universität wien

DIPLOMARBEIT / DIPLOMA THESIS

Titel der Diplomarbeit / Title of the Diploma Thesis

„A simulation study of person fit in the Rasch model“

verfasst von / submitted by

Richard Artner BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Magister der Naturwissenschaften (Mag. rer. nat.)

Wien, 2016 / Vienna, 2016

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	298
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Diplomstudium Psychologie
Betreut von / Supervisor:	Univ.-Prof. i.R. Mag. Dr. Klaus D. Kubinger

Danksagung

Ich danke meinem Diplomarbeitsbetreuer Univ.-Prof. Mag. Dr. Klaus D. Kubinger für das in mich gelegte Vertrauen und die guten Einfälle bezüglich des Simulationsdesign. Vor allem bedanke ich mich für das Halten der besten Lehrveranstaltung die ich in meiner gesamten Studienlaufbahn jemals besucht habe, nämlich das Fachliteraturseminar in psychologischer Diagnostik. Niemals zuvor und niemals danach habe ich jemanden in vergleichbarer Klarheit über statistische Zusammenhänge sprechen hören. Die Erklärungen von Univ.-Prof. Mag. Dr. Klaus D. Kubinger waren niemals komplizierter als unbedingt notwendig und für viele Dinge dich in meinem Statistikstudium zu diesem Zeitpunkt bereits gelernt hatte entwickelte ich aufgrund dieser Lehrveranstaltung ein intuitiveres Verständnis.

Ein großer Dank gilt außerdem Mag. Jan Steinfeld für dessen Unterstützung bei der Programmierung der Simulationsstudie. Ohne den Rat des Mag. Jan Steinfeld hätte ich die Daten nicht in einem ersten Schritt abgespeichert und eine exakte Reproduzierbarkeit wäre somit nicht gewährleistet. Mag. Jan Steinfeld war außerdem so freundlich einen Blick auf meinen R Code zu werfen und alle meine Fragen dazu zufriedenstellend zu beantworten.

Besonderer Dank gilt meiner Familie, die mich immer unterstützte und stets an mich glaubte.

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 one should choose is still not entirely clear, despite the rich body of literature on person fit including numerous simulation studies. 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. This study further analyzed the decrease in Specificity of the Andersen Likelihood-Ratio test, with the median of the raw score as an internal criterion, in case of person misfit and the positive effect of the removal of suspicious persons 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 persons and a higher number of items. Ht, OUFIT, and INFIT show huge deviations between nominal and actual Specificity levels. The results further showed that person misfit has a huge negative impact on the Specificity of the Andersen Likelihood-Ratio test. However, the removal of suspicious persons with C* worked quite good and the nominal Specificity can almost be respected if the Specificity level of C* was set to 0.95.

Keywords: Rasch model, Andersen's Likelihood-Ratio test, Specificity, Sensitivity, power, simulation study, person fit, item response theory, aberrant responding

Abstract – Deutsch

Die Validierung individueller Testergebnisse im Rasch-Modell (1-PL Modell) ist von großer Wichtigkeit. Obschon es viele wissenschaftliche Arbeiten, insbesondere Simulationsstudien, zu diesem Thema gibt, ist es nicht eindeutig welchen Personen-Fit Index man zur Validierung heranziehen soll. Aus diesem Grund wurde eine Simulationsstudie durchgeführt in der fünf bekannte Personen-Fit Indexe anhand ihrer Spezifizität und ihrer Sensitivität in verschiedenen Szenarien miteinander verglichen wurden. Außerdem wurde in dieser Studie die Verringerung der Spezifizität des Likelihood-Ratio-Test nach Andersen, mit dem Median des Rohwerts als internes Teilungskriterium, bei Vorliegen von Personen-Misfit analysiert. Zusätzlich wurde der positive Effekt des Entfernens von verdächtigen Personen mit dem Index C* auf die Spezifizität des Likelihood-Ratio-Tests nach Andersen untersucht. Die drei nicht parametrischen Indexe Ht, C* und U3 schnitten marginal besser ab als die parametrischen Indexe OUTFIT und INFIT. Alle Indexe schnitten mit einer größeren Anzahl an Personen und einer größeren Anzahl an Items besser ab. Personen-Misfit hatte einen starken negativen Einfluss auf die Spezifizität des Likelihood-Ratio-Test nach Andersen. Das Entfernen von verdächtigen Personen mit dem Index C* war sehr effektiv und die nominale Spezifizität konnte beinahe eingehalten werden wenn für die Spezifizität des Index C* 0.95 gewählt wurde.

Schlüsselwörter: Rasch-Modell, Likelihood-Ratio-Test nach Andersen, Spezifizität, Sensitivität, Teststärke, Simulationsstudie, Personen-Fit, Probabilistische Testtheorie

Table of Contents

I. Introduction 1

II. Theoretical part 2

1. Glossary & Notation 2

2. The Rasch model 3

Assumptions and Properties 4

Item fit & Andersen´s Likelihood-Ratio test 6

3. Person fit 6

4. Person Fit Indices 8

C* and U3 8

Ht 8

OUTFIT & INFIT 9

III. Simulation study 10

1. Aim of the study 10

2. Design 10

General style of programming 10

Distribution of model parameters 11

3. Scenarios 11

Guessing 12

Cheating 1 12

Cheating 2 12

Careless 13

Distorting 1 13

Distorting 2 14

Fatigue 1 14

Fatigue 2 14

4. Methods of comparisons 15

5. What exactly was simulated? 17

6. Results 20

Simulation A – Critical values 20

Simulation B - Area under the ROC curve 21
Simulation B - Specificity of the indices 25
Simulation B – Critical values in case of aberrant response 29
Simulation B - Specificity of the ALR test in case of underfit 32
Simulation C - Specificity of the ALR test before and after removal of flagged persons in case of underfit 34
Simulation C - Specificity of the ALR test in case of fatigue 1 and 2 35
Accuracy of the results 36
7. Discussion 42
Types of person misfit 42
Deviations from the nominal Specificity 43
Andersen Likelihood-Ratio test 44
Accuracy of the results 44
8. Summary 45
IV. Directories 46
1. References 46
2. List of tables 48
3. List of figures 48
V. Lebenslauf 50

I. Introduction

Psychometric tests are used in a large array of fields and for various reasons, because they provide some sort of guidance in the assessment of certain characteristics. A test can help in the estimation of some person characteristic (e.g. cognitive skills) or it can support the estimation of the suitability for a specific task, position, job, school or university. Every performance test has a certain way of interpreting the answers given by a person, which always involves some sort of numerical quantification. The biggest class of performance test consists of items, which can be questions or tasks, where the response of a person on each item is quantified. If this quantification is dichotomous a certain response (e.g. “The correct one”) is quantified with 1 and all other possible responses are quantified with 0. In the case of an item which has eight possible answers one of which is correct and the task to select exactly one of these eight answers, selecting the correct item yields a 1 and selecting any of the other seven answers leads to a 0. In the case of a dichotomous quantification we get a response vector the size of the number of items for each person consisting of one´s and zero´s[1]. It may or may not be of interest which particular items are answered correctly, depending on the model frame work (*Theoretical part*).

Each psychometric test claims that there is some sort of correlation between the response vector of a person and the person characteristic of interest. However there are always a multitude of variables who influence the response behavior of a test. A math test in school for instance is seen as a useful tool to assess the mathematical skills of the students in some particular subject area of mathematics. The performance of a student certainly is influenced by his mathematical skills, but it is also influenced by his alertness during the test, his reading skills in case of word problems, and his ability to perform under stress, just to name a view. Furthermore there are certain behaviors that systematically distort the response vector and if undetected lead to wrong assumptions about the characteristic of interest. These behaviors include cheating, guessing, and careless responding, distorting behavior, fatigue and a low level of motivation. We view those behaviors as different types of person misfit and we have a great interest in detecting them.

In this work we try to model certain types of person misfit in the item response theory (latent trait theory) framework and compare the performance of five indices in detecting aberrant responding persons by conducting a simulation study. The theoretical part describes the

[1] Since it does occur that no response is given to a certain item, a third possible value beside zero and one, indicating a missing response, is of need (e.g. -99).

assumptions and properties of the item response theory framework. It further characterizes person misfit in more detail and explain how we can recognize it with the help of the Guttman scale. Lastly the five person fit indices used in the simulation study are discussed in detail.

The second part of the work describes the structure of the simulation and its scenarios, and presents the results. The results are then analyzed and compared with other simulation studies in the literature.

II. Theoretical part

1. Glossary & Notation

ALR test	Andersen´s Likelihood Ratio test (Andersen, 1973)
Cumulative distribution function (CDF)	A function with values between zero and one which gives the probability that a random variable takes a value smaller or equal to the input.
Null hypothesis (H_0)	Assumption(s) about the probabilistic distribution(s) of one or more random variables.
Overfit	A (unrealistic) good fit of the data to a certain model.
P-value	For a given test statistic, it is the probability that its value or an even more extreme value arises, if H_0 is true.
Rasch model	A psychometric probabilistic model with one item parameter and one latent trait parameter for dichotomous data named after the Dane Georg Rasch
Sensitivity	The probability to reject the H_0, if it is wrong (1 - Type-II-risk).
Specificity	The probability of not rejecting H_0, if it is true (1 - Type-I-risk).
Type-I-error	The wrongful rejection of H_0.
Type-II-error	The wrongful maintenance of H_0.
Type-I-risk	The probability of wrongly rejecting H_0.
Type-II-risk	The probability of failing to reject the H_0, if it is wrong.
Underfit	A poor fit of the data to a certain model.
(p – Quantile)	The smallest value of a random variable where the CDF of that value is p; 0<=p<=1

I	Number of items, $\{i = 1,\ldots,I\}$
N	Number of persons, $\{n = 1,\ldots,N\}$
ξ_v	Latent ability trait of person v.
σ_i	Item difficulty parameter for item i.
X_{nj}	The response of person n to item i; $X_{ni} = 1$ is a correct response and $X_{ni} = 0$ is an incorrect response
$X_{n.}$	The response vector of person n, $\{X_{n1}, X_{n2},\ldots,X_{nI}\}$
$X_{.i}$	The response vector of item i, $\{X_{1i}, X_{2i},\ldots,X_{Ni}\}$
r_n	The number of correct responses of person n, the sum of the elements of $X_{n.}$, that is. We call this value the raw score of person n.
p_i	The number of persons with correct responses on item i, the sum of $X_{.i}$ that is.
$r_{n,m}$	$r_{n,m}$ is the sum of all items where persons n & m both answered correctly.

2. The Rasch model

The Rasch model is the most prominent model in the item response theory (IRT). The basic assumption of IRT is that we cannot observe traits of interest directly. Therefore these traits are called latent. What we can do is infer from discrete responses of a person, particularly the answers to the items of a test, to the individual characteristic of the latent ability trait of interest (e.g. In IRT we cannot measure the ability to memorize visual images directly, but we can test how many pictures are recognized correctly and infer to the characteristic of this ability trait.)

In this work we only consider dichotomous IRT models, with only two possible realizations for the answer to a given item, namely right or wrong. In these models, for each person the response to each individual item is a Bernoulli random variable (1 = correct response, 0 = incorrect response). The probability for these random variables is a function of the latent ability trait and one or more item characteristics as parameters.

In the case of the Rasch model, we have only one item parameter, namely the item difficulty parameter. Formula (1) shows the probability function for the random variables[2].

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

If we want to know the probability that person n answers item i correctly, we plug in 1 for x_{ni} and get formula (2). We can easily see, that the probability is smaller than 0.5 if the exponent of the exponential function is smaller than 0. It is important to note, that a centralization of the parameters (i.e. A constant is added or subtracted such that the expected value of the parameter is 0.) leads to an easy interpretation of them.

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

If we want to know the probability that person n answers item i incorrectly, we plug in 0 for x_{ni} and get formula (3).

$$P(-|\xi_n, \sigma_i) = \frac{1}{1 + e^{(\xi_n - \sigma_i)}} \quad (3)$$

Assumptions and Properties

The Rasch model has important assumptions. If we have a test where the Rasch model holds, we can view these assumptions as important properties which allow some auxiliary methods of comparison which wouldn't be valid in the classical test theory. These properties are the reason why a tedious construction of a test, where the Rasch model holds, is often worth the effort. Let us now have a look at these assumptions:

The Rasch model holds, if and only if all items measure the manifestation of the same latent trait (=person parameter) and the probability of a correct response only depends on the latent trait beside the item difficulty. That means, that two persons with the same latent trait have the

[2] The Rasch model can be seen as a log-linear regression for a Bernoulli distributed dependent variable and two metric independent variables (item difficulty, person ability).

same probability for a correct response on each item. This property is often called one-dimensionality.

Furthermore we have so called local stochastic independence for items and persons. That means that the answer to an item has no influence on the answering of subsequent items. More technical, for any person the partial correlation between the answer on item j and the answer on another item k with respect to the person parameter and the item difficulty is zero, for every j and k. This property also holds for the persons. The answers of a person k, does not influence the answers of another person j. So the partial correlation between the answer of person j and the answer of another person k for a certain item with respect to the item difficulty and the person parameters is zero, for every j and k. It can be shown that in case of local stochastic independence the raw score r_n is a sufficient statistic for the person parameter[3]. We can then conclude that two persons have the same person parameter if they both answered k out of I items correctly, even though they answered different items correctly. This conclusion might be invalid if the test was developed according to the classical test theory!

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 according to their difficulty is always the same, for each possible population subgroup (e.g. Item j is harder than item k for every person even though both are easy for someone with a high person parameter and both are hard for someone with a low person parameter.) Furthermore the relative difference in difficulty between two items is the same in each possible population subgroup. If this property doesn´t hold for an item, we have a so called differential item functioning (DIF), a different relative difficulty in two subgroups that is. If there is no DIF for the items of a test, a comparison of the parameter of two persons 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, with different persons answering them that is. The estimation of difficulty and person parameters is therefore sample-independent.

A more detailed analyses and proofs of these properties can be found in Fischer´s work (Fischer, 1974).

[3] A statistic is sufficient in respect to a certain parameter if it contains all important information for a parameter estimation. E.g. If I want to estimate the probability, that a coin lands on its head I just need to know how often it showed tail in my sample (and of course the number of throws) and not the exact sequence of throws. More general: The sum of i.i.d Bernoulli distributed random variables is a sufficient statistic for their parameter, if you know the sample size.

Item fit & Andersen´s Likelihood-Ratio test

So when does the Rasch model apply for a test with I items? It applies if the formula (1) and all assumptions are correct for each and every item. The goodness of fit for individual items is generally called “item fit”. A global test is the (Andersen, 1973). To test whether the items of a test conform the Rasch model or not, the Andersen Likelihood-Ratio (ALR) test for two or more groups is often used. The ALR 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 we can enlarge the Likelihood of our data if we allow for different item difficulty parameters in the subgroups. If the assumption of specific objectivity is violated, these improvement will be higher as in the case of one or more DIF. A more technical and detailed description of the ALR test is omitted in this work and can for example be found in Andersen´s paper (Andersen, 1973).

3. Person fit

Another basic assumption in the Rasch model, not explicitly mentioned in the last chapter is that a person tries his best to solve each item. If the behavior of a person violates one or more assumptions of the Rasch model, we call it “person misfit”. This among others includes, cheating and distorting, inattentive or careless behavior. In order to quantify, specify and measure the type and magnitude of person misfit we have to introduce the Guttman scale (Mokken, 1971), which is named after the Israeli mathematician Louis Guttman.

Instead of a continuous probabilistic model let us now consider a simple deterministic approach. A person answers an item correctly if his person parameter is higher or equal a certain value and answers it incorrectly if his person parameter is lower than that value. In this deterministic model we know that a person 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 1 being the easiest and I being the most difficult item. 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 items follows the perfect Guttman scale and we know that a person answered item k wrong, we can conclude that this person also answered the items k+1 to I wrong.

Let us denote the response vector of a person the Guttman vector of that person. For a four item test we have 2^4=16 possible response Guttman 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 probabilistic Rasch model we obviously cannot expect the response vector of a person to always correspond with the perfekt Guttman scale. That being sad, given a raw score of k, the perfect Guttman scale is the most likely Guttman vector. Let us for example take a Rasch conform test with 10 items and a person with a raw score of 6. Let us take a look at six possible Guttman 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°

It is easy to verify with the help of the formulas (2), and (3) and the property of local stochastic independence, that:

P(V1) > P(V2) > P(V3) > P (V4) > P (V5) > > P (V6) ..... with P standing for Probability.

Without any loss of generality we can conclude that in the Rasch model, “strong” deviations from the perfect Guttman scale are unlikely. If we want to test if the Rasch model holds (itemfit as well as personfit) we can therefore look at the Guttmann vectors of the persons and compare them to the perfect Guttman scale. The greater the difference between the actual person response vectors and the perfekt Guttman scale, the less likely the results. If the actual values are very close to the perfect Guttman scale, we label them as an “overfit”. On the other hand we label strong deviations from the perfect Guttman scale as an “underfit”. If someone talks about person misfit he usually means a model underfit. In this work however overfit and underfit are both seen as a potential model misfit, and it should always be clear of we talk about the former or the later. The non-parametric C* and U3 presented in the next chapter measure the magnitude of deviation from the perfect Guttman scale on a continuous scale.

4. Person Fit Indices

C* and U3

C* and U3 (Flier, 1980 and 1982) are non-parametric person fit indices. Non-parametric means that no model parameters are estimated (item difficulty and person parameter in our case). C* was developed by Harnisch & Linn (1981) and U3 was developed by Flier (1982). They belong to the family of group-based Guttmann error statistics (Meijer & Sijtsma, 2001). Each index in this family satisfies the general equation (4), for some weight w_i. As we can see in formula (5) C* uses the proportion of persons which gave correct answers to item i as the weight for item i. U3 uses a more complicated weight, which includes the natural logarithm (formula (6)).

$$G_n = \frac{\sum_{i=1}^{r_n} w_i - \sum_{i=1}^{I} X_{n,i} * w_i}{\sum_{i=1}^{r_n} w_i - \sum_{i=I-r_n+1}^{I} w_i} \tag{4}$$

$$C_n^* = \frac{\sum_{i=1}^{r_n} \frac{p_i}{N} - \sum_{i=1}^{I} X_{n,i} * \frac{p_i}{N}}{\sum_{i=1}^{r_n} \frac{p_i}{N} - \sum_{i=I-r_n+1}^{I} \frac{p_i}{N}} \tag{5}$$

$$U3_n = \frac{\sum_{i=1}^{r_n} \ln\left(\frac{\frac{p_i}{N}}{1-\frac{p_i}{N}}\right) - \sum_{i=1}^{I} X_{n,i} * \ln\left(\frac{\frac{p_i}{N}}{1-\frac{p_i}{N}}\right)}{\sum_{i=1}^{r_n} \ln\left(\frac{\frac{p_i}{N}}{1-\frac{p_i}{N}}\right) - \sum_{i=I-r_n+1}^{I} \ln\left(\frac{\frac{p_i}{N}}{1-\frac{p_i}{N}}\right)} \tag{6}$$

Both indices take values between 0 and 1. 0 in case of the perfect Guttman score, 1 in case of the reversed Guttman score. The higher the value of C* and U3, the stronger the model underfit.

Ht

Another non-parametric person fit index was proposed by Sijtsma is Ht (Sijtsma, 1986). Let us rank persons (increasingly) according to their total score r_n, formula (7) then gives the index value for person n. This value is the sum of the covariances between person n and all other persons who processed the test, normed by the sum of the maximum of the covariances.

It therefore can take values between minus infinity and 1, although negative values can only be obtained by an absurdly high level of person underfit resulting in negative covariances. The higher the value of Ht the stronger the model overfit. It has been shown, that Ht can take the value zero even in case of a non-perfect Guttman scale vector and therefore it does not belong to the category of group-based Guttmann error statistics, since this statistic cannot be written in the form of equation (4) (Sijtsma, 1986).

$$H_n^T = \frac{\sum_{n \neq \mathrm{m}} \left(\frac{r_{n,m}}{I} - \frac{s_n * s_m}{I^2} \right)}{\sum_{n>m} \left(\frac{r_m}{I} - \frac{s_n * s_m}{I^2} \right) + \sum_{n<m} \left(\frac{r_n}{I} - \frac{s_n * s_m}{I^2} \right)} \tag{7}$$

OUTFIT & INFIT

OUTFIT & 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 residuals. The residual for a person and an items is the difference between the observed and the expected response. Formula (8) shows how the standardized residual for person n and item i is computed. Based on these residuals Wright & Masters proposed the OUTFIT mean squared error (formula (9)) and the INFIT mean squared error (formula (10)) (Wright & Masters, 1990). The former is the average of the sum of the squared residuals (i. e. unweighted), whereas the latter weights the sum of the squared residuals by the variance of the response. The index values from the formulas (9) & (10) can be standardized with the Wilson-Hilferty transformation. With this transformation OUTFIT and INFIT asymptotically follow a student t distributed variable with infinite degrees of freedom (which is equivalent to a normal distributed variable with mean zero and variance one), 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 & Masters work (Wright & Masters, 1990).

$$z_{ni} = \frac{\left(X_{ni} - \mathrm{E}(X_{ni})\right)}{\sqrt{Var(X_{ni})}} \tag{8}$$

$$OUTFIT_n = \frac{\sum_{i=1}^{I} {z_{ni}}^2}{N} \tag{9}$$

$$INFIT_n = \frac{\sum_{i=1}^{I} Var(X_{ni}) * {z_{ni}}^2}{\sum_{i=1}^{I} Var(X_{ni})} \quad (10)$$

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

III. Simulation study

1. Aim of the study

Which index do I take, if I want to analyze potential person misfit. The answer is not entirely clear if we look at past studies in this field. The two main issues regarding past research are the method of comparison between different indices and the way person misfit was operationalized. This study has the purpose to shed (some more) light on the detection skills of certain indices for person misfit in the Rasch model. It also takes a close look at the influence of certain parameters (e.g. number of items, number of persons[4]). It further analyses the need for good indices in a specific problem and the support they can offer in this case.

2. Design

General style of programming

As a programming language R was chosen (R Development Core Team, 2008). It is the most used language in the scientific community when it comes to statistics and psychometrics and it is completely open source, so everyone can use it for free and therefore reproduce results rather easily. Furthermore there exist already written functions (in the form of packages) regarding person fit and the Rasch model which can be used and extended for our purpose.

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

[4] In the following chapter real life concepts like objects and phenomena (e.g. person, 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.

author upon request. Furthermore exact reproducibility is established since binary matrices are generated and stored in a first step, loaded and analyzed later on.

Distribution of model parameters

The item difficulty parameters were chosen nonrandom and equally spaced over the interval [-2.5, 2.5]. For 25 Items we therefore get 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]. With this sequence of item difficulty parameters a non-adaptive performance test with increasing item difficulty is modeled later on.

The latent ability of persons 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.

3. Scenarios

Four parameters were varied to produce different scenarios. The number of items was either 25 or 50, the number of persons was either 100 or 500 and the percentage of persons who responded aberrantly was either 5% or 30%. 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 Gutmann error sensitive indices.

Aberrant response scenarios where generated in the following way. In a first step for each person and each item the probability of a correct response was computed according to the Rasch model and the corresponding person ability and item difficulty. In a second step persons 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 described by the type of aberrant response. More precisely, the selection procedure followed a random sample without replacement with the size as a product of the number of persons and the portion of aberrant response (e.g. 500*0.3=150) and certain ability depending weights for the persons. In the final step response vectors where generated with the realization of (number of items) independent Bernoulli distributed random variables with the probability of a person giving a correct response generated in the first two steps.

Guessing

There is no reason to suspect that the ability of a person has a high impact on whether he guesses if he doesn´t know the answer to an item or not. Therefore persons where chosen randomly with equal probability. The probabilities for responding correctly were altered in a way which models a multiple choice test which has exactly five wrong and one right answer to each item, namely each probability less than 1/6 was replaced by 1/6. Therefore even persons 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/she has in general more to gain from cheating as someone with a higher ability. Therefore in this scenario, the lower the ability of a person the higher the probability of getting chosen as a cheater. More specifically, the probability of getting chosen decreased in a linear fashion from the person with the lowest ability to the person with the highest ability. In the case of 100 persons that means that the person with the lowest ability is twice as likely as the person with the second lowest, three times as likely as the person with the third lowest and 100 times as likely as the person with the highest ability to get chosen as a cheater.

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 person and each item probabilities where 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 where he/she knows the answer.

Cheating 2

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

Careless

Just like in *Guessing* there is good reason to assume that carless behavior is fairly independent of the latent ability[5]. Therefore persons where chosen randomly with equal probability. The Rasch model has the underlying assumption that a person tries his best to perform as good as possible on the test. Sloppy calculations on a power achievement test for math skills for instance lead to an underestimation of the latent math trait of interest. In this scenario the probabilities for correctly responding to the items were 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 most likely gives correct answers to the easiest items and intentional wrong answers to items with medium difficulty. Here difficulty is meant as a subjective measure for that particular person. Mathematically this subjective estimation of an items difficulty is the difference between the latent trait of the person and the item difficulty parameter.

Since persons with a high ability have more room to distort the estimation of their ability downward 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. That means for example, that the person with the lowest ability has half the chance of getting chosen as the one with the second lowest and a third of the chance as the one with the third lowest ability.

For each distorting person the probability of a correct response to an item was changed to 0, if the difference of the person's ability and the item difficulty was lower than 1.1[6]. This models a person who actively answers all items wrong, where his/her probability of correctly responding is lower than 75%. The response vectors of those persons tend towards the perfect Gutmann score 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/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.

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

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

Distorting 2

The only difference to *Distorting 1* is that the cut off value for the difference of the person's ability and the item difficulty is changed to 1.74[7]. This mimics a person who actively answers all items wrong, where his probability of correctly responding is lower as 85%. 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 and magnitude of fatigue is assumed in this scenario. Every person therefore had the same probability to get chosen as someone experiencing fatigue. For each of those aberrant 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 25 items that means 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).

Fatigue 1 is modeled as a sudden performance loss due to fatigue. The magnitude of performance loss stays the same from the starting point to the end of the test. The magnitude of the performance loss was a 30% decrease of the probability of a correct response. That means that the solving probabilities computed under the Rasch model were multiplied by 0.7.

Fatigue 2

The selection of persons experiencing fatigue and the starting point of the fatigue were chosen in exactly the same way as in *Fatigue 1*. The difference lies in the effect of fatigue on the performance. Instead of a sudden strong performance loss of 30% which stayed 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 25 items that means 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.

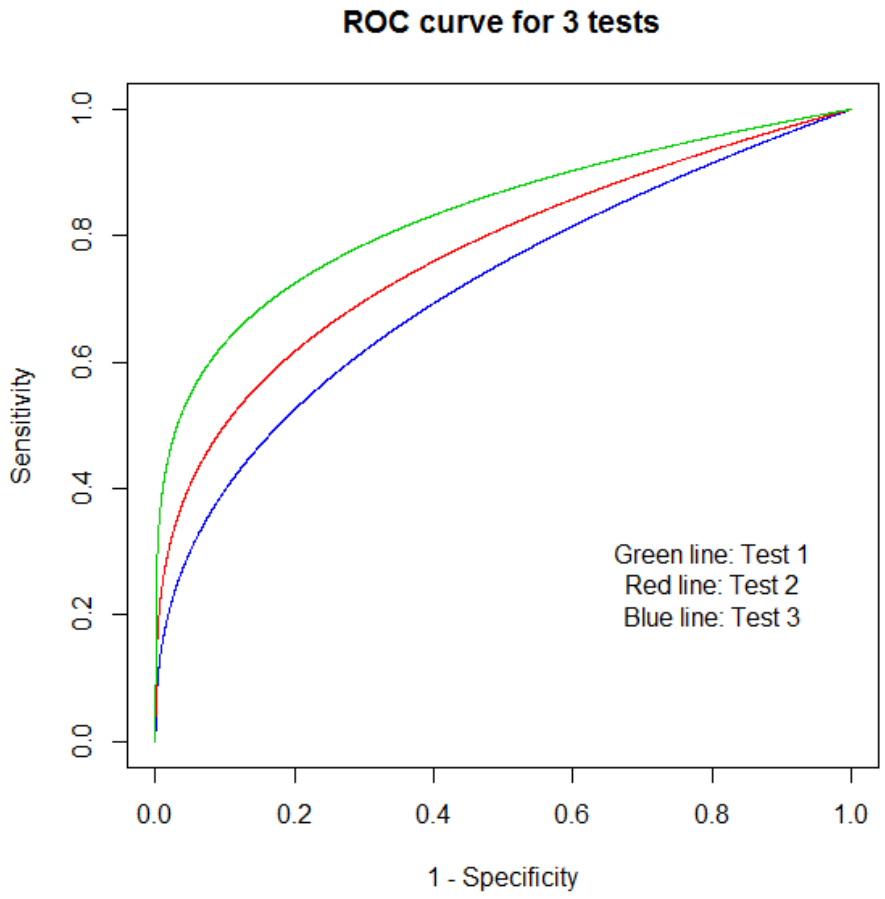
[7] $\frac{e^{1.74}}{1+e^{1.74}} = \sim 0.8501$

4. Methods of comparisons

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 where H0 is wrong (Sensitivity = 0.95), test B only 82% (Sensitivity = 0.82), and if both tests rightfully maintain the H0 on average in 95% of the cases where the H0 is right, than it certainly is better to use test A, if we allow our probability of wrongly rejecting H0 to be 0.05. 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, test B a Sensitivity of 0.76, if we set the Specificity to 0.99. If you only want to reject the H0 wrongly with a probability of 1%, you now should use Test B, since it is better at detecting cases in which the H1 applies.

Example 2: Let us now assume that test A always has a higher Sensitivity than test B if they have the same 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 probability of wrongly rejecting H0 on the abscissa and the Sensitivity on the ordinate. The ROC curve is a non-decreasing function and it always lies above the 45 degree line (otherwise the test is weaker as random guessing!). As an illustration three ROC curve were made up (graph 1). Test 1 has a higher Sensitivity than Test 2 and Test 3 for any given Specificity since the green line always lies above the red and the blue line. The comparison of different tests gets trickier if they have intersecting ROC curves as seen in Example 1.



Graph 1: The ROC curves of three made up tests

Is it now fair to conclude that test A is better than test B since its ROC lies completely above the ROC curve of B, if we further assume that both test 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 invers correlated and depending on the situation there 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 pre-simulation with no aberrant responses satisfying a Specificity of 0.95 and 0.99 were computed. This enables us to estimate how good we can estimate critical values for our tests which satisfy the chosen Specificity regardless of the magnitude and frequency of deviations of H0.

5. What exactly was simulated?

To answer the questions of interest regarding the performance of five person fit indices and the influence of person misfit on a global model 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 where estimated for each test and for each combination of the number of items and the number of persons. These estimations were used as critical values in simulation B. For Ht, C* and U3 the theoretical distribution of the index under the null hypotheses (namely: 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. For OUTFIT and INFIT it is claimed that the distribution under the null hypothesis is asymptotically student t distributed with infinite degrees of freedom. Since the sample sizes (100 and 500 persons) are far from infinite using asymptotic quantiles can lead to strong deviations from the expected Specificity. Therefore 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)

<table>
<tr><td></td><td>Persons</td><td colspan="2">100</td><td colspan="2">500</td></tr>
<tr><td></td><td>Items</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) are taken.
• Afterwards those values are averaged over the 1000 iterations and to be used as critical values in the second simulation.</td></tr>
<tr><td>C*</td></tr>
<tr><td>U3</td></tr>
<tr><td>OUTFIT</td></tr>
<tr><td>INFIT</td></tr>
</table>

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

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 are taken) and the Sensitivity which correspond to a Specificity of 0.05 in the Simulation B.
- Critical values (once again the empirical quantiles are taken) and the Sensitivity which correspond to a Specificity of 0.01 in the Simulation B.

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 (e.g. The higher the value of C*, the stronger the underfit of that person. To obtain a Specificity of 0.95 we therefore take the 0.95 quantile as the critical value in case of an aberrant response that produces underfit and the 0.05 quantile in case of an aberrant response that produces overfit. In the case of underfit we reject the null hypothesis if the C* value of a person is higher than the respective critical value, and in the case of overfit we reject the null hypothesis if the C* value of a person is lower than the respective critical value.)

<table>
<tr><td></td><td>Persons</td><td colspan="4">100</td><td colspan="4">500</td></tr>
<tr><td></td><td>Items</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>Deviation</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="5">Type of misfit</td><td>Careless</td><td colspan="4" rowspan="3">• 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="4" rowspan="3">The same procedure as in the framed box. Additionally the p-value of the ALR 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><td colspan="4" rowspan="2">The same procedure as in the framed box, although this time the direction of the five one-sided tests is reversed and the respective critical values from the first simulation are used.</td></tr>
<tr><td>Distorting 1</td></tr>
<tr><td></td><td>Distorting 2</td><td colspan="8"></td></tr>
</table>

Table 2: Simulation B - Computation of the area under the ROC curve, Specificities, Sensitivities, critical values, and the ALR test in some cases.

For four scenarios (*Careless*, *Cheating 1*, *Cheating 2*, *Guessing* all with 25 items, 500 persons and 30% aberrant responses) additionally an ALR test was computed. For the ALR test each sample was divided in two groups according to a median (the 50% quantile) split of the raw score, and the computed p-value was stored.

In simulation C (table 3) the Specificity of the ALR test (criterion: median split of raw score) before and after the removal of suspicious respondents was investigated in eight scenarios, namely *Careless*, *Cheating 1*, *Cheating 2*, *Guessing* for 500 persons, 25 items, with 5% or 30% aberrant response. We will see in the result section why C* was the index of choice in this simulation. Additionally the influence on the Specificity of the ALR test of the two non-detectable scenarios *Fatigue 1* and *Fatigue 2* with 500 persons, 25 items and 30% aberrant response was investigated.

<table>
<tr><td></td><td>Persons</td><td colspan="2">500</td></tr>
<tr><td></td><td>Items</td><td colspan="2">25</td></tr>
<tr><td></td><td>Deviation</td><td>5%</td><td>30%</td></tr>
<tr><td rowspan="5">Type of misfit</td><td>Careless</td><td colspan="2" rowspan="3">• 2000 iterations
• Step 1: The p-value of the ALR test with two groups generated by the median split of the raw score is computed. Thereby all 500 persons are used.
• Step 2: Computation of the C* index for each person and removal of suspicious persons (Specificity = 0.95). The number of removed persons gets saved.
• Step 3: Step 1 is repeated for all persons 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><td colspan="2" rowspan="2">• 800 iterations
• The p-value of the ALR test with two groups generated by the median split of the raw score is computed. Thereby all 500 persons are used.</td></tr>
<tr><td>Fatigue 1</td></tr>
<tr><td></td><td>Fatigue 2</td><td colspan="2"></td></tr>
</table>

Table 3: Simulation C - Specificity of the ALR test before and after the removal of suspicious respondents.

6. Results

Simulation A – Critical values

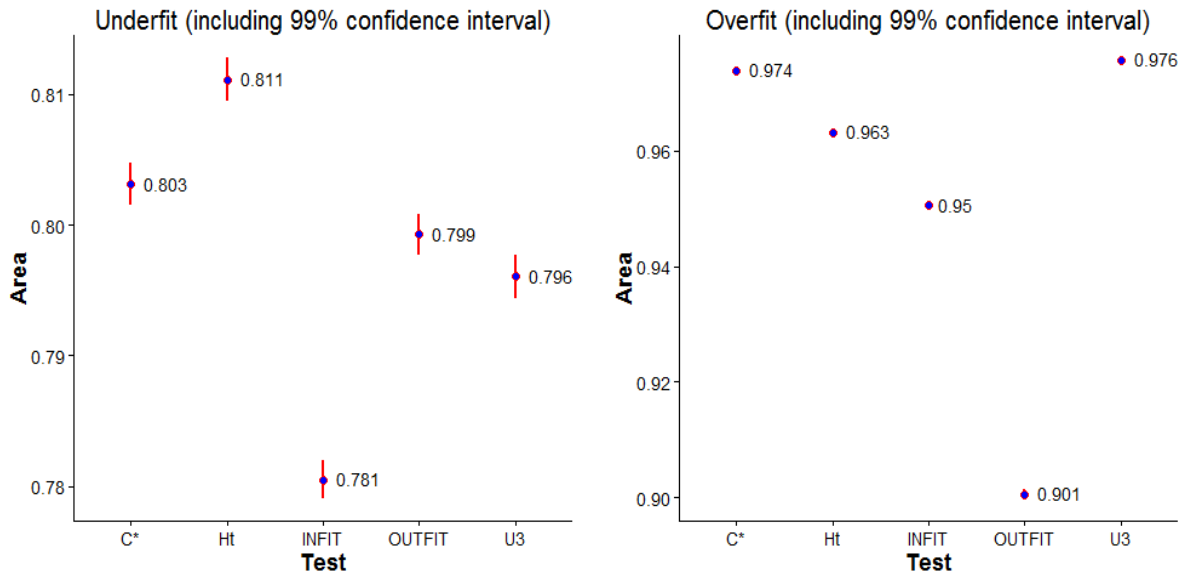
The estimated quantiles in each cell (table 4), are the unweighted average, the so called sample mean, of 1000 empirical quantiles for the respective test and scenario. These averages were rounded to three digits. We can see that, the estimations of C* and U3 differ in the third comma digit at max. A result that was to be expected, since we saw how closely they are related to each other in the theoretical part. The differences between quantiles for OUTFIT and INFIT on the other hand, differ quite strong from each other. Furthermore they are far from symmetric, as they would be if they follow a symmetric distribution (student t distribution in this case). For instance, the 1% quantile for OUTFIT (100 persons, 25 items) is -1.833, the 99% quantile is 2.529.) This gives an absolute difference of 0.696, which indicates a strong deviation from the postulated asymptotic distribution (the limit distribution should be a standard normal distribution). The precision of these estimates will be discussed in a later chapter (*Accuracy of the results*). These estimates were taken as the critical values for the respective scenarios und Specificities in Simulation B.

	Test.quantile	100 persons, 25 items	100 persons, 50 items	500 persons, 25 items	500 persons, 50 items
1	Ht.99	0.723	0.651	0.702	0.631
2	Ht.95	0.648	0.577	0.643	0.572
3	Ht.05	0.236	0.287	0.244	0.292
4	Ht.01	0.042	0.179	0.109	0.213
5	C*.99	0.458	0.359	0.419	0.340
6	C*.95	0.319	0.277	0.319	0.278
7	C*.05	0.028	0.069	0.032	0.073
8	C*.01	0.003	0.034	0.002	0.041
9	U3.99	0.454	0.358	0.419	0.342
10	U3.95	0.318	0.276	0.319	0.278
11	U3.05	0.029	0.070	0.034	0.074
12	U3.01	0.003	0.036	0.002	0.044
13	OUTFIT.99	2.529	2.520	2.338	2.315
14	OUTFIT.95	1.601	1.627	1.579	1.586
15	OUTFIT.05	-1.317	-1.391	-1.323	-1.392
16	OUTFIT.01	-1.833	-2.034	-1.826	-1.963
17	INFIT.99	2.217	2.237	2.075	2.095
18	INFIT.95	1.442	1.453	1.416	1.428
19	INFIT.05	-1.690	-1.627	-1.654	-1.591
20	INFIT.01	-2.382	-2.426	-2.305	-2.276

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

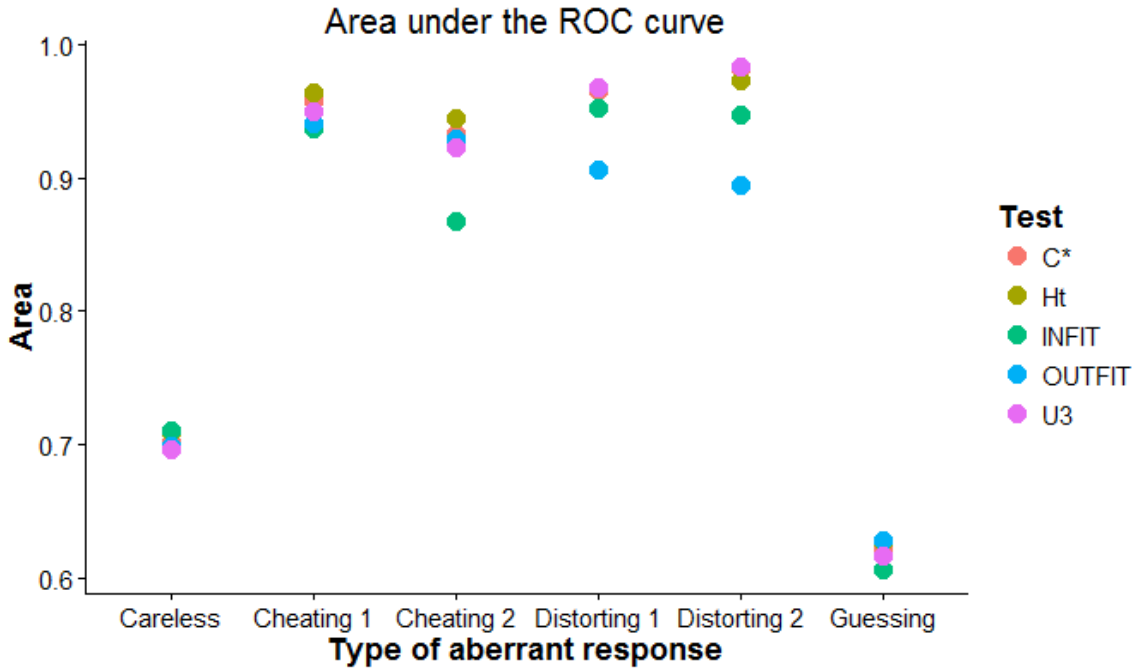
Simulation B - Area under the ROC curve

Let us now have a look at the results of our main criterion of comparison, the area under the ROC curve. Graph 2 shows that Ht has the highest area under the ROC curve and INFIT the lowest in case of underfit. Their estimated difference is quite small (0.03), and the overall performance of these five indices in case of underfit very much alike. In the case of overfit the estimated area of OUTFIT is 0.049 lower as the area of the second worst index INFIT and 0.075 behind the best index U3. C* and Ht are better than OUTFIT and INFIT in the case of underfit as well as in the case of overfit. The ability to detect aberrant responding persons varies less in the case of overfit., since the 99% confidence intervals for the area under the ROC curve are smaller in the case of overfit, even though the underfit estimates come from 64 000 and the overfit estimates from 32 000 iterations.



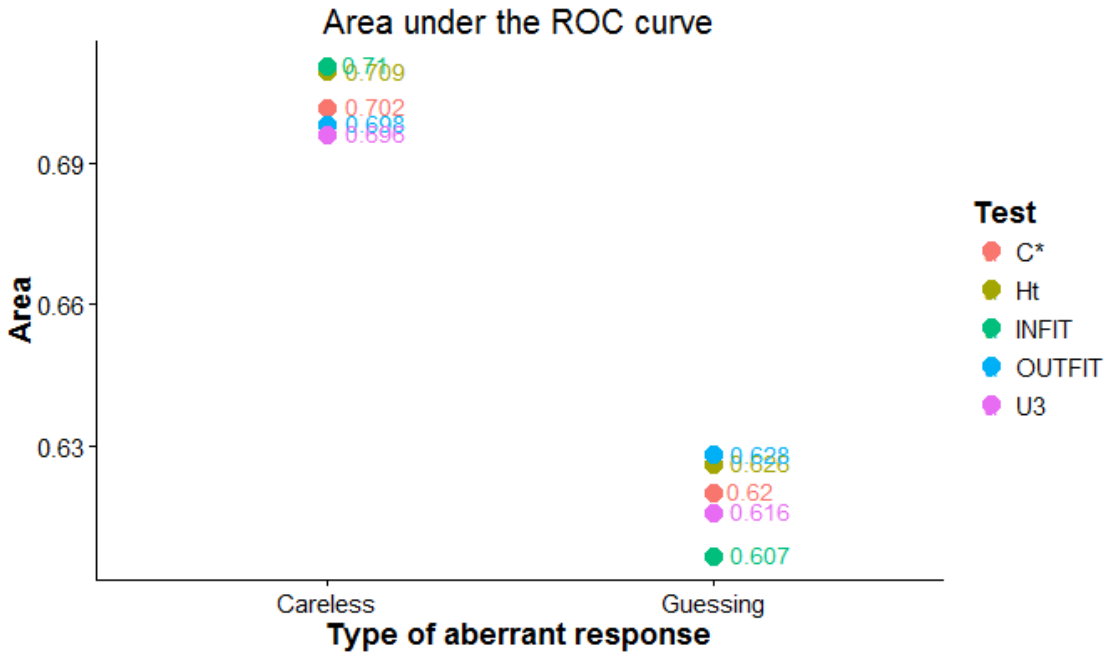
Graph 2: A comparison of five person-fit indices, in their ability to detect aberrant responding examines in case of underfit and overfit, by using the area under the ROC curve as the criterion.

The following graphs show a more detailed analysis for the performance of our five indices. We can see that *Guessing* and *Careless* are the hardest to detect (Graph 3). For a clearer picture of the relative performance of our indices we therefore put them in one graph (Graph 4), and *Cheating 1*, *Cheating 2*, *Distorting 1*, *Distorting 2* in another on (Graph 5).

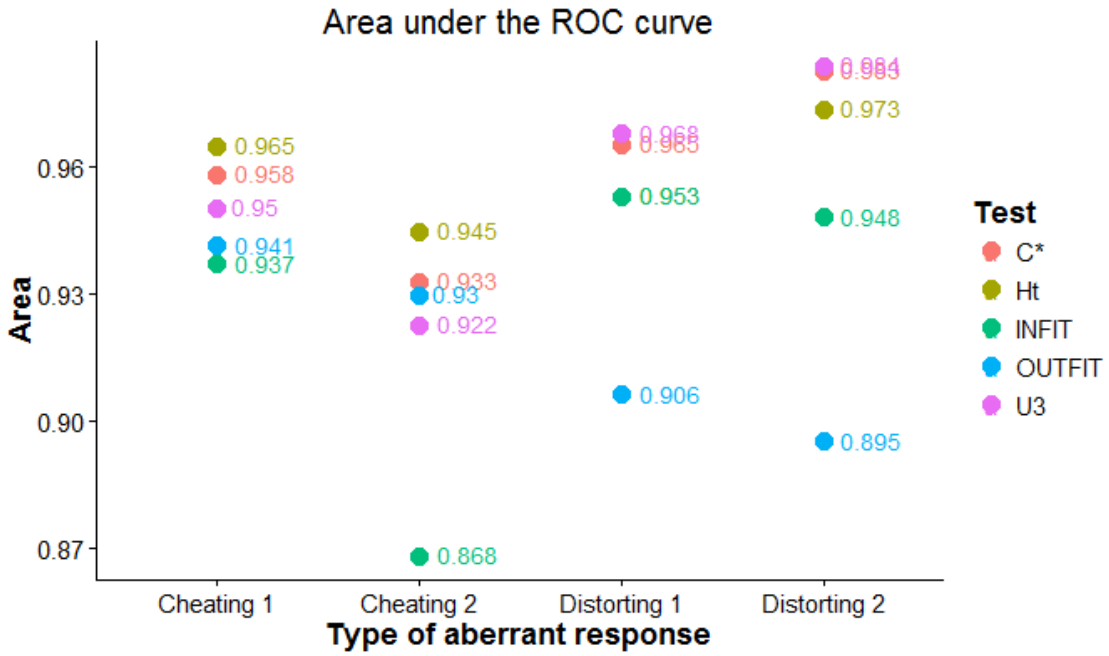


Graph 3: A comparison of five person-fit indices, in their ability to detect careless, cheating, distorting and guessing persons.

The differences in performance are small in the case of *Careless* and *Cheating* (Graph 4). Their estimated area under the ROC curve only varies by 0.014 in the case of *Careless* and 0.021 in the case of *Guessing*.



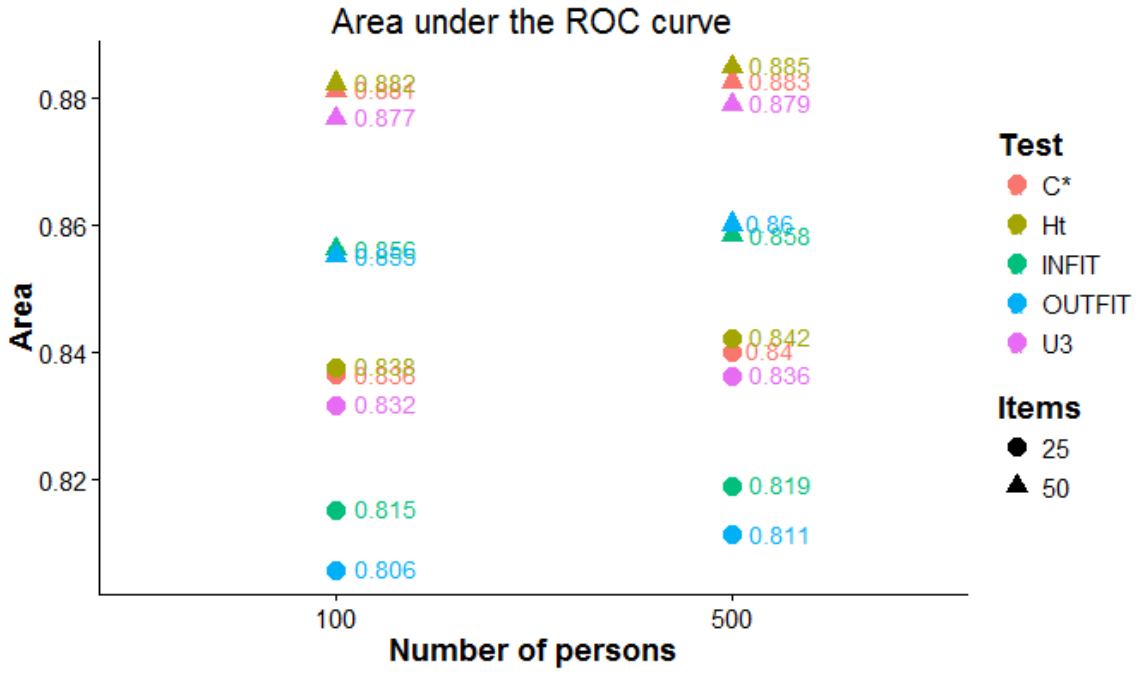
Graph 4: A comparison of five indices, in their ability to detect careless and guessing persons.



Graph 5: A comparison of five indices, in their ability to detect two types of cheating and two types of distorting persons.

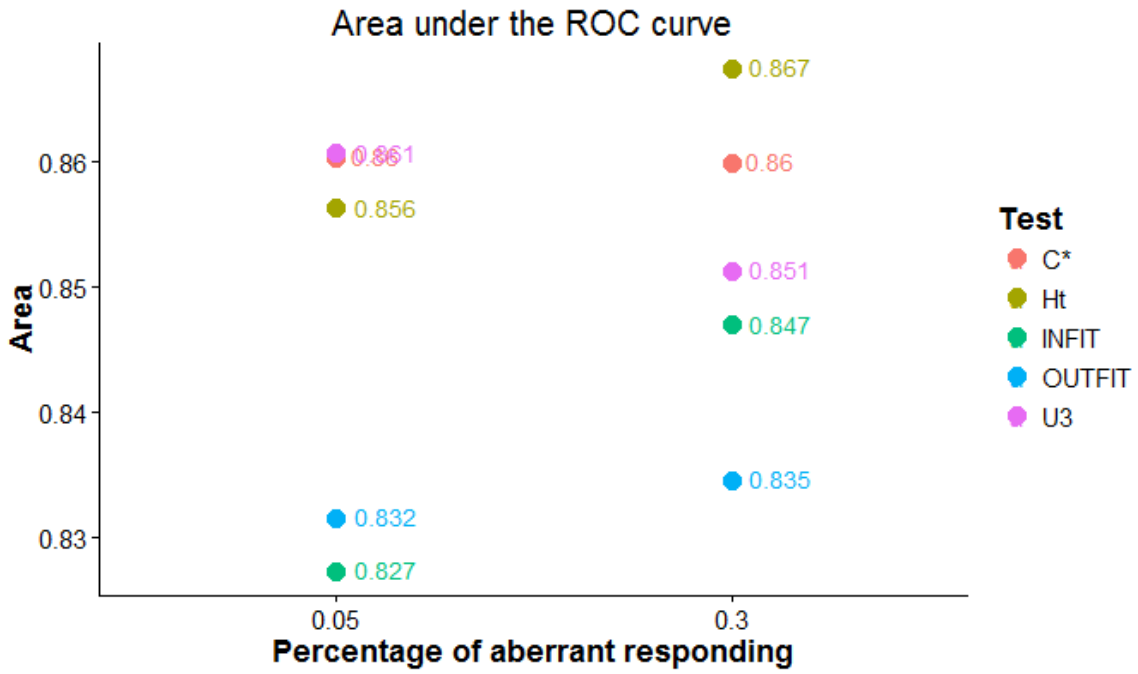
Ht performs best in case of *Cheating 1 & 2*, second best in case of *Distorting 1* and third best in case of *Distorting 2*. OUTFIT performs mediocre in case of *Cheating 1 & 2,* and worst in case of *Distorting 1 & 2* by substantial margins. INFIT in contrast performs mediocre in case of *Distorting 1 & 2,* and worst in case of *Distorting 1 & 2*, with a substantial margin in case of *Cheating 2*. C* performs slightly better than U3 in case of *Cheating 1 & 2*, equally good in case of *Distorting 1*, and slightly worse in case of *Distorting 2*.

The influence of the number of items, the number of persons and the percentage of aberrant responding persons on the ability to detect person misfit can be seen in the following two graphs. We can detect person misfit better if the test has 50 instead of 25 items, since triangles always lie above same colored circles for 100 and 500 persons (Graph 6). The ability to detect person misfit also increases with the number of persons per test, since each unique point given by shape and color is higher for 500 persons than 100 persons. Furthermore there seems to be no interaction between the influence of the number of items and the number of person. The effect of the number of items is quite large (about 5-7% increase in area), the effect of the number of persons quite small (less than 1% increase in area). The influence of the number of items and the number of persons is pretty much the same for all five tests.



Graph 6: A comparison of five indices, in their ability to detect aberrant responding persons in four conditions given by the number of persons and the number of items.

The influence of the percentage of aberrant responding persons on the other hand seems to depend on the index (Graph 7). The performance increases with the percentage of aberrant responding persons for INFIT and Ht, decreases for U3, and stays pretty much the same for OUTFIT and C*.



Graph 7: A comparison of five indices, in their ability to detect aberrant responding persons in two conditions given by the percentage of aberrant responding persons.

Simulation B - Specificity of the indices

Now that we analyzed our main criterion, the area under the ROC curve, we take a look at the adherence of the two chosen Specificity levels 0.95 and 0.99 for each index. Table 5 shows the actual Specificity for each scenario 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. The values for Ht, OUTFIT and INFIT on the other hand strongly deviate from 0.95 in many scenarios. Medium sized deviations are marked yellow, strong deviations are marked red[8]. Ht tends to produce more type-I-errors, since it´s Specificity values are mainly smaller as 0.95. This increased risk of type-I-errors is particularly strong in the case of 30% aberrant responding persons. It is strongest in the scenarios *Cheating 1 & 2* and *Distorting 1 & 2*. Some deviations are shockingly high with Specificities as low as 74.81% in the case of *Cheating 1* with 500 persons, 50 items and 30% aberrant responding persons.

[8] The magnitude of deviation is measured in the relative deviation from the probability for a type-I-error (0.05 in this case). Specificity values of 0.9 and 0.975 are therefore consider equally strong deviations since the former corresponds to a 100% increase, and the later to a 100% decrease of the probability for a type-I-error.

In contrast to Ht, OUTFIT and INFIT produce less type-I-errors, since it´s Specificity values are always higher as 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 & 2* and *Distorting 1 & 2*, and in case of 30% aberrant responding persons. Specificity levels in these case are mostly higher as 0.98 and sometimes even higher as 0.995.

Specificity should be close to 0.95	Scenario: Type of misfit. Number of persons. Number of items. Proportion of aberrant responding persons	Ht	C*	U3	OUTFIT	INFIT
1	Guessing.100.25.0.05	0.9530	0.9547	0.9553	0.9563	0.9563
2	Cheating1.100.25.0.05	0.9483	0.9551	0.9554	0.9645	0.9669
3	Cheating2.100.25.0.05	0.9512	0.9554	0.9555	0.9608	0.9616
4	Careless.100.25.0.05	0.9516	0.9548	0.9550	0.9572	0.9585
5	Guessing.500.25.0.05	0.9499	0.9512	0.9512	0.9534	0.9541
6	Cheating1.500.25.0.05	0.9452	0.9510	0.9509	0.9622	0.9641
7	Cheating2.500.25.0.05	0.9475	0.9511	0.9511	0.9584	0.9600
8	Careless.500.25.0.05	0.9493	0.9513	0.9515	0.9547	0.9561
9	Guessing.100.50.0.05	0.9504	0.9548	0.9551	0.9577	0.9571
10	Cheating1.100.50.0.05	0.9418	0.9551	0.9554	0.9679	0.9695
11	Cheating2.100.50.0.05	0.9468	0.9550	0.9552	0.9639	0.9647
12	Careless.100.50.0.05	0.9500	0.9550	0.9557	0.9603	0.9602
13	Guessing.500.50.0.05	0.9486	0.9509	0.9508	0.9541	0.9547
14	Cheating1.500.50.0.05	0.9395	0.9510	0.9509	0.9663	0.9683
15	Cheating2.500.50.0.05	0.9438	0.9509	0.9509	0.9612	0.9624
16	Careless.500.50.0.05	0.9473	0.9511	0.9512	0.9564	0.9576
17	Guessing.100.25.0.3	0.9477	0.9549	0.9555	0.9665	0.9683
18	Cheating1.100.25.0.3	0.8817	0.9530	0.9557	0.9915	0.9946
19	Cheating2.100.25.0.3	0.9112	0.9530	0.9569	0.9854	0.9888
20	Careless.100.25.0.3	0.9410	0.9539	0.9535	0.9731	0.9762
21	Guessing.500.25.0.3	0.9450	0.9516	0.9514	0.9646	0.9665
22	Cheating1.500.25.0.3	0.8749	0.9489	0.9518	0.9924	0.9945
23	Cheating2.500.25.0.3	0.9067	0.9482	0.9524	0.9860	0.9888
24	Careless.500.25.0.3	0.9384	0.9508	0.9503	0.9718	0.9745
25	Guessing.100.50.0.3	0.9392	0.9539	0.9542	0.9696	0.9720
26	Cheating1.100.50.0.3	0.7609	0.9526	0.9558	0.9955	0.9978
27	Cheating2.100.50.0.3	0.8498	0.9525	0.9567	0.9908	0.9935
28	Careless.100.50.0.3	0.9274	0.9549	0.9545	0.9793	0.9807
29	Guessing.500.50.0.3	0.9376	0.9507	0.9505	0.9688	0.9705
30	Cheating1.500.50.0.3	0.7481	0.9486	0.9520	0.9965	0.9976
31	Cheating2.500.50.0.3	0.8443	0.9470	0.9523	0.9916	0.9933
32	Careless.500.50.0.3	0.9245	0.9510	0.9499	0.9775	0.9797
33	Distorting1.100.25.0.05	0.9348	0.9533	0.9530	0.9626	0.9595
34	Distorting2.100.25.0.05	0.9353	0.9531	0.9531	0.9626	0.9594
35	Distorting1.500.25.0.05	0.9333	0.9503	0.9499	0.9584	0.9557

Specificity should be close to 0.95	Scenario: Type of misfit. Number of persons. Number of items. Proportion of aberrant responding persons	Ht	C*	U3	OUTFIT	INFIT
36	Distorting2.500.25.0.05	0.9350	0.9509	0.9500	0.9577	0.9550
37	Distorting1.100.50.0.05	0.9355	0.9521	0.9524	0.9644	0.9623
38	Distorting2.100.50.0.05	0.9355	0.9514	0.9518	0.9644	0.9616
39	Distorting1.500.50.0.05	0.9336	0.9490	0.9515	0.9618	0.9592
40	Distorting2.500.50.0.05	0.9346	0.9489	0.9508	0.9610	0.9584
41	Distorting1.100.25.0.3	0.8290	0.9497	0.9521	0.9892	0.9764
42	Distorting2.100.25.0.3	0.8298	0.9483	0.9514	0.9885	0.9749
43	Distorting1.500.25.0.3	0.8240	0.9482	0.9500	0.9867	0.9753
44	Distorting2.500.25.0.3	0.8258	0.9472	0.9506	0.9864	0.9741
45	Distorting1.100.50.0.3	0.8115	0.9479	0.9505	0.9927	0.9873
46	Distorting2.100.50.0.3	0.8143	0.9432	0.9498	0.9921	0.9853
47	Distorting1.500.50.0.3	0.8044	0.9458	0.9512	0.9916	0.9860
48	Distorting2.500.50.0.3	0.8063	0.9410	0.9517	0.9911	0.9848

Table 5: Actual Specificity values for each test and each scenario, if the respective critical values from simulation A, that correspond to a Specificity of 0.95, are used.

Table 6 shows the actual Specificity for each scenario and each index, if the respective critical values for a Specificity of 0.99 from Simulation A are taken. The values for C* and U3 lie close to 0.99, although the precision is somewhat lower as in table 5. They are a little bit too high in scenarios with 100 persons. These small deviations may be linked to the precision of the critical values estimated in Simulation A (See chapter *Accuracy of the results* for a closer look). The type of aberrant response and the percentage of person misfit do not seem to influence the accuracy of the Specificity values.

The values for Ht, OUTFIT and INFIT strongly deviate from the specified Specificity of 0.99 in many scenarios. Just as in the case of a Specificity value of 0.95, Ht tends to produce more type-I-errors, and the this deviation is particularly strong in the case of 30% aberrant responding persons, and in the scenarios *Cheating 1 & 2* and *Distorting 1 & 2*. Some deviations are once again shockingly high with Specificities as low 92.96% in the case of *Distorting 1* with 500 persons, 25 items and 30% aberrant responding persons. OUTFIT and INFIT produce less type-I-errors, since it´s Specificity values are always higher as 0.99. Just like in the case of Ht, the deviations are strongest in the scenarios *Cheating 1 & 2* and *Distorting 1 & 2*, and in case of 30% aberrant responding persons. Specificity levels in these case are mostly higher as 0.997 and sometimes even higher as 0.999.

Specificity should be close to 0.99	Scenario: Type of misfit. Number of persons. Number of items. Proportion of aberrant responding persons	Ht	C*	U3	OUTFIT	INFIT
1	Guessing.100.25.0.05	0.9957	0.9956	0.9955	0.9942	0.9947
2	Cheating1.100.25.0.05	0.9953	0.9954	0.9952	0.9960	0.9961
3	Cheating2.100.25.0.05	0.9956	0.9956	0.9954	0.9954	0.9954
4	Careless.100.25.0.05	0.9955	0.9953	0.9951	0.9944	0.9949
5	Guessing.500.25.0.05	0.9912	0.9912	0.9911	0.9917	0.9916
6	Cheating1.500.25.0.05	0.9908	0.9912	0.9911	0.9941	0.9939
7	Cheating2.500.25.0.05	0.9908	0.9912	0.9910	0.9934	0.9930
8	Careless.500.25.0.05	0.9912	0.9913	0.9912	0.9922	0.9921
9	Guessing.100.50.0.05	0.9948	0.9950	0.9947	0.9940	0.9945
10	Cheating1.100.50.0.05	0.9943	0.9952	0.9950	0.9965	0.9966
11	Cheating2.100.50.0.05	0.9949	0.9954	0.9950	0.9957	0.9960
12	Careless.100.50.0.05	0.9948	0.9951	0.9950	0.9950	0.9952
13	Guessing.500.50.0.05	0.9905	0.9911	0.9911	0.9919	0.9917
14	Cheating1.500.50.0.05	0.9892	0.9911	0.9912	0.9948	0.9949
15	Cheating2.500.50.0.05	0.9899	0.9912	0.9913	0.9938	0.9936
16	Careless.500.50.0.05	0.9905	0.9911	0.9913	0.9925	0.9923
17	Guessing.100.25.0.3	0.9957	0.9957	0.9955	0.9966	0.9968
18	Cheating1.100.25.0.3	0.9944	0.9953	0.9953	0.9995	0.9997
19	Cheating2.100.25.0.3	0.9947	0.9952	0.9955	0.9988	0.9993
20	Careless.100.25.0.3	0.9956	0.9956	0.9953	0.9974	0.9979
21	Guessing.500.25.0.3	0.9908	0.9913	0.9911	0.9948	0.9946
22	Cheating1.500.25.0.3	0.9853	0.9906	0.9910	0.9995	0.9995
23	Cheating2.500.25.0.3	0.9869	0.9902	0.9911	0.9987	0.9987
24	Careless.500.25.0.3	0.9903	0.9911	0.9908	0.9963	0.9963
25	Guessing.100.50.0.3	0.9942	0.9951	0.9948	0.9965	0.9969
26	Cheating1.100.50.0.3	0.9776	0.9944	0.9947	0.9997	0.9999
27	Cheating2.100.50.0.3	0.9848	0.9944	0.9947	0.9991	0.9996
28	Careless.100.50.0.3	0.9935	0.9952	0.9947	0.9980	0.9983
29	Guessing.500.50.0.3	0.9889	0.9909	0.9908	0.9955	0.9954
30	Cheating1.500.50.0.3	0.9522	0.9900	0.9910	0.9998	0.9998
31	Cheating2.500.50.0.3	0.9699	0.9897	0.9910	0.9993	0.9993
32	Careless.500.50.0.3	0.9872	0.9910	0.9908	0.9972	0.9972
33	Distorting1.100.25.0.05	0.9891	0.9856	0.9857	0.9947	0.9941
34	Distorting2.100.25.0.05	0.9895	0.9859	0.9860	0.9945	0.9939
35	Distorting1.500.25.0.05	0.9863	0.9879	0.9880	0.9927	0.9916
36	Distorting2.500.25.0.05	0.9867	0.9879	0.9880	0.9928	0.9914
37	Distorting1.100.50.0.05	0.9907	0.9933	0.9940	0.9959	0.9955
38	Distorting2.100.50.0.05	0.9906	0.9930	0.9935	0.9955	0.9951
39	Distorting1.500.50.0.05	0.9865	0.9903	0.9907	0.9931	0.9925
40	Distorting2.500.50.0.05	0.9867	0.9903	0.9906	0.9931	0.9924
41	Distorting1.100.25.0.3	0.9494	0.9854	0.9855	0.9994	0.9971
42	Distorting2.100.25.0.3	0.9504	0.9849	0.9851	0.9994	0.9969

Specificity should be close to 0.99	Scenario: Type of misfit. Number of persons. Number of items. Proportion of aberrant responding persons	Ht	C*	U3	OUTFIT	INFIT
43	Distorting1.500.25.0.3	0.9296	0.9879	0.9880	0.9991	0.9956
44	Distorting2.500.25.0.3	0.9322	0.9878	0.9879	0.9991	0.9955
45	Distorting1.100.50.0.3	0.9520	0.9916	0.9933	0.9995	0.9987
46	Distorting2.100.50.0.3	0.9531	0.9908	0.9934	0.9996	0.9986
47	Distorting1.500.50.0.3	0.9334	0.9884	0.9905	0.9990	0.9979
48	Distorting2.500.50.0.3	0.9343	0.9875	0.9907	0.9990	0.9977

Table 6: Actual Specificity values for each test and each scenario, if the respective critical values from simulation A, that correspond to a Specificity of 0.99, are used.

Tables 5 and 6 show a clear influence of the type of aberrant response and the percentage of average responding persons on the deviation of the chosen Specificity level for Ht, OUFIT and INFIT. Specificity values for C* and U3 on the other seem to be independent of the type and the amount of aberrant response.

Simulation B – Critical values in case of aberrant response

We just saw, that the actual Specificity (and of course the inverse correlated Sensitivity) from Ht, OUTFIT and INFIT deviate from the specified level, and in some scenarios the magnitude of deviation is huge. We now take a look at the critical values (arithmetic mean of the 95% quantiles for each of the 2000 iterations) for each underfit scenario and compare them to the estimated values in Simulation A (Table 7). Each of the same colored columns should contain (almost) the same values. If we look at Ht we can see this is clearly not the case. The critical values for certain scenarios are always smaller as the respective estimation from Simulation A. This makes perfect sense, since we reject the null hypotheses for small values of Ht, and we saw in Table 5, that the actual Specificity levels for Ht are too low. Moreover we can recognize certain patterns. We can see, that the critical values are always smaller in case of 30% as in 5% aberrant response for a given number of persons and items. We also recognize that values for *Cheating 1* are smaller than *Cheating 2,* which again are smaller than values for *Guessing* and *Careless,* for a given number of persons and items. This is of course the explanation, for the size of the deviations in Specificity levels and their relation with the type of underfit and percentage of aberrant responding persons that we have seen in Table 5.

Critical values for a Specificity of 0.95	Scenario: Type of misfit. Number of persons. Number of items. Proportion of aberrant responding persons	Ht	C*	U3	OUTFIT	INFIT
1	**Simulation A (100, 25)**	**0.2360**	**0.3190**	**0.3180**	**1.6010**	**1.4420**
2	Guessing.100.25.0.05	0.2342	0.3192	0.3176	1.5791	1.4282
3	Guessing.100.25.0.3	0.2241	0.3211	0.3199	1.4631	1.3078
4	Cheating1.100.25.0.05	0.2273	0.3188	0.3175	1.4699	1.3111
5	Cheating1.100.25.0.3	0.1717	0.3244	0.3199	0.9051	0.7219
6	Cheating2.100.25.0.05	0.2314	0.3185	0.3171	1.5242	1.3677
7	Cheating2.100.25.0.3	0.1876	0.3241	0.3180	1.1004	0.9239
8	Careless.100.25.0.05	0.2329	0.3190	0.3181	1.5703	1.4082
9	Careless.100.25.0.3	0.2147	0.3230	0.3221	1.3618	1.2036
10	**Simulation A (500, 25)**	**0.2440**	**0.3190**	**0.3190**	**1.5790**	**1.4160**
11	Guessing.500.25.0.05	0.2427	0.3187	0.3186	1.5487	1.3864
12	Guessing.500.25.0.3	0.2355	0.3187	0.3186	1.4163	1.2529
13	Cheating1.500.25.0.05	0.2360	0.3189	0.3187	1.4444	1.2802
14	Cheating1.500.25.0.3	0.1823	0.3216	0.3183	0.8314	0.6593
15	Cheating2.500.25.0.05	0.2391	0.3188	0.3188	1.4918	1.3255
16	Cheating2.500.25.0.3	0.1986	0.3226	0.3177	1.0361	0.8618
17	Careless.500.25.0.05	0.2419	0.3185	0.3183	1.5353	1.3671
18	Careless.500.25.0.3	0.2269	0.3198	0.3203	1.3126	1.1465
19	**Simulation A (100, 50)**	**0.2870**	**0.2770**	**0.2760**	**1.6270**	**1.4530**
20	Guessing.100.50.0.05	0.2845	0.2769	0.2756	1.5911	1.4341
21	Guessing.100.50.0.3	0.2726	0.2792	0.2784	1.4443	1.2725
22	Cheating1.100.50.0.05	0.2765	0.2767	0.2756	1.4539	1.2833
23	Cheating1.100.50.0.3	0.2072	0.2805	0.2769	0.7050	0.5211
24	Cheating2.100.50.0.05	0.2808	0.2766	0.2758	1.5192	1.3453
25	Cheating2.100.50.0.3	0.2280	0.2808	0.2762	0.9675	0.7798
26	Careless.100.50.0.05	0.2834	0.2766	0.2756	1.5606	1.3969
27	Careless.100.50.0.3	0.2650	0.2787	0.2784	1.2910	1.1301
28	**Simulation A (500, 50)**	**0.2920**	**0.2780**	**0.2780**	**1.5860**	**1.4280**
29	Guessing.500.50.0.05	0.2900	0.2781	0.2781	1.5525	1.3915
30	Guessing.500.50.0.3	0.2805	0.2783	0.2785	1.3752	1.2138
31	Cheating1.500.50.0.05	0.2823	0.2779	0.2781	1.4090	1.2456
32	Cheating1.500.50.0.3	0.2144	0.2800	0.2774	0.6164	0.4486
33	Cheating2.500.50.0.05	0.2858	0.2780	0.2780	1.4719	1.3129
34	Cheating2.500.50.0.3	0.2350	0.2814	0.2772	0.8828	0.7051
35	Careless.500.50.0.05	0.2888	0.2780	0.2779	1.5272	1.3630
36	Careless.500.50.0.3	0.2718	0.2783	0.2790	1.2381	1.0701

Table 7: Comparison of the critical values for each of the five indices and each underfit scenario, which lead to a Specificity of 0.95, with the respective critical values from Simulation A.

In case of OUTFIT and INFIT, we reject the null hypotheses for high index values. The actual critical values in each scenario are always lower than their respective values from Simulation A and therefore explain that the actual Specificity levels for OUTFIT and INFIT are too low (Table 5). Just as in the case of Ht, we can clearly recognize the systematic influence of the type of response and the number of aberrant responding persons on the size of the critical value for a given number of persons and items.

If we take a close look at C* and U3, we can see that the critical values are smaller (always in case of C*, almost always in the case of U3) in case of 30% as in 5% aberrant response for a given number of persons and items. We also see that critical values for *Cheating 1 & 2* are a little bit higher as the critical values for *Guessing* and *Careless*, for a given number of persons and items. A small influence of the type and the percentage of aberrant responding persons on the critical values seems to be the case even for C* and U3, although the magnitude does not raise concern as we have seen in table 5.

Table 8 shows the estimated critical values for each overfit scenario. Once again we can spot patterns for Ht, OUTFIT and INFIT which show a clear relatedness between the size of the deviation and the type of aberrant behavior as well as the percentage of aberrant responding persons.

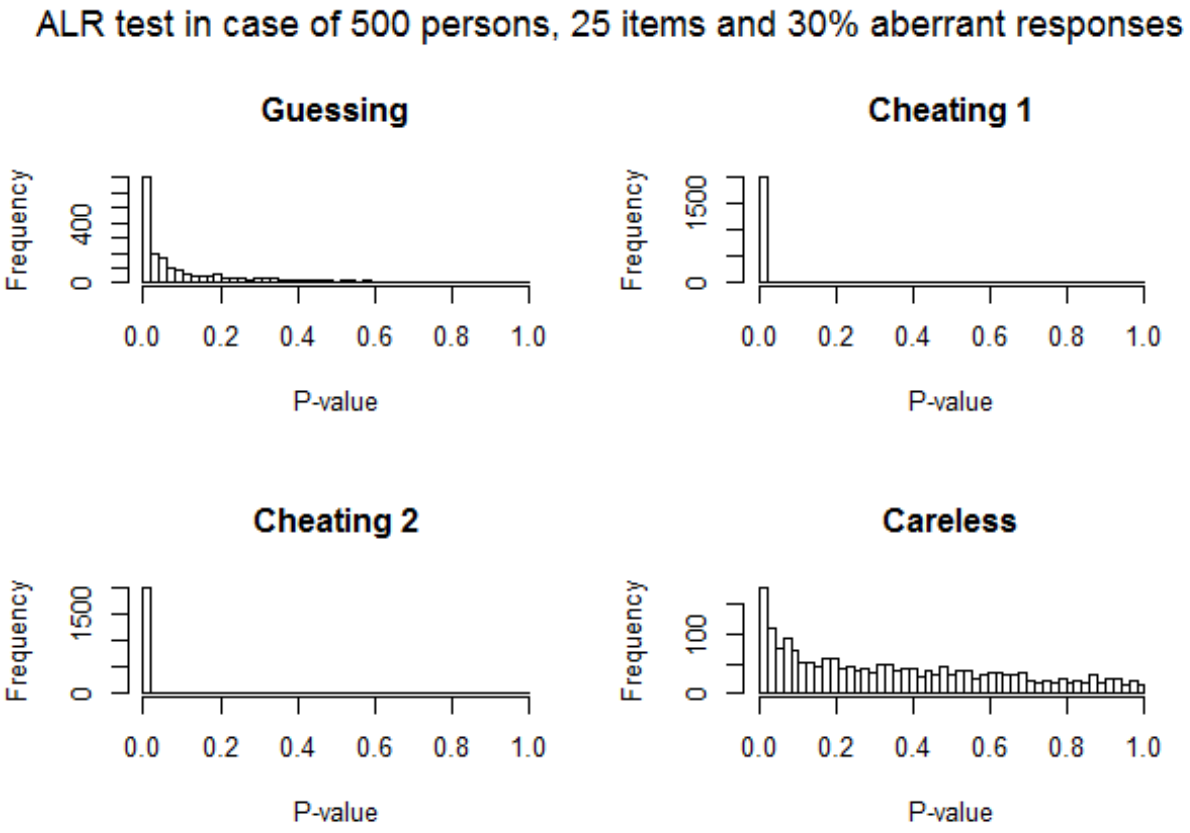
Critical values for a Specificity of 0.95	Scenario: Type of misfit. Number of persons. Number of items. Proportion of aberrant responding persons	Ht	C*	U3	OUTFIT	INFIT
1	**Simulation A (100, 25)**	**0.6480**	**0.0280**	**0.0290**	**-1.3170**	**-1.6900**
2	Distorting1.100.25.0.05	0.6603	0.0281	0.0289	-1.2502	-1.6392
3	Distorting1.100.25.0.3	0.7254	0.0259	0.0276	-0.9012	-1.3942
4	Distorting2.100.25.0.05	0.6605	0.0282	0.0290	-1.2480	-1.6402
5	Distorting2.100.25.0.3	0.7243	0.0252	0.0272	-0.9149	-1.4151
6	**Simulation A (500, 25)**	**0.6430**	**0.0320**	**0.0340**	**-1.3230**	**-1.6540**
7	Distorting1.500.25.0.05	0.6557	0.0318	0.0336	-1.2632	-1.6059
8	Distorting1.500.25.0.3	0.7188	0.0307	0.0335	-0.8993	-1.3271
9	Distorting2.500.25.0.05	0.6547	0.0321	0.0336	-1.2678	-1.6118
10	Distorting2.500.25.0.3	0.7164	0.0304	0.0339	-0.9181	-1.3486
11	**Simulation A (100, 50)**	**0.5770**	**0.0690**	**0.0700**	**-1.3910**	**-1.6270**
12	Distorting1.100.50.0.05	0.5890	0.0684	0.0696	-1.3005	-1.5412
13	Distorting1.100.50.0.3	0.6529	0.0656	0.0681	-0.8315	-1.1020
14	Distorting2.100.50.0.05	0.5893	0.0680	0.0692	-1.3021	-1.5513
15	Distorting2.100.50.0.3	0.6512	0.0633	0.0678	-0.8520	-1.1427
16	**Simulation A (500, 50)**	**0.5720**	**0.0730**	**0.0740**	**-1.3920**	**-1.5910**
17	Distorting1.500.50.0.05	0.5838	0.0722	0.0744	-1.2981	-1.5046
18	Distorting1.500.50.0.3	0.6451	0.0705	0.0741	-0.8122	-1.0317
19	Distorting2.500.50.0.05	0.5831	0.0722	0.0741	-1.3051	-1.5121
20	Distorting2.500.50.0.3	0.6441	0.0683	0.0743	-0.8313	-1.0640

Table 8: Comparison of the critical values for each of the five indices and each overfit scenario, which lead to a Specificity of 0.95, with the respective critical values from Simulation A.

Simulation B - Specificity of the ALR test in case of underfit

Graph 8 shows the distribution of p-values for the ALR test with two groups generated by a median split of the raw score. The test is often used to decide whether the Rasch model holds or does not. Since we generated our data in a way which models a test with 25 items where the Rasch model holds, but with 30% aberrant responding persons, we do not want to reject the H_0. Ideally the p-values of the ALR test would be equally distributed over the interval [0, 1] as in the case of Rasch model conform data without aberrant response (For a detailed analysis of the Specificity of the ALR test have a look at the diploma thesis of 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 aberrant responding persons. Graph 8 shows the distribution of p-values in case of *Guessing*, *Cheating 1*, *Cheating 2* and *Careless*. They are extremely right skewed in the case of *Guessing* and *Careless*, and essentially zero in the case of *Cheating 1* and *Cheating 2*.

These are unpleasant results, since we clearly cannot decide whether items are Rasch model conform in the case of aberrant responding persons (producing an underfit). If we allow the probability of the type-I-error of the ALR 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 8: P-value distribution for the ALR test with a median split of the raw score as a criterion in case of underfit.

As we have seen in Simulation A, those four types of aberrant responses can be detected fairly well with the index C*. In Simulation C we therefore analyze the potential support of C*, if we are testing for Rasch model conformity of the items with the ALR test in the case of aberrant responding persons. Suspicious persons (Specificity of C* set to 0.95) will be removed and the ALR test will be computed for the remaining persons. Since the influence of the aberrant responses on the distribution of the p-values of the ALR test was so strong, Simulation C will also analyze *Guessing*, *Cheating 1*, *Cheating 2* and *Careless* in the case of 5% aberrant response.

Simulation C - Specificity of the ALR test before and after removal of flagged persons in case of underfit

The results for the probability of rejecting the null hypotheses with the ALR test can be seen in Table 9. If we compare the values in case of 30% aberrant responding persons with the results from Simulation B, we see that they are essentially equal (The difference is 0.003 for *Guessing*, 0 for *Cheating 1 & 2* and 0.0025 for *Careless*). We further see that even in the case of 5% aberrant responding persons, the probability for a type-I-error is strongly elevated in the case of *Cheating 1 & 2*. The removal of suspicious persons 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 the ALR values after the removal of suspicious persons are closest to 0.05 for *Cheating 1 & 2* in case of 30% aberrant responding persons, even though these scenarios are the most problematic if we use all 500 persons. This makes perfect sense, since the area under the ROC curve is the highest in case of *Cheating 1 & 2* as we have seen in Simulation B.

As we saw in Simulation B, *Guessing* is the hardest to detect (Graph 3) and therefore fewer persons as in *Cheating 1 & 2* and *Careless* were removed. In the case of 30% persons with *Guessing* behavior we still have a 166% (0.133/0.05=2.66) increased probability of a type-I-error after the removal of suspicious persons. In order to obtain a 0.05 probability of a type-I-error the Specificity level of C* has to be lowered. The downside of this lowering will be addressed in the *Discussion* section.

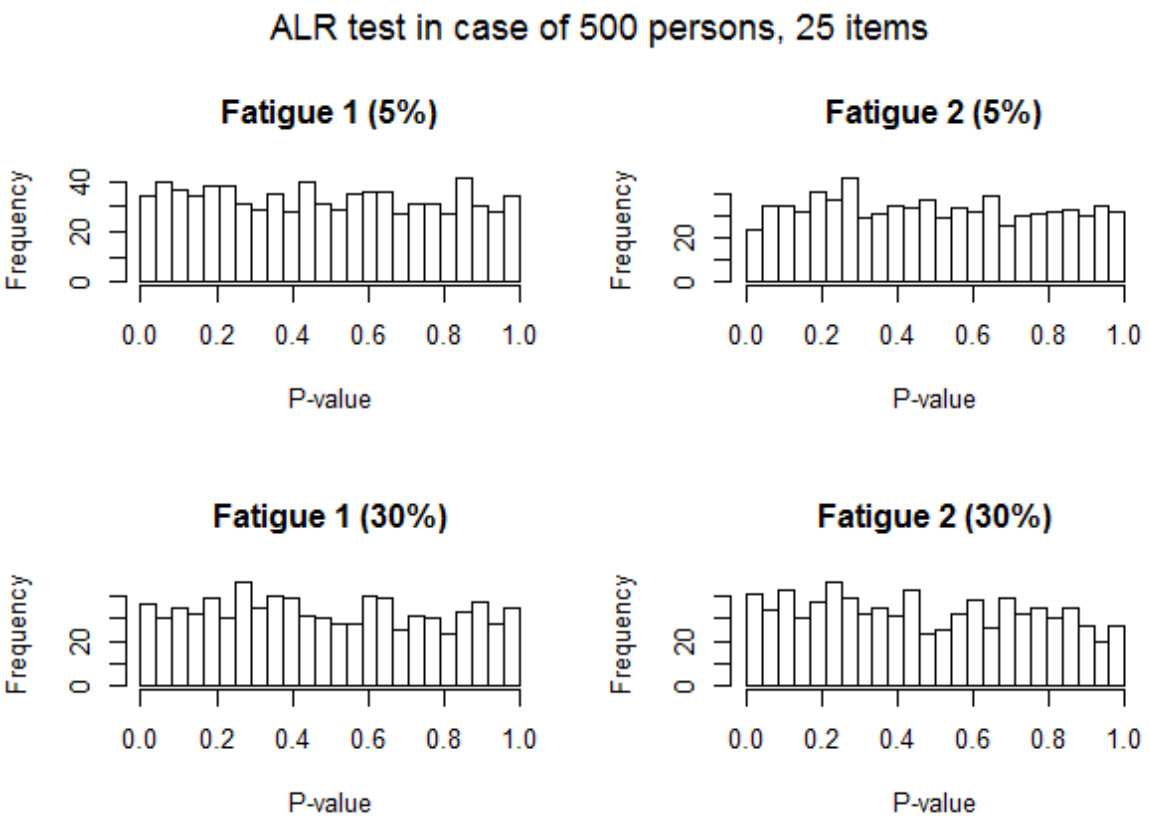
The average number of removed persons are a product of the actual Specificity for a chosen level of Specificity and the associated Sensitivity for a certain scenario. For instance the Sensitivity for C* in case of *Cheating 1* with 30% aberrant responding persons is 0.7967, the actual Specificity (Table 5) 0.9489 for the critical value from Simulation A related to a Specificity of 0.95. We therefore expect 0.7967*500*0.3 + (1-0.9489)*500*0.7 =137.39 persons to be removed, which is not far from the actual value 136.411.

Probability for the type-I-error should be close to 0.05	Scenario: Type of misfit. Proportion of aberrant responding persons for 500 people and 25 items	ALR test with all persons	ALR test without suspicious persons	Average number of persons removed
1	Guessing.0.05	0.062	0.052	26.674
2	Cheating1.0.05	0.300	0.058	42.962
3	Cheating2.0.05	0.128	0.053	40.819
4	Careless.0.05	0.062	0.054	28.242
5	Guessing.0.3	0.489	0.133	39.102
6	Cheating1.0.3	1.000	0.050	136.411
7	Cheating2.0.3	1.000	0.040	125.632
8	Careless.0.3	0.162	0.081	46.838

Table 9: Elevated risk of the type-I-error in case of aberrant response before and after the removal of suspicious persons.

Simulation C - Specificity of the ALR test in case of fatigue 1 and 2.

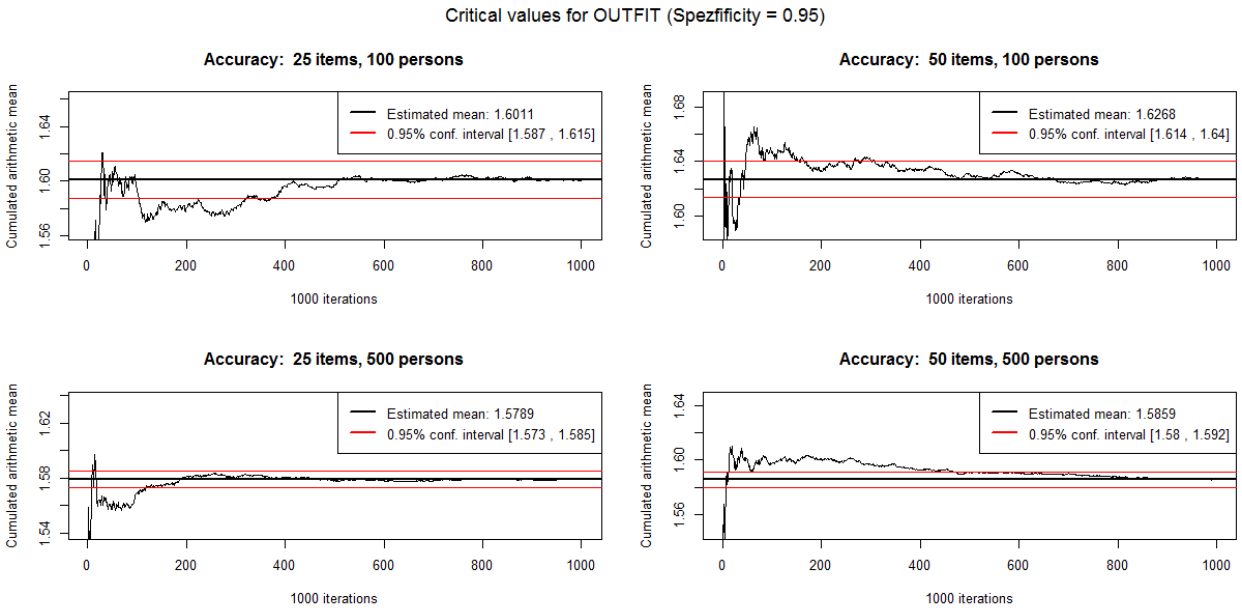
If people experience fatigue at some point in the test the estimation of their latent ability trait will certainly be too low. The estimation of the Rasch model conformity of the items on the other hand, seems to be unaffected of persons experiencing fatigue as we can see in Graph 9. The distribution of p-values seem to be uniform 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 probability of the type-I-error of the ALR 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 94.88%, 96.5%, 94.88% and 94.38% for *Fatigue 1 (5%), Fatigue 2 (5%)*, *Fatigue 1 (30%)* and *Fatigue 2 (30%)*. We can therefore conclude, that the probability of a type-I-error is unaffected of this sort of aberrant response.



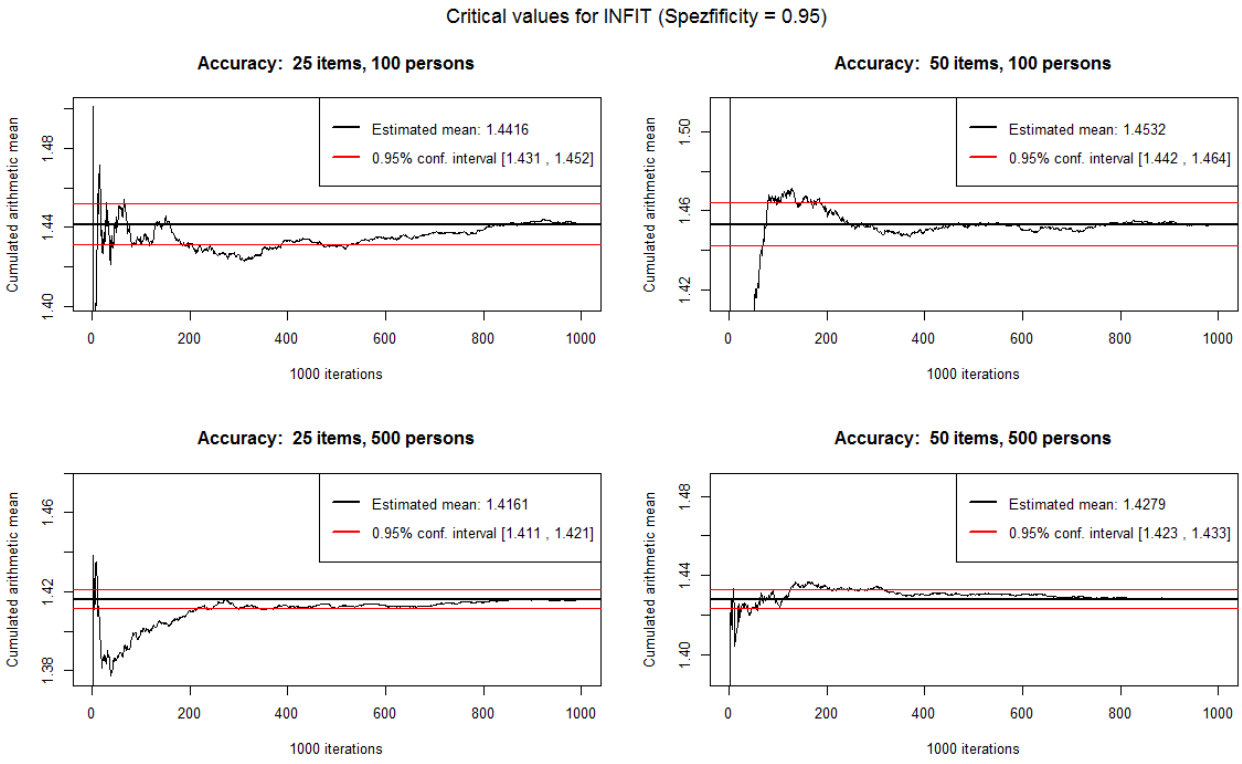
Graph 9: P-value distribution for the ALR test with a median split of the raw score as a criterion in case of fatigue.

Accuracy of the results

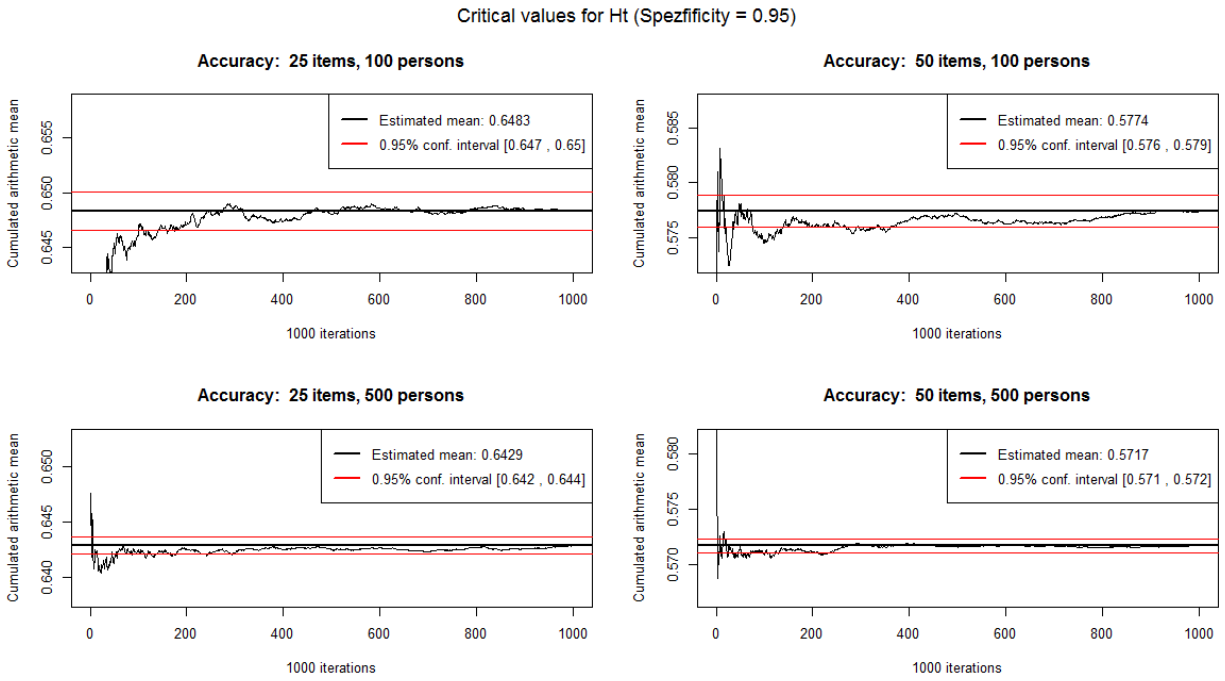
In Simulation A we estimated the 0.01, 0.05, 0.95 and 0.99 quantiles for the five indices in four scenarios depending on the number of items and the number of persons. The graphs 10-14 show the progress of the estimation for the 0.95 quantile. The first thing we note, is that the confidence intervals (95%) for the estimations are smaller in case of 500 person for each index. This was to be expected, since one additional iteration with 500 persons is not so different from five additional iterations with 100 persons each. The precision for 25 and 50 items on the other hand, is about equal. In all cases we can see that the scattering gets less as the number of iterations grow, a clear sign of convergence to the true value. In the case of 100 persons we note it takes about 400-700 iterations for the curve to vary only slightly around the estimated mean. After 800 iterations no substantial shift happens in any of those 20 estimations. If we take another at Table 7, we see that in the case of OUTFIT, INFIT, and Ht most critical values lie way outside the 95% confidence interval.



Graph 10: The progress of the estimation of the 95% quantile for the index OUTFIT and the respective number of items and persons is shown.

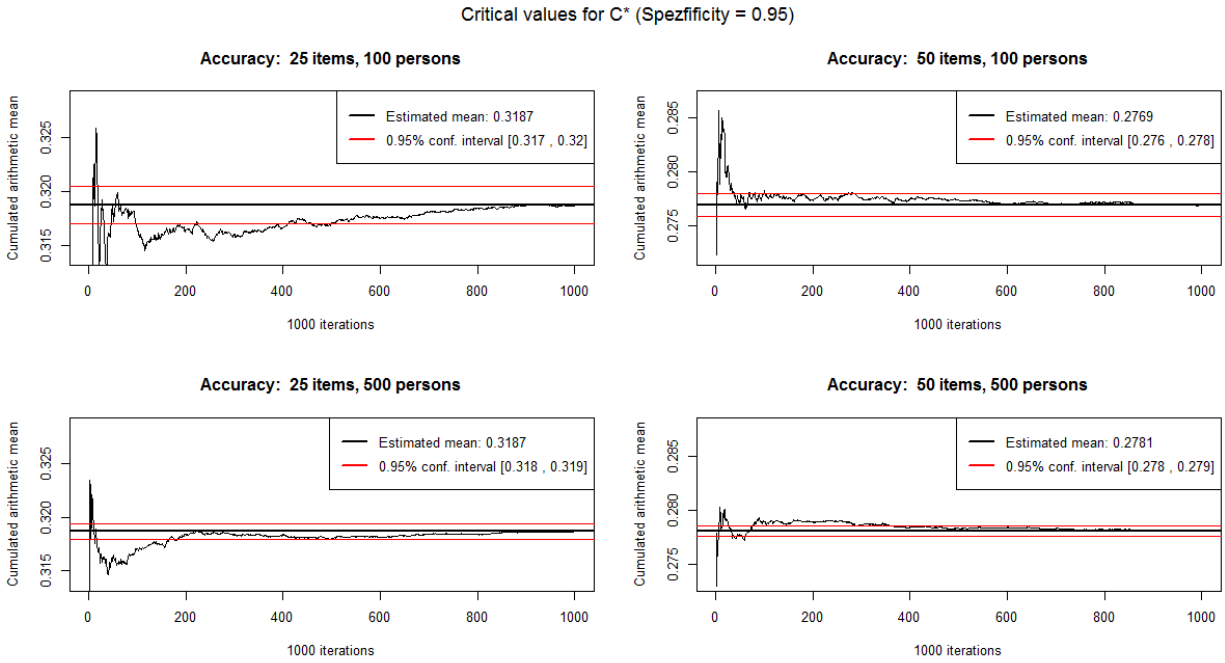


Graph 11: The progress of the estimation of the 95% quantile for the index INFIT and the respective number of items and persons is shown.

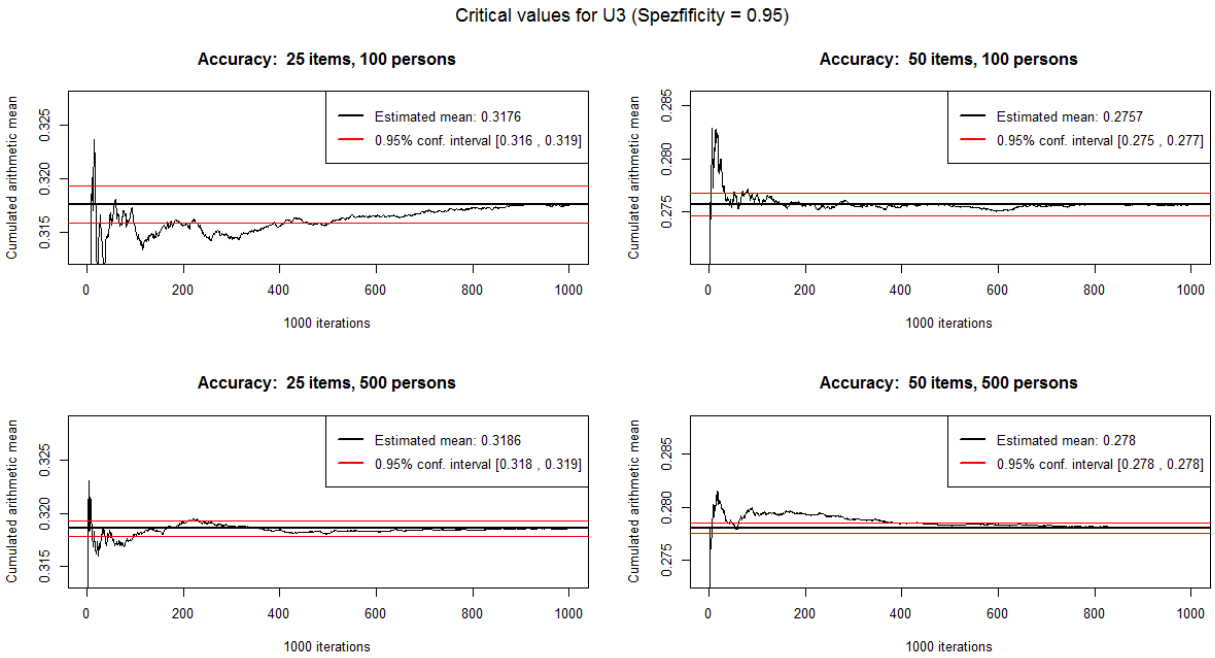


Graph 12: The progress of the estimation of the 95% quantile for the index Ht and the respective number of items and persons is shown.

In the case of C* the estimations are essentially the same in case of 100 or 500 persons, but they clearly differ in case of 25 or 50 items (Graph 13). If the true 95% quantiles of C* and U3 depend ever so slightly on the number of persons, or not at all, cannot be answered with these estimations, but we can say that it isn´t of any practical importance.

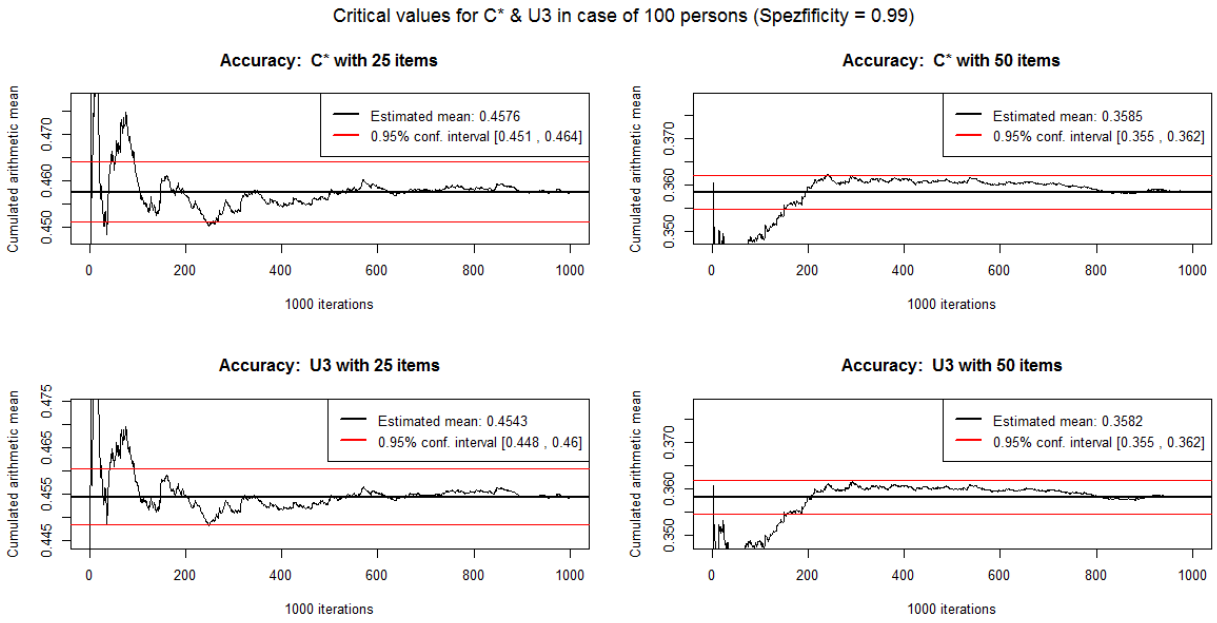


Graph 13: The progress of the estimation of the 95% quantile for the index C and the respective number of items and persons is shown.*



Graph 14: The progress of the estimation of the 95% quantile for the index U3 and the respective number of items and persons is shown.

Although the actual Specificity values for C* and U3 where close to the chosen nominal values, they were a bit too high in the case of underfit, 100 persons and a nominal value of 0.99 (chapter: *Simulation B - Specificity of the indices*). Could this come from the fact that our estimations for the 99% quantile were to high (Remark: For C* and U3 we reject the null hypotheses if the value is bigger as the critical value, if we test for underfit.) In Graph 15, we see that the estimations for the 99% quantile, have wider confidence intervals as the estimations for the 95% quantile (Graphs 13 and 14). We also can also see that the variation of the cumulative arithmetic mean clearly gets smaller with an increasing number of iterations.



Graph 15: The progress of the estimation of the 99% quantile for the indices C and U3 with 100 persons and the respective number of items is shown.*

Table 10 shows the estimated critical values for U3 and C* in case of 100 persons. Those values that lie outside of the respective confidence interval (Graph 15) are marked red. Every single value, that is marked red lies above the respective confidence interval. Since the four plots in Graph 15, do not show patterns which could be seen as signs of overestimations, person misfit (underfit) probably influences the right tail of the distribution of index values of U3 and C* in case of 100 persons. What also speaks for this theory, is the fact that the estimated values are lower (and therefore further away from the estimations in Simulation A) in case of 30% aberrant responding persons as in case of 5% for all scenarios (Table 10).

Specificity = 0.99	Scenario: Type of misfit. Number of persons. Number of items. Proportion of aberrant responding persons	U3	C*
1	Guessing.100.25.0.05	0.4437	0.4486
2	Cheating1.100.25.0.05	0.4464	0.4504
3	Cheating2.100.25.0.05	0.4459	0.4492
4	Careless.100.25.0.05	0.4484	0.4525
5	Guessing.100.50.0.05	0.3566	0.3566
6	Cheating1.100.50.0.05	0.3539	0.3535
7	Cheating2.100.50.0.05	0.3544	0.3539
8	Careless.100.50.0.05	0.3538	0.3540
9	Guessing.100.25.0.3	0.4263	0.4283
10	Cheating1.100.25.0.3	0.4236	0.4300
11	Guessing.100.25.0.05	0.4271	0.4354
12	Cheating1.100.25.0.05	0.4437	0.4486
13	Cheating2.100.25.0.05	0.4464	0.4504
14	Careless.100.25.0.05	0.4459	0.4492
15	Guessing.100.50.0.05	0.4484	0.4525
16	Cheating1.100.50.0.05	0.3566	0.3566

Table 10: Critical values for C and U3 indices and each underfit scenario with 100 persons, which lead to a Specificity of 0.99.*

The precision of Simulation B is very high, since we had 2000 iterations for each scenario and each test. We had a total of 240 (5 indices x 48 scenarios) different computations for the actual Specificities, Sensitivities, critical values and the area under the ROC curve. To highlight the precision of these estimates, 99% confidence intervals where computed for the actual Specificities (for the nominal values 0.95 and 0.99) and the area under the ROC curve. The longest of these 2000 confidence intervals was 0.008175 for the area under the ROC curve, 0.004313 for the Specificity at a nominal value of 0.95 and 0.001961 for the Specificity at a nominal value of 0.99. The confidence intervals were computed is if the estimations are student t distributed, since we averaged 2000 independent and identical distributed random variables[9]. In order to assess how good these approximations are, confidence intervals were computed via bootstrap for some cases. The results turned out to be equal in the first five decimal placed.

The precision of the estimation of the Specificity of the ALR test in Simulation C for underfit and 30% aberrant responding persons is very high (We already saw that the estimations in Simulation B are essentially equal to the estimations in Simulation C; see chapter: *Simulation C - Specificity of the ALR test before and after removal of flagged persons in case of underfit*). In case of 5 % aberrant responding persons we have wider confidence intervals for the estimations. The width of the 95% confidence interval for *Guessing*, *Cheating 1*, *Cheating 2*

[9] We make use of the so called central limit theorem.

and *Careless* are 0.0254, 0.0208, 0.0246, and 0.0255 respectively. This precision is not sufficient, if we want to tell whether the ALR test is 20 % robust in case of *Guessing* and *Careless* behavior, since the both got an actual Specificity estimation of 0.938 (chapter: *Simulation C - Specificity of the ALR test before and after removal of flagged persons in case of underfit*)[10].

7. Discussion

Types of person misfit

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. In a three-parameter logistic model (3-PL), which includes a guessing parameter, guessing behavior would not be seen as a person misfit. The way *Guessing* und *Careless* were modeled in this simulation is similar to most simulation studies. Rupp (2013) wrote a review paper about person fit in which he summarized all simulation studies in this regard. Rupp also categorizes the types of misfit that were modeled in the past. Rupp further says: "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 persons provide a lower score than would be expected based on the chosen model) and (2) *spuriously high scores* (i.e. when persons 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 his real life counterpart.

Cheating 1 & *2* was modeled somewhat different from other simulation studies, but the biggest difference can be found in *Distorting 1* & *2*, which obviously fall in the category (1) *spuriously low scores*. 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. The author of this work wonders why such a behavior should occur in real life. Tendeiro and Meijer (2014) modeled (1) *spuriously low scores* by choosing persons

[10] A test is a % robust if the actual type-I-risk does not divert more than a % from the nominal type-I-risk. In case of a = 0.2, and a nominal Specificity of 0.95, the actual Specificity has to be between 0.94 and 0.96. For more details take a look at Rasch & Guiard (2004).

with a person parameter higher than 0.5 and enough correct answers and changing a certain number of randomly chosen correct responses with a probability of 80% into incorrect ones. 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) difficult items wrong and easy items correct in order to avoid suspicion. Maybe such a behavior was not modeled in the past since it produces a model overfit instead of an underfit.

Deviations from the nominal Specificity

Since OUTFIT and INFIT performed slightly worse than our three non-parametric indices in our main criterion there is no need to further analyze the differences between the actual type-I-risks and the chosen ones (see chapter: *Simulation B - Area under the ROC curve*). Ht on the other hand, performed slightly better than C* and U3 in our main criterion. In this simulation the actual type-I-risk was found to be shockingly high in case of 30% aberrant responding persons and the scenarios *Cheating 1* & *2* and *Distorting 1* & *2*. The next paragraph compares the results of this work and other simulation studies which use the index Ht.

Karabatsos (2003) and Zhang & 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 misfit and the percentage of aberrant responding persons. 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 & 4 of their work. However, they do not discuss the fact that these values vary quite strong. 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 (depending among others on the number of items and the type of response). In each scenario 1000 persons were simulated, and for Ht the cut off values were the respective (1%, 5%, and 10% empirical quantiles) of all persons who did not respond aberrantly. Once again, the dependence of the empirical quantiles on the type of misfit and the number of persons who respond aberrant was not examined and discussed. One work that compared nominal and empirical type-I-error rates for Ht was the simulation study from Tendeiro & Meijer (2014). They

report that the actual type-I-error for Ht, averaged across all experimental conditions was 0.94 for a nominal value of 0.95 which is nowhere near the magnitude of elevated type-I-risk found in this simulation. They derived the critical values for the nominal Specificity 0.95 by simulating scores of 10000 persons without aberrant behavior. This may be problematic, since they simulated 100 datasets with 1000 persons (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 persons is not the same as averaging the quantiles of 100 datasets containing 1000 persons (Remark: In this simulation we saw that Sensitivities were higher for 500 persons than 100.)

Andersen Likelihood-Ratio test

In Simulation C, suspicious persons were removed with the index C* and the respective critical values for a Specificity of 0.95. This removal led to a strong increase of the actual Specificity of the ALR test, particularly in the case of *Cheating 1 & 2*. With a lower Specificity level for C* the improvement would be even better, since additional people with a response vector containing Guttman errors will be removed and the remaining persons will behave on average even more Rasch model conform (Remark: If the deviation from the perfect Guttman scale comes from the aberrant response behavior or from chance does not matter for the ALR test!). If we want to test whether the Rasch model holds for a test we obviously worry about type-II-errors as well. If we want to detect items with a DIF we need a high Sensitivity of the ALR test. If we remove persons with C* at a low nominal Specificity we increase the Specificity but decrease the Sensitivity of the ALR test. This simulation study clearly shows the need to remove persons 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* further research, investigating both types of errors for the ALR test, is of need.

Accuracy of the results

This work contains a rather large section which deals with the accuracy of the results. The author of this work thinks that this is a prerequisite of every quality simulation study. In his review paper Rupp (2014) says that: "The numbers of replications for simulation studies in statistics often seem to be chosen rather arbitrarily, either because the number appears "appealingly simply" (e.g. 100, 250, 500, 1000), or because time constraints prevent authors from running more replications." Some just choose a number of replications and report the confidence intervals of the estimations (e.g. Karabatsos (2003) simulated only one dataset containing 500 persons for 60 different scenarios and reported 95% confidence intervals for the area under the ROC curve in different scenario compositions for each test.)

The accuracy of estimations for a given number of iterations depends on the variation (Remark: The variation is mostly measured in the form of the second centralized moment the so called variance.) and simple rules of thumb like: “Use 1000 iterations and your estimations will be fairly accurate” can never work for all kinds of research questions and all kinds of variables.

8. Summary

In this simulation study the performance of three non-parametric (Ht, U3, and C*) and two parametric (OUTFIT and INFIT) person fit indices was analyzed. The detection ability (measured via the area under the ROC curve) of the indices was best in case of strong underfit (*Cheating 1 & 2*) and strong overfit (*Distorting 1 & 2*) and detection rates increased with the number of persons and the number of items.

In the case of Ht, OUTFIT, and INFIT the distribution of index values strongly depends on the type of aberrant response and the number of aberrant responding persons. Because of this fact, we cannot know the actual Specificity and Sensitivity for a certain critical value and therefore these three indices are of no practical use. U3 and C* seem to satisfy a nominal Specificity value fairly well and their overall performance in this study is almost as good as the performance of Ht and even slightly better than the performance of OUFIT and INFIT. C* performed a little bit better than U3 in case of underfit and equally good in case of overfit. Therefore C* can be suggested as the person fit index of choice if real life data is to be analyzed. The detection ability of C* was equally good in case of 5% and 30% aberrant responding persons.

In simulation C a practical application of person fit analysis was investigated. Namely, the reduction of the type-I-risk for the ALR test with the median split of the raw score as an internal criterion. If aberrant responding persons (producing a model underfit) are not removed, the actual type-I-risk of the ALR test is strongly elevated (e.g. 100% for *Cheating 1 & 2* in case of 30% aberrant responding persons). This elevated type-I-risk can be (almost) brought back to normal if we remove persons with C* such that persons with no aberrant response behavior are only removed with a probability of 95%. The improvement in the type-I-risk of the ALR test works particularly well in the case of *Cheating 1 & 2*, which are easier to detect than *Guessing* and *Careless*.

The scenarios *Fatigue 1 & 2* do influence the response behavior in a way which can neither be label as overfit or underfit. This kind of aberrant response seems to have no impact on the ALR test.

IV. Directories

1. References

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

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.

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.

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). Simulation des Risikos 1. Art und der Teststärke von vier verschiedenen Modelltests für das Rasch-Modell

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.

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.

2. List of tables

Table 1: Simulation A - Computation of critical values for the five person fit indices. 17
Table 2: Simulation B - Computation of the area under the ROC curve, Specificities, Sensitivities, critical values, and the ALR test in some cases. 19
Table 3: Simulation C - Specificity of the ALR test before and after the removal of suspicious respondents. 20
Table 4: Estimated quantiles (0.01, 0.05, 0.95 and 0.99) for the five indices in four different scenarios. 21
Table 5: Actual Specificity values for each test and each scenario, if the respective critical values from simulation A, that correspond to a Specificity of 0.95, are used. 27
Table 6: Actual Specificity values for each test and each scenario, if the respective critical values from simulation A, that correspond to a Specificity of 0.99, are used. 29
Table 7: Comparison of the critical values for each of the five indices and each underfit scenario, which lead to a Specificity of 0.95, with the respective critical values from Simulation A. 30
Table 8: Comparison of the critical values for each of the five indices and each overfit scenario, which lead to a Specificity of 0.95, with the respective critical values from Simulation A. 32
Table 9: Elevated risk of the type-I-error in case of aberrant response before and after the removal of suspicious persons. 35
Table 10: Critical values for C* and U3 indices and each underfit scenario with 100 persons, which lead to a Specificity of 0.99. 41

3. List of figures

Graph 1: The ROC curves of three made up tests 16
Graph 2: A comparison of five person-fit indices, in their ability to detect aberrant responding examines in case of underfit and overfit, by using the area under the ROC curve as the criterion. 22

Graph 3: A comparison of five person-fit indices, in their ability to detect careless, cheating, distorting and guessing persons. 22
Graph 4: A comparison of five indices, in their ability to detect careless and guessing persons. 23
Graph 5: A comparison of five indices, in their ability to detect two types of cheating and two types of distorting persons. 23
Graph 6: A comparison of five indices, in their ability to detect aberrant responding persons in four conditions given by the number of persons and the number of items. 24
Graph 7: A comparison of five indices, in their ability to detect aberrant responding persons in two conditions given by the percentage of aberrant responding persons. 25
Graph 8: P-value distribution for the ALR test with a median split of the raw score as a criterion in case of underfit. 33
Graph 9: P-value distribution for the ALR test with a median split of the raw score as a criterion in case of fatigue. 36
Graph 10: The progress of the estimation of the 95% quantile for the index OUTFIT and the respective number of items and persons is shown. 37
Graph 11: The progress of the estimation of the 95% quantile for the index INFIT and the respective number of items and persons is shown. 37
Graph 12: The progress of the estimation of the 95% quantile for the index Ht and the respective number of items and persons is shown. 38
Graph 13: The progress of the estimation of the 95% quantile for the index C* and the respective number of items and persons is shown. 38
Graph 14: The progress of the estimation of the 95% quantile for the index U3 and the respective number of items and persons is shown. 39
Graph 15: The progress of the estimation of the 99% quantile for the indices C* and U3 with 100 persons and the respective number of items is shown. 40

V. **Lebenslauf**

Richard Artner, BSc

Geboren 1989 in Wien

(Akademische) Ausbildung:

2003-2008	**Matura** an der höheren technischen Lehranstalt Wien Donaustadt, Zweig: Datenverarbeitung und Organisation
Seit 2010	Diplomstudium Psychologie an der Universität Wien Vordiplom am 28.02.2012
2011-2014	**Bachelor of Science** – Hauptuniverstität Wien Studium: Statistik
Seit 2014	Magisterstudium Statistik – Hauptuniverstität Wien
2015	Pflichtpraktikum (240 Stunden) an der Universtität Wien – Institut für Methodenlehre

Sprachkenntnisse:

Deutsch	Muttersprache
Englisch	Sehr gute Kenntnisse in Wort und Schrift

Programme:

Programmiersprache R, Statistikprogramm SPSS, MS Office

(Wissenschaftliche) Interessen:

Wahrscheinlichkeitstheorie, Extremwertstatistik, Sozialpsychologie, Psychometrie