Response Time-based Treatment of Omitted Responses in Computer-based Testing

Andreas Frey[1,2], Christian Spoden[1], Frank Goldhammer[3,4], & S. Franziska C. Wenzel[5]

[1]Friedrich Schiller University Jena, Germany

[2]Centre for Educational Measurement (CEMO) at the University of Oslo, Norway

[3]German Institute for International Educational Research (DIPF), Germany

[4]Centre for International Student Assessment (ZIB), Germany

[5]Goethe University Frankfurt, Germany

This is a post-peer-review, pre-copyedit version of an article published in Behaviormetrika.

The final authenticated version is available online at:

http://dx.doi.org/10.1007/s41237-018-0073-9

Author Notes

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Correspondence should be addressed to: Andreas Frey, Friedrich Schiller University Jena, Am Planetarium 4, 07743 Jena, Germany. E-mail: andreas.frey@uni-jena.de

Andreas Frey is now at Goethe University Frankfurt, Germany. Christian Spoden is now at the German Institute for Adult Education, Leibniz Centre for Lifelong Learning, Bonn, Germany.

Abstract

A new response time-based method for coding omitted item responses in computer-based testing is introduced and illustrated with empirical data. The new method is derived from the theory of missing data problems of Rubin and colleagues and embedded in an item response theory framework. Its basic idea is using item response times to statistically test for each individual item whether omitted responses are missing completely at random (MCAR) or missing due to a lack of ability and thus not at random (MNAR) with fixed type-1 and type-2 error levels. If the MCAR hypothesis is maintained, omitted responses are coded as not administered (NA), and as incorrect (0) otherwise. The empirical illustration draws from the responses given by $N = 766$ students to 70 items of a computer-based ICT-skills test. The new method is compared with the two common deterministic methods of scoring omitted responses as 0 or as NA. In result, response time thresholds from 18 to 58 seconds were identified. With 61 %, more omitted responses were recoded into 0 than into NA (39 %). The differences in difficulty were larger when the new method was compared to deterministically scoring omitted responses as NA compared to scoring omitted responses as 0. The variances and reliabilities obtained under the three methods showed small differences. The paper concludes with a discussion of the practical relevance of the observed effect sizes, and with recommendations for the practical use of the new method as a method to be applied in the early stage of data processing.

Keywords: testing, computer-based testing, missing data, response time, item response theory

Response Time-Based Treatment of Omitted Responses in Computer-Based Testing

Empirical test data typically contains missing item responses. The treatment of these missing responses is not trivial. Up to now, the treatment of missing responses is frequently based on one or on a mixture of three *deterministic approaches*. Under these deterministic approaches, the same procedure is applied for each missing response[1]. The first deterministic approach is listwise deletion. Here, if one or more responses are missing, the complete data of a test taker are excluded from the analysis. Listwise deletion does not only imply a severe loss of information but is also likely to result in biased ability estimates (Little & Rubin, 2002). The Task Force on Statistical Inference of the American Psychological Association (Wilkinson & Task Force on Statistical Inference, American Psychological Association, Science Directorate, 1999) considers listwise deletion to be among the worst methods for dealing with missing data. The second deterministic approach considers all missing responses as incorrect answers. This approach is based on the assumption that a lack of ability is the cause for all missing responses. Even though this assumption might be appropriate for some missing responses, it penalizes test takers in an unjustified way if they skipped an item before cognitively processing it (e.g., skipped by mistake, rapid non-responding; Lord, 1974). The third deterministic approach is treating missing responses as not administered. This approach is based on the assumption that all missing responses appeared randomly and are completely unrelated to ability. Obviously, even though some responses may be missing randomly in a data set, it is also possible that some missing responses are due to systematic effects such as a lack of ability. A deterministic treatment of missing responses as not administered increases the proportion of correct answers and thus results in overestimation of ability if missing responses are at least partly due to a lack of ability (e.g., Holman & Glas, 2005). To sum up, deterministic approaches do not offer sound solutions for the scoring of missing responses.

In order to avoid the problems of the deterministic approaches, several *more complex approaches* for handling missing responses were suggested in the last decades (see Enders, 2010 for an overview). Reasonable methods to deal with missing responses embrace full information maximum likelihood estimation (e.g., Arbuckle, 1996; Enders & Bandalos, 2001), multiple imputation techniques (e.g., Rubin, 1987), and model-based methods (e.g., Glas & Pimentel, 2008; Holman & Glas, 2005; Köhler, Pohl & Carstensen, 2015; Mislevy & Wu, 1996; O'Muircheartaigh & Moustaki, 1999; Rose, von Davier, & Xu, 2010). Although psychometrically sophisticated, these approaches exhibit three major disadvantages. First, they are relatively complex and quite demanding for daily testing practice. One needs extensive psychometric training to accurately apply these complex methods. Second, the flexible model structures can be very useful to statistically model missingness in a given data set but are problematic for establishing a measurement model applicable across several assessments of a measurement instrument. Aspects such as linking and equating across assessment cycles and definition of proficiency levels might be challenging or even impossible if using the complex approaches. Thus, for application areas where a scale with stable measurement properties is needed, it would be better to first apply a method to code missing responses and to apply the actual measurement model afterwards. Third, the complex approaches are based on assumptions that can be problematic. With regard to model-based methods for handling missing responses, Köhler et al. (2015) showed that the assumptions of a uni-dimensional latent missing propensity variable and of normality for the bivariate distribution of the latent missing propensity variable and the latent ability variable (the model-based approaches typically make these assumptions) could be violated. When using an empirically derived bivariate distribution, different conclusions on the best way to treat missing responses result. Reflecting the three mentioned and other disadvantages, hardly any operational testing programs currently use the more complex

approaches, but still apply deterministic approaches to deal with missing responses first and then scale the data with the actual psychometric measurement model.

The advent of computer-based testing, however, opens up new possibilities for significantly improving the handling of missing responses. When computers are used to deploy tests, collateral information can be collected that makes it possible to better understand which mechanism caused a person to not respond to a test item. Such an understanding of the underlying mechanisms of missing responses are a prerequisite of handling them appropriately on an item- and person-specific level. Very promising collateral information easy to gather during computer-based assessments is the times test takers are spending to answer individual test items. Only tentative steps had been taken so far to utilize item response times when handling missing responses. In the computer-based large-scale assessment PIAAC (Programme for the International Assessment of Adult Competencies), for example, all missing responses on the cognitive items with response times less than five seconds were treated as not administered (OECD, 2016). Missing responses with response times equal than or larger than five seconds are coded as incorrect. The idea behind this coding rule is that it is not possible to cognitively process any of the test items when working less than five seconds on it. The idea that an item cannot be processed if it was presented too short is compelling, but the relatively arbitrary threshold of five seconds for all test items may not the best choice. In order to provide a more differentiated and empirically anchored solution, Weeks, von Davier and Yamamoto (2016) suggested a method using logistic regression to identify item-specific response time thresholds and use these to inform the coding of missing responses. The method was illustrated with empirical data stemming from PIAAC 2012. The reported results underline that the general five second-rule, which ignores item-specific collateral information, is most likely too simple (see also Goldhammer, Martens, & Lüdtke, 2017). It is too simple because the response time thresholds varied between items and were considerably higher

than five seconds (20 to 30 seconds on average). Despite these interesting results, it should be noted that the Weeks et al. (2016) approach was developed against a pragmatic background and therefore does not root in a theory of missing data. Furthermore, it does not take into account that the proportion of actually given responses to an item is typically much greater than the proportion of missing responses. Depending upon the magnitude of the difference in available data points between these two categories for a specific item (imagine the extreme cases of [a] one case with a missing response and all others with given responses to an item, and [b] 50 % with missing and 50 % with given responses to an item), the precision of the identified threshold varies between items. Lastly, the approach does not quantify the error probabilities connected with coding the missing responses as not administered or as incorrect. Nevertheless, this is necessary to justify that the procedure does not systematically penalize individual test takers if different booklets are used.

The present paper starts at the point where the Weeks at al. methods ends. It presents and illustrates a response time-based, item-specific method to handle missing responses. The suggested method is derived from the theoretical framework of missing data problems of Rubin and colleagues (e.g., Rubin, 1976; Little & Rubin, 2002). It accounts for the possibility that the ratios of observed to missing responses can vary between items and provides full control of type-1 and type-2 error levels. Due to its strict theoretical derivation and the consideration of type-1 and type-2 error levels, the proposed method goes beyond approaches focusing on rapid-guessing behavior (for an overview see Lee & Chen, 2011). The identification of reaction time thresholds is not determined by relatively soft criteria such as the visual inspection of response time distributions, item surface features such as the amount of reading required (for a comparison of criteria for threshold determination, see Kong, Wise, & Bhola, 2007), or general rules of thumb (e.g., threshold set at 10 % of the average time test takers used to answer an item as derived from

Wise & Ma, 2012 and applied in Wise & Kingsbury, 2016) but by statistical testing based on Rubin's framework for missing data problems. The suggested method can be regarded as a combination of the reaction time approach used in the analysis of rapid-guessing behavior and Rubin's approach, with the objective of an item-specific coding of observed omitted responses as an early step in data preparation.

The rest of the paper is organized as follows. First, the basic concepts of Rubin's framework for missing data problems are outlined. Then, three different types of missing responses common in test data are described: Each type is assigned to one or more dominant missing response mechanisms according to the framework of Rubin. Based on the concepts presented until then, the three objectives of the present paper are specified. The next section is devoted to a detailed description of the new method. After the formal description of the method, its application is illustrated using an empirical data set. The paper closes with a brief summary, a discussion of the practical relevance of the differences obtained with the new method compared to common ways to deal with missing responses, and the practical implications which can be derived from the findings.

Basic Concepts of Rubin's Framework for Missing Data Problems

Rubin and colleagues proposed an influential classification system for missing data problems. Most of the work currently done in the area of missing data originated from the concepts of this classification system. The authors are distinguishing three *missing data mechanisms*. These are describing potential causes for the occurrence of missing values in empirical data. In other words, the missing data mechanisms are specifying how the probability of a missing value relates to the observed data.

The first missing data mechanism is called *missing completely at random* (MCAR). Data are MCAR if the probability of missing values for a variable Y is independent of (a) the values of

Y and (b) of any other variable in the data set. Under the assumption of MCAR, missing responses are completely due to random processes. This is the case, for example, if a test taker skips a test item (intentionally or by mistake) before cognitively processing it, or when an incomplete booklet design (e.g., Frey, Hartig & Rupp, 2009; Rutkowski, Gonzales, von Davier, & Zhou, 2014) is used to randomly assign each test taker a subset of the complete set of available test items. If missing responses are due to the MCAR mechanism, the observed responses are a simple random sample of the hypothetically complete set of responses. Therefore, no bias in the parameter estimates of interest is to be expected, given that the assumption of MCAR holds. Several methods had been proposed to test the MCAR assumption (Chen & Little, 1999; Diggle, 1989; Kim & Bentler, 2002; Little, 1988). The basic idea behind about all of the proposed MCAR-tests is that the test takers exhibiting missing values and the test takers without missing values are both random draws from the same underlying population. Hence, the distribution of the variable of interest, the moments of this distribution, as well as the relationship of the variable of interest with other measured variables, are the same for both groups of test takers (with missing values; without missing values). Based on this consideration, a straightforward procedure is to statistically test whether the mean of the variable of interest (e.g., a unidimensional ability variable derived from non-missing responses) or other measured variables are the same for test takers with missing values on the variable of interest compared to test takers without missing values by one or several *t*-tests. If no significant difference results, the assumption of MCAR can be maintained. Several potential issues need to be considered when using the *t*-test to statistically test the MCAR assumption. Some of which are the possibility of largely unequal group sizes (typically, a very small number of individuals with missing values in one group and a large number of individuals with valid values in the other group), the related aspect of variance

heterogeneity between the two groups, and variations in the type-2 error level between different tests (cf. Enders, 2010).

The second missing data mechanism is called *missing at random* (MAR). It is connected with less strict assumptions than MCAR. Data are considered to be MAR if the probability of missing values on a variable Y is related to another variable in the data set but not to the variable Y itself given the observed data set. Thus, after controlling for the other measured variables, the propensity of missing values for the variable Y and the values of Y are unrelated under the MAR assumption. In testing situations, MAR can occur, for example, if a high level of content related anxiety (e.g., math anxiety) is related to a high probability to skip test items tapping the respective content area (e.g., mathematics). According to Rubin (1976), MAR data are ignorable, making it possible to use maximum likelihood estimation and multiple imputation. A practical problem of MAR is that its assumptions are generally regarded as being not testable in a statistical sense (Enders, 2010). However, since data satisfying the condition of MCAR are also MAR by definition, testing for MCAR will also provide evidence for MAR even though not all occurrences of MAR cases will be detected.

The third missing data mechanism is called *missing not at random* (MNAR). Data are MNAR if the probability of missing data on the variable Y relates to the values of Y itself, even after controlling for the other variables in the data set. In test settings, this might occur, if a test taker seriously tries to solve an item but fails and decides to tick no response for that item. Without knowing the unobserved values there is no straightforward way to verify that data are MNAR. However, again, hypotheses can be formulated in a way that rejecting the assumption that data are MCAR lead to the conclusion that they are MAR or MNAR which can be helpful in several situations.

Types of Missing Item Responses

Prior to scaling and analyzing test data it has to be decided how to code missing items responses as one important aspect of data processing. Rubin`s classification system for missing data problems can be particular helpful for determining how different types of missing responses should be coded. Three different types of missing responses are typically differentiated for test data.

The first type is responses *missing by design*. These missing responses result, for example, when using an incomplete booklet design for distributing the items to the test takers. As long as the planned incompleteness is based on random processes (such as a random assignment of booklets to test takers), responses missing by design can be regarded as MCAR. Hence, they are ignorable and can be coded as not administered, which means neglecting them in the scaling process.

The second type of missing responses are a series of missing values at the end of the test. These are referred to as *not reached.* They are also coded as not administered for most power tests. This is straightforward, as long as the time needed to answer the test items is not a part of the construct at stake. If time is crucial for the construct of interest, however, other methods are more appropriate (as in speeded performance tests; e.g., Goldhammer & Kroehne, 2014; van der Linden, 2009).

Difficult decisions have to be made with regard to missing responses of the third type. These are missing responses that appear within the test followed by one or more valid responses. They are called *omitted responses* (OR). For many tests, OR are deterministically coded as incorrect. Nevertheless, this is probably an inappropriate method for at least some of the observed OR. If OR occurred because an item was skipped before it was cognitively processed, for example, the missing values would be unrelated to the construct of interest and would thus satisfy

the conditions of MAR or MCAR. In this case, scoring OR as incorrect would penalize test takers who have not even tried to solve the item and would therefore be inappropriate for most test situations (e.g., Ludlow & O'Leary, 1999). Based on results from simulated data which indicated less biased estimates of item parameters and person abilities under this procedure (de Ayala, Plake, & Impara, 2001; Finch, 2008), it was also argued that OR should be treated as not administered. However, the difficulties with this deterministic procedure are obvious (e.g., Ludlow & O'Leary, 1999): If test takers get this procedure to know, the most rational test taking strategy would imply to respond only to few carefully selected items that are easy to solve. Thereby, very high or even perfect scores can easily be achieved as long as the test takers are able to solve at least some of the items. Empirical findings underline the notion that neither coding OR as incorrect nor as not administered is appropriate. Typically, low to moderate negative correlations between the number of OR in a test and the estimated ability are observed. In PIAAC, correlations of -.37 and -.32 resulted between the number of OR and student competences in literacy and numeracy, respectively (Weeks at al., 2016). For the large-scale assessment PIRLS 2011 (Progress in International Reading Literacy Study), Robitzsch (2016) reported a correlation of -.37 between the tendency to omit a response and reading competence for fourth-graders.

Please insert Table 1 around here

To conclude, a deterministic approach seems not to be appropriate to code OR. For some of the OR, treating the respective item as not administered seems appropriate and for others treating them as an incorrect response. An overview of the three types of missing responses, their underlying missing data mechanism, and the typical scoring is shown in Table 1.

Study Objectives

As shown before, OR are the type of missing responses for which a deterministic coding strategy is obviously problematic. The focus of the rest of the paper is on this type of missing responses. Using item-specific response times is a promising approach to derive a more appropriate treatment of OR than deterministically coding all OR as incorrect (OR = 0) or all OR as not administered (OR = NA). The observed response times can be used to separate item-specific OR for which the assumption of MCAR holds from those for which this assumption needs to be rejected. Because MCAR responses are ignorable for likelihood-based inferences, such a separation would give a justification to code the former as not administered and the latter as incorrect (assuming that missingness reflects low ability). A straightforward statistical MCAR test using response times can be build up by assuming that OR caused by skipping an item before cognitively processing it are likely to be associated with short response times, and MNAR responses are likely to be associated with longer response times which are needed to read and understand the item. The paper aims at the following three objectives:

1. Specify an item-specific method based on Rubin's theory of missing data and embedded in item response theory, which is capable to separate OR for which the assumption of MCAR holds from MNAR item responses using response time information.

2. Illustrate the applicability of the new method with empirical data.

3. Examine the effects of the new method on item parameter estimates, ability estimates, variances, and reliabilities in comparison to deterministic methods.

Proposed Method

The general purpose of the proposed method is deriving well-founded decisions which OR are due to a lack of ability (and therefore should be coded as incorrect) and which are due to

skipping before they were cognitively processed (and therefore should be coded as not administered) by utilizing item response times. The method bases on three core assumptions:

1. OR being MCAR are associated with relatively short response times.
2. OR being MNAR are associated with relatively long response times.
3. For each item, a response time threshold exists that is separating OR being MCAR from OR being MNAR.

According to Rubin's theory of missing data, the assumption that OR for an item $i \in \{1,2,\dots,I\}$ are MCAR can be maintained if the average ability (derived from the responses to the other test items) of test takers with a valid response to item *i* (μ_0) is equal to the average ability (derived from the responses to the other test items) of test takers with OR for item *i* (μ_1). Based on the empirical evidence that the number of OR and ability are generally negatively correlated (see above), this can be differentiated for achievement tests by assuming that test takers with a valid response to item *i* will have a higher average ability than test takers with OR for item *i*. The following pair of hypotheses express this formally:

$$H_0: \mu_0 - \mu_1 \leq 0$$

$$H_1: \mu_0 - \mu_1 > 0$$

To decide upon these hypotheses, a *t*-test for two independent samples is applicable. For the method proposed here, first, a *t*-test is carried out considering all test takers with OR for item *i* to calculate $\hat{\mu}_1$ as an unbiased estimate of μ_1 and all test takers with valid responses to item *i* to calculate $\hat{\mu}_0$. Subsequently, $\hat{\mu}_1$ is calculated for one or more subsets from the set of test takers with OR for item *i* and $\hat{\mu}_0$ for the rest of the sample, respectively. The subsets to calculate $\hat{\mu}_1$ are assembled by considering only test takers with OR for item *i* whose response time rt_i for this item was smaller than the maximum response time $rt_{i,max}$ for item *i* across the sample, reduced by a fixed time interval rt_Δ. For the calculation of the average ability $\hat{\mu}_1$, all test takers $j \in$

$\{1,2,\ldots,N\}$ are used, for which $rt_{ji} \leq rt_{i,max} - s \cdot rt_{\Delta}$ applies, with $s \in \{0,1,\ldots,S\}$ denoting the steps. Thus, first the average ability of all test takers with OR for item *i* is calculated, then the average ability of the test takers with a response time being at least rt_{Δ} smaller than the maximum response time for item *i*, then the average ability of the test takers with a response time being at least $2 \cdot rt_{\Delta}$ smaller than the maximum response time for item *i*, and so on. For the calculation of $\hat{\mu}_0$, the average ability of the remaining test takers (all test takers not included in the calculation of $\hat{\mu}_1$) was computed. Figure 1 shows three hypothetical distributions obtained for decreasing response times.

Please insert Figure 1 around here

At each step s, the null hypothesis given above is tested with an independent samples *t*-test. The procedure continues until a non-significant difference results or until the number of subjects in the OR group gets too small. For the common case that the ability difference is significant at the onset of the procedure and gets non-significant at a certain step, the current level of rt_i at this step constitutes the critical response time $rt_{i,crit}$ for this item. As a result, OR for test takers with a response time of $rt_{ji} \leq rt_{i,crit}$ for item *i* can be regarded as MCAR. These OR are coded as not administered. For the test takers with $rt_{ji} > rt_{i,crit}$, the MCAR assumption is rejected and the alternative of MNAR adopted. OR are coded as incorrect for this group of test takers.

The proposed method differs from a sequence analysis in that the sample is known beforehand and is not supplemented during its application. Although several hypotheses are tested one after another, they are substantially different from each other (differences in average ability between groups with the groups being formed differently according to reaction time). An adjustment of the type-1 and type-2 error levels is therefore not indicated.

Statistical Test

When applying the *t*-test for two independent samples as described in the previous section, the following three aspects need to be taken into account:

1. The number of test takers with OR can become small, especially for small $rt_{i,crit}$ values.

2. Sample sizes for test takers with OR and test takers without OR for item *i* will typically differ largely. As a result, heterogeneous variances have to be expected.

3. Statistical power may vary heavily between items and between $rt_{i,crit}$ levels.

These three aspects must and can be considered with well-documented statistical techniques.

The issue of small sample sizes can directly be accounted for by reducing s only as long as an approximately normal sampling distribution of the ability mean can be assumed. As a general rule of thumb, a minimum group sample size of $n = 30$ is deemed appropriate for most situations. According to the central limit theorem, this sample size leads to an approximately normal sampling distribution of the mean (the crucial requirement for the *t*-test), independent of the distribution of the analyzed variable in the population. Anyhow, the shape of the sampling distribution of the mean depends on both the shape of the distribution of the analyzed variable in the population and the sample size. If the variable of interest is normally distributed in the population (as is typically the case for abilities measured with tests), substantially smaller sample sizes are sufficient. Theoretically, any sample size would lead to a normal sampling distribution of the mean if the population distribution is perfectly normal. Thus, for empirical applications where one can reasonably assume a normal distribution without outliers for the analyzed variable on the population level, a sample size of $n \geq 10$ would generally produce an approximately normal sampling distribution of the mean, thereby meeting a core requirement of the independent

samples *t*-test. To be sure, one can check whether the empirical ability distributions do not contain outliers, are unimodal, approximately symmetric, and not skewed.

The second aspect arises from decreasing the sample size in the group of test takers with OR for an item during the proposed method: As the samples in the OR group decreases, the difference to the sample size in the second group increases. This also increases the probability that the variances differ between the two groups. The resulting variance heterogeneity can be accounted for by using a test statistic that considers the two group-specific variances separately. The Welch-test (Welch, 1947) does this by calculating the test statistic

$$t_w = \frac{\hat{\mu}_0 - \hat{\mu}_1}{\sqrt{\frac{\hat{\sigma}_0^2}{n_0} + \frac{\hat{\sigma}_1^2}{n_1}}} \quad (1)$$

where $\hat{\mu}_0$, $\hat{\sigma}_0^2$, and n_0 are the mean, the variance and the sample size of the group of test takers with a valid response to item *i*. $\hat{\mu}_1$, $\hat{\sigma}_1^2$, and n_1 are the same quantities for the group of test takers with OR for item *i*.

However, applying the Welch-test does not automatically lead to constant power of the statistical tests, which is necessary in the present case to draw conclusion that are comparable across items. To achieve this, a statistical test according to the Neyman and Pearson-tradition (e.g., Neyman & Pearson, 1933) leading to a binary decision (H_0 or H_1) must be applied. Therefore, the type-1 and type-2 error probabilities α and β are set a-priori. Since the sample sizes n_0 and n_1 are given by the data, the minimum effect size for a significant effect can be calculated. For the present case, in which sample sizes and variances can differ between the two compared groups, Hedges g (Hedges, 1981) is an appropriate effect size measure:

$$g = \frac{\hat{\mu}_0 - \hat{\mu}_1}{\sqrt{\frac{(n_0 - 1)\,\hat{\sigma}_0^2 + (n_1 - 1)\,\hat{\sigma}_1^2}{n_0 + n_1 - 2}}} \tag{2}$$

Additionally to the type-1 and type-2 error rates, the calculation of the minimum effect size is determined by two quantities when referring to the t distribution: The degrees of freedom (df) and the non-centrality parameter (NCP) of the t distribution. While under variance homogeneity, the df are simply computed by $n_0 + n_1 - 2$, the Satterthwaite correction (Satterthwaite, 1946) can be applied to determine the corrected degrees of freedom df_{corr} of the t distribution under heterogeneous variances by:

$$df_{corr} = \frac{\left(\frac{\hat{\sigma}_0^2}{n_0} + \frac{\hat{\sigma}_1^2}{n_1}\right)^2}{\frac{\hat{\sigma}_0^4}{{n_0}^2(n_0 - 1)} + \frac{\hat{\sigma}_1^4}{{n_1}^2(n_1 - 1)}} \tag{3}$$

The NCP is then computed by referring to the two group-specific variances according to:

$$NCP = \frac{g_{crit}}{\sqrt{\frac{\hat{\sigma}_0^2}{n_0} + \frac{\hat{\sigma}_1^2}{n_1}}} \tag{4}$$

By inserting the quantities for samples sizes, group variances, type-1 and type-2 error, the critical effect size measure g_{crit}, referring to values from the t distribution, can be found with the uniroot minimization. With the empirical standardized mean difference between the two groups g_{emp} and the critical effect size g_{crit}, a decision between null and alternative hypothesis can be made. If $g_{emp} \leq g_{crit}$ the H_0 is maintained; if $g_{emp} > g_{crit}$ the H_1 is adopted. Thereby, the critical response time $rt_{i,crit}$ that differentiates test takers with OR due to MCAR from the other test takers is identified with fixed type-1 and type-2 error rate. The method is item-specific so that items differing in the amount of time needed to process and answer them can be treated easily. Regarding type-1 and type-2 errors levels, in most applications, incorrect coding of both OR = 0

and OR = NA will be similarly problematic. Therefore, for empirical applications of the new method, it is appropriate to assign the same value to α and β errors.

Recoding of Omitted Responses

In a next step, the information from the statistical test is used to recode the response variables. The responses of test taker j to item i are recoded into x_{ji}^{*} according to the principle:

$$x_{ji}^{*} = \begin{cases} 0 \; if \; x_{ji} = 0 \\ 1 \; if \; \; x_{ji} = 1 \\ 0 \; if \; x_{ji} = \text{OR and } rt_{ji} > \; rt_{i,crit} \\ \text{NA} \; if \; x_{ji} = \text{OR and } rt_{ji} \leq \; rt_{i,crit} \end{cases} \tag{5}$$

The valid responses 0 and 1 are not altered. OR are recoded item-wise into 0 or NA depending upon the time the test taker needed to respond to the respective item (rt_{ji}) in relation to the critical response time $rt_{i,crit}$ for that item. If the individual's response time is smaller or equal to the critical response time, the OR is considered to be MCAR. Ignorability is given in this case so that OR are correspondingly coded as not administered. Otherwise, in the cases when the individual response time exceeds the critical response time, the OR is considered to be MNAR (non-ignorable) and coded as 0. This rationale is applied to all items included in the item pool.

Iterative Process

The proposed method uses item response theory models (IRT; e.g., de Ayala, 2009) for placing the test takers on an ability scale based on their responses given to the test items. Therefore, one or several item parameter variables and an ability parameter variable are estimated. IRT scaling is often accomplished in a two-step procedure where item parameters are estimated in the first step and abilities are consecutively estimated in the second step with item parameters fixed at the values from step one. In the proposed method, both steps, item parameter and ability estimation, depend on the current state of recoding of the OR of all items except item

i. As the item parameter(s) of item *i* are probably inaccurate due to the yet unclear coding of the OR, the responses for the particular item are not used for ability and item parameter estimation.

Obviously, the estimation of abilities is also less accurate for items that were analyzed at the beginning of the procedure when about no recoding of OR was realized so far (so that the OR had to be deterministically coded to a default value; typically as incorrect) and gets more accurate the more items had been recoded before. To circumvent this problem, the procedure described above is repeated until the matrix **D** with *N* lines and *I* columns, containing the responses of all test takers to all presented items after recoding, does not change across two consecutive iterations (i.e., no additional OR recoding did take place). At the beginning of each iteration, the item ordering is randomized. Figure 2 shows a flowchart including all steps of the proposed method.

Please insert Figure 2 around here

Empirical Illustration

In this section, the presented new response times-based method to deal with OR is illustrated. The illustration serves two purposes. First, it underlines the empirical applicability of the method and shows how to embed it in test calibration processes. Second, it examines not-yet known effects of the new method on item and ability parameter estimates and on the variance and the reliability of the measured ability. Regarding the second aspect, the new method is compared with the two common deterministic methods to (a) code OR as incorrect and (b) code OR as not administered. The illustration strives to answer four research questions:

1. Which differences in item difficulty estimates can be observed when the new method for dealing with OR is used compared to coding OR as incorrect or coding OR as not administered?

Based on empirical findings, we can expect that test takers with a low ability tend to omit items with a higher probability compared to test takers with a high ability. Thus, applying the new method will result in coding some OR as NA that would have been coded as incorrect under the deterministic rule OR = 0, and in coding some OR as incorrect that would have been coded as not administered under the deterministic rule OR = NA. Thus, for the first case it can be expected that lower difficulties will result for some items for the new method compared to the deterministic method OR = 0. Correspondingly, it can be expected that some item difficulties will be higher for the new method compared to the deterministic method OR = NA. Any of the two deterministic methods can be considered as appropriate only if they lead to the same item difficulty estimates as the new method. The larger the differences of the item difficulties are the "less correct" is the respective deterministic method.

Since the new method employs scaling with an IRT model, potential effects of the new method on the difficulty estimates will be reflected by corresponding effects on the ability estimates. Anyhow, in order to understand which impact the new method can have on the test results that are directly relevant for the test takers, the differences between the three methods are examined with the second research question:

2. Which differences in ability parameter estimates can be observed when the new method for dealing with OR is used compared to coding OR as incorrect or coding OR as not administered?

Using the consideration from above, it can be expected that the new method will lead to higher ability estimates for some test takers when compared to the deterministic method OR = 0 and to lower ability estimates for some test takers when compared to the deterministic method OR = NA.

While deriving assumptions for differences between the three methods regarding difficulty and ability parameter estimates is straightforward, predictions regarding the effect of the new method on more aggregated statistics such as the ability variance and the reliability are more difficult. Therefore, these effects are examined in an explorative manner with the following two research questions:

3. Which differences in the ability variance can be observed when the new method for dealing with OR is used compared to coding OR as incorrect or coding OR as not administered?

4. Which differences in the reliability can be observed when the new method for dealing with OR is used compared to coding OR as incorrect or coding OR as not administered?

The examination of the four research questions uses data from the calibration study of a computer-based test measuring student skills in applying information and communication technologies (ICT-skills; Wenzel et al., 2016). This test uses simulations-based items varying largely in their complexity and response time.

Method

Participants.

The examined sample consisted of N = 766 students (46 % female; mean age: 15.2 years, SD = 0.57) from two German federal states (Baden-Württemberg: 392 students; Rhineland-Palatinate: 374 students). In these two federal states, schools were asked if they are willing to participate. 33 volunteering schools, equipped with enough suitable computers, were selected to participate in the calibration study. The majority of the tested students attend the ninth grade (71 %). Prior to the testing date, written declarations of consent that their children can participate in the study were collected from the parents. The students received a feedback regarding their test performance and no further compensation.

Materials and Procedure.

The presentation of the ICT-skills test items took place on school computers. The responses gathered during testing and log-data records were transferred to a file server using a secure web-protocol. The log-data comprised the response times of the students to each presented item. The complete item pool consisted of 70 dichotomously scored items (0 = incorrect; 1 = correct). Because presenting all 70 items to each student would have resulted in very long test sessions, 11 test compositions were assembled. Each of these compositions contained about half of the available items. The items were balanced across the set of compositions with respect to expected response times, the underlying cognitive processes necessary to solve an item (access, manage, integrate, evaluate, create; see International ICT Literacy Panel, 2002) and the type of ICT application addressed (e.g., e-mail application, folder structure, presentation software, spreadsheets, text processors, web browsers). Each student had to answer one randomly assigned test composition. Single items were presented 256 to 329 times (M = 291, SD = 19). The students had 60 minutes to answer the items including time of a small introduction to familiarize themselves with the testing system at the start. Wenzel et al. (2016) give additional information regarding the original study, the sample, and the testing instrument. Results regarding validity can be found in Engelhardt, Naumann, Goldhammer, Frey, Wenzel, Hartig, and Horz (in press).

Since the new method to deal with OR is typically carried out as an early step of data preparation and, therefore, prior to the usual scaling and item selection steps, all 70 items were analyzed here. In each of the three research conditions, the OR in the data were treated with one of the three methods compared (new response time-based method, OR = incorrect, OR = not administered). Sequences of missing responses at the end of the compositions (not reached) were coded as not administered. Response times were defined as the time from the first presentation of the item until a “next” button was pressed.

Implementation of the Proposed Method.

When applying the new method to deal with OR, item parameters were determined according to the Rasch model. Abilities were estimated with Warm's (1989) weighted maximum likelihood estimator. Both the estimation of Rasch item parameters and the ability estimation were executed by functions from the "TAM"-package (Robitzsch, Kiefer, & Wu, 2018) in the R computing language (R Core Team, 2017). For identification purposes, the ability mean was constrained to zero. The function pwr.t2n.test from the "pwr"-package (Champely, 2017) was modified to handle heterogeneous variances for the independent samples *t*-test. These functions were called from R-code that was written anew to carry out the new method.

At the beginning of the procedure, the OR were coded as incorrect. As described above, this default setting was iteratively tested by the new method. The error probabilities were set to $\alpha = \beta = .10$ and the step width to rt_{Δ} = 1 seconds.

Results

The number of observed OR per item ranged from 0 to 34 with a median of 7.5 (3.2 % of the item responses possible). The average response time (valid responses and OR) per item ranged from 40.91 seconds to 241.04 seconds (M = 105.21; SD = 39.87). Individual test takers skipped 0 to 17 items (M = 0.86; SD = 1.74). The data set thereby contains a relative small amount of OR. The number of OR per test taker showed a correlation of -.42 with ability and is thus in the range of the findings reported for PIRLS and PIAAC, as reported above. The algorithm of the new response time-based method converged after three iterations. It resulted in reasonable response time thresholds ranging from 18 to 58 seconds given the type and content of the ICT-skills items. As can be expected due to the small proportion of OR in the data set, only for 11 items were thresholds necessary. For the other items, the average ability was not significantly lower for students with an OR to the respective item compared to students with a

valid response to that item. In total, 660 OR were recoded. 260 (39 %) of these OR were identified to be MCAR and thus treated as not administered. For 400 (61 %) OR, the MCAR assumption was rejected and the MNAR assumption adopted. These OR were recoded into incorrect responses.

The first research question focusses on the difference between the difficulty estimates obtained under the three compared methods. The average item difficulty using the new method is $M = 0.42$ ($SD = 1.63$), $M = 0.44$ ($SD = 1.63$) for the condition OR = 0, and $M = 0.36$ ($SD = 1.66$) for the condition OR = NA. Figure 3 shows the differences in difficulty parameter estimates for each of the 70 items between the new method and the deterministic coding of OR = 0 (panel A) and between the new method and the deterministic coding of OR = NA (panel B). In line with the result that more OR were coded as incorrect by the new method than as not administered, more and somewhat larger item difficulty differences were observed when comparing with OR = NA (max. difference: 0.34 logits) than with OR = 0 (max. difference: -0.23 logits). Interestingly, larger differences resulted for easy items than for difficult items, especially in the OR = NA conditions. This underlines that applying the same deterministic recoding to NA for all items is most likely inappropriate for dealing with OR.

Please insert Figure 3 around here

Research question 2 focusses on the differences in ability estimates caused by the choice of method for dealing with OR. Since the average ability was constrained to be zero for identification purposes, the mean ability differs not between the compared methods. Potential average differences between the methods can be seen in the difficulty estimates reported above. Note that, when the item parameters would have been constrained to 0 for identification, the

average differences reported for the item difficulties would have resulted for ability. On the individual level, applying the new method changed ability estimates up to 2.010 logits when compared to OR = 0 and up to -0.68 logits when compared to OR = NA as reference method. Overall, the differences were relatively small while the differences between the new method were a bit larger when compared to deterministically treating OR as not administered compared to treating OR as incorrect (see Figure 4).

Please insert Figure 4 around here

The research questions 3 and 4 focus on the effects of the new method on ability variance and reliability compared to the two deterministic methods to deal with OR. Table 2 shows the variances and reliabilities obtained with the three methods. The differences on this high level of aggregation are small with differences on the second decimal. The results for the new method range between the two deterministic methods.

Please insert Table 2 around here

Discussion

A new method to deal with OR applicable for computer-based ability tests is introduced. The method uses item response times to statistically test for each individual item of a test whether observed OR are MCAR or MNAR with fixed type-1 and type-2 error levels. If the hypothesis of MCAR is maintained, OR are coded as not administered. If the hypothesis of MCAR is rejected and the alternative of MNAR adopted, OR are coded as incorrect. The method is derived from the theory of missing data of Rubin and colleagues.

In the empirical illustration, the proposed method was successful in identifying item-specific critical response time thresholds significantly differentiating OR being MCAR from OR being MNAR. About 61 % of the OR were coded as incorrect and the rest as not administered. Therefore, the frequently used deterministic approach of scoring all OR as incorrect seems to be a bit “more correct” than scoring OR as not administered. Of course, this might be different for other data sets. Anyhow, that does not change the fact both deterministic approaches are inappropriate to handle OR. Different processes are likely to cause OR in one data set. Response times help to differentiate these processes and to decide whether they are due to random processes (because they were presented in too short a time) or are connected to a lack of ability.

In direct comparison, the observed differences between the three compared methods were relatively small but systematic. With up to 2.2 times the standard deviation, effects on the ability estimates were much larger on an individual level. Thus, the choice of the method to deal with the OR has a relevant impact on the test result for some test takers. Furthermore, it is well possible that the occurrence of non-ignorable versus ignorable OR is systematically linked to individual characteristics of the test takers. In this case, specific groups of persons are likely to benefit more than others from applying deterministic methods to deal with OR. In how far and which group-related results (e.g., from large-scale assessments) are affected remains an interesting empirical question for future research. Currently we do not know if the issues connected with the coding of OR are a substantial problem with regard to empirical findings based on test data. Lastly, the results show that statistics on the scale level such as the variance and the reliability are only slightly affected be the choice of method for handling OR. However, more experiences with different kinds of data sets are needed to draw conclusions on systematic effects on these statistics.

From a practical point of view, we recommend using the new method as an early step of test data analysis. It should be carried out after the data have been prepared and checked for plausibility and before the actual IRT scaling. If one wants to check whether the new method was successful, model-based approaches can be applied between the execution of the new method and the actual IRT scaling. Based on the considerations and results of Robitzsch (2016) we recommend the two-dimensional model proposed by Mislevy and Wu (1996) for this purpose. This model can be used to examine (a) whether non-ignorable NA-codes remain and (b) whether the missing propensity depends on the unobserved values themselves.

Since the new method incorporates statistical inference, it will work best for large data sets with a large proportion of OR. Nevertheless, the data used in for illustration purposes in the present paper underlines that the method is also useful for medium sized data sets with a relatively small OR proportion. If the basic idea should be used for even smaller data sets, the Mann–Whitney *U*-test or the Fisher-Pitman permutation test for the equality of means are nonparametric alternatives to the *t*-test. Since the new method means making decisions about individual test takers based on group statistics and have a certain probability of being incorrect on the individual level (in the illustration $\alpha = \beta \leq .10$), we recommend the new method for situations where reporting is carried out on the group level only. For example, large-scale assessments such as PISA, PIRLS, TIMSS, or PIAAC are an appropriate field of application for the new method. We are not yet recommending the method for high-stakes tests from which decisions for single individuals are derived, where missing are a much smaller issue anyway. Even though the error probabilities are known, incorrect recodings can and will occur that might make a more or less large difference for individuals.

Summing up, the suggested method proved to be a promising easy method to deal with OR in computer-based testing. Future studies will show in how far it might be suitable as a standard

procedure especially for large-scale assessments. The application of the method to larger data sets is currently in progress.

References

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling* (pp. 243–277). Mahwah, NJ: Lawrence Erlbaum Publishers.

Champely, S. (2017). *pwr: Basic functions for power analysis* [software]. R package version 1.2-1.

Chen, H. Y., & Little, R. J. A. (1999). A test of missing completely at random for generalised estimating equations with missing data. *Biometrika, 86,* 1–13.

De Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. NY: The Guildford Press.

De Ayala, R. J., Plake, B. S., Impara, J. C. (2001). The impact of omitted responses on the accuracy of ability estimation in item response theory. *Journal of Educational Measurement, 38,* 213–234.

Diggle, P. J. (1989). Testing for random dropouts in repeated measurement data. *Biometrics, 45,* 1255–1258.

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling, 8,* 430–457.

Engelhardt, L., Naumann, J., Goldhammer, F., Frey, A., Wenzel, S. F. C., Hartig, K., & Horz, H., (in press). Convergent evidence for validity of a performance-based ICT skills test. *European Journal of Psychological Assessment*.

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement, 45,* 225–245.

Frey, A., Hartig, J., & Rupp, A. (2009). Booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice, 28*, 39–53.

Glas, C. A. W., & Pimentel, J. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological Measurement, 68*, 907–922.

Goldhammer, F., & Kroehne, U. (2014). Controlling Individuals' Time Spent on Task in Speeded Performance Measures: Experimental Time Limits, Posterior Time Limits, and Response Time Modeling. *Applied Psychological Measurement*, *38*, 255–267

Goldhammer, F., Martens, T., & Lüdtke O. (2017). Conditioning factors of test-taking engagement in PIAAC: an exploratory IRT modelling approach considering person and item characteristics. *Large-scale Assessments in Education*, *5*(1).

Hedges, L.V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*, 107–128.

Holman, R., & Glas, C. A. W. (2005). Modelling non-ignorable missing data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology*. *58*, 1–17.

International ICT Literacy Panel. (2002). *Digital transformation: A framework for ICT literacy*. Princeton, NJ. Retrieved October 23, 2014 from http://www.ets.org/research/policy_research_reports/publications/report/2002/cjik

Kim, K. H., & Bentler, P. M. (2002). Tests of homogeneity of means and covariance matrices for multivariate incomplete data. *Psychometrika*, *67*, 609–624.

Köhler, C., Pohl, S., & Carstensen, C. H. (2015). Taking the missing propensity into account when estimating competence scores: Evaluation of item response theory models for

nonignorable omissions. *Educational and Psychological Measurement, 75*, 850–874.

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

Lee, Y.-H., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling*, *53*, 359-379.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83,* 1198–1202.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: Wiley.

Ludlow, L. H., & O'Leary, M. (1999). Scoring omitted and not-reached items: Practical data analysis implications. *Educational and Psychological Measurement, 59*, 615–630.

Mislevy, R. J., & Wu, P. K. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing* (RR-96-30-ONR). Princeton, NJ: Educational Testing Service.

Neyman, J., & Pearson, E. S. (1933). On the testing of statistical hypotheses in relation to probability a priori. *Proceedings of the Cambridge Philosophical Society, 29,* 492–510.

OECD (2016). *Technical report of the Survey of Adult Skills (PIAAC)* (2nd edition). Paris: Author.

O'Muircheartaigh, C., & Moustaki, I. (1999). Symmetric pattern models: A latent variable approach to item non-response in attitude scales. *Journal of the Royal Statistical Society, Series A, 162*, 177–194.

R Core Team (2017). *R: A language and environment for statistical computing* [software]. Vienna: R Foundation for Statistical Computing.

Robitzsch, A. (2016). Zu nichtignorierbaren Konsequenzen des (partiellen) Ignorierens fehlender Item Responses im Large-Scale Assessment [On non-negligible consequences of (partially) ignoring missing item responses in large- scale assessments]. In B. Suchań, C. Wallner-Paschon, & C. Schreiner (Eds.), *PIRLS & TIMSS 2011 - die Kompetenzen in Lesen, Mathematik und Naturwissenschaft am Ende der Volksschule: Österreichischer Expertenbericht* (pp 55-64). Graz: Leykam.

Robitzsch, A., Kiefer, T., & Wu, M. (2018). *TAM: Test Analysis Modules* [software]. R package version 2.9-35.

Rose, N., von Davier, M., & Xu, X. (2010). *Modeling non-ignorable missing data with IRT* (ETS Research Report No. 10-11). Princeton, NJ: Educational Testing Service.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley & Sons.

Rutkowski, L., Gonzales, E., von Davier, M., & Zhou, Y. (2014). Assessment design for international large-scale assessments. In L. Rutkowski, M. von Davier & D. Rutkowski, D. (Hrsg.), *Handbook of international large-scale assessment: background, technical issues, and methods of data analysis* (pp. 75–95). Boca Raton: CRC Press.

Satterthwaite, F. E. (1946). An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin, 2*, 110–114. doi:10.2307/3002019

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. Journal of Educational Measurement, 46, 247–272.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450.

Weeks, J. P., von Davier, M., & Yamamoto, K. (2016). Using response time data to inform the coding of omitted responses. *Psychological Test and Assessment Modeling, 58*, 671–701.

Welch, B. L. (1947). The generalization of "Student's" problem when several different population variances are involved. *Biometrika, 34*, 28–35. doi:10.1093/biomet/34.1-2.28

Wenzel, S. F. C., Engelhardt, L., Hartig, K., Kuchta, K., Frey, A., Goldhammer, F., Naumann, J., & Horz, H. (2016). Computergestützte, adaptive und verhaltensnahe Erfassung Informations- und Kommunikationstechnologie-bezogener Fertigkeiten (ICT-Skills) [Computerized adaptive and behaviorally oriented measurement of Information and communication technology-related skills (ICT-skills)]. In Bundesministerium für Bildung und Forschung (BMBF) Referat Bildungsforschung (Hrsg.), *Forschungsvorhaben in Ankopplung an Large-Scale Assessments* (pp. 161-180). Niestetal: Silber Druck.

Wilkinson, L., & Task Force on Statistical Inference, American Psychological Association, Science Directorate. (1999). Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, *54*, 594–604. http://dx.doi.org/10.1037/0003-066X.54.8.594.

Wise, S. L., & Kingsbury, G. G. (2016). Modeling student test-taking motivation in the context of an adaptive achievement test. *Journal of Educational Measurement, 53*, 86–105.

Wise, S. L., & Ma, L. (2012, April). *Setting response time thresholds for a CAT item pool: The normative threshold method*. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, Canada.

Footnotes

[1]At this point, we discuss missing responses at a general level. A distinction between the types “missing by design”, “not reached”, and “omitted” is introduced later in the text in the section “Types of Missing Item Responses”.

Table 1

Type of missing response, dominant missing data mechanism and typical scoring applied in achievement tests

Type	Dominant Missing Data Mechanism	Typical Scoring
Missing by design	MCAR	Not administered
Not reached	MCAR	Not administered
Omitted	MNAR, MCAR, MAR	Incorrect

Note. MCAR = missing completely at random; MNAR = missing not at random; MAR = missing at random.

Table 2

Variances and reliabilities of the ability estimates for three different methods to deal with omitted responses (OR)

Method	Variance	Reliability
OR = incorrect	0.861	.681
OR = not administered	0.846	.659
New response time-based method	0.858	.673

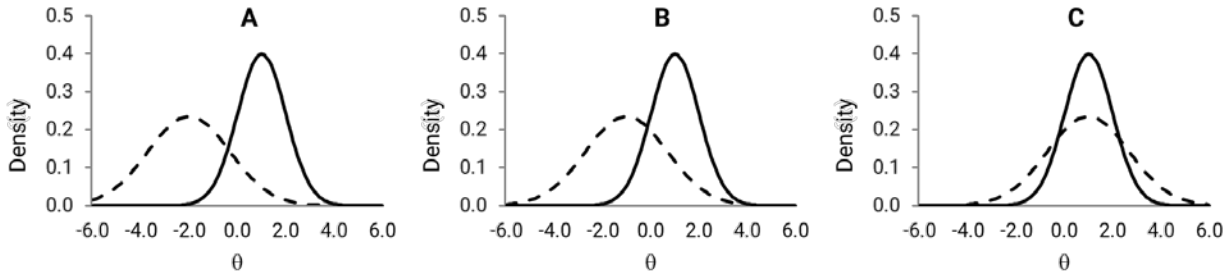

Figure 1. Hypothetical densities for the ability θ of test takers with a valid response to the analyzed item (solid lines) compared to test takers with an omitted response for this item (dashed lines). The abilities are estimated based on the responses given to the other items of the test with the analyzed item excluded. Panel A shows all test takers with an omitted response to the analyzed item, panel B test takers with an omitted response and a medium response time for this item, and panel C test takers with an omitted response and a small response time for this item.

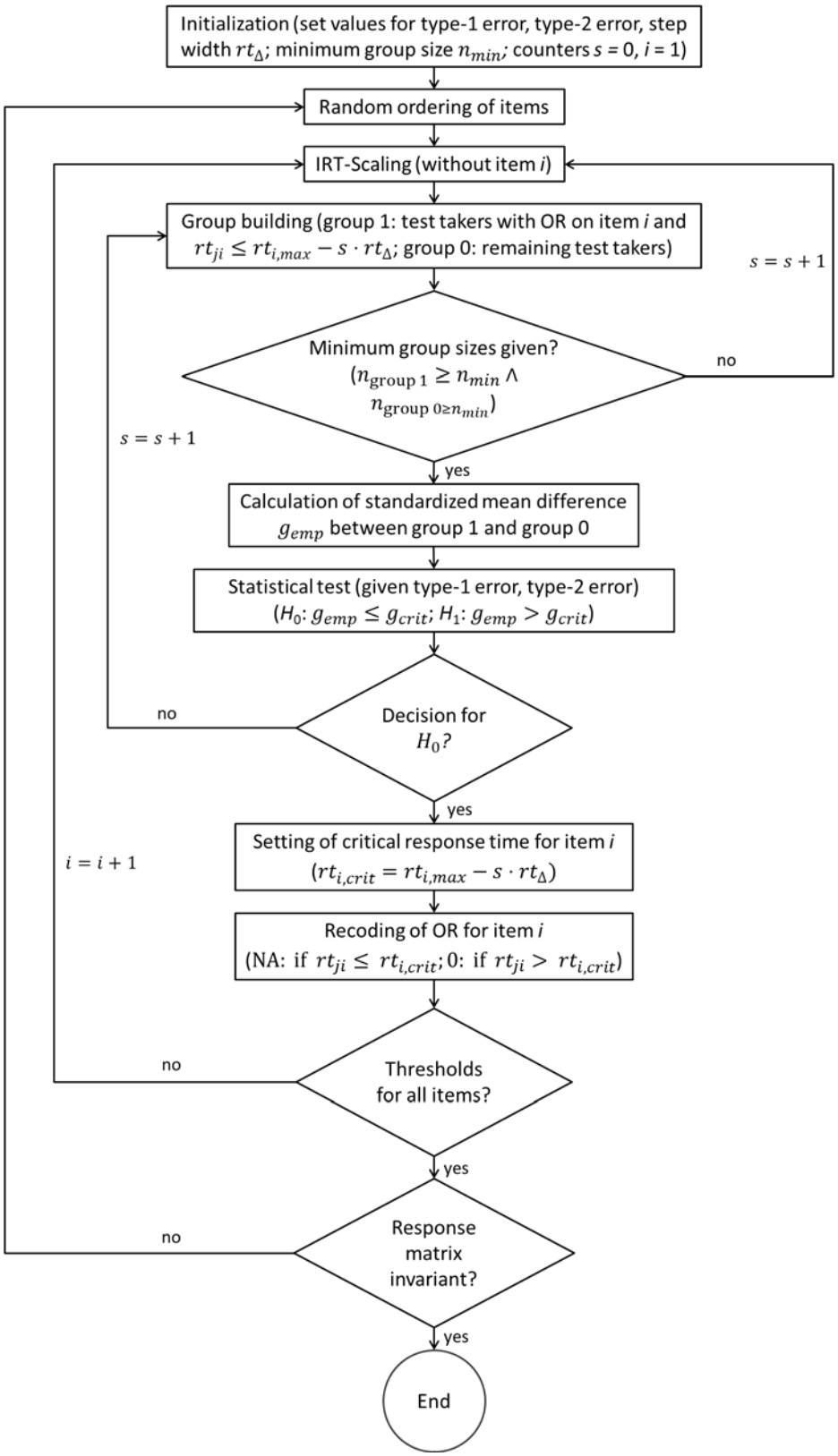

Figure 2. Flowchart of the new method for recoding omitted responses into not administered (NA) or incorrect (0) based on item response times.

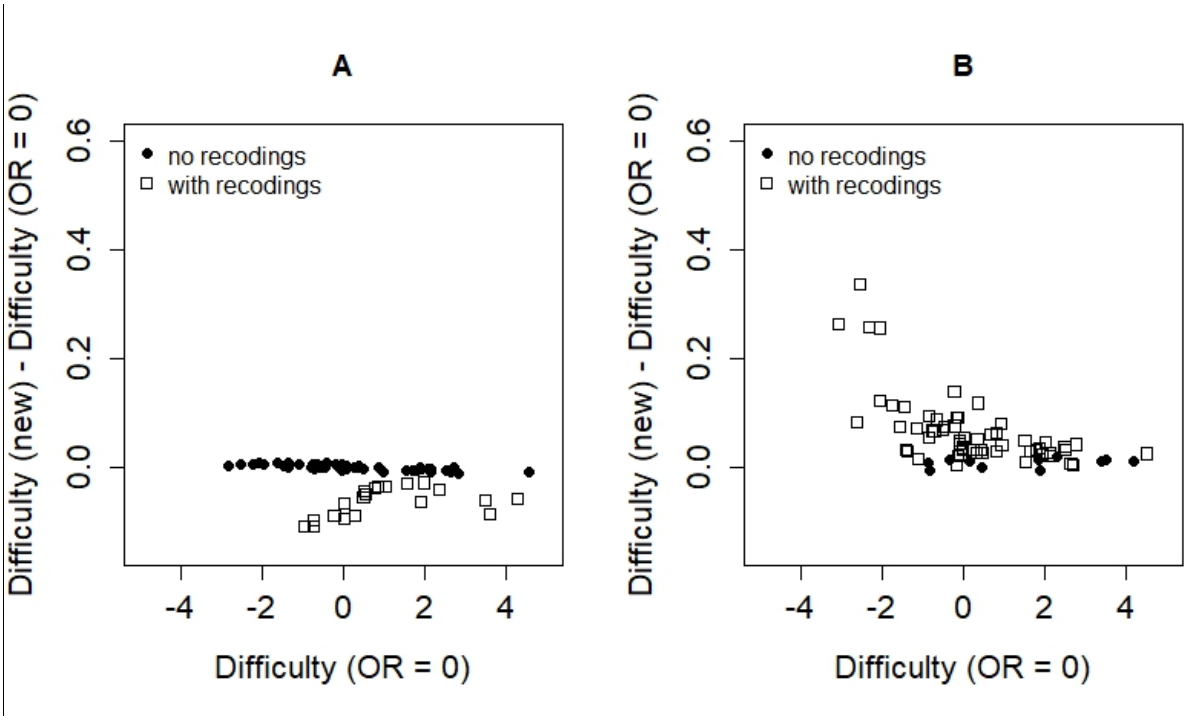

Figure 3. Differences between item difficulty estimates for 70 items obtained with the new response time-based method and scoring all omitted responses as incorrect (OR = 0) or as not administered (OR = NA) by item difficulty. Items with no recoding of ORs are displayed with filled dots; items with recoding of ORs are displayed with unfilled squares.

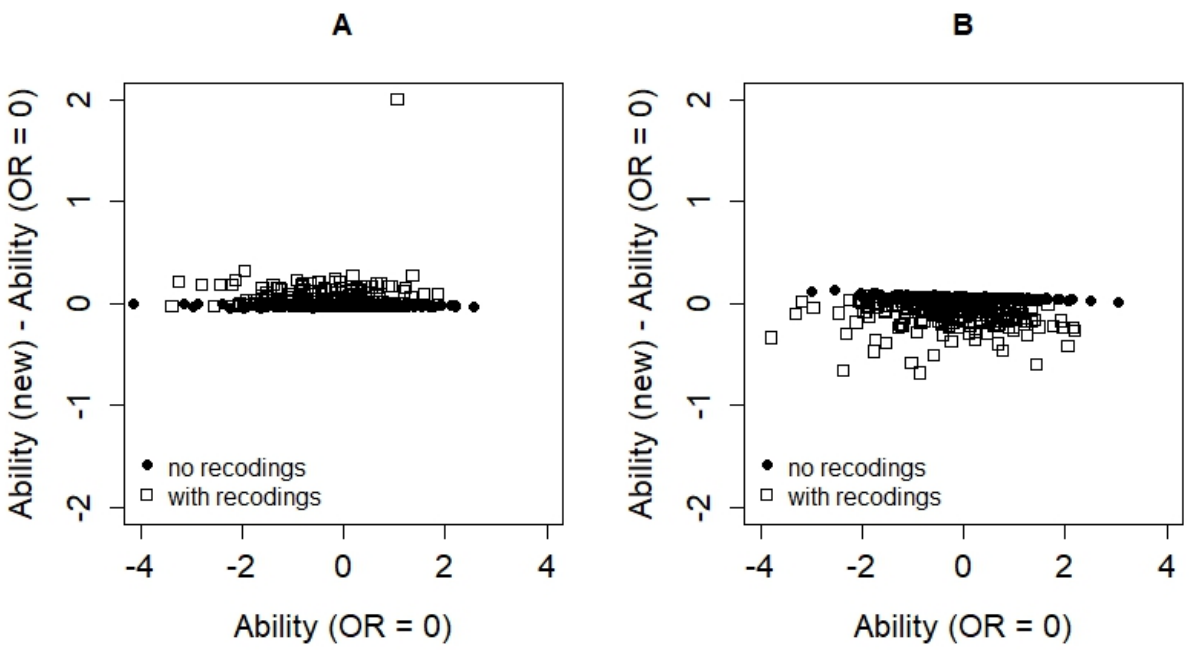

Figure 4. Differences between ability estimates for N = 766 students obtained with the new response time-based method and scoring all omitted responses as incorrect (OR = 0) or as not administered (OR = NA) by ability level. Items with no recoding of ORs are displayed with filled dots; items with recoding of ORs are displayed with unfilled squares.