro, 2015), *eRm* (Mair, Hatzinger & Maier, 2015), *pROC* (Robin et. al., 2011) and *Truncated normal*

distribution (Trautmann et al., 2014). The complete R code including all non-basic functions, the simulation design, the code for the analysis (tables and graphs) and the exact execution of the simulation are available from the author upon request. Furthermore, exact reproducibility was established since binary matrices were generated and stored in a first step, and loaded and analyzed later on.

The *item difficulty parameters* were chosen non-randomly and are equally spaced over the interval [-2.5, 2.5] in order to model a non-adaptive performance test with increasing item difficulty. For 25 items this corresponds to the following difficulties (rounded to three digits): [-2.500, -2.292, -2.083, -1.875, -1.667, -1.458, -1.250, -1.042, -0.083, -0.625, -0.417, -0.208, 0, 0.208, 0.417, 0.625, 0.083, 1.042, 1.250, 1.458, 1.667, 1.875, 2.083, 2.292, 2.500].

The *latent ability of respondents* was chosen randomly according to a truncated normal distribution over the interval [-3, 3] with a mean of 0 and a standard deviation of 1.5.

Four parameters were varied to produce different scenarios. The *number of items* (I) was either 25 or 50, the *number of respondents* (N) was either 100 or 500 and the *proportion of respondents who responded aberrantly* (N_{AR}) was either 0.05 or 0.3. Furthermore eight different types of aberrant response behaviors were generated. The primary focus in developing those types of person-misfit was to model real-life misfit as realistically as possible. *Guessing*, *Cheating 1*, *Cheating 2*, *Careless* produce a model underfit. *Distorting 1* and *Distorting 2* produce a model overfit. *Fatigue 1* and *Fatigue 2* produce small model deviations which are neither exclusively an overfit nor an underfit. Therefore they cannot be detected with Guttman error sensitive indices.

Aberrant response scenarios were generated in the following way: In a first step, for each respondent and each item the probability of a correct response was computed according to equation (1) by plugging in the corresponding person ability and item difficulty. In a second step, respondents were chosen randomly (not necessarily with equal probability) and the respective probability of a correct response to a certain item was altered according to certain rules specified for each type of aberrant response. More technically: The selection procedure followed a random sample without replacement with the size as a product of the number of respondents and the portion of aberrant response (e.g. 500*0.3=150) as well as certain ability depending weights for the respondents. In the final step, response vectors were generated with the realization of [number of items] independent Bernoulli distributed random variables with the probability of a respondent giving a correct response generated in the first two steps.

Types of person-misfit

***Guessing*:** There is no reason to suspect that the ability of a person has a high impact on whether he or she guesses, in case of an item where the answer is not known by him or her, or not. For this reason, respondents were randomly chosen with equal probability. The probabilities for responding correctly were altered in a way that models a multiple choice test which has exactly five wrong and one right answer to each item, thus, proba-

bilities less than 1/6 were replaced by 1/6. Therefore, even respondents with a low ability parameter had a one in six chance to answer the most difficult items right.

***Cheating 1*:** If a person has a low ability he or she has in general more to gain from cheating as someone with a higher ability. Hence, in this scenario, the lower the ability of a respondent, the higher the probability of getting chosen as a cheater. More specifically, the probability of getting chosen decreased in a linear fashion from the respondent with the lowest ability to the respondent with the highest ability. In case of N=100, the respondent with the lowest ability parameter was chosen with a probability of 100/(101*100/2), the respondent with the second lowest ability parameter was chosen with a probability of 99/(101*100/2), and the respondent with the highest ability parameter was chosen with a probability of 1/(101*100/2). In case of N=500, the respective probabilities were 500/(501*500/2), 499/(501*500/2) and 1/(501*500/2).

In past studies, cheating behavior was often modeled by a deterministic imputation of correct responses to some items (e.g. Karabatsos, 2003). Since the act of cheating (e.g. looking stuff up in the internet, copying from the seatmate) seldom guarantees to produce the right answer to an item, a probabilistic model was chosen. For each cheating respondent and each item, probabilities were generated according to a truncated normal distribution on the interval [0.6, 1] with a mean of 0.8 and a standard deviation of 0.1. Whenever these probabilities were greater than their respective probabilities computed according to the Rasch model, the latter were replaced by the former. This procedure of choosing the maximum of those two probabilities is necessary to realistically model real life cheating behavior since we can assume that no one cheats on items for which he or she knows the answer.

***Cheating 2*:** This scenario differed from *Cheating 1* only in the parameters of the truncated normal distribution. The interval was [0.8, 1], the mean 0.9 and the standard deviation 0.1. The act of cheating therefore increased the probability of a correct response even stronger as in *Cheating 1*.

***Careless*:** The Rasch model has the underlying assumption that a person tries his or her best to perform as well as possible on the test and, therefore, careless behavior is to be counted as person-misfit. Sloppy calculations on a power achievement test for math skills, for instance, lead to an underestimation of the latent math trait of interest. Just like in *Guessing* there is good reason to assume that careless behavior is fairly independent of the latent ability[5]. Hence, respondents were chosen randomly with equal probability. The probabilities for correctly responding to the items were then reduced by 20% (i.e. each probability was multiplied by 0.8).

***Distorting 1*:** If someone actively tries to distort the estimation of the latent ability downward without drawing suspicions, he or she most likely gives correct answers to the easiest items and intentionally wrong answers to items with medium difficulty. Here, difficulty is meant as a subjective measure for that particular person. Mathematically,

[5] If the latent ability trait of interest happens to be “accuracy”, “preciseness”, “exactness” or of that sort this assumption is obviously violated.

this subjective estimation of an item's difficulty is the difference between the latent trait of the person and the item difficulty parameter.

Since respondents with a high ability have more room to distort the estimation of their ability downwards, the probability of getting chosen was modeled in an increasing linear fashion from the person with the lowest ability to the person with the highest ability. In case of N=100 the respondent with the lowest ability parameter was chosen with a probability of 1/(101*100/2), the respondent with the second lowest ability parameter was chosen with a probability of 2/(101*100/2) and the respondent with the highest ability parameter was chosen with a probability of 100/(101*100/2).

For each distorting respondent, the probability of a correct response to an item was changed to 0 if the difference of the respondent's ability and the item difficulty was lower than 1.1. This models a person who actively answers all items wrong, where his or her probability of correctly responding is lower than 75%[6]. The response vectors of those respondents tend towards the perfect Guttman scale since the easiest items are answered correctly with a high probability, medium and hard items are answered wrong with (almost) certainty (Remark: In case of a multiple choice test, a person may not be able to answer a difficult item wrong with certainty if he or she does not know the answer to it.). In any case, it is safe to assume that the probability of a correct response is lower than predicted according to the Rasch model if the person tries to answer the item wrong. This aberrant response behavior therefore produces a model overfit.

***Distorting 2*:** The only difference to *Distorting 1* was that the cut off value for the difference of the respondent's ability and the item difficulty was 1.74 instead of 1.1. This mimics a person who actively answers all items wrong, where his or her probability of correctly responding is lower as 85%[7]. The magnitude of distortion is therefore stronger as in *Distorting 1*.

***Fatigue 1*:** Everyone can experience fatigue and no relation between ability and the probability as well as the magnitude of fatigue is assumed in this scenario. Therefore, every respondent had the same probability to get chosen as someone experiencing fatigue. For each of these [$N*N_{AR}$] respondents it was randomly chosen at which item fatigue set in. The start of the fatigue was not before 50% and not after 80% of the items were completed. All items which fulfilled these requirement had equal probabilities of getting selected as the starting point of fatigue (e.g. For I=25 that meant that the items 12, 13, 14, 15, 16, 17, 18, 19, 20 all had a 1/9 probability of getting selected as the starting point).

In this scenario, a sudden performance loss due to fatigue was modeled. The magnitude of the performance loss was a 30% decrease of the probability of a correct response, and

---

[6] $\frac{e^{1.1}}{1+e^{1.1}} =\sim 0.7503$

[7] $\frac{e^{1.74}}{1+e^{1.74}} =\sim 0.8501$

it stayed constant from the starting point (item) to the end of the test. That means that the probabilities computed under the Rasch model were multiplied by 0.7.

***Fatigue 2*:** The selection of respondents experiencing fatigue and the starting point of the fatigue were chosen in exactly the same way as in *Fatigue 1*. Their difference lies in the effect of fatigue on the performance. Instead of a sudden strong performance loss of 30% which stays constant until the end of the test, a smooth decrease of performance was modeled. The progression of fatigue was modeled in a linear fashion. At the starting item the performance loss was 10%, and at the last item it was 50%. For I=25 that meant that the decrease of the probability of solving items 12, 13, 14, 15, 16, 17, 18, 19, 20 was 10, 15, 20, 25, 30, 35, 40, 45, 50 percent respectively, if the twelfth item was chosen as a starting point.

Methods of comparison

Which test is better (in a statistical sense)? This question is, in general, not trivial to answer. The classical Neyman-Pearson test concept searches for the most powerful test for a chosen specificity.

Example 1: If test A detects on average 87% of the cases for which the null hypothesis (H_0) is wrong (sensitivity = 0.87) and test B only 82% (sensitivity = 0.82), with a 5% probability of wrongfully rejecting H_0 (type-I-risk) for both tests, than test A is to be favored over test B if the specificity is chosen to be 0.95. The question whether test A or test B is "better", gets tricky if we further assume that test A has a sensitivity of 0.71 and test B a sensitivity of 0.76 if the type-I-risk happens to be 0.01. If the specificity is chosen to be 0.99, test B is to be favored over test A since it is better at detecting cases in which H_0 is wrong.

Example 2: Assume that test A always has a higher sensitivity than test B for any given level of specificity. In this case, the receiver operator characteristic (ROC) curve of test A always lies above the ROC curve of test B. The ROC curve is a simple two dimensional plot with the specificity on the abscissa and the sensitivity on the ordinate which has its origins in signal detection theory (Petersen, Birdsall, & Fox, 1954). The ROC curve obviously is a non-decreasing function which always lies above the 45 degree line (otherwise the test is worse than random guessing!).

In our first example test A and test B have intersecting ROC curves. One good criterion to assess whether test A or test B performs better overall is to compare the areas under their respective ROC curves. The area under the ROC curve is often called AUC (Hanley, & McNeil, 1982) and it has been shown that it is a linear transformation of the GINI index (Hand, & Till, 2001). Furthermore, the AUC of a test is also equivalent to the Wilcoxon rank-sum statistic as well as the Mann-Whitney test statistic and can be interpreted as the probability that the test will rank a randomly chosen instance where H_0 is incorrect higher as a randomly chosen instance in which H_0 is correct (Hanley, & McNeil, 1982).

Is it fair to conclude that test A is better than test B in the second example since its ROC lies completely above the ROC curve of B, under the additional assumption that both tests are equally hard to conduct? Sadly no, because one additional property of test A is needed, namely the knowledge of the corresponding critical values for each value of specificity. If it is unknown which critical value leads to which specificity and which sensitivity, the test is hard to implement, since it is tough to classify results obtained with a certain critical value. Specificity and sensitivity are always inverse correlated and depending on the situation their importance varies. Because of this possible scenario, the method of comparison of the five indices was twofold:

- The main criterion was the area under the ROC curve (a value between 0.5 and 1).
- Additionally, the specificity and sensitivity for critical values obtained in a presimulation with no aberrant responses satisfying a specificity of 0.95 and 0.99 were computed. This enabled the estimation of the differences between the actual and the nominal specificity values for our indices.

Description of all executed simulations

To answer the questions of interest regarding the performance of the five person-fit indices and the influence of person-misfit on Andersen´s LR test a sequential simulation design was implemented. That means that results obtained in a simulation affect or determine the setup of subsequent simulations. The complete analysis breaks down into three different simulations (Table 1, Table 2 and Table 3). In Simulation A the 0.01, 0.05, 0.95, and 0.99 quantiles were estimated for each test and for each combination of the number of items and the number of respondents. These estimations were used as critical values in Simulation B. For Ht, C* and U3 the theoretical distribution of the index under H_0 (The Rasch model is correct for each person and each item) is not known and, therefore, the estimation of critical values with Simulation A (Table 1) a necessity.

Table 1:
Simulation A - Computation of critical values for the five person-fit indices

<table>
<tr><td></td><td>N</td><td colspan="2">100</td><td colspan="2">500</td></tr>
<tr><td></td><td>I</td><td>50</td><td>25</td><td>50</td><td>25</td></tr>
<tr><td rowspan="5">Test</td><td>Ht</td><td colspan="4" rowspan="5">• 1000 iterations
• At each iteration the empirical quantiles (0.01, 0.05, 0.95, and 0.99) were taken.
• Afterwards those values were averaged over the 1000 iterations in order to be used as critical values in Simulation B.</td></tr>
<tr><td>C*</td></tr>
<tr><td>U3</td></tr>
<tr><td>OUTFIT</td></tr>
<tr><td>INFIT</td></tr>
</table>

OUTFIT and INFIT asymptotically follow a standard normal distribution but the sample sizes in this simulation (100 and 500 respondents) are far from infinite, and using asymptotic quantiles can therefore lead to strong deviations from the expected specificity. Hence, empirically derived critical values were used for OUTFIT and INFIT too. This way of computing specificities and sensitivities is what Rupp calls "best method with highest precision" in his review paper (Rupp, 2013).

In Simulation B (Table 2) for each scenario and each test the following was computed:

- The area under the ROC curve
- The sensitivity and specificity for the respective critical values which should correspond to a specificity of 0.05 obtained via Simulation A.
- The sensitivity and specificity for the respective critical values which should correspond to a specificity of 0.01 obtained via Simulation A.
- Critical values (once again the empirical quantiles were taken) and the sensitivity which correspond to a specificity of 0.05 in the Simulation B.
- Critical values (once again the empirical quantiles were taken) and the sensitivity which correspond to a specificity of 0.01 in the Simulation B.

Table 2:
Simulation B - Computation of the area under the ROC curve, specificities, sensitivities, critical values, and Andersen´s Likelihood-Ratio test in some cases

<table>
<tr><td></td><td>N</td><td colspan="4">100</td><td colspan="4">500</td></tr>
<tr><td></td><td>I</td><td colspan="2">50</td><td colspan="2">25</td><td colspan="2">50</td><td colspan="2">25</td></tr>
<tr><td></td><td>N_{AR}</td><td>5%</td><td>30%</td><td>5%</td><td>30%</td><td>5%</td><td>30%</td><td>5%</td><td>30%</td></tr>
<tr><td rowspan="6">Type of misfit</td><td>Careless</td><td colspan="6" rowspan="4">• 2000 iterations
• Estimation of the area under the ROC curve for each test (Ht, C*, U3, OUTFIT, INFIT).
• Computation of specificity and sensitivity for the respective critical values obtained in the first simulation for each test.
• Computation of the sensitivity and critical values for two levels of specificity (0.95 and 0.99) for each test.</td><td colspan="2" rowspan="4">The same procedure as in the framed box. Additionally the p-value of Andersen´s LR test with a median split of the raw score was computed.</td></tr>
<tr><td>Cheating 1</td></tr>
<tr><td>Cheating 2</td></tr>
<tr><td>Guessing</td></tr>
<tr><td>Distorting 1</td><td colspan="8" rowspan="2">The same procedure as in the framed box, although this time the direction of the five one-sided tests was reversed and the respective critical values from Simulation A were used.</td></tr>
<tr><td>Distorting 2</td></tr>
</table>

In the case of *Distorting 1* and *Distorting 2*, the direction of the five (one-sided) tests was reversed since they tend to produce a model overfit instead of a model underfit. Therefore, not the same critical values from Simulation A were used. In the case of C*, for instance, the higher the value of C*, the stronger the underfit of that person. Hence, in order to obtain a specificity of 0.95, the 0.95 quantile of C* was used as the critical value in case of an aberrant response that produces underfit and the 0.05 quantile in case of overfit. In the case of underfit, H_0 was rejected if the C* value of a person happened to be higher than the respective critical value. In the case of overfit, H_0 was rejected if the C* value happened to be lower than the respective critical value.

In addition, Andersen´s LR test was computed for four scenarios (*Careless*, *Cheating 1*, *Cheating 2*, and *Guessing* with I=25, N=500, and N_{AR}=0.3). In each iteration the respective sample of N respondents was divided into two groups according to a median (the 50% quantile) split of the raw score. The p-value of Andersen´s LR test (for two groups) was then computed and stored.

In Simulation C (Table 3), the specificity of Andersen´s LR test (criterion: median split of the raw score) before and after the removal of suspicious respondents was investigated in eight scenarios, namely *Careless*, *Cheating 1*, *Cheating 2*, and *Guessing* with I=25, N=500 and either N_{AR}=0.3 or N_{AR}=0.05. The result section will show why C* was the index of choice in this simulation. Additionally, the influence of the two non-detectable scenarios *Fatigue 1* and *Fatigue 2* on the specificity of Andersen´s LR test with I=25, N=500 and N_{AR}=0.3 was investigated.

Table 3:
Simulation C - Specificity of Andersen´s Likelihood-Ratio test before and after the removal of suspicious respondents

<table>
<tr><td></td><td>N</td><td colspan="2">500</td></tr>
<tr><td></td><td>I</td><td colspan="2">25</td></tr>
<tr><td></td><td>N_{AR}</td><td>5%</td><td>30%</td></tr>
<tr><td rowspan="5">Type of misfit</td><td>Careless</td><td colspan="2" rowspan="4">• 2000 iterations
• Step 1: The p-value of Andersen´s LR test with two groups generated by the median split of the raw score was computed. Thereby all 500 respondents were used.
• Step 2: Computation of the C* index for each person and removal of suspicious respondents (specificity = 0.95). The number of removed respondents was saved.
• Step 3: Step 1 is repeated for all respondents who were not removed in Step 2.</td></tr>
<tr><td>Cheating 1</td></tr>
<tr><td>Cheating 2</td></tr>
<tr><td>Guessing</td></tr>
<tr><td>Fatigue 1</td><td colspan="2" rowspan="2">• 800 iterations
• The p-value of Andersen´s LR test with two groups generated by the median split of the raw score was computed. Thereby all 500 respondents were used.</td></tr>
<tr><td>Fatigue 2</td></tr>
</table>

Results

Simulation A – Critical values

The estimated quantiles in each cell (Table 4) are the unweighted average, the so-called sample mean, of 1000 empirical quantiles for the respective test in the four different scenarios. These averages were rounded to three digits and taken as the critical values for the respective scenarios und specificities in Simulation B.

The estimations of C* and U3 differ in the third comma digit at max. The quantiles for OUTFIT and INFIT, on the other hand, differ quite strongly from each other. Furthermore, they are far from symmetric and a symmetric distribution (e.g. a Student t-distribution) would be a poor fit. For instance, the 1% quantile for OUTFIT (N=100, I=25) is -1.833, the 99% quantile is 2.529.) The distributions of OUTFIT and INFIT seem to deviate quite strongly from their asymptotic distribution (standard normal).

Table 4:
Estimated quantiles (0.01, 0.05, 0.95 and 0.99) for the five indices in four different scenarios

			Person-fit index				
N	I	Quantile	Ht	C*	U3	OUTFIT	INFIT
100	25	99%	0.723	0.458	0.454	2.529	2.217
		95%	0.648	0.319	0.318	1.601	1.442
		5%	0.236	0.028	0.029	-1.317	-1.690
		1%	0.042	0.003	0.003	-1.833	-2.382
100	50	99%	0.651	0.359	0.358	2.520	2.237
		95%	0.577	0.277	0.276	1.627	1.453
		5%	0.287	0.069	0.070	-1.391	-1.627
		1%	0.179	0.034	0.036	-2.034	-2.426
500	25	99%	0.702	0.419	0.419	2.338	2.075
		95%	0.643	0.319	0.319	1.579	1.416
		5%	0.244	0.032	0.034	-1.323	-1.654
		1%	0.109	0.002	0.002	-1.826	-2.305
500	50	99%	0.631	0.340	0.342	2.315	2.095
		95%	0.572	0.278	0.278	1.586	1.428
		5%	0.292	0.073	0.074	-1.392	-1.591
		1%	0.213	0.041	0.044	-1.963	-2.276

Simulation B - Area under the ROC curve

Table 5 and 6 show that the performance differences between the five indices are quite small. In case of N_{AR}=0.05, Ht, C* and U3 are best in case of underfit, C* and U3 in case of overfit (Table 5). In case of N_{AR}=0.3, Ht is best in case of underfit, C* and U3 in case of overfit (Table 6). C* and U3 perform very much alike. In case of N_{AR}=0.3, C* performs slightly better than U3 in case of underfit and slightly worse in case of overfit. *Guessing* and *Careless* are the hardest to detect for all five indices. No confidence interval estimations are given, since they are very small and every substantial difference in performance (e.g. a 0.02 point difference) can be considered as statistically significant with a type-I-risk of 1%. As will be shown in the next section, it is not necessary to take too close at the performance differences of the five indices due to severe deviations between the nominal and the actual specificity values in the case of Ht, OUTFIT and INFIT.

Person-misfit can be detected better if the test has 50 instead of 25 items and with 500 instead of 100 respondents (Table 5 & 6). There seems to be no interaction between the influence of the number of items and the number of respondents. The effect of the number of items is rather large (about 5-7% increase in area), the effect of the number of respondents quite small (less than 1% increase in area). The influence of the number of items and the number of respondents is pretty much the same for all five tests.

The influence of the number of aberrant responding respondents does depend on the index. In case of underfit, for instance, the performance increases with N_{AR} for INFIT and Ht, decreases for U3, and does not change for OUTFIT as well as C* (Table 5 & 6).

Simulation B - Specificity of the indices

Table 7 shows the actual specificity for each scenario with N_{AR}=0.3 and each index, if the respective critical values for a specificity of 0.95 from Simulation A are taken. The values for C* and U3 lie close to 0.95 in each and every of the 48 scenarios. On the other hand, the values for Ht, OUTFIT and INFIT strongly deviate from 0.95 in many scenarios. Medium sized deviations are marked with one plus sign, strong deviations with two[8]. Ht tends to produce more type-I-errors, since its specificity values are mainly smaller as 0.95. The deviations are the strongest in the scenarios *Cheating 1, Cheating 2, Distorting 1* and *Distorting 2*. In case of *Cheating 1* with N=500 and I=50, Ht only has a specificity of 74.81% which corresponds to a 403.8% increase of the type-I-risk.

In contrast to Ht, OUTFIT and INFIT produce less type-I-errors as expected, since its specificity values are always higher than 0.95. The sensitivity values of these two indices are therefore decreased. Just like in the case of Ht, the deviations are strongest in the scenarios *Cheating 1, Cheating 2, Distorting 1* and *Distorting 2* with specificity levels mostly higher than 0.98 and sometimes even higher than 0.995.

[8] The magnitude of deviation is measured in the relative deviation from the nominal type-I-risk (0.05 in this case). Hence, specificity values of 0.9 and 0.975 are considered equally strong deviations since the former corresponds to a 100% increase, and the later to a 100% decrease of the type-I-risk.

Table 5:
Area under the ROC curve for all scenarios with N_{AR}=0.05 and all five tests

Type of aberrant response	N	I	Person-fit index Ht	C*	U3	OUTFIT	INFIT
Guessing	100	25	0.6127	0.6096	0.6060	0.6121	0.5951
	100	50	0.6424	0.6374	0.6335	0.6474	0.6253
	500	25	0.6098	0.6066	0.6038	0.6125	0.5938
	500	50	0.6414	0.6356	0.6317	0.6495	0.6228
Cheating 1	100	25	0.9302	0.9327	0.9341	0.9133	0.8977
	100	50	0.9786	0.9797	0.9800	0.9525	0.9486
	500	25	0.9347	0.9362	0.9376	0.9160	0.9002
	500	50	0.9812	0.9820	0.9822	0.9580	0.9507
Cheating 2	100	25	0.8931	0.9004	0.9066	0.9074	0.8216
	100	50	0.9549	0.9610	0.9638	0.9366	0.8504
	500	25	0.8936	0.8991	0.9054	0.9055	0.8233
	500	50	0.9567	0.9609	0.9635	0.9391	0.8489
Careless	100	25	0.6675	0.6670	0.6670	0.6671	0.6777
	100	50	0.7267	0.7280	0.7293	0.7233	0.7402
	500	25	0.6762	0.6755	0.6759	0.6758	0.6856
	500	50	0.7309	0.7311	0.7319	0.7261	0.7444
Average in case of underfit:			0.8019	0.8027	0.8033	0.7964	0.7704
Distorting 1	100	25	0.9270	0.9415	0.9414	0.8734	0.9238
	100	50	0.9764	0.9867	0.9873	0.9357	0.9647
	500	25	0.9350	0.9487	0.9495	0.8817	0.9285
	500	50	0.9804	0.9888	0.9904	0.9423	0.9681
Distorting 2	100	25	0.9563	0.9692	0.9671	0.8581	0.9183
	100	50	0.9893	0.9963	0.9953	0.9248	0.9506
	500	25	0.9634	0.9748	0.9742	0.8676	0.9236
	500	50	0.9917	0.9972	0.9974	0.9311	0.9530
Average in case of overfit:			0.9649	0.9754	0.9753	0.9018	0.9413

Table 6:
Area under the ROC curve for all scenarios with N_{AR}=0.3 and all five tests

Type of aberrant response	N	I	Person-fit index Ht	C*	U3	OUTFIT	INFIT
Guessing	100	25	0.6067	0.6010	0.5968	0.6045	0.5877
	100	50	0.6405	0.6312	0.6252	0.6425	0.6172
	500	25	0.6111	0.6048	0.6007	0.6097	0.5910
	500	50	0.6439	0.6337	0.6280	0.6462	0.6198
Cheating 1	100	25	0.9550	0.9351	0.9136	0.9260	0.9257
	100	50	0.9885	0.9796	0.9656	0.9647	0.9691
	500	25	0.9583	0.9373	0.9173	0.9296	0.9298
	500	50	0.9905	0.9821	0.9702	0.9700	0.9729
Cheating 2	100	25	0.9417	0.9026	0.8717	0.9182	0.8783
	100	50	0.9843	0.9661	0.9428	0.9532	0.9176
	500	25	0.9472	0.9061	0.8780	0.9198	0.8828
	500	50	0.9860	0.9667	0.9473	0.9575	0.9224
Careless	100	25	0.6869	0.6741	0.6644	0.6727	0.6790
	100	50	0.7455	0.7288	0.7146	0.7198	0.7354
	500	25	0.6911	0.6775	0.6678	0.6766	0.6828
	500	50	0.7487	0.7307	0.7171	0.7239	0.7384
Average in case of underfit:			0.8204	0.8036	0.7888	0.8022	0.7906
Distorting 1	100	25	0.9234	0.9392	0.9442	0.8656	0.9366
	100	50	0.9743	0.9849	0.9881	0.9340	0.9794
	500	25	0.9288	0.9444	0.9503	0.8752	0.9411
	500	50	0.9780	0.9871	0.9908	0.9417	0.9817
Distorting 2	100	25	0.9513	0.9640	0.9677	0.8504	0.9408
	100	50	0.9869	0.9942	0.9961	0.9268	0.9753
	500	25	0.9581	0.9694	0.9735	0.8650	0.9451
	500	50	0.9899	0.9955	0.9975	0.9366	0.9772
Average in case of overfit:			0.9613	0.9723	0.9760	0.8994	0.9596

Table 7:
Actual specificity values for each test and all scenarios with N_{AR}=0.3, if the respective critical values from Simulation A, that correspond to a specificity of 0.95, are used. One plus sign indicates that the actual type-I-risk deviates by more than 20% from 0.05, two plus signs indicate a deviation by more than 100%

			Person-fit index				
Type of aberrant response	N	I	Ht	C*	U3	OUTFIT	INFIT
Guessing	100	25	0.9477	0.9549	0.9555	0.9665+	0.9683+
	100	50	0.9392	0.9539	0.9542	0.9696+	0.9720+
	500	25	0.9450	0.9516	0.9514	0.9646+	0.9665+
	500	50	0.9376+	0.9507	0.9505	0.9688+	0.9705+
Cheating 1	100	25	0.8817++	0.9530	0.9557	0.9915++	0.9946++
	100	50	0.7609++	0.9526	0.9558	0.9955++	0.9978++
	500	25	0.8749++	0.9489	0.9518	0.9924++	0.9945++
	500	50	0.7481++	0.9486	0.9520	0.9965++	0.9976++
Cheating 2	100	25	0.9112+	0.9530	0.9569	0.9854++	0.9888++
	100	50	0.8498++	0.9525	0.9567	0.9908++	0.9935++
	500	25	0.9067+	0.9482	0.9524	0.9860++	0.9888++
	500	50	0.9376+	0.9507	0.9505	0.9688+	0.9705+
Careless	100	25	0.9410	0.9539	0.9535	0.9731+	0.9762++
	100	50	0.9274+	0.9549	0.9545	0.9793++	0.9807++
	500	25	0.9384	0.9508	0.9503	0.9718+	0.9745+
	500	50	0.9245+	0.9510	0.9499	0.9775++	0.9797++
Average in case of underfit:			0.8982	0.9518	0.9532	0.9799	0.9822
Distorting 1	100	25	0.8290++	0.9497	0.9521	0.9892++	0.9764++
	100	50	0.8115++	0.9479	0.9505	0.9927++	0.9873++
	500	25	0.8240++	0.9482	0.9500	0.9867++	0.9753++
	500	50	0.8044++	0.9458	0.9512	0.9916++	0.9860++
Distorting 2	100	25	0.8298++	0.9483	0.9514	0.9885++	0.9749+
	100	50	0.8143++	0.9432	0.9498	0.9921++	0.9853++
	500	25	0.8258++	0.9472	0.9506	0.9864++	0.9741+
	500	50	0.8063++	0.9410	0.9517	0.9911++	0.9848++
Average in case of overfit:			0.8181	0.9464	0.9509	0.9898	0.9805

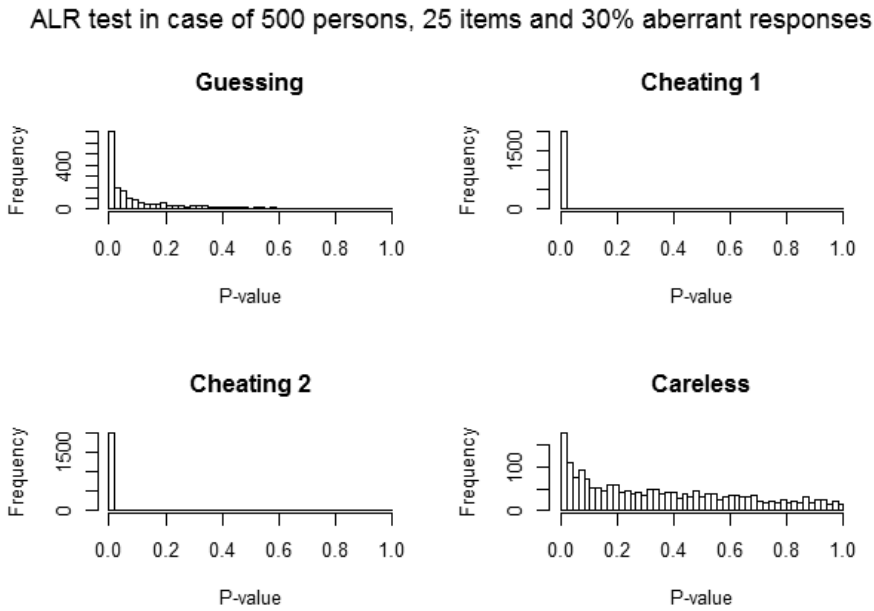
The actual specificity values in case of N_{AR}=0.05 as well as all scenarios where the respective critical values for a specificity of 0.99 from Simulation A are taken can be found in the appendix of this work (Table 9, 10 & 11). In case of a nominal specificity of 0.99 Ht, OUTFIT and INFIT once again do not satisfy the nominal type-I-risk. Ht produces too many type-I-errors, particularly in case of N_{AR}=0.3. Once again some deviations are shockingly high with specificities as low as 92.96% in the case of *Distorting 1* with N=500, I=25 and N_{AR}=0.3. OUTFIT and INFIT produce less type-I-errors, since their specificity values are always higher than 0.99. Just like in the case of Ht, the deviations are strongest in the scenarios *Cheating 1, Cheating 2, Distorting 1* and *Distorting 2*, and in case of N_{AR}=0.3 with specificity levels mostly higher than 0.997 and sometimes even higher than 0.999.

The values for C* and U3 lie close to 0.99 (Appendix: Table 10 & 11), although the precision is somewhat lower than in Table 5. They are a bit too high in underfit scenarios with 100 respondents. In order to check if these small deviations can be linked to the precision of the critical values estimated in Simulation A, the actual critical values are compared with the confidence interval for the respected estimation in Simulation A (Appendix: Table 12). In only two out of 16 scenarios, the critical values from Simulation B for C* as well as for U3 lie within the respective 95% confidence interval. Interestingly, specificity values for C* and U3 seem to be independent of the type and the amount of aberrant response (Appendix: Table 10 & 11). The appendix additionally contains the estimated critical values for all scenarios and all indices (Appendix: Table 13 & 14). The values are systematically influenced by the type of aberrant response and the value of N_{AR} in case of Ht, OUTFIT and INFIT.

Simulation B - Specificity of Andersen´s LR test in case of underfit

Graph 1 shows the distribution of p-values for Andersen´s LR test with two groups generated by a median split of the raw score in case of *Guessing*, *Cheating 1*, *Cheating 2* and *Careless*. Since the data was generated in a way which models a test with 25 items where the Rasch model holds but with 30% aberrant responses, we do not want to reject H_0. Ideally, the p-values of Andersen´s LR test would be equally distributed over the interval [0, 1] as in the case of Rasch model conform data without aberrant response (A detailed analysis of the specificity of Andersen´s LR test was conducted by Futschek (2014)). The stronger the deviation of the actual p-value distribution from the uniform distribution on the interval [0, 1], the stronger the influence of the particular type of aberrant response. The distributions of p-values are extremely right skewed in the case of *Guessing* and *Careless*, and the p-value is essentially zero with probability one in the case of *Cheating 1* and *Cheating 2* (Graph 1).

These are unpleasant results, since we clearly cannot decide whether items are Rasch model conform in the case of respondents that produce a model underfit by showing certain aberrant responses (e.g. cheating). If we allow the probability of the type-I-error of Andersen´s LR test to be 0.05, the actual specificity values (Number of p-values greater than 0.05 divided by number of all p-values) are only 51.4%, 0%, 0% and 84.05% for *Guessing*, *Cheating 1*, *Cheating 2* and *Careless*.



Graph 1:
P-value distribution of Andersen´s Likelihood-Ratio test (criterion: median split of the raw score) in case of underfit

Fortunately, those four types of aberrant responses can be detected fairly well with the index C* (Tables 5 & 6). Simulation C therefore analyzed the potential support of C*, if Rasch model conformity is to be tested with Andersen´s LR test in the case of aberrant responses. Suspicious respondents, with the specificity of C* being set to 0.95, were removed and Andersen´s LR test was computed for the remaining respondents. Because of the strong influence of the aberrant responses on the distribution of the p-values of Andersen´s LR in case of N_{AR}=30%, Simulation C also analyzed *Guessing*, *Cheating 1*, *Cheating 2* and *Careless* in case of N_{AR}=5%.

Simulation C - Specificity of Andersen´s LR test before and after removal of flagged respondents in case of underfit

Table 8 shows the probability of rejecting the null hypotheses with Andersen´s LR for all underfit scenarios with N=500 and I=25. The values in case N_{AR}=0.3 are essentially equal to the results in Simulation B (The difference is 0.003 for *Guessing*, 0 for *Cheating 1* and *2* and 0.0025 for *Careless*). The type-I-risk is strongly elevated even in the case of N_{AR}=0.05 and *Cheating 1* and *2*. The removal of suspicious respondents with the index C* works fairly well, since the probability of a type-I-error is reduced in each and every scenario. One important thing to note is that - after the removal of suspicious re-

spondents - the type-I-risks of Andersen´s LR test are closest to 0.05 for *Cheating 1* and *2* with N_{AR}=0.3, even though these scenarios are the most problematic ones if all 500 respondents are used. This makes perfect sense, since the area under the ROC curve is highest in case of *Cheating 1* and *2,* and therefore most respondents producing aberrant responses were removed[9].

Guessing is the hardest to detect (Graph 3) and, therefore, fewer respondents as in *Cheating 1* and *2* and *Careless* were removed. In the case of 30% respondents with *Guessing* behavior, we still have a 166% (0.133/0.05=2.66) increased type-I-risk after the removal of suspicious respondents. In order to obtain a type-I-risk of 0.05 the specificity level of C* has to be lowered which of course has a downside in case of a true H_0.

Table 8:

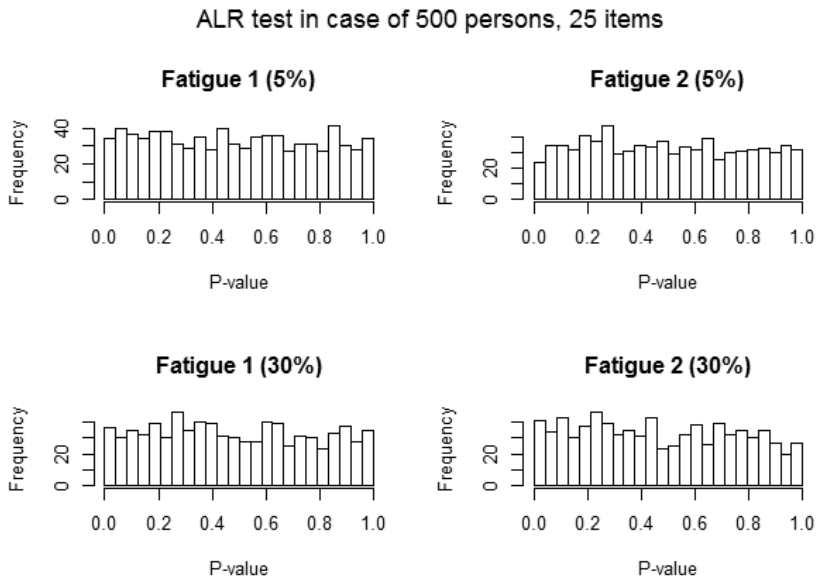
Elevated type-I-risk of Andersen´s Likelihood-Ratio test (criterion: median split of the raw score) in case of N=500 and I=25 before and after the removal of suspicious respondents with C* and the respective critical values from Simulation A corresponding to a specificity of 0.95

N_{AR}	Type of aberrant response	Andersen´s LR test with all respondents	Andersen´s LR test without suspicious respondents	Average number of respondents who were removed
5%	Guessing	0.062	0.052	26.674
	Cheating 1	0.300	0.058	42.962
	Cheating 2	0.128	0.053	40.819
	Careless	0.062	0.054	28.242
30%	Guessing	0.489	0.133	39.102
	Cheating 1	1.000	0.050	136.411
	Cheating 2	1.000	0.040	125.632
	Careless	0.162	0.081	46.838

[9] The average number of respondents who were removed is a product of the actual specificity for a chosen level of specificity and the associated sensitivity for a certain scenario. For instance, the sensitivity for C* in case of *Cheating 1* and N_{AR}=0.3 is 0.7967, the actual specificity (Table 5) 0.9489 for the critical value from Simulation A corresponding to a specificity of 0.95. Therefore, 0.7967*500*0.3 + (1-0.9489)*500*0.7 =137.39 respondents are expected to be removed, which is not far from the actual value: 136.411

Simulation C - Specificity of Andersen´s LR test in case of fatigue 1 and 2.

If people experience fatigue at some point in the test the estimation of their latent ability trait will certainly be too low. The estimation of the Rasch model conformity of the items, however, seems to be unaffected of respondents experiencing fatigue (Graph 2). The distribution of p-values does look uniformly distributed over the interval [0, 1]. Even in the two cases of 30% aberrant response, no deviation from uniformity can be spotted. If we allow the type-I-risk of Andersen´s LR test to be 0.05, the actual specificity values are 94.88%, 96.5%, 94.88% and 94.38% for *Fatigue 1 (N_{AR}=5%), Fatigue 2 (N_{AR}=5%)*, *Fatigue 1 (N_{AR}=30%)* and *Fatigue 2 (N_{AR}=30%)*. It can therefore be concluded, that the type-I-risk is unaffected by this sort of aberrant response.



Graph 2:
P-value distribution of Andersen´s Likelihood-Ratio test (criterion: median split of the raw score) in case of fatigue for N_{AR}=0.05 as well as N_{AR}=0.3

Discussion

The simulated scenarios in this study have a high degree of "realism" and closely model real life phenomena (e.g. cheating). *Guessing* was treated as a person-misfit since there is only one item parameter, namely the item difficulty. Some logistic models include a guessing parameter, particularly the three parameter (3-PL) and the difficulty plus guessing (D+G) parametric logistic models (Birnbaum, 1968; Kubinger & Draxler, 2007). Within these models guessing behavior would not be seen as a person-misfit, and they

are particularly adequate in case of a "1 out of X" multiple choice format. The downside is the loss of the specific objectivity property.

The way *Guessing* and *Careless* were modeled in this simulation is similar to most simulation studies. For a comparison take a look at Rupp´s (2013) review paper in which almost all simulation studies regarding person-fit up to this point are summarized and categorized. Rupp writes that: "However, despite the relatively large array of labels for aberrant responding, there are really only two types of statistical score effects that are effectively created, which are (1) *spuriously low scores* (i.e. when respondents provide a lower score than would be expected based on the chosen model) and (2) *spuriously high scores* (i.e. when respondents provide a higher score than would be expected based on the chosen model)." Although one can easily think of a behavior where the probability for a correct answer rises for some items and decreases for some other items in such a way that the expected number of correct responses corresponds to the expected number given the latent person parameter and the item difficulties, this categorization seems to be a good way not to confuse a certain modeled behavior with its real life counterpart. *Guessing, Cheating 1, Cheating 2, Fatigue 1* and *Fatigue 2* create spuriously high scores while *Careless, Distorting 1* and *Distorting 2* create spuriously low scores.

Cheating 1 and *2* were modeled somewhat different from other simulation studies, but the biggest difference can be found in *Distorting 1* and *Distorting 2*. Karabatsos (2003) modeled *creative examines* by choosing the person parameter from a uniform distribution over the interval [0.5, 2] and imputing incorrect responses for the 18% easiest items. Such a behavior is obviously easy to detect, but the author of this work wonders why such a behavior should occur in real life. Tendeiro and Meijer (2014) modeled *spuriously low scores* by choosing respondents with a person parameter higher than 0.5 and enough correct answers and changing a certain number of randomly chosen correct responses into incorrect ones with a probability of 80%. Once again, it is hard to imagine how such a behavior should arise in real life. If someone wants to distort the estimation of his person parameter downwards in a smart way, he will most likely answer medium (relative to his parameter) and difficult items wrong and easy items correct in order to avoid suspicion. This distorting behavior produces a model overfit instead of an underfit and that may be the reason why such a behavior has not been modeled before.

The findings of this study do differ quite substantially from other simulation studies. On average, Ht had the largest area under the ROC curve, but C* and U3 only performed marginally worse. However, the type-I-risk of Ht depends on the type and the prevalence of aberrant response behavior and in many cases it differs by more than 100% from the nominal type-I-risk. Karabatsos (2003) as well as Zhang and Walker (2008) compared the area under the ROC curve of different indices, but they did not analyze the dependence of the critical values on the type of misfit and the percentage of respondents showing aberrant behavior. In Karabatsos study Ht performed best with the advantage over U3 and C* being bigger as in this study. Dimitrov and Smith (2006), clearly influenced by the work of Karabatsos (2003), also compared Ht with some parametric person-fit indices by estimating the area under the ROC curve. They list the critical values of Ht in different scenarios (number of items, type of aberrant response) corresponding to specificity values of 0.95 and 0.99 in the tables 3 and 4 of their work. However, they do not

discuss the fact that these critical values depend strongly on the type of aberrant response in the case of Ht. One can only wonder if they view their tables as a useful tool to choose the right critical value for a chosen nominal specificity. It is not possible to know how many people in a sample show aberrant response as well as the type of misfit and therefore such a table cannot be used in practice. St-Onge and colleagues (2011) compared the sensitivities of two parametric person-fit indices with U3 and Ht for certain specificity values, namely 0.9, 0.95, and 0.99. They used 100 repetitions for each scenario. In each scenario, which depended among others on the number of items and the type of response, 1000 respondents were simulated and the cut off values for the indices were the respective 1, 5 and 10% empirical quantiles of all respondents who did respond model conform. Sadly, the dependence of the empirical quantiles on the type of misfit and the number of persons with aberrant response behavior was neither examined nor discussed. One work that compared nominal and empirical type-I-error rates for Ht was the simulation study from Tendeiro and Meijer (2014). It reports that the actual type-I-risk for Ht, averaged across all experimental conditions, was 6% for a nominal value of 5% which is nowhere near the magnitude of elevated type-I-risk found in this simulation. Tendeiro and Meijer derived the critical values for a nominal specificity of 0.95 by simulating scores of 10000 respondents without aberrant behavior. This may be problematic, since they later simulated 100 datasets with 1000 respondents (some of them responding aberrant) for each scenario in order to compare the sensitivities of the indices. Taking a quantile in a dataset with 10000 respondents is not the same as averaging the quantiles of 100 datasets containing 1000 respondents.[10]

The removal of suspicious respondents with the index C* and the respective critical values corresponding to a specificity of 0.95 led to a strong increase of the actual specificity of Andersen´s LR test, particularly in the case of *Cheating 1* and *2*. A lower specificity level of C* would result in an even lower type-I-risk of Andersen´s LR test, since additional people with a high number of Guttman errors would be removed[11]. However, when it comes to testing the Rasch model type-II-errors, that is failing to detect a DIF for some item, are of concern as well. Lowering the nominal specificity of C* results in a decreased sensitivity of Andersen´s LR test. This simulation study clearly shows the need to remove respondents with suspicious behavior from the sample and it recommends the use of the index C*. In order to answer the question of the "optimal" specificity level for C*, when using the index as a "screening device" offering support for Andersen´s LR test, further research investigating both types of errors, is of need.

[10] In this simulation we saw that sensitivities were higher for N=500 than for N=100.

[11] If deviations from the perfect Guttman scale come from the aberrant response behavior or from chance does not matter for Andersen´s LR test

References

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, *38*(1), 123-140.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*.

Dimitrov, D. M., & Smith, R. M. (2006). Adjusted Rasch person-fit statistics. *Journal of Applied measurement*, *7*(2), 170.

Emons, W. H., Meijer, R. R., & Sijtsma, K. (2002). Comparing simulated and theoretical sampling distributions of the U3 person-fit statistic. *Applied Psychological Measurement*, *26*(1), 88-108.

Hand, D. J., & Till, R. J. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine learning*, *45*(2), 171-186.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29-36.

Jorge N. Tendeiro (2015). PerFit: Person Fit. R package version 1.3.1

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16*(4), 277-298.

Kubinger, K. D., & Draxler, C. (2007). A comparison of the Rasch model and constrained item response theory models for pertinent psychological test data. In *Multivariate and mixture distribution Rasch models* (pp. 293-309). Springer New York.

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

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*(2), 107-135.

Mokken, R. J. (1971). A theory and procedure of scale analysis: With applications in political research (Vol. 1). Walter de Gruyter.

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests. Grundlagen und Anwendungen.* Bern: Hans Huber.

Futschek, K. (2014). Actual type-I-and type-II-risk of four different model tests of the Rasch model. *Psychological Test and Assessment Modeling*, *56*(2), 168-177.

Harnisch, D. L., & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement*, 133-146.

Heike Trautmann, Detlef Steuer, Olaf Mersmann and Björn Bornkamp (2014). truncnorm: Truncated normal distribution. R package version 1.0-7.

Peterson, W. W., Birdsall, T. G., & Fox, W. C. (1954). The theory of signal detectability. *Information Theory, Transactions of the IRE Professional Group on*, *4*(4), 171-212.

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

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

Rupp, A. A. (2013). A systematic review of the methodology for person fit research in item response theory: Lessons about generalizability of inferences from the design of simulation studies. *Psychological Test and Assessment Modeling*, *55*(1), 3.

Sato, T. (1975). The construction and interpretation of S-P tables. Tokyo: Meiji Tosho.

Sijtsma, K. (1986). A coefficient of deviance of response patterns. *Kwantitatieve Methoden*, *7*(22), 131-145.

St-Onge, C., Valois, P., Abdous, B., & Germain, S. (2011). Accuracy of person-fit statistics: A Monte Carlo study of the influence of aberrance rates. *Applied Psychological Measurement*, 0146621610391777.

Tendeiro, J. N., & Meijer, R. R. (2014). Detection of invalid test scores: The usefulness of simple nonparametric statistics. *Journal of Educational Measurement*, *51*(3), 239-259.

Van Der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology*, *13*(3), 267-298.

Wright, B. D., & Masters, G. N. (1990). Computation of OUTFIT and INFIT Statistics. *Rasch Meas Trans*, *3*(4), 84-85.

Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez and Markus Müller (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12, p. 77. DOI: 10.1186/1471-2105-12-77

Zhang, B., & Walker, C. M. (2008). Impact of missing data on person – model fit and person trait estimation. *Applied Psychological Measurement*, *32*(6), 466-479.

Appendix

Table 9:
Actual specificity values for each test and all scenarios with N_{AR}=0.05, if the respective critical values from Simulation A, that correspond to a specificity of 0.95, are used. A plus sign indicates that the actual type-I-risk deviates by more than 20% from 0.05

			Person fit index				
Type of aberrant response	N	I	Ht	C*	U3	OUTFIT	INFIT
Guessing	100	25	0.9530	0.9547	0.9553	0.9563	0.9563
	100	50	0.9504	0.9548	0.9551	0.9577	0.9571
	500	25	0.9499	0.9512	0.9512	0.9534	0.9541
	500	50	0.9486	0.9509	0.9508	0.9541	0.9547
Cheating 1	100	25	0.9483	0.9551	0.9554	0.9645+	0.9669+
	100	50	0.9418	0.9551	0.9554	0.9679+	0.9695+
	500	25	0.9452	0.9510	0.9509	0.9622+	0.9641+
	500	50	0.9395	0.9510	0.9509	0.9663+	0.9683+
Cheating 2	100	25	0.9512	0.9554	0.9555	0.9608+	0.9616+
	100	50	0.9468	0.9550	0.9552	0.9639+	0.9647+
	500	25	0.9475	0.9511	0.9511	0.9584	0.9600+
	500	50	0.9438	0.9509	0.9509	0.9612+	0.9624+
Careless	100	25	0.9516	0.9548	0.9550	0.9572	0.9585
	100	50	0.9500	0.9550	0.9557	0.9603+	0.9602+
	500	25	0.9493	0.9513	0.9515	0.9547	0.9561
	500	50	0.9473	0.9511	0.9512	0.9564	0.9576
Average in case of underfit:			0.9478	0.9530	0.9532	0.9597	0.9608
Distorting 1	100	25	0.9348+	0.9533	0.9530	0.9626+	0.9595
	100	50	0.9355+	0.9521	0.9524	0.9644+	0.9623+
	500	25	0.9333+	0.9503	0.9499	0.9584	0.9557
	500	50	0.9336+	0.9490	0.9515	0.9618+	0.9592
Distorting 2	100	25	0.9353+	0.9531	0.9531	0.9626+	0.9594
	100	50	0.9355+	0.9514	0.9518	0.9644+	0.9616+
	500	25	0.9350+	0.9509	0.9500	0.9577	0.9550
	500	50	0.9346+	0.9489	0.9508	0.9610+	0.9584
Average in case of overfit:			0.9347	0.9511	0.9516	0.9616	0.9589

Table 10:
Actual specificity values for each test and all scenarios with N_{AR}=0.3, if the respective critical values from Simulation A, that correspond to a specificity of 0.99, are used. One plus sign indicates that the actual type-I-risk deviates by more than 20% from 0.05, two plus signs indicate a deviation by more than 100%

			Person fit index				
Type of aberrant response	N	I	Ht	C*	U3	OUTFIT	INFIT
Guessing	100	25	0.9957+	0.9957+	0.9955+	0.9966+	0.9968+
	100	50	0.9942	0.9951+	0.9948	0.9965+	0.9969+
	500	25	0.9908	0.9913	0.9911	0.9948	0.9946
	500	50	0.9889	0.9909	0.9908	0.9955+	0.9954+
Cheating 1	100	25	0.9944	0.9953+	0.9953+	0.9995++	0.9997++
	100	50	0.9776++	0.9944	0.9947	0.9997++	0.9999++
	500	25	0.9853	0.9906	0.9910	0.9995++	0.9995++
	500	50	0.9522++	0.9900	0.9910	0.9998++	0.9998++
Cheating 2	100	25	0.9947	0.9952+	0.9955+	0.9988++	0.9993++
	100	50	0.9848+	0.9944	0.9947	0.9991++	0.9996++
	500	25	0.9869	0.9902	0.9911	0.9987++	0.9987++
	500	50	0.9699++	0.9897	0.9910	0.9993++	0.9993++
Careless	100	25	0.9956+	0.9956+	0.9953+	0.9974+	0.9979++
	100	50	0.9935	0.9952+	0.9947	0.9980++	0.9983++
	500	25	0.9903	0.9911	0.9908	0.9963+	0.9963+
	500	50	0.9872	0.9910	0.9908	0.9972+	0.9972+
Average in case of underfit:			0.9864	0.9929	0.9930	0.9979	0.9981
Distorting 1	100	25	0.9494++	0.9854	0.9855	0.9994++	0.9971+
	100	50	0.9520++	0.9916	0.9933	0.9995++	0.9987++
	500	25	0.9296++	0.9879	0.9880	0.9991++	0.9956+
	500	50	0.9334++	0.9884	0.9905	0.9990++	0.9979++
Distorting 2	100	25	0.9504++	0.9849+	0.9851	0.9994++	0.9969+
	100	50	0.9531++	0.9908	0.9934	0.9996++	0.9986++
	500	25	0.9322++	0.9878	0.9879	0.9991++	0.9955+
	500	50	0.9343++	0.9875	0.9907	0.9990++	0.9977++
Average in case of overfit:			0.9418	0.9880	0.9893	0.9993	0.9973

Table 11:
Actual specificity values for each test and all scenarios with N_{AR}=0.05, if the respective critical values from Simulation A, that correspond to a specificity of 0.99, are used. One plus sign indicates that the actual type-I-risk deviates by more than 20% from 0.05, two plus signs indicate a deviation by more than 100%

			Person fit index				
Type of aberrant response	N	I	Ht	C*	U3	OUTFIT	INFIT
Guessing	100	25	0.9957+	0.9956+	0.9955+	0.9942	0.9947
	100	50	0.9948	0.9950+	0.9947	0.9940	0.9945
	500	25	0.9912	0.9912	0.9911	0.9917	0.9916
	500	50	0.9905	0.9911	0.9911	0.9919	0.9917
Cheating 1	100	25	0.9953+	0.9954+	0.9952+	0.9960+	0.9961+
	100	50	0.9943	0.9952+	0.9950+	0.9965+	0.9966+
	500	25	0.9908	0.9912	0.9911	0.9941	0.9939
	500	50	0.9892	0.9911	0.9912	0.9948	0.9949
Cheating 2	100	25	0.9956+	0.9956+	0.9954+	0.9954+	0.9954+
	100	50	0.9949	0.9954+	0.9950+	0.9957+	0.9960+
	500	25	0.9908	0.9912	0.9910	0.9934	0.9930
	500	50	0.9899	0.9912	0.9913	0.9938	0.9936
Careless	100	25	0.9955+	0.9953+	0.9951+	0.9944	0.9949
	100	50	0.9948	0.9951+	0.9950+	0.9950+	0.9952+
	500	25	0.9912	0.9913	0.9912	0.9922	0.9921
	500	50	0.9905	0.9911	0.9913	0.9925	0.9923
Average in case of underfit:			0.9928	0.9933	0.9913	0.9941	0.9942
Distorting 1	100	25	0.9891	0.9856	0.9857	0.9947	0.9941
	100	50	0.9907	0.9933	0.9940	0.9959+	0.9955+
	500	25	0.9863	0.9879	0.9880	0.9927	0.9916
	500	50	0.9865	0.9903	0.9907	0.9931	0.9925
Distorting 2	100	25	0.9895	0.9859	0.9860	0.9945	0.9939
	100	50	0.9906	0.9930	0.9935	0.9955+	0.9951+
	500	25	0.9867	0.9879	0.9880	0.9928	0.9914
	500	50	0.9867	0.9903	0.9906	0.9931	0.9924
Average in case of overfit:			0.9883	0.9893	0.9896	0.9940	0.9933

Table 12:
Critical values for C* and U3 indices and each underfit scenario with N=100, which lead to a specificity of 0.99. An asterisk indicates that the value lies outside the 95% confidence interval estimated in Simulation A

I	Type of aberrant response	N_{AR}	U3	C*
25	Guessing	5 %	0.444*	0.449*
		30 %	0.426*	0.428*
	Cheating 1	5 %	0.446*	0.450*
		30 %	0.424*	0.430*
	Cheating 2	5 %	0.446*	0.449*
		30 %	0.427*	0.435*
	Careless	5 %	0.448	0.453
		30 %	0.427*	0.427*
The 95% confidence intervals for the critical values from Simulation A in case of I=25, N=100 and a nominal specificity of 0.99			[0.448, 0.460]	[0.451, 0.646]
50	Guessing	5 %	0.357	0.357
		30 %	0.343*	0.343*
	Cheating 1	5 %	0.354*	0.354*
		30 %	0.345*	0.347*
	Cheating 2	5 %	0.354*	0.354*
		30 %	0.347*	0.350*
	Careless	5 %	0.448*	0.453*
		30 %	0.345*	0.343*
The 95% confidence intervals for the critical values from Simulation A in case of I=50, N=100 and a nominal specificity of 0.99			[0.355, 0.362]	[0.355 0.362]

Table 13:
Comparison of the critical values for each of the five indices and each scenario with N_{AR}=0.3, which lead to a specificity of 0.95, with the respective critical values from Simulation A

			Person fit index				
N	I	Type of aberrant response	Ht	C*	U3	OUTFIT	INFIT
100	25	Guessing	0.2241	0.3211	0.3199	1.4631	1.3078
	25	Cheating 1	0.1717	0.3244	0.3199	0.9051	0.7219
	25	Cheating 2	0.1876	0.3241	0.3180	1.1004	0.9239
	25	Careless	0.2147	0.3230	0.3221	1.3618	1.2036
Simulation A (100 , 25)			0.2360	0.3190	0.3180	1.6010	1.4420
	50	Guessing	0.2726	0.2792	0.2784	1.4443	1.2725
	50	Cheating 1	0.2072	0.2805	0.2769	0.7050	0.5211
	50	Cheating 2	0.2280	0.2808	0.2762	0.9675	0.7798
	50	Careless	0.2650	0.2787	0.2784	1.2910	1.1301
Simulation A (100 , 50)			0.2870	0.2770	0.2760	1.6270	1.4530
500	25	Guessing	0.2355	0.3187	0.3186	1.4163	1.2529
	25	Cheating 1	0.1823	0.3216	0.3183	0.8314	0.6593
	25	Cheating 2	0.1986	0.3226	0.3177	1.0361	0.8618
	25	Careless	0.2269	0.3198	0.3203	1.3126	1.1465
Simulation A (500 , 25)			0.2440	0.3190	0.3190	1.5790	1.4160
	50	Guessing	0.2805	0.2783	0.2785	1.3752	1.2138
	50	Cheating 1	0.2144	0.2800	0.2774	0.6164	0.4486
	50	Cheating 2	0.2350	0.2814	0.2772	0.8828	0.7051
	50	Careless	0.2718	0.2783	0.2790	1.2381	1.0701
Simulation A (500 , 50)			0.2920	0.2780	0.2780	1.5860	1.4280
100	25	Distorting 1	0.7254	0.0259	0.0276	-0.9012	-1.3942
		Distorting 2	0.7243	0.0252	0.0272	-0.9149	-1.4151
Simulation A (100 , 25)			0.6480	0.0280	0.0290	-1.3170	-1.6900
	50	Distorting 1	0.6529	0.0656	0.0681	-0.8315	-1.1020
		Distorting 2	0.6512	0.0633	0.0678	-0.8520	-1.1427
Simulation A (100 , 50)			0.5770	0.0690	0.0700	-1.3910	-1.6270
500	25	Distorting 1	0.7188	0.0307	0.0335	-0.8993	-1.3271
		Distorting 2	0.7164	0.0304	0.0339	-0.9181	-1.3486
Simulation A (500 , 25)			0.6430	0.0320	0.0340	-1.3230	-1.6540
	50	Distorting 1	0.6451	0.0705	0.0741	-0.8122	-1.0317
		Distorting 2	0.6441	0.0683	0.0743	-0.8313	-1.0640
Simulation A (500 , 50)			0.5720	0.0730	0.0740	-1.3920	-1.5910

Table 14:
Comparison of the critical values for each of the five indices and each scenario with N_{AR}=0.05, which lead to a specificity of 0.95, with the respective critical values from Simulation A

			Person fit index				
N	I	Type of aberrant response	Ht	C*	U3	OUTFIT	INFIT
100	25	Guessing	0.2342	0.3192	0.3176	1.5791	1.4282
	25	Cheating 1	0.2273	0.3188	0.3175	1.4699	1.3111
	25	Cheating 2	0.2314	0.3185	0.3171	1.5242	1.3677
	25	Careless	0.2329	0.3190	0.3181	1.5703	1.4082
Simulation A (100 , 25)			0.2360	0.3190	0.3180	0.2360	0.3190
	50	Guessing	0.2845	0.2769	0.2756	1.5911	1.4341
	50	Cheating 1	0.2765	0.2767	0.2756	1.4539	1.2833
	50	Cheating 2	0.2808	0.2766	0.2758	1.5192	1.3453
	50	Careless	0.2834	0.2766	0.2756	1.5606	1.3969
Simulation A (100 , 50)			0.2870	0.2770	0.2760	0.2870	0.2770
500	25	Guessing	0.2427	0.3187	0.3186	1.5487	1.3864
	25	Cheating 1	0.2360	0.3189	0.3187	1.4444	1.2802
	25	Cheating 2	0.2391	0.3188	0.3188	1.4918	1.3255
	25	Careless	0.2419	0.3185	0.3183	1.5353	1.3671
Simulation A (500 , 25)			0.2440	0.3190	0.3190	0.2440	0.3190
	50	Guessing	0.2900	0.2781	0.2781	1.5525	1.3915
	50	Cheating 1	0.2823	0.2779	0.2781	1.4090	1.2456
	50	Cheating 2	0.2858	0.2780	0.2780	1.4719	1.3129
	50	Careless	0.2888	0.2780	0.2779	1.5272	1.3630
Simulation A (500 , 50)			0.2920	0.2780	0.2780	0.2920	0.2780
100	25	Distorting 1	0.6603	0.0281	0.0289	-1.2502	-1.6392
		Distorting 2	0.6605	0.0282	0.0290	-1.2480	-1.6402
Simulation A (100 , 25)			0.6480	0.0280	0.0290	0.6480	0.0280
	50	Distorting 1	0.5890	0.0684	0.0696	-1.3005	-1.5412
		Distorting 2	0.5893	0.0680	0.0692	-1.3021	-1.5513
Simulation A (100 , 50)			0.5770	0.0690	0.0700	0.5770	0.0690
500	25	Distorting 1	0.6557	0.0318	0.0336	-1.2632	-1.6059
		Distorting 2	0.6547	0.0321	0.0336	-1.2678	-1.6118
Simulation A (500 , 25)			0.6430	0.0320	0.0340	0.6430	0.0320
	50	Distorting 1	0.5838	0.0722	0.0744	-1.2981	-1.5046
		Distorting 2	0.5831	0.0722	0.0741	-1.3051	-1.5121
Simulation A (500 , 50)			0.5720	0.0730	0.0740	0.5720	0.0730