Identifying careless responders in routine outcome monitoring data

G. Franz
S1143301
Master Thesis Clinical Psychology
Supervisor: Dr. J. M. Conijn
Institute of Psychology
Universiteit Leiden
01-07-2016

Abstract

In routine outcome monitoring (ROM) self-report instruments are often used to assess clients' symptoms of psychopathology. Careless responding, due to, for example, lack of motivation, concentration problems or insufficient language skills, can seriously undermine the validity of ROM self-report data.

In a simulation study we examined how well 11 post-hoc statistical indices are able to detect careless responding. The indices included person-fit measures, outlier statistics and consistency measures. Furthermore we determined the effects of careless responding on scale totals and ability parameters, averaged on group level. By applying cutoffs from the simulation study to a real-life ROM data set (*N* = 3,483), the prevalence of careless responding in ROM data was determined.

We found that person-fit indices and Mahanalobis distances worked well for different base rates and types of careless responding. Random responding affected scale totals and ability parameter estimates for those respondents with manipulated answers but for the group as a whole the effects were very small. Depending on the index and the chosen cutoff for classifying respondents as careless, prevalence rates for careless responding up to 33% were found.

More evidence is needed to determine whether these high prevalence rates are due to the lengthy ROM procedure used. It is also possible that the indices show aberrant but meaningful answer patterns by respondents with atypical symptom profiles.

1 Introduction

1.1 Routine outcome monitoring

In health care as well as in mental health care there is a growing demand to ensure that the therapies offered work as intended and are worth their money. The question whether a therapy works can be answered on different levels: Howard, Moras, Brill, Martinovich, and Lutz (1996) stated that “it is not sufficient for the practitioner to know that a particular treatment can work (efficacy) or does work (effectiveness) on average” (p. 1060). What clinicians need to know is whether a certain treatment is working for this client at this time. To this end clinicians can systematically and regularly assess the outcome of psychotherapies during their course. Such feedback systems monitoring clients’ progress during psychotherapies are known as routine outcome monitoring (ROM, de Beurs et al., 2011).

ROM systems have by now been implemented in many countries and in many different settings in mental health care (Trauer, 2010). Studies showed that implementing ROM can significantly improve treatment success (Boswell, Kraus, Miller, & Lambert, 2015; Carlier et al., 2012). ROM can provide feedback on the development of clients’ wellbeing to clinicians as well as clients. Furthermore, it can provide managers and policy makers with information for their decision making (Trauer, 2010). ROM data in aggregated form (e.g. aggregated over departments or over institutions) can also be used for benchmarking purposes to compare the quality of mental health care (Hoenders et al., 2014; Nugter & Buwalda, 2012).

For ROM purposes clinician-rated scales such as the Health of the Nation Outcome Scales (Wing et al., 1998) as well as self-report instruments such as the Brief Symptom Inventory (BSI, Derogatis & Melisaratos, 1983) are implemented to assess clients’ psychological problems and symptoms of psychopathology. Often a combination of disorder specific and general distress scales is used. Some mental health institutions use a battery of instruments at intake and at the end of treatment whereas others use fewer scales at frequent intervals (de Beurs et al., 2011).

An example of an elaborate ROM procedure is the assessment done collaboratively by Leiden University Medical Center and the Mental Health Care Centre Rivierduinen. From 2002 on the two institutions implemented a procedure to assess all patients referred to them for outpatient treatment of mood, anxiety and somatoform disorders. During the intake, patients are assessed with a standardized diagnostic interview and observer-rated as well as self-report instruments (van Noorden et al., 2012). Some general distress scales such as the

BSI are used for all patients; disorder-specific scales are added depending on the patients' diagnosis. In total the intake-session takes about two hours (de Beurs et al., 2011). If outpatient treatment is started, the assessment is repeated every three to four months. Patients that do not speak Dutch well enough, are illiterate or suffer from serious cognitive impairments do not participate in the ROM procedure. The data collected is primarily used for diagnostics and to evaluate a patient's course of psychotherapy (van Noorden et al., 2012). In anonymized form it is also used for research.

1.2 Careless Responding

When using self-report instruments the underlying assumption is that respondents answer the questions as truthfully and accurately as possible (Marjanovic, Holden, Struthers, Cribbie, & Greenglass, 2014). Regarding ROM, Boswell et al. (2015) stated that „these systems and their usefulness in treatment are predicated on accurate self-reporting of levels of disturbance and corresponding changes." (p. 12). Unfortunately, this accuracy often might be lacking. Respondents may show a certain response style such as acquiescence (i.e., the tendency to agree with statements) or the tendency of preferring extreme answer possibilities (Osborne & Blanchard, 2011). Respondents might choose answers based on a social desirability bias. Or they might - especially with longer instruments - grow tired, lose concentration and become inclined to answer carelessly (Barnette, 1999; Emons, 2008). The latter is known as careless response, random response or inconsistent response (Huang, Curran, Keeney, Poposki, & DeShon, 2012). Characteristic for these response patterns is that test-takers "respond without reference to the content of the items" (Baer, Ballenger, Berry, & Wetter, 1997, p. 139). Nichols, Greene, & Schmolck (1989) call such response behavior content nonresponsivity.

A number of studies examined the prevalence of careless responding in survey data - often in the context of personality assessment inventories - and the consequences this response behavior can have on study results. Meade and Craig (2012) studied careless responding in a sample of undergraduate psychology students (N = 386) that completed a long online survey of mainly personality measures. Using latent profile analysis they found that 11% of the participants could be classified as extremely inattentive. This group consisted of two distinct classes of careless response: A small group (2% of all participants) answered using long strings of identical answers in a row. The other group (9% of all respondents) answered in a random way. Across their series of studies on careless behaviour, Maniaci and

Rogge (2014) reported “that roughly 3–9% of respondents exhibited problematic levels of inattention” (p. 80). Several other studies found that over 50% of participants admitted to responding in a random way on one or more items of a personality inventory (Baer et al., 1997; Berry et al., 1992). Credé (2010) found that base rates of careless response as low as 5% can have a substantial impact on observed correlations.

1.3 Detection Methods

The different approaches to detect careless responding can be divided in two main groups: direct (or item based) screening methods and statistical (or post-hoc) screening methods (DeSimone, Harms, & DeSimone, 2015; Meade & Craig, 2012). In direct screening methods, items (or even whole scales) are added to a questionnaire with the purpose of detecting careless response. One can ask participants directly how much effort they put into answering and how reliable they estimate their answers to be (Meade & Craig, 2012). Another way is to use validity scales that contain items that every conscientious responder either should or should not endorse such as “I can remember a time when I spoke to someone who wore glasses” (Fervaha & Remington, 2013, p. 1356). Another method uses pairs of items that ask the same (or the opposite) thing, so-called semantic synonyms (or antonyms) where conscientious responders are expected to give matching answers (DeSimone et al., 2015).

Many validity scales were developed for large personality assessment inventories. For the Minnesota Multiphasic Personality Inventory, for example, there exist several validity scales. Some contain items that less than 10% of the subjects endorse (infrequency scales), others scales with pairs of items that should be answered in the same way (inconsistency scales) (Berry, et al., 1992; Tellegen, 1988). The development and normative testing of such a validity scale is time-consuming and expensive (Marjanovic, Struthers, Cribbie, & Greenglass, 2014)(Marjanovic, Struthers, Cribbie, & Greenglass, 2014).

Statistical methods to detect aberrant response patterns use among others outlier statistics or measures of consistency. For questionnaires administered by computer completion time can also be used as an indicator for careless response (Meade & Craig, 2012).

1.4 Study aim

The focus of research on careless response has so far been in the context of personality assessment inventories as these usually are lengthy and therefore prone to careless response

(Maniaci & Rogge, 2014). Until now it has not been researched if and to what extent careless responses from mental health care patients lead to biased ROM data and how such careless responses could be detected. It can be assumed that using lengthy ROM assessment batteries comes with a considerable risk of careless response that might invalidate the data. This study aims to answer three research questions (RQs) regarding careless response in ROM data:

RQ 1: Which statistical indices are useful for detecting careless responders in ROM data?

RQ 2: What effects might careless response have on the results of the self-report instruments, i.e., on scale totals and latent trait estimates?

RQ 3: Is there evidence for careless response in ROM data?

These questions were addressed by means of a simulation study (RQ1 and RQ2) and a real-data application (RQ3) to ROM data (N = 3,483). The simulation study used item parameters estimated in the empirical ROM data set. To address the first RQ, several statistical indices were calculated post-hoc for the simulated datasets and used to predict careless responding in logistic regressions models. Predictive power of the indices then was compared. To address the second RQ, scale totals and latent trait estimates of data sets with and without simulated careless responding were compared. To address the third RQ, optimal cutoff points derived from the simulation study were then applied to the real ROM data set to estimate prevalence of careless responding in ROM data.

2 Methods

2.1 Participants

This study uses ROM data from outpatients referred to Leiden University Medical Center and the Mental Health Care Centre Rivierduinen for treatment of mood, anxiety and somatoform disorders. Included are all 3,543 patients that had their baseline ROM measurement between 01.01.2005 and 31.12.2008 and completed the BSI (Derogatis & Melisaratos, 1983), the Mood and Anxiety Symptoms Questionnaire (MASQ - Watson et al., 1995), and the Dimensional Assessment of Personality Pathology – Short Form (DAPP-SF - van Kampen, de Beurs, & Andrea, 2008). In addition to the questionnaire data demographic data such as housing situation, employment status, and country of birth was collected. Although the data was collected using touch-screen computers (van Noorden, et al., 2012), 60 participants still had missing item score values. It was decided to only include complete datasets in this study. This led to a final sample size of 3,483 patients (64.6 % woman) with age ranging from 17 to 91 (M_{Age} = 39.02, *SD* = 12.69). 80% of the patients were born in the Netherlands. 7.6 % of the sample only completed primary education, 29.4 % lower secondary education, 34.6 % higher secondary education and 17.7 % higher professional or university education. The data was anonymized; according to the Ethical Review Board of the Leiden University Medical Center it may be used for research purposes (van Noorden et al., 2012).

2.2 Measurement Instruments

The BSI is a shortened version of the Symptom Checklist (SCL-90) and consists of 53 items divided in nine subscales that reflect main symptom domains of psychiatric disorders such as somatization, anxiety or depression (Derogatis & Melisaratos, 1983; De Beurs & Zitman, 2006). Subjects indicate to what degree they experienced the problem described in each item during the past week. Answers are scored on a 5-point Likert scale ranging from 0 (“not at all”) through 4 (“extremely”). 49 of the 53 items are assigned to a subscale; the subscales consist of four to seven items each. Subscales include among others somatization, obsessive-compulsive, depression and phobic anxiety (Derogatis & Melisaratos, 1983)(Derogatis & Melisaratos, 1983). The total sore on the BSI reflects the general degree of psychopathology (de Beurs, den Hollander-Gijsman, Helmich, & Zitman, 2007).

The MASQ was developed as an instrument to measure anxiety and mood disorder symptoms following the tripartite model of anxiety and depression (Watson et al., 1995). It consists of 90 items asking about symptoms of depression and anxiety disorders on a 5-point Likert scale ranging from 1 (“not at all”) to 5 (“extremely”). 76 of these 90 items are assigned to one of five subscales. They measure anhedonic depression (22 items), anxious arousal (17 items) and general distress, the latter divided into three subscales general distress depression (12 items), general distress anxiety (11 items) and general distress mixed (14 items). 15 items in total are reversely scored, nearly all of which belong to the anhedonic depression subscale. The majority of items in the anhedonic depression subscale describe positive feelings. When being rescored, these items represent lack of positive affect (de Beurs er al., 2007).

The DAPP-SF is the short form of the DAPP - Basic Questionnaire and consists of 136 items(de Beurs, Rinne, van Kampen, Verheul, & Andrea, 2009) (de Beurs, Rinne, van Kampen, Verheul, & Andrea, 2009). Its 18 subscales reflect different dimensions of disordered personality (van Kampen et al., 2008) and consist of six to ten items each. DAPP-SF subscales include among others identity problems, social avoidance, narcissism, and self-harm. The subscales can be combined to four second-order factors: emotional dysregulation, dissocial behavior, inhibitedness, and compulsivity (de Beurs et al., 2009). The DAPP-SF can be used as a screening-tool for personality disorders.

2.3 Indices for careless response

For every participant, a total of 11 screening indices were calculated to detect careless response. Most of these indices were calculated per subscale. Of the in total 279 items belonging to the three scales 261 are assigned to one of the 32 subscales. The indices calculated per subscale therefore made use only of these 261 items.

2.3.1 Statistical synonyms (Syn) and antonyms (Ant)

Statistical synonyms and antonyms use the concept of semantic synonyms / antonyms but instead of adding a-priori items to a scale that have the same (or opposite) item content, pairs of items that are post-hoc identified as highly (positively or negatively) correlated are used. Conscientious respondents are expected to answer both items in a pair similarly (or dissimilarly in case of antonyms). Thus, for Syn and Ant each, 30 pairs of items (using 44 unique items for Syn and using 23 unique items for Ant) were identified that had high positive (Syn) or negative (Ant) correlations to define statistical synonyms and antonyms. The schematic in Figure 1 depicts the calculation of Syn; the procedure for the calculation of Ant is similar. For both indices the correlations between the items were calculated using the original item scores of the 279 items before reversed coding. For the synonyms the 30 highest correlations ranged from .88 to .72, for the antonyms the 30 lowest ranged from -.68 to -.54. For the simulation study, only those pairs could be used where both items were elements of the 261 items assigned to a subscale. Therefore, 29 items pairs (using 42 unique items) were used to calculate Syn, 12 item pairs (using 12 unique items) to calculate Ant in the simulation. For the real-data application all pairs were used.

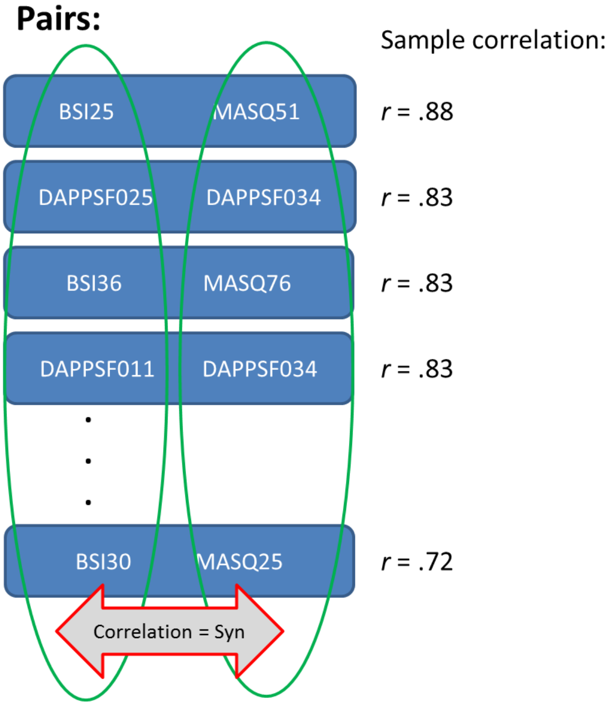

Figure 1. Calculation of Syn

The within-person correlation between those synonymous pairs and the within-person correlation between those antonymous pairs were then calculated (Meade & Craig, 2012;

DeSimone et al., 2015). Conscientious responding should lead to high absolute correlations; lower correlations for the synonyms and higher correlations for the antonyms are taken as an indication of more careless response. For participants with no variance on the item scores of the first or second half of the pairs, Syn or Ant could not be calculated.

2.3.2 Long string indices

Respondents answering carelessly may give the same answer for several items in a row (e.g. ticking the “Somewhat agree”-box on a Likert scale without reading the item). By determining the length of strings of consecutive identical answers such a response behavior might be detected (DeSimone et al., 2015). For every participant, the maximum length of a string of consecutive identical answers (Lmax) and the average length of consecutive identical answers (Lmean) were calculated (Meade & Craig, 2012). To compute Lmax and Lmean, the original item scores before recoding were used in the order that the items were presented.

2.3.3 Inter-item standard deviation (ISD)

The answers of a conscientious respondent on a unidimensional subscale should be consistent with each other. Marjanovic et al. (2014) introduced the ISD as a measure for that consistency. For all unidimensional subscales, per participant the SD of his or her answers on that subscale were calculated and then averaged over the subscales. Higher ISD values are taken as indicative of a more careless response.

2.3.4 Even-odd consistency (EO)

For this measure of answer consistency, all unidimensional subscales were divided into an even and odd part (i.e., the even-numbered items and the odd-numbered items apart). For every even- and odd-scale the average score was calculated. Then the correlation between the even- and odd-averages was calculated and corrected for test-length using the Spearman–Brown prophesy formula (DeSimone et al., 2015; Meade & Craig, 2012). Lower correlations indicate more careless response.

2.3.5 Mahanalobis distance (Ma)

Outliers, i.e., answers that differ strongly from the answers given by the majority of respondents can be an indication of careless response (DeSimone et al., 2015). Mahanalobis distances, i.e., multivariate outlier statistics, were calculated for every unidimensional subscale. The distances then were averaged over all subscales (Meade & Craig, 2012; Zijlstra, van der Ark, & Sijtsma, 2011). High values are taken as taken as indicative of a more careless response.

2.3.6 Item-Based Outlier Score (O)

This simple outlier score introduced by Zijlstra, van der Ark, and Sijtsma (2007) is based on the rank order of the answer categories per item, with the modal answer having rank 0, the next popular answer rank 1 etc. Per unidimensional subscale the rank numbers of the answers chosen by each participant were added and then averaged over all subscales. Higher O values indicate a choice of more unpopular answer categories and thus possibly careless response.

2.3.7 Person-fit indices

IRT based person-fit indices such as the statistic lz (Drasgow, Levine, & Williams, 1985)(Drasgow, Levine, & Williams, 1985) can be used to identify aberrant response. The lz-value is the standardized log-likelihood of a vector of answers under the estimated unidimensional IRT model (Conijn, Emons, & Sijtsma, 2014)(Conijn, Emons, & Sijtsma, 2014). Per unidimensional subscale and participant, the lz-values were calculated using the Graded Response Model (GRM, Embretson & Reise, 2000) and then averaged over all subscales. Lower log-likelihood values indicate bad person-fit and possibly careless response.

Additionally a non-parametric person-fit approach was used: The number of Guttman errors (Gu; Emons, 2008). A Guttman error occurs when a respondent endorses a difficult answer without endorsing an easier one. For every unidimensional subscale the number of Guttman errors was calculated; the number of errors then was averaged over the subscales. To account for the fact that the number of possible Guttman errors is dependent on the length of the test and the subscale total, as an additional index normed Guttman errors (nGu; Emons, 2008) were calculated for every unidimensional subscale and then averaged over the subscales. For subjects with a perfect score on a subscale (i.e., choosing only the highest or only the lowest answer category for all items of a subscale nGu could not be calculated and thus the average was calculated disregarding that subscale. For both types of Gutmann indices higher values are taken as an indication of more careless response.

2.3.8 Model Assumptions

The indices calculated per subscale as well as the Item Response Theory (IRT) model used to conduct the simulation study requires the subscales to be unidimensional, i.e., measuring only one trait. The ratio of the first to the second eigenvalue of the polychoric correlation matrix was examined as a heuristic test for sufficient unidimensionality for all subscales (Embretson & Reise, 2000, p. 228). Furthermore a non-graphical scree test was conducted (Raîche, Walls, Magis, Riopel, & Blais, 2013)(Raîche, Walls, Magis, Riopel, & Blais, 2013). For all BSI subscales and some of the DAPP-SF subscale multidimensionality was not a problem as both tests recommended one factor solutions. For those subscales where the tests favored using more than one factor the second eigenvalue was considerably lower (by factor 3 minimum) than the first. Conijn et al. (2014) found that small violations of unidimensionality are tolerable for detecting aberrant behaviour using IRT methodology. Therefore it was concluded that the scales used here are suitable for IRT analysis.

2.3.9 Recoding of indices

To make the comparison of the different indices easier it was decided that all indices should have the same orientation, i.e., higher values should indicate more careless response. Therefore the values of lz, EO, and Syn and were multiplied with -1.

2.4 Analyses

The research design used here was partly based on the one implemented by Meade and Craig (2012).

2.4.1 Simulation study design

To evaluate the performance of the indices under different conditions, a simulation study was conducted using the GRM. For an item with five ordered categories, as is the case for all items of the scales used in the current study, the GRM is defined by four category-threshold parameters ($\beta_1 – \beta_4$) indicating the popularity of an answer category plus a slope parameter α. The first step in the simulation study was to estimate the parameters for each of the subscales in the ROM data using the R package mirt (Chalmers, 2012) (Chalmers, 2012). With these parameters, 200 replicated samples of N = 5000 each were simulated: First, for every simulated participant, theta-values per subscale were drawn from a multivariate normal distribution taking into account the correlations of the subscale theta parameters of the ROM sample participants. Then item responses were simulated using the probabilities derived from the participants' theta values and the GRM item parameters.

Into the datasets careless responses were inserted using a 3 x 6 factorial design. Factor Baserate simulated 5%, 10% or 20% of the 5000 respondents responding carelessly. Factor Type simulated six different types of careless response. The first three types of careless responding all represented random responding varying in their degree and distribution across the scales:

- Random responses in 25% of the 261 items (i.e., 65 items) - condition Ran25

- Random responses in 50% of the 261 items (i.e., 130 items) - condition Ran50
- Random responses in 25% of all items but situated in the second half of the questionnaire (i.e., 65 items of items 131 – 261) - condition Ran25sec

The distribution of random responses was based on previous research suggesting that careless responding probably does not occur on all items of a questionnaire but only on a part of the questionnaire (Meade & Craig, 2012) and that that there may be more careless responses on items that occur later in a test (de Beurs et al., 2012). Item scores to be replaced were chosen randomly within the limits of the conditions. Random responding was operationalized by drawing responses from a uniform distribution.

A further two conditions represented careless response behavior of using the same response in a row regardless of item content. To simulate such consecutive identical answers (or longstring answers) several strings of responses with different length were inserted into the dataset at randomly chosen positions. DeSimone et al. (2015) suggested that strings of answers longer than six to fourteen items in a row might be indicative of careless response. Therefore in one condition on average 25% (corresponding to 65 items) of items were replaced by chains of answers (condition Lstr25). In total 10 chains were inserted of which six were of length seven and four of length six. In another condition, 10 chains (four of length fourteen and six of length thirteen) were inserted (condition Lstr50), on average leading to replacement of 50% (corresponding to 130 items) of items. For the insertion of the chains 10 items were sampled from items 1 to 260 (without replacement). For each of the sampled items that item’s answer category was then repeated the number of times given by the chosen string length. If a string exceeded the questionnaire (e.g. by trying to insert a string of length thirteen after item 255) the string was cut at the last item. This method led to some variation in the total number of items inserted. In a test run using item scores of 50,000 simulated participants, on average 65 (corresponding to 25% - condition Lstr25) and 130 items (corresponding to 50% - condition Lstr50) were inserted by this method. The procedure did not control for the overlap of strings.

In reality a dataset would probably contain a mixture of different types of careless response. Therefore the sixth condition in Factor Type combined all of the above mentioned five types of careless responding. Each type of careless responding was represented by 20% of the careless responders.

2.4.2 Evaluating the usefulness of the indices (RQ1)

To test the performance of the indices and thereby answering RQ1, in a first step per condition a logistic regression was estimated for every replication sample predicting the probability of there being careless response. This was done separately for every index. Subsequently the results were averaged per condition. Classification tables were calculated giving the frequencies of true versus predicted random / non-random response and thus showing the predictive power of each logistic regression model. As the prevalence of the simulated random response was relatively low it was decided to use each condition's base rate as cutoff when dichotomizing the fitted probability values (Agresti, 2007)[1].

In a second step four measures were calculated that enabled a comparison of the predictive power of the logistic regression models per index and condition. First, per condition the sensitivity (i.e., the percentage of true positives), second the specificity (i.e., the percentage of true negatives) of each index were calculated and compared. Third, for every regression model Cohen's coefficient kappa (Cohen, 1960)(Cohen, 1960) was calculated for the predicted classification and the true classification and averaged per condition. Fourth, for every regression model Tjur's (2009) coefficient of determination *D* was calculated and averaged per condition. Tjur's *D* measures the amount of variance explained by a logistic regression model, is easy to understand, shares many characteristics with the coefficient of determination R^2 and has advantages over other pseudo R^2-measures: Looking at the distribution of the predicted probabilities for the fail and the success group (i.e., those with the predictor being 0 vs. the predictor being 1) *D* is simply the difference in mean of the two distributions.

2.4.3 Examining effects of careless responding (RQ2)

For all samples the indices described were being calculated as well as the scale totals and the participants' theta values. To calculate participants' scale totals, subscale mean scores were calculated for every subscale and subsequently averaged per scale. To answer RQ2 the difference between the scale totals and the theta values of the careless responses and those of the original ones (i.e., the simulated responses before the careless responses were inserted) were calculated for every person and then averaged per sample and per condition.

[1] When optimal cutoffs were determined later for the samples, it turned out that for all indices the probabilities for the optimal cutoffs found were not far from the base rate which justifies using the base rate as cutoff instead of the often used .5 in the performance analysis.

2.4.4 Prevalence of careless responding in real ROM data set (RQ3)

To answer RQ3, in a first step the correlations of the indices calculated for every person of the ROM sample were examined. In a second step optimal cutoff criteria from the simulation study were applied to the real ROM data set. With the mix conditions being the most realistic one, these conditions were used to choose cutoffs for the indices that optimized sensitivity and specificity. A Receiver Operating Characteristics (ROC) analysis was conducted to see which cutoff would maximize sensitivity and specificity(Swets, 1973) (Swets, 1973).

ROC curves plot the true positive rate (i.e., sensitivity) against the false positive rate (i.e., 1 – specificity). Beginning and end of the curve correspond to the situation that none (sensitivity of 0) or all (specificity of 0) of the respondents are classified as careless. The points within are a tradeoff between identifying most careless responders without classifying many conscientious responders as careless.

For that purpose, Youden (1950) introduced an index J which is calculated as sensitivity + specificity – 1. This index – sometimes also called informedness - gives the vertical distance between the diagonal and the ROC curve in a ROC diagram thus indicating the degree to which a classification works better than chance (Powers, 2011)(Powers, 2011). For every logistic model of the mix-conditions informedness was calculated for all possible cutoffs and the cutoff with maximum informedness value determined. If there were ties, the one with a higher sensitivity value was chosen. If there still were ties, the lowest cutoff was chosen. Per index and base rate condition the optimal cutoffs were averaged and subsequently applied to the real ROM data set to determine the extent of careless response.

3 Results

3.1 Preliminary analysis of indices in simulation samples

In a preliminary analysis the behavior of the indices was investigated by computing the average change in the index values between the original samples and the manipulated samples per condition. Understanding how indices behave when, for example, the base rate of careless responding changes, can contribute to assessing the suitability of an index in a certain ROM context. Table A1 (in the appendix) shows these average ‘index biases’, separately for careless responders, conscientious responders and the whole sample.

Three groups of indices can be identified: For most indices (ISD, EO, Syn, Ant, Lmax, Lmean) adding careless responses to a sample did not change the index values for the conscientious responders. Those of careless responders increased, independently of Factor Baserate, those of the whole sample increased slightly. This increase was higher for higher base rates. Exceptions were indices Lmax and Lmean that in the random conditions (i.e., Ran25, Ran50, and Ran25sec) actually decreased for the careless responders.

Indices lz and Ma belong to a second group of indices where some sort of standardization took place so that the average index values for the whole group were about zero in all conditions. Here there is a positive index bias for the careless responders that decreases as base rate increases, and a much smaller negative bias for the conscientious responders that becomes smaller (i.e., closer to zero) with decreasing base rate. For the third group of indices (Gu, nGu, and O) there is a positive index bias for the careless responders, a small positive one for the conscientious responders and one for the whole group. These biases increase with base rate.

3.2 Performance of indices for detection of careless response (RQ1)

In order to answer RQ 1 logistic regression models were fitted for all indices and conditions so that predictive performance of the indices could be compared. To measure the performance of the indices sensitivity, specificity , Cohen’s Kappa and Tjur’s *D* were used. Figure 2 shows informedness (as a summary measure for sensitivity and specificity) per condition and index, averaged over Factor Baserate.

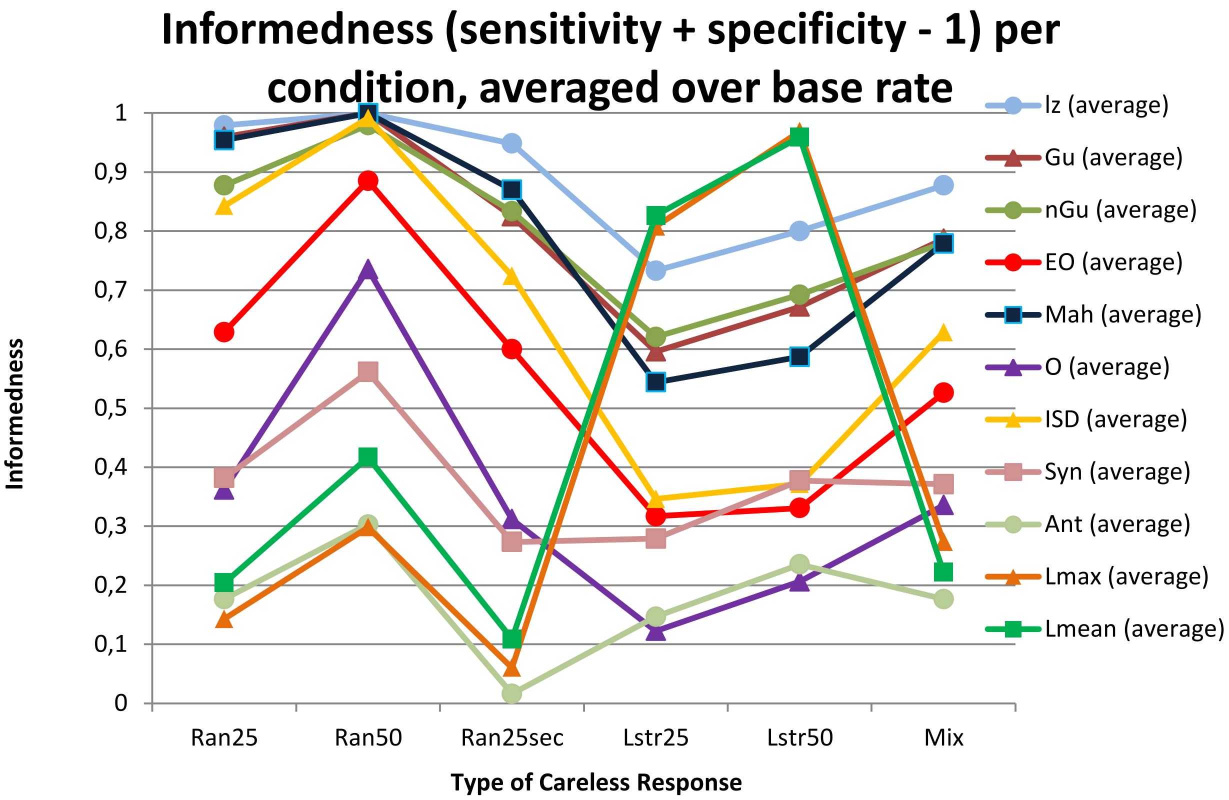

Figure 2. Informedness (sensitivity + specificity - 1) per condition and index, averaged over base rate

Figure 3 shows Tjur's *D* and Figure 4 shows Kappa, both per condition and index, averaged over Factor Baserate.

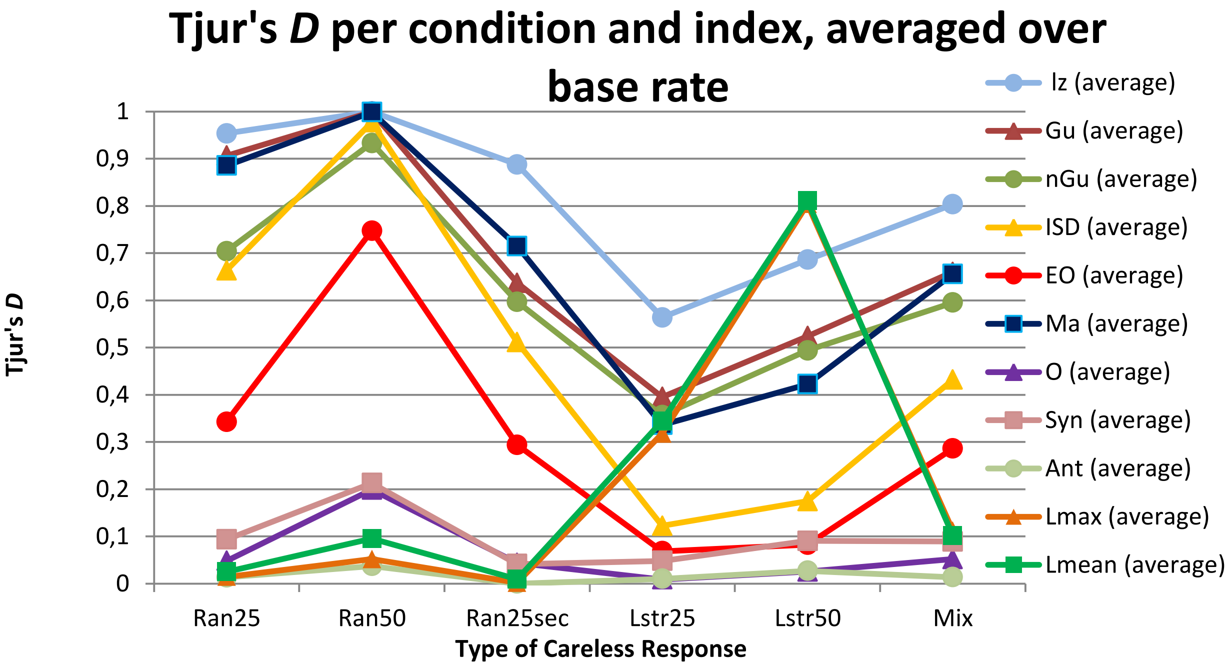

Figure 3. Tjur's *D* per condition and index, averaged over base rate

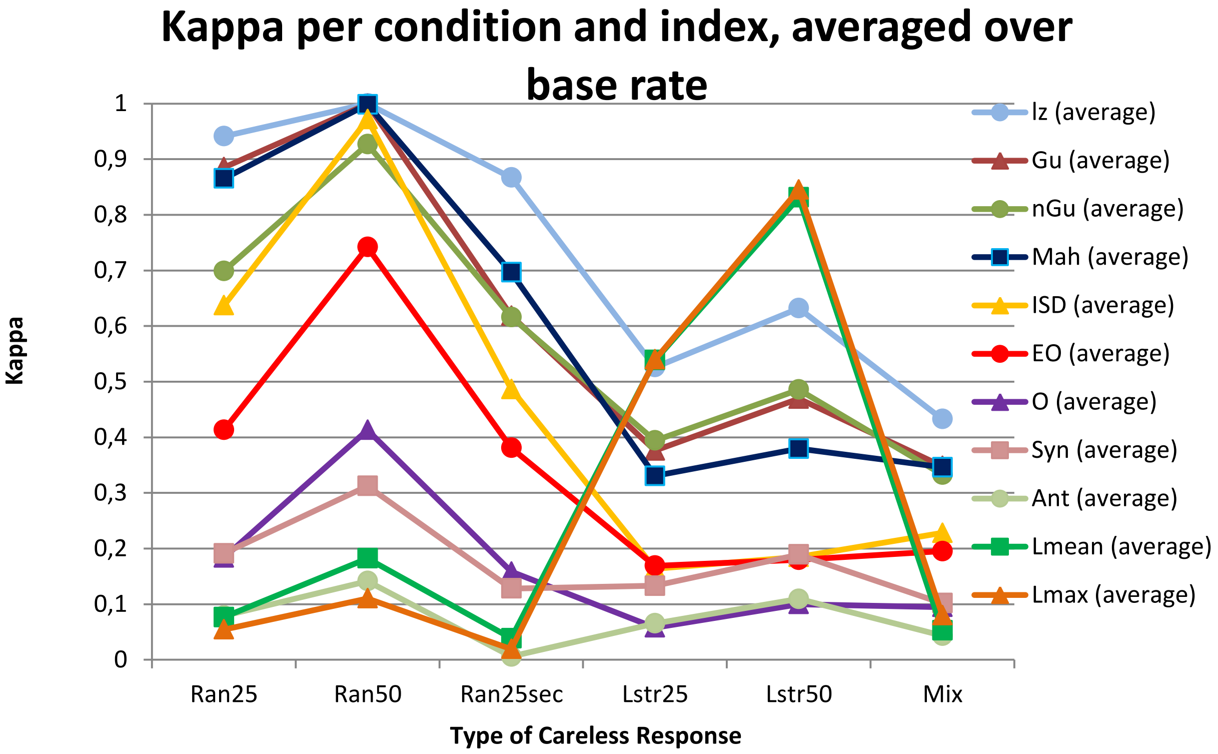

Figure 4. Kappa per condition and index, averaged over base rate

3.2.1 Effect of Factor Type

As can be seen from the graphs the general picture concerning the performance of the indices in the different conditions is more or less the same for all measures of performance. Table 1 shows informedness values, Table 2 Tjur's *D* per index and condition.

Table 1. Average informedness values per index and condition

Index	Base rate	Ran25	Ran50	Ran25sec	Lstr25	Lstr50	Mix
lz	.05	0.98	1.00	0.96	0.74	0.81	0.88
	.1	0.98	1.00	0.95	0.74	0.80	0.88
	.2	0.98	1.00	0.94	0.72	0.78	0.87
Gu	.05	0.96	1.00	0.83	0.60	0.68	0.79
	.1	0.96	1.00	0.83	0.60	0.67	0.79
	.2	0.96	1.00	0.82	0.59	0.66	0.79
nGu	.05	0.88	0.98	0.83	0.62	0.70	0.78
	.1	0.88	0.98	0.83	0.62	0.69	0.78
	.2	0.88	0.98	0.83	0.62	0.68	0.78
ISD	.05	0.84	0.99	0.72	0.35	0.38	0.63
	.1	0.84	0.99	0.72	0.35	0.37	0.63
	.2	0.84	0.99	0.72	0.35	0.37	0.63
EO	.05	0.63	0.88	0.60	0.32	0.33	0.53
	.1	0.63	0.89	0.60	0.32	0.33	0.53
	.2	0.63	0.88	0.60	0.32	0.33	0.53
Ma	.05	0.95	1.00	0.90	0.56	0.63	0.79
	.1	0.96	1.00	0.88	0.55	0.60	0.78
	.2	0.95	1.00	0.83	0.52	0.54	0.77
O	.05	0.36	0.74	0.31	0.13	0.22	0.34
	.1	0.36	0.74	0.31	0.12	0.21	0.34
	.2	0.36	0.73	0.31	0.12	0.19	0.33
Syn	.05	0.38	0.57	0.27	0.28	0.38	0.37
	.1	0.38	0.56	0.27	0.28	0.38	0.37
	.2	0.38	0.56	0.27	0.28	0.38	0.37
Ant	.05	0.18	0.30	0.02	0.15	0.24	0.18
	.1	0.18	0.30	0.01	0.15	0.23	0.18
	.2	0.18	0.30	0.01	0.14	0.23	0.18
Lmax	.05	0.14	0.30	0.06	0.83	0.97	0.27
	.1	0.14	0.30	0.06	0.80	0.97	0.28
	.2	0.14	0.30	0.06	0.80	0.97	0.28
Lmean	.05	0.20	0.42	0.11	0.83	0.96	0.22
	.1	0.21	0.42	0.11	0.83	0.96	0.22
	.2	0.20	0.42	0.11	0.82	0.96	0.22

For Factor Type distinct patterns of performance behavior can be discerned for three groups of conditions: First the three random conditions (i.e., Ran25, Ran50, Ran25sec), second the two longstring conditions (i.e., Lstr25 and Lstr50) and last the Mix condition. In the random conditions indices lz, Gu, nGu, Ma, and ISD show high performance, followed by EO. O and Syn work poorer in detecting careless response and Lmean, Lmax, and Ant show low performance. Performance in condition Ran50 is better than in condition Ran25; for Ran25sec the indices perform as well or slightly worse than in condition Ran25. All indices beside Lmean and Lmax show lower performance in the longstring conditions than in the random conditions. In de Mix condition the order of performance is the same as for the random conditions.

Table 2.Average values of Tjur's *D* per index and condition

Index	Base rate	Ran25	Ran50	Ran25sec	Lstr25	Lstr50	Mix
lz	.05	0.95	1.00	0.88	0.53	0.68	0.79
	.1	0.96	1.00	0.89	0.57	0.69	0.81
	.2	0.96	1.00	0.89	0.60	0.69	0.81
Gu	.05	0.88	1.00	0.57	0.35	0.50	0.63
	.1	0.91	1.00	0.64	0.40	0.53	0.66
	.2	0.93	1.00	0.70	0.43	0.54	0.68
nGu	.05	0.64	0.91	0.51	0.29	0.45	0.54
	.1	0.71	0.94	0.60	0.36	0.50	0.60
	.2	0.77	0.95	0.68	0.42	0.53	0.65
ISD	.05	0.59	0.97	0.45	0.09	0.14	0.39
	.1	0.67	0.98	0.52	0.12	0.18	0.44
	.2	0.73	0.98	0.57	0.16	0.20	0.47
EO	.05	0.26	0.69	0.21	0.04	0.05	0.23
	.1	0.34	0.75	0.29	0.07	0.08	0.29
	.2	0.42	0.80	0.37	0.10	0.12	0.34
Ma	.05	0.86	1.00	0.71	0.31	0.44	0.64
	.1	0.89	1.00	0.73	0.34	0.44	0.66
	.2	0.91	1.00	0.70	0.35	0.39	0.66
O	.05	0.02	0.09	0.02	0.00	0.02	0.03
	.1	0.04	0.18	0.04	0.01	0.03	0.05
	.2	0.08	0.32	0.07	0.01	0.04	0.08
Syn	.05	0.05	0.14	0.02	0.03	0.05	0.05
	.1	0.09	0.21	0.04	0.04	0.09	0.09
	.2	0.14	0.29	0.06	0.07	0.14	0.13
Ant	.05	0.01	0.02	0.00	0.00	0.01	0.01
	.1	0.01	0.03	0.00	0.01	0.02	0.01
	.2	0.02	0.06	0.00	0.02	0.04	0.02
Lmax	.05	0.01	0.03	0.00	0.17	0.72	0.09
	.1	0.01	0.05	0.00	0.31	0.81	0.12
	.2	0.02	0.08	0.00	0.47	0.88	0.14
Lmean	.05	0.01	0.05	0.00	0.21	0.74	0.08
	.1	0.02	0.09	0.01	0.34	0.82	0.11
	.2	0.04	0.15	0.02	0.49	0.88	0.12

When comparing the different performance measures, there are minor discrepancies in the ordering of the indices: In the Lstr25 condition, the kappa values (shown in Table 3) for lz are as high as for Lmean and Lmax, the informedness values for lz are lower as for Lmean and Lmax, Tjur’s *D* for lz is higher than Lmean and Lmax. Kappa values for lz, Gu, nGu and Ma in the Mix condition are lower than in the Lstr50 condition, the informedness values of these indices in the Mix condition are higher than those in the Lstr50 condition but lower than those in the Ran25 condition.

Table 3. Average kappa values per index and condition

Index	Base rate	Ran25	Ran50	Ran25sec	Lstr25	Lstr50	Mix
lz	.05	0.91	1.00	0.82	0.39	0.52	0.39
	.1	0.95	1.00	0.88	0.54	0.65	0.45
	.2	0.97	1.00	0.91	0.65	0.73	0.46
Gu	.05	0.82	1.00	0.48	0.25	0.34	0.28
	.1	0.90	1.00	0.63	0.38	0.48	0.36
	.2	0.94	1.00	0.75	0.51	0.59	0.41
nGu	.05	0.56	0.88	0.46	0.26	0.35	0.26
	.1	0.72	0.93	0.63	0.40	0.49	0.34
	.2	0.82	0.96	0.75	0.53	0.61	0.39
ISD	.05	0.49	0.95	0.34	0.09	0.10	0.16
	.1	0.65	0.98	0.49	0.16	0.18	0.23
	.2	0.77	0.99	0.63	0.25	0.27	0.30
EO	.05	0.27	0.62	0.25	0.09	0.10	0.13
	.1	0.42	0.76	0.38	0.16	0.17	0.20
	.2	0.55	0.84	0.51	0.25	0.26	0.26
Ma	.05	0.79	1.00	0.62	0.22	0.29	0.28
	.1	0.88	1.00	0.71	0.33	0.39	0.36
	.2	0.92	1.00	0.76	0.43	0.46	0.40
O	.05	0.10	0.25	0.09	0.03	0.06	0.06
	.1	0.18	0.41	0.15	0.05	0.10	0.09
	.2	0.27	0.58	0.24	0.09	0.14	0.13
Syn	.05	0.11	0.19	0.07	0.07	0.11	0.06
	.1	0.18	0.31	0.12	0.13	0.18	0.10
	.2	0.28	0.44	0.20	0.20	0.28	0.15
Ant	.05	0.04	0.07	0.00	0.03	0.06	0.02
	.1	0.07	0.13	0.01	0.06	0.10	0.04
	.2	0.12	0.22	0.01	0.10	0.17	0.06
Lmax	.05	0.03	0.05	0.01	0.35	0.74	0.05
	.1	0.05	0.10	0.02	0.56	0.87	0.08
	.2	0.09	0.18	0.03	0.70	0.93	0.11
Lmean	.05	0.04	0.10	0.02	0.37	0.72	0.03
	.1	0.07	0.17	0.03	0.54	0.85	0.05
	.2	0.12	0.28	0.06	0.70	0.92	0.07

Looking at sensitivity and specificity values separately, in most cases sensitivity and specificity values are similar (Table A2 in the appendix shows average sensitivity and specificity values per index and condition). The exceptions are Lmean and Lmax in the random conditions: For them sensitivity is much higher (e.g., 71.6 in condition Ran25 low) than specificity (e.g., 48.6 in condition Ran25 low). That means that these indices can detect random responders relatively good, but in doing so they also classify many conscientious responders as careless.

Remarkable are the high informedness values for lz in the Ran50 condition. A number of samples show informedness values of 1. This is caused by complete separation, a situation where the outcome variable is perfectly determined by the predictor (Albert & Anderson, 1984) (Albert & Anderson, 1984). A closer inspection of the confusion tables for all indices, conditions and samples revealed that this not only occurs with lz values but also with Gu and Ma. Table 4 shows the number of samples where complete separation occurs. Only condition Ran50 is affected. Although this might seem ideal in terms of prediction, in cases like these logistic models cannot be estimated as there is no maximum likelihood estimate.

Table 4. Number of samples per condition with complete separation

Type	Base rate	lz	Gu	Ma
ran50	low	193	161	142
ran50	mid	183	134	106
ran50	hi	163	111	74

3.2.2 Effect of Factor Baserate

As a consequence of using the base rate of careless response as a cutoff for the logistic regression, the informedness values for the low, mid and high base rate conditions differed very little and it seemed reasonable to pool the values. Kappa values increase with higher base rate. For Tjur this is also generally the case although there are few exceptions. To illustrate the effect of Factor Baserate on performance, Figure 5 depicts Tjur’s *D* per Factor Type and Factor Baserate for 5 indices. The effect of base rate for the other indices or for Kappa is not displayed because the effect was very similar. As the graph shows it depends on index and condition how far the values for the low, mid, and high base rate conditions lie apart.

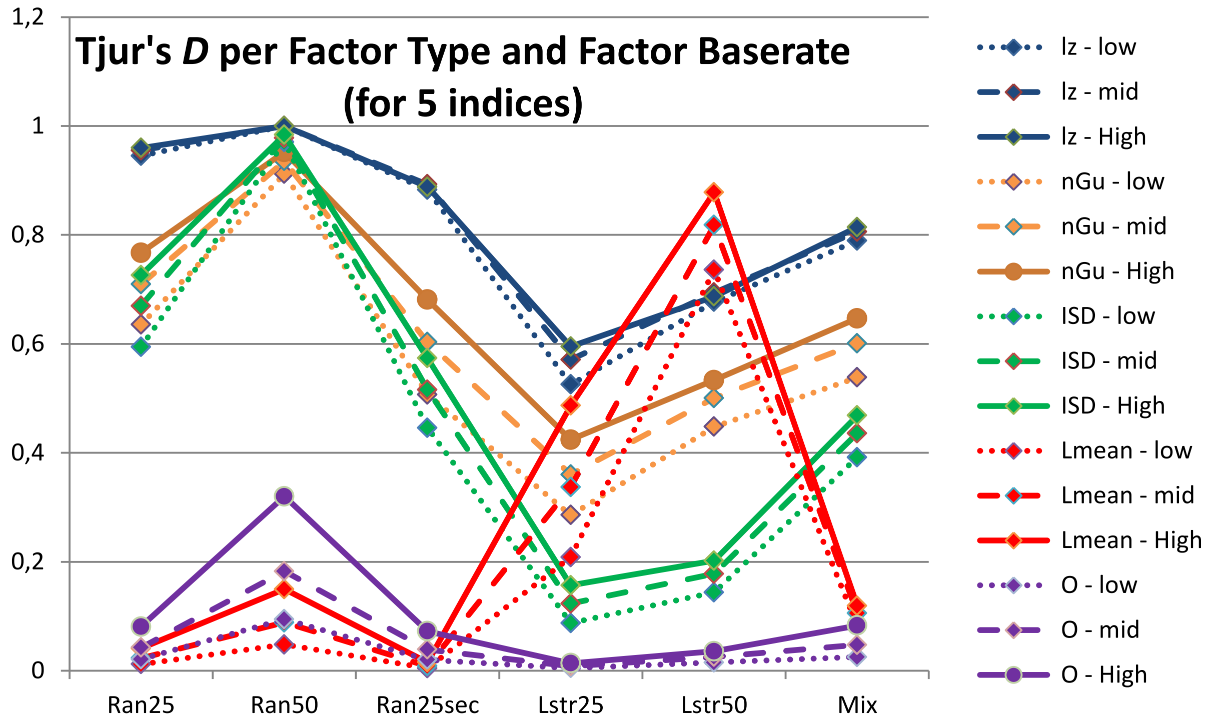

Figure 5. Tjur's *D* per Factor Type and Factor Baserate for 5 indices

3.3 Effect of careless response (RQ2)

3.3.1 Effect on scale and subscale total scores

The subscale and scale total scores were calculated for the manipulated and the original samples and the differences between the two were calculated and averaged per condition. Table 5 shows the average scale total bias for the BSI, MASQ, and DAPP-SF per condition, separately for the random responders only and the complete sample with conscientious and random responders together. Subscale totals are not shown.

Scale totals

In the Ran25 and Ran50 conditions the scale totals for all three scales were on average overestimated for the careless responders. The difference was larger in the Ran50 condition than in the Ran25 condition (e.g., average BSI total bias of 0.20 in the Ran25 low condition and of 0.40 in the Ran50 low condition). Bias was highest for the BSI totals and lowest for the MASQ totals. For condition Ran25sec, only the DAPP-SF total was affected because careless responses were inserted in the second half of the items (i.e., items131-261) where only DAPP-SF items are located.

For the longstring conditions there was no clear pattern of bias and even the bias for the careless responders was very small (range -0.06 – 0.03). In the Mix conditions there was again an overestimation of the scale totals for the careless responders which was highest for the BSI totals and lowest for the MASQ totals. In the Mix conditions the bias was somewhat smaller than in the Ran25 conditions.

Table 5. Average scale total bias per condition, separately for careless responders and the complete sample

Type	Base rate	BSI		MASQ		DAPP	
		Careless responders	All responders	Careless responders	All responders	Careless responders	All responders
Ran25	low	0.20	0.01	0.10	0.01	0.13	0.01
	mid	0.20	0.02	0.010	0.01	0.13	0.01
	high	0.20	0.04	0.10	0.02	0.13	0.03
Ran50	low	0.40	0.02	0.19	0.01	0.25	0.01
	mid	0.40	0.04	0.20	0.02	0.25	0.03
	high	0.40	0.08	0.20	0.04	0.25	0.05
Ran25sec	low	0	0	0	0	0.25	0.01
	mid	0	0	0	0	0.25	0.03
	high	0	0	0	0	0.25	0.05
Lstr25	low	0.00	0.00	-0.03	-0.00	0.02	0.00
	mid	0.00	0.00	-0.03	-0.00	0.02	0.00
	high	0.00	0.00	-0.03	-0.01	0.02	0.00
Lstr50	low	-0.01	-0.00	-0.06	-0.00	0.03	0.00
	mid	-0.00	-0.00	-0.06	-0.00	0.03	0.00
	high	-0.01	-0.00	-0.06	-0.01	0.03	0.01
Mix	low	0.12	0.01	0.04	0.00	0.13	0.01
	mid	0.12	0.01	0.04	0.00	0.13	0.01
	high	0.12	0.02	0.04	0.01	0.13	0.03

Note. Subscale and scale totals were calculated as item score means using item scores 0 – 4.

Factor Baserate had no influence on the bias when only the careless responders were considered. When the whole sample was examined bias grew with the base rate. For example, the bias was 0.02, 0.04, and 0.08 for the three increasing base rates in the Ran50 condition.

Subscale Totals

At subscale level, in the Random and the Mix conditions most BSI, MASQ, and DAPP-SF subscales showed overestimation of the subscale score. The exceptions were the anhedonic depression subscale and three DAPP-SF subscales. Those subscales showed a negative bias and are those with a relative high subscale total: In the real ROM data the MASQ scale total, for example, was 1.60, whereas the anhedonic depression subscale total was 2.43. In the longstring conditions a number of subscales totals showed overestimation, others underestimations with no clear pattern discernible. This pattern of bias can explain the scale total biases around zero which were discussed in the previous section.

3.3.2 Effect on latent trait estimates

Latent trait estimates theta were estimated per person and subscale. For these estimates a normal distribution with a mean of 0 and a SD of 1 were assumed. Therefore the average theta value in any sample – manipulated or not – is zero and consequently the average theta bias (i.e. theta values of manipulated sample minus theta values of original samples, averaged over subscales and samples) as well. Nevertheless adding careless responses changes people's theta values. Thus theta biases were calculated per sample separately for conscientious and careless responders and averaged per condition and scale. The results can be found in Table 6.

Table 6. Average theta bias for manipulated samples per condition and scale, separately for conscientious and careless responders

Type	Baserate	BSI		MASQ		DAPP	
		Consc. Responders	Careless responders	Consc. Responders	Careless responders	Consc. Responders	Careless responders
Ran25	low	-0.01	0.18	-0.00	0.08	-0.01	0.14
	mid	-0.02	0.17	-0.001	0.07	-0.02	0.14
	high	-0.04	0.15	-0.02	0.07	-0.03	0.12
Ran50	low	-0.02	0.35	-0.01	0.15	-0.01	0.27
	mid	-0.04	0.33	-0.02	0.14	-0.03	0.26
	high	-0.07	0.30	-0.03	0.13	-0.06	0.23
Ran25sec	low	0	0	0	0	-0.01	0.27
	mid	0	0	0	0	-0.03	0.25
	high	0	0	0	0	-0.06	0.22
Lstr25	low	0.00	0.00	0.00	-0.03	-0.00	0.05
	mid	0.00	0.00	0.00	-0.03	-0.01	0.04
	high	-0.00	0.00	0.00	-0.03	-0.01	0.04
Lstr50	low	0.00	-0.01	0.00	-0.08	-0.00	0.07
	mid	0.00	-0.00	0.00	-0.07	-0.01	0.07
	high	0.00	-0.00	0.02	-0.07	-0.01	0.06
Mix	low	-0.01	0.11	-0.01	0.02	-0.01	0.16
	mid	-0.01	0.10	-0.00	0.02	-0.02	0.15
	high	-0.02	0.09	-0.001	0.02	-0.03	0.13

The theta value biases for the random responders show a similar pattern as the scale total biases: In the random conditions there is an overestimation of latent trait estimates, with the biases for the BSI biggest (e.g. 0.35 for condition Ran50 low) and those for the MASQ smallest (e.g., 0.15 for condition Ran50 low). For the longstring conditions there are some positive and some negative biases, all close to zero. In the Mix condition the same pattern as in the random conditions is discernible but the biases are somewhat smaller.

Other than with the scale total biases the theta values for the conscientious responders show bias as well. Here a slight underestimation of latent trait estimates can be found. The effect of Factor Baserate also differs: For the random and the mix condition the random responders show a larger positive bias and the conscientious responders a larger negative bias (i.e. further away from zero) with a smaller base rate.

Looking at the subscales in the random and the Mix conditions, again there is negative bias for the MASQ anhedonic depression subscale and some DAPP-SF subscales (those with negative subscale total bias, but also some others). In the longstring conditions there again are some subscales with positive and some with negative latent trait estimate bias without a clear pattern.

3.4 Evidence for carless response in ROM data (RQ3)

3.4.1 Preliminary analysis of indices calculated from real ROM data

In a first step the careless response indices were calculated from the real ROM data. Table 7 gives an indication of the distribution of the indices. For most of the indices the distributions are fairly normal with a tail to the right.

Table 7. Descriptives of indices calculated from ROM data

Index	Minimum	Maximum	Mean	Std. Deviation
lz [a]	-1.30	2.66	-0.26	0.42
Gu	121.50	3007.50	849.04	343.30
nGu	0.05	0.63	0.20	0.08
ISD	0.19	1.61	0.89	0.18
EO [a]	-0.99	-0.21	-0.87	0.06
Ma	2.55	30.03	8.15	3.10
O	152.50	735.00	321.13	81.47
Syn [a]	-0.95	0.24	-0.58	0.20
Ant	-0.98	0.38	-0.56	0.25
Lmax	3	143	7.96	6.16
Lmean	1.19	9.33	1.49	0.30

Notes: *N* = 3843

[a] lz, EO, and Syn were multiplied with -1 so that for all indices higher values indicate more careless response.

Table 8 shows the Pearson correlation between the indices. The lz-values, the number of Guttman errors and the Mahalanobis-distances are highly correlated (*r* > 0.75), ISD and O to a smaller degree (0.50 < *r* < .82). The two longstring indices are correlated .81 but both are not substantially correlated with other indices with exception of ISD (*r* = -.39 and -.42 for Lmax and Lmean, respectively). Psychological synonyms and antonyms are neither highly correlated with the other indices.

Table 8. Correlations between indices

Index	Lz	Gu	nGu	ISD	EO	Ma	O	Syn	Ant	Lmax
lz [a]	1									
Gu	0.90	1								
nGu	0.87	0.75	1							
ISD	0.74	0.82	0.53	1						
EO [a]	0.49	0.51	0.40	0.55	1					
Ma	0.91	0.94	0.80	0.81	0.52	1				
O	0.42	0.52	0.31	0.50	0.25	0.65	1			
Syn [a]	-0.04	-0.14	0.01	-0.27	-0.04	-0.11	-0.19	1		
Ant	-0.01	-0.01	-0.04	-0.17	-0.04	-0.11	-0.31	0.22	1	
Lmax	-0.01	-0.13	0.25	-0.39	-0.06	-0.07	-0.16	0.19	0.07	1
Lmean	0.01	-0.13	0.33	-0.42	-0.03	-0.07	-0.22	0.26	0.12	0.81

[a] lz, EO, and Syn were multiplied with -1 so that for all indices higher values indicate more careless response.

3.4.2 ROC analysis of Mix condition

A ROC analysis was conducted per sample to determine which cutoff optimizes sensitivity and specificity simultaneously. These cutoffs were then applied to the real ROM data to determine the prevalence of careless responding (RQ3). This was done for the Mix conditions only, separately for Factor Baserate, as the Mix conditions simulate careless response in the most realistic way. Figure 6 depicts the ROC curves of 10 samples of the Mix Low condition for the indices lz, ISD and Lmax, giving an illustration of the sensitivity and specificity values that belong to the optimal cutoff for a high (lz), middle (ISD) and low (Lmax) performance index. The curves show some variance between the samples but the difference between the indices is much more obvious. This indicates that the prevalence figures probably will not vary much between samples either but more between indices used.

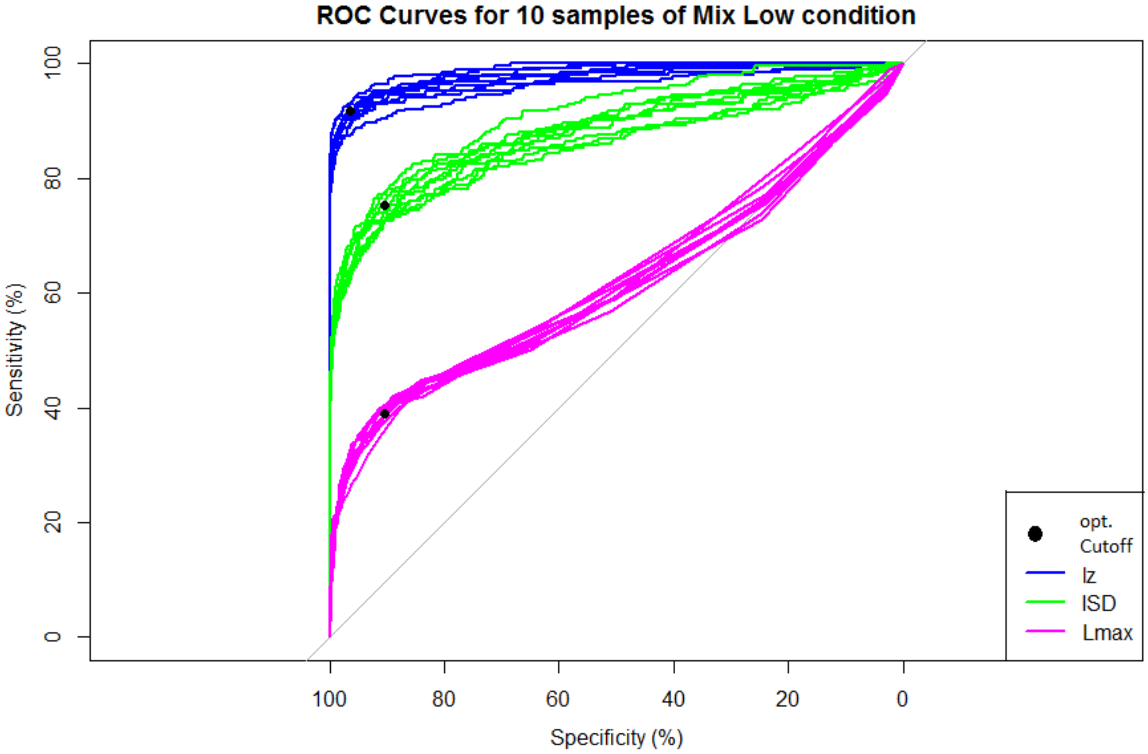

Figure 6. ROC curves for 10 samples of condition Mix Low

The graph also illustrates that when using Lmax high values of sensitivity can only be achieved by accepting very low values of specificity. In that case the Lmax curve approaches (or even cuts) the diagonal, indicating a performance no better than chance.

The optimal cutoff for every sample and index were calculated and then averaged per index and base rate. Table 9 shows the cutoff values of the indices in the three Mix conditions.

Table 9. Optimal cutoff values derived from ROC analysis for the Mix conditions

Base rate	lz	Gu	nGu	ISD	EO	Ma	O	Syn	Ant	Lmax	Lmean
low	-0.03	1091.5	0.23	1.05	-0.77	10.19	326.2	-0.31	-0.46	9.59	1.62
mid	-0.08	1091.8	0.23	1.05	-0.77	9.82	325.3	-0.31	-0.47	9.61	1.61
high	-0.16	1088.0	0.24	1.05	-0.77	9.17	324.9	-0.31	-0.47	9.37	1.61

Cutoff values for the indices show little effect of base rate for most indices. For lz cutoff values increase slightly with base rate, for Ma they decrease. Looking at the probability cutoffs one can see an effect of base rate. Cutoff values here increase with base rate.

Table 10. Prevelance of careless response in ROM data set when applying optimal cutoffs from Mix conditions

Base rate	lz	Gu	nGu	ISD	EO	Ma	O	Syn	Ant	Lmax	Lmean
low	23.72%	19.78%	26.36%	19.52%	5.37%	21.07%	38.10%	11.11%	31.47%	20.27%	16.80%
mid	26.96%	19.78%	25.98%	19.32%	5.71%	23.94%	38.56%	11.05%	32.82%	20.27%	17.86%
high	33.42%	20.18%	25.29%	19.58%	5.83%	29.95%	38.87%	11.05%	32.18%	20.27%	17.86%

Applying the cutoffs to the real ROM data set resulted in the prevalence rates shown in Table 10. The rates vary strongly between the indices and range from 5.4% (EO) to 33.4 % (lz) which seems unrealistically high. Because of the unexpected results it was decided post hoc to also use the null distribution of the indices of the original samples to estimate the prevalence of careless responding.

3.4.3 Estimating the prevalence of careless responding using alternative cutoffs

To find alternative cutoffs the indices calculated for the original samples (before careless responses had been inserted) were combined to form the null distribution of the indices. Values in the real ROM data set above the 95th percentile (and the 99th to be conservative) of the null distribution were regarded as careless response. Table 11 gives the 95th and 99th percentile values for the null distribution of each index. These cutoffs are higher than those found by ROC analysis.

Table 11. Careless response cutoff values based on null distribution percentiles

Percentile	lz	Gu	nGu	ISD	EO	Ma	O	Syn	Ant	Lmax	Lmean
95th	-0.015	1102.5	0.242	1.073	-0.679	10.627	448.5	0.024	0.343	12	1.729
99th	0.095	1211.5	0.269	1.124	-0.571	11.715	519.5	0.145	0.606	19	2.039

Based on these cutoffs additional prevalence values were calculated which can be found in Table 12. The prevalence figures shown here are lower than those derived by way of ROC analysis – de 99th percentile based ones especially.

Table 12. Careless response prevalence values based on null distribution percentiles

Percentile	lz	Gu	nGu	ISD	EO	Ma	O	Syn	Ant	Lmax	Lmean
95th	22.71%	18.86%	23.86%	16.42%	1.12%	17.48%	8.64%	0.72%	0.09%	11.00%	10.48%
99th	16.94%	12.72%	16.25%	9.99%	0.23%	11.60%	2.35%	0.09%	0.00%	3.47%	3.82%

There is still a great variation in prevalence figures when using different indices. For the 95th percentile lz and nGu show the highest prevalence figures (22.7% and .23.9%, respectively), followed by Gu, Ma, and ISD (18.9% , 17.5%, and 16.4% , respectively). Using Lmax and Lmean cutoffs results in a prevalence rate of just over 10%, the rate for using the other indices lie under 10%. When using the 99th percentile cutoffs prevalence rates range from 16.9% (lz) to 0.0% (Ant). Using lz, nGu, Gu, Ma or ISD results in prevalence rates over or around 10%; for the others indices rates under 5% are achieved.

4 Discussion

ROM has become an important tool in assessing the outcome of psychotherapies as well as comparing the quality of mental health care across, for example, different mental health services or different target groups. Therefore it is necessary that the data collected is meaningful and careless responders can be identified. The aim of this study is threefold: First we investigated the usefulness of different statistical indices for identifying careless responders in ROM data. Furthermore, the effects of careless responding on scale totals and latent trait estimates were examined. Finally, optimal cutoffs for the indices were determined and then applied to the real ROM data set in order to estimate the prevalence of careless responding. We found that several indices (lz, Gu, nGu, Ma) showed good performance in identifying careless responders. Inserting random responses led to an overestimation of scale totals as well as theta values for the careless responders, but had a much smaller effect on the whole sample (in case of scale totals) or conscientious responders (in the case of theta values). Applying the findings of the simulation study to the real ROM data prevalence figures varied, depending on index or method used, from under 5% up to 33%.

4.1 Performance of indices for detection of careless response (RQ1)

Several studies have used and compared indices to detect careless responding (e.g., Meade & Craig, 2012; Huang et al., 2012; Niessen, Mbeijer, & Tendeiro, 2016), but none so far has compared all the indices presented here. Most of the findings of this study regarding the performance of the indices match those of the previous studies. For example, Meade and Craig (2012) found that Mahanalobis distances showed very high sensitivity and specificity values for totally random (100% of items replaces by random responses) and partially random (25% of items replaces by random responses) conditions when the random responses came from a uniform distribution. Even-odd consistency (called EO in our study) worked well in the totally random conditions but not at all in the partially random condition. Drawing random responses from a normal random distribution radically change the performance of the indices with Mahanalobis distances only peforming reasonably well in the partially random condition. The actual sensitivity and specificity values Meade and Craig (2012) find are not comparable with those found in our study because they use the standard cutoff for the fitted probability of

.5 instead of the base rate as in this study. This also explains why they find a clear effect of base rate (better performance with higher base rates) where this study does not.

Zijlstra et al. (2011) compared the performance of six different outlier statistics, among which Mahanalobis distances, item-based outlier statistic (called O in this study), the intraindividual variance (called ISD in this study) and the number of Guttman errors. Consistent with out results, they found that O and ISD performed badly when used to detect random responses, but Gu and Ma showed good sensitivity and specificity values. Emons (2008) as well as Niessen et al. (2016) found that indices Gu and lz showed similar and relatively high detection rates for carelessness and inattention, which matches our findings. None of the studies simulated longstring responses; Huang et al. (2012) and Niessen et al. (2016) used a longstring index to detect careless responding produced by the respondents as required in the instruction. Consistent with our results their sensitivity values for detecting careless responses with longstring indices were relatively low.

EO, Syn and Ant were used by many studies, but did not show satisfying results in both our study and the previous studies. A closer look at the mechanics of all three measures shows some flaws that have thus far not been discussed. To compute Syn and Ant, pairs of high (positive or negative) correlation are selected. As the correlation function is symmetrical it is arbitrary if item1 / item2 or item2 / item1 are chosen as pair. When calculating a respondent's correlation between item scores of the first half and the second half of the pairs, switching item1 and item2 will nevertheless usually change the result. Also it is not clear whether using a correlation is the best measure to capture respondents answering similarly. If item2 has (roughly) item scores twice as high as those of item1 they will be highly correlated. If item3 and item4 are answered similarly but sometimes item3 gets higher items scores and sometimes item4 they will have low correlations but resemble the original psychological synonyms much better than the pair item1 / item2. As alternative indices that still use the principles behind Syn and Ant, pairs with low (or high) average absolute differences in item scores could be chosen. Further studies would need to examine whether these revised Syn and Ant indices would deliver satisfactory results.

For EO again there is a methodological problem: As with split-half reliability the values of the subscale totals depend on the split used. Using an even-odd split is just one method of splitting, which is chosen arbitrarily. For a conscientious responder changing the order of two items should not deliver different results. It nevertheless influences the EO value for that person. When using subscales with uneven numbers of items it will also make a difference if sumscores or average scores are used for the subscale totals as the odd “half” is one item longer. This choice of calculation method therefore may also bias the results. Meade and Craig (2012) reasoned that EO did not work well because their questionnaire only contained five subscale with 20 items each. As this study used in total 32 subscales this cannot be the only explanation for the mediocre performance of EO here. Another factor for the performance of EO could be the length of the subscales: The shorter a subscale the more influential the choice of split will be on the subscale totals. In our study BSI and DAPP-SF have relatively short subscale (four – ten items) which might be too short for a consistent working of EO. Further studies could look into the performance of EO when using a large number of sufficiently long subscales.

4.2 Effect of careless response (RQ2)

This study found that inserting random responses resulted in biased scale totals: Random responders as a group showed on average considerable scale total bias, with bias highest for BSI and lowest for the MASQ. The sample as a whole (i.e., careless and conscientious responders together) showed small scale total bias that increased with base rate. In the longstring conditions no clear bias pattern appeared. For ability parameter estimates, especially the careless responders showed biased parameter averages. Conscientious responders were affected as well, but to a much smaller degree, with bias depending on base rate. The bias was higher for the BSI and lowest for the MASQ.

We found that the effect of careless responding differed for the different scales. For all scales the distribution of the item scores was skewed; the mean item score was lower than the middle score option (in the real ROM data set the scale means were 1.19 for the BSI, 1.60 for the MASQ and 1.50 for the DAPP-SF using item score 0 - 4). Huang, Liu, and Bowling (2015) showed that in situations where scale means depart from scale midpoints a small prevalence of careless responding can distort correlations between measures. By the same mechanism the differences in bias for the scales can be explained: When item scores are replaced by random responses, the average of the inserted item scores should lie around the middle score option. The scale means of the manipulated sample thus will be dragged towards the middle score. In our case, with all scale means below the middle score option, scale means should increase; this effect should increase with base rate, which was what our study found. That the scale means in the manipulated samples of the longstring conditions did not show much change, can be explained by the way longstring behavior was simulated: A respondent's answer on a randomly chosen item was repeated several times, which means that the more popular item options were repeated with a higher probability, thus (on average) leaving the scale means unbiased.

4.3 Evidence for careless response in ROM data (RQ3)

Using optimal cutoffs obtained by ROC analysis for the indices performing best (i.e., lz, Gu, nGu and Ma) we found, depending on base rate and index used, prevalence rates of careless responders of 21 – 33%. Using the 95th percentile of the null distribution as cutoff this study found prevalence rates of 16 – 23 % and even lower ones when using the 99th percentile (11.6% using Ma, 16.9% using lz).

The prevalence figures derived from ROC analysis were much higher than the numbers proposed by Meade and Craig (2012): They found that 10 – 12% of their sample responded in a way that could be regarded as seriously careless. Conijn, Emons, De Jong, and Sijtsma (2015) reported misfit detection rates of around 11 – 14 % in clinical samples. These numbers are very similar to our results using the (conservative) 99th percentile of the null distribution as cutoff and might be a more realistic estimation than the 21 – 33% when optimal cut-points were used.

The studies comparing the performance of indices detecting careless response employed different strategies regarding finding good cutoffs. Meade and Craig (2012) simply reported sensitivity and specificity values for the .5 cutoff. Huang et al. (2012) used a fixed .95 and .99 specificity value to find a suitable cutoff and report the corresponding sensitivity values. By that they could easily compare the sensitivity values of all indices studied. Niessen et al. (2016) used rational cutoffs (i.e., those that had been suggested in earlier studies or were based on a scree-like plot) as well as empirical cutoffs.

Whether cutoffs derived from simulation studies lead to reliable prevalence figures depends to a great degree on the methods used in the simulation: Is the simulation a good match for the phenomenon in real life (Niessen et al., 2016)? Using random responses from a uniform random distribution on all items of a self-report instrument (as done by Meade and Craig (2012) in one of their conditions) is surely not realistic. This study tried to avoid that pitfall by mixing different types of careless response simulation in the condition used for determining the cutoffs. Nevertheless the question remains whether careless responders in real life data choose answers truly randomly. It is possible that drawing from a normal random distribution (thus giving preference to the middle item scores over the more extreme ones) simulates careless behavior better.

Although our study clearly could show that inserting careless responses leads to enlarged index values for most indices and conditions and that the indices can be used to detect careless responders in the simulation, it is not clear whether suspicious index values in real life data sets are necessarily caused by careless responding. Most of the indices used here simply detect aberrant answer patterns. The person-fit indices mark respondents that endorse “difficult” items (i.e. items describing severe symptoms) while not endorsing more “easy” items. Mahanalobis distances identify outliers – answer patterns that are strikingly different than those of other respondents. That these aberrant answer patterns are caused by careless responding is just one possible (and surely plausible) interpretation. They could also be caused by respondents with an atypical symptom profile. In recent years, lz-values have been used in several studies to find such atypical symptom profiles: Conrad et al. (2010) used lz-values to identify a group of patients with an atypical diagnostic pattern of suicide; Wanders, Wardenaar, Penninx, and Meijer (2015) identified patients showing atypical depression symptomatology by using lz-values. Here the interpretation is that the answers given by these atypical respondent groups are unusal but still meaningful. If on the other hand aberrant response patterns are interpreted as caused by careless responding, test scores cannot be regarded as meaningful. The high prevalence figures could thus also be caused by a group of patients with valid responses, but showing atypical symptomology.

Simulating longstring responses realistically was a challenge for this study. Simply inserting longstrings of fixed length seemed to be too deterministic (the length of that string in most cases would have been the Lmax value). Therefore by not controlling for overlap a certain variance was ensured. By continuing existing item score answers the simulation methods tried to mimic a situation where a respondent repeats a previous answer several times. As a consequence the longstring conditions show lower total score bias than the other conditions. Length and frequency of the inserted longstrings were chosen in a way that (on average) they enlarged Lmax and Lmean. It is not surprising that the longstring indices managed to detect careless responses in the longstring conditions well. That using longstring indices for determining the real life prevalence of careless response results in considerably lower figures than when using lz or Gu could have two reasons: Longstring answer behavior is relatively rare (as Meade and Craig (2012) as well as Niessen et al. (2016) state) or the simulation method is not a good match for reality.

4.4 Limitations and Strengths

This study examined a larger number of indices suitable for detecting careless response than previous studies. These indices included IRT person-fit measures as well as more traditional indices such as Mahalanobis distances and those derived from established inconsistency scales. The data used came from a large ROM sample and consisted of answers to self-report instruments consisting of more than 260 items in total which made sure the detection methods had enough power (but possibly with the downside of detecting inconsequential aberrations). Different types of careless responding were simulated, and the findings of the simulation study then applied to the real ROM data set. The variety of methods that were employed helped in getting a broad picture of the matter at hand.

4.4.1 Generalizability

Leiden University Medical Center and the Mental Health Care Centre Rivierduinen - where the ROM data for this study were collected - use a very elaborate ROM intake procedure consisting of several self-report and observer rated instruments. This probably leads to a more frequent occurrence of careless responding than with clients only filling in for example the 21 item version of the Depression Anxiety Stress Scale. Especially the findings regarding prevalence of careless responding are therefore not generalizable to all ROM contexts. In addition to the ROM procedure used, the population in question probably also has an effect on the prevalence of careless response. Regarding the study where the data used here originated, de Beurs et al. (2011) stated that “the patients monitored in this study form a representative sample of the patients typically seen in clinical practice (p. 10).” Nevertheless patients in specialized mental health care (such as the patients in this study) usually show more and more serious symptoms than those in primary mental health care, with as a consequence usually more cognitive problems and difficulty in concentration. As these problems are one of the reasons for careless responding, lower prevalence rates can be expected in ROM data from primary health care settings.

Whereas the prevalence of careless responding is affected by choice of procedure or population, the performance differences of the indices in detecting careless responders on the other hand should be the same in other ROM situations using self-report questionnaires. The results of this study show no indication that in low prevalence situations different indices should be used than in high prevalence situation. Whether it is feasible to detect careless responding in very low prevalence situations could be further examined.

4.4.2 Methodological Caveats

Self-report instruments using Likert scales such as BSI, MASQ, and DAPP-SF produce ordinal data. It cannot be assumed that the distance between answering 2 (moderately) and 3 (quite a bit) on the MASQ is the same as between 3 and 4 (extremely). Strictly speaking the use of descriptive statistics as means and SDs is not appropriate for ordinal data. Spearman rank order correlations should be used instead of Pearson correlations. Due to the complexity of the computation for the simulation this study did not always follow these optimal procedures.

For the analysis of the real ROM data set, the item scores were used as given, that is 0 – 4 for the BSI and 1 – 5 for MASQ and DAPP-SF. For the simulation only item scores 0 – 4 were used. This difference in score range affected EO, Syn, Ant and the longstring indices. When complications associated with this approach were discovered, for the sake of consistency the indices were calculated again for the real ROM data set, this time using item scores 0 – 4 for all scales. Correlations between old and new indices were high (ranging from .83 for EO to .99 for both longstring indices). Therefore it was decided that it was not necessary to repeat the prevalence calculations with the new indices.

The analysis of the scale total and the ability parameter biases showed different results for the three scales. This already indicates that the findings here are not simply generalizable to other scales. Whilst using all item scores as a whole set and averaging over all subscales had the advantage of using a sufficient number of items (in total) and sufficient subscales, it also might have prohibited to show the differences in performance regarding the scales. BSI consists of nine very small subscales (four to seven items), MASQ of 5 subscales with lengths ranging from 11 to 22 items. As already mentioned the difference between mean item score and middle item score plays a role in determining the bias. Subscales containing many reversed items (from MASQ and DAPP-SF) showed different mean item scores than the rest of the subscales. Calculating the indices separately for the scales might provide more information under which conditions detection of careless response works best. For the simulation study the BSI, MASQ, and DAPP-SF were assumed to be answered in that particular order. This of course has consequences for the longstring indices (as they were calculated over the whole set of items regardless of scale), but also for the Ran25sec condition. Here only DAPP-SF items were involved. Future studies should try to vary that order, especially as it was not clear whether the original data set always abided by that order.

4.4.3 Recommendations

This study focused on the effect of careless responding on the group levels (i.e. the bias for the careless responder, the conscientious responders and for the whole group). This is of interest when using ROM data for research or benchmarking purposes. When ROM data is used for individual diagnostic purposes careless responding of one patient could have a much bigger effect on that patient's total score and subsequently the treatment decisions made. Further studies should therefore include that perspective.

It lies in the nature of ROM data they consist of repeated measures. It exceeded the scope of this study to take that into account – here only the intake measurements were used. Interesting follow-up questions could be whether careless response indices are independent across measurement points, what effect careless response has on gain scores, and whether atypical development patterns (strong improvement in some subscales, strong deterioration in others, e.g.) can give additional insight as to the plausibility of the data.

As this study tried to find index cutoffs that could be applied to the real-life data set, it was decided to only use one index at a time as a predictor in the logistic regression. The combination of successful indices might lead to a better identification of careless responders. Therefore, further investigation of other classification methods and better models using multiple predictors seems promising. This study as well as that of Meade and Craig (2012) encountered problems with samples showing perfect separation. Future simulation studies should try to overcome this by avoiding conditions with high numbers of randomly inserted items scores and use conditions with mixed types of careless responding instead if they wish to use models for inferential purposes.

To better be able to make this distinction more "external verification" might help, i.e. additional information by supplementing the questionnaires. Measurement of response time (as suggested by Niessen et al., 2016), self-report carelessness or diligence, repetition of items (possibly with slightly different wording), use of screener items (Berinsky, Margolis, & Sances, 2014) (Berinsky, Margolis, & Sances, 2014) or the Conscientious Responders Scale introduced by Marjanovic et al. (2014) are possible additions that might (or might not) support the interpretation of aberrant responses as careless. Further studies should investigate whether these additions improve the classification success whilst they are still easy enough to being implemented.

4.5 Application of results to ROM practice

This study used different methods to determine cutoffs for detecting careless responders which in turn led to sensitivity and specificity values. When choosing a method for finding suitable cutoffs in daily ROM practice the chosen method must fit the context of the detection of careless responders: As there is always a pay-off between sensitivity and specificity one has to decide what hurts less: Throwing out valid responses or leaving in careless ones. Researcher conducting a study with a random sample for inferential purposes should be reluctant to throw results out lightly because of possible careless responding. Here being strict and choosing the 99th percentile of the null distribution might ensure that only few valid responses are discarded. When on the other hand ROM data is used for monitoring individual patients it might be reasonable to use lower cutoffs and flag responses as "potentially careless". Before acting on the ROM data, clinicians should talk about the results with their clients, and possible inconsistencies can be discussed. The optimal cutoffs from the ROC analysis might be suitable here as they optimize sensitivity and specificity simultaneously. When ROM data is used for benchmarking purposes a middle ground might be useful: As large amounts of data are collected over a longer period of time a more lenient approach (such as using the 95th of the null distribution) than with a one-time trial might be useful. This of course requires a consistent approach combined with transparency of the procedure.

When choosing indices for detecting careless responders in ROM practice, not only performance issues have to be considered but also the practicality of implementing such a routine. Although lz, Gu and nGu show a very high performance of detecting careless responders in the simulation, their use requires some expert knowledge (and is also more time-consuming for large data sets) as they usually are not readily available in statistical applications such as SPSS. Ma, ISD, and EO can more easily be implemented. Ma shares another disadvantage with lz, Gu and nGu: For these indices always the whole dataset is needed for calculation. Whether that is practical depends on how often a screening should take place. Screening when a certain set of ROM data is needed for a study or when ROM data has to be delivered to a benchmarking agency twice a year still seems feasible. If new ROM data has to be screened weekly or daily to support clinicians in their decision making regarding a therapy, this might be a problem: For these indices the values depend on the whole distribution. Calculating indices again will probably change the values for the old data set. Someone classified as careless responders might become a conscientious responder by just adding more respondents and thus changing the distribution. The choice is therefore to either use less sensitive, but easy to applicate indices (ISD, EO) or a powerful, relatively easy-to-use index (Ma) which is dependent on the distribution.

Indices depending on the distribution also have the problem that they are susceptible to changes in prevalence. In a ROM context it can easily be imagined that external circumstances change entailing changes in prevalence of careless responding: In the Netherlands the re-imbursement of different treatments by the health insurance or the own-risk sum patients have to pay before they are entitled to re-imbursement is regulated by law. This law is adapted regularly which causes changes in patients' health seeking behavior. A higher own-risk regulation might lead to patient cohorts with a more serious symptomology as others might postpone treatment. This in turn can increase the cohort's prevalence of careless responding. The increase (or decrease) in prevalence of careless responding cannot be noticed by for example the Ma average of the sample (as this is always zero) but only by the distribution of the values. Before implementing a careless-response-detection-procedure, these mechanisms should be considered carefully.

4.6 Conclusion

This study is the important first step in examining methods to identify careless responders in ROM data. There is an indication that a lengthy ROM procedure such as the one used here leads to a considerable prevalence of careless responding. Indices lz, nGu, Gu, and Ma showed very good performance in the simulation study. How good they perform in real-life situation and whether simple screening procedures can be implemented has to be examined further.

References

Agresti, A. (2007). *An Introduction to Categorical Data Analysis* (2nd ed.). Hoboken, New Jersey: Wiley.

Albert, A., & Anderson, J. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika, 71*, 1-10.

Baer, R. A., Ballenger, J., Berry, D. T., & Wetter, M. W. (1997). Detection of random responding on the MMPI-A. *Journal of Personality Assessment, 68*, 139-151.

Barnette, J. J. (1999). Nonattending respondent effects on internal consistency of self-administered surveys: A Monte Carlo simulation study. *Educational and Psychological Measurement, 59*, 38–46.

Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2014). Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American Journal of Political Science, 58*, 739–753.

Berry, D. T., Wetter, M. W., Baer, R. A., Larsen, L., Clark, C., & Monroe, K. (1992). MMPI-2 random responding indices: Validation using a self-report methodology. *Psychological Assessment, 4*, 340-345.

Boswell, J. F., Kraus, D. R., Miller, S. D., & Lambert, M. J. (2015). Implementing routine outcome monitoring in clinical practice: Benefits, challenges, and solutions. *Psychotherapy Research, 25*(1), 6-19.

Carlier, I. V., Meuldijk, D., van Vliet, I. M., van Fenema, E., van der Wee, N. J., & Zitman, F. G. (2012). Routine outcome monitoring and feedback on physical or mental health status: Evidence and theory. *Journal of Evaluation in Clinical Practice, 18*, 104–110.

Chalmers, P. R. (2012). mirt: A multidimensional Item Response Theory package for the R environment. *Journal of Statistical Software, 48*, 1-29.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Conijn, J. M., Emons, W. H., & Sijtsma, K. (2014). Statistic lz-based person-fit methods for noncognitive multiscale measures. *Applied Psychological Measurement, 38*, 122–136.

Conijn, J., Emons, W. H., De Jong, K., & Sijtsma, K. (2015). Detecting and explaining aberrant responding to the Outcome Questionnaire–45. *Assessment, 22*, 513–524.

Conrad, K. J., Bezruczko, N., Chan, Y.-F., Riley, B., Diamond, G., & Dennis, M. L. (2010). Screening for atypical suicide risk with person fit statistics among people presenting to alcohol and other drug treatment. *Drug and Alcohol Dependence, 106*, 92–100.

Credé, M. (2010). Random responding as a threat to the validity of effect size estimates in correlational research. *Educational and Psychological Measurement, 70*, 596–612.

de Beurs, E., & Zitman, F. G. (2006). De Brief Symptom Inventory (BSI): De betrouwbetrouwbaarheid en validiteit van een handzaam alternatief voor de SCL-90 [The Brief Symptom Inventory (BSI): reliability and validity of a practical alternative to SCL-90]. *Maandblad Geestelijke Volksgezondheid, 61*, 120-141.

de Beurs, E., Barendregt, M., Flens, G., van Dijk, E., Huijbrechts, I., & Meerding, W. J. (2012). Vooruitgang in de behandeling meten - Een vergelijking van vragenlijsten voor zelfrapportage [Measuring treatment progress – A comparison of self-report questionnaires]. *Maandblad Geestelijke Volksgezondheid, 67*, 259-264.

de Beurs, E., den Hollander-Gijsman, M. E., Helmich, S., & Zitman, F. G. (2007). The tripartite model for assessing symptoms of anxiety and depression: Psychometrics of the Dutch version of the mood and anxiety symptoms questionnaire. *Behaviour Research and Therapy, 45*, 1609–1617.

de Beurs, E., den Hollander-Gijsman, M. E., van Rood, Y. R., van der Wee, N. J., Giltay, E. J., van Noorden, M. S., . . . Zitman, F. G. (2011). Routine outcome monitoring in the Netherlands: Practical eperiences with a web-based strategy for the assessment of treatment outcome in clinical practice. *Clinical Psychology and Psychotherapy, 18*, 1–12.

de Beurs, E., Rinne, T., van Kampen, D., Verheul, R., & Andrea, H. (2009). Reliability and validity of the Dutch Dimensional Assessment of Personality Pathology-Short Form (DAPP-SF), a shortened version of the DAPP-Basic Questionnaire. *Journal of Personality Disorders, 23*, 308–326.

Derogatis, L. R., & Melisaratos, N. (1983). The Brief Symptom Inventory: An introductory report. *Psychological Medicine, 13*, 595-605.

DeSimone, J. A., Harms, P. D., & DeSimone, A. J. (2015). Best practice recommendations for data screening. *Journal of Organizational Behavior, 36*, 171–181.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*, 67-86.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Earlbaum.

Emons, W. H. (2008). Person-fit analysis of polytomous items. *Applied Psychological Measurement, 32*, 224–247.

Fervaha, G., & Remington, G. (2013). Invalid responding in questionnaire-based research: Implications for the study of schizotypy. *Psychological Assessment, 25*, 1355–1360.

Hoenders, R. H., Bos, E. H., Bartels-Velthuis, A. A., Vollbehr, N. K., van der Ploeg, K., de Jonge, P., & de Jong, J. T. (2014). Pitfalls in the assessment, analysis, and interpretation of routine outcome monitoring (ROM) data: Results from an outpatient clinic for integrative mental health. *Administration and Policy in Mental Health and Mental Health Services Research, 41*, 647-659.

Howard, K. I., Moras, K., Brill, P. L., Martinovich, Z., & Lutz, W. (1996). Evaluation of psychotherapy: Efficacy, effectiveness, and patient progress. *American Psychologist, 51*, 1059-1064.

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of business and psychology, 27*, 99–114.

Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology, 100*, 828–845.

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality, 48,* 61–83.

Marjanovic, Z., Holden, R., Struthers, W., Cribbie, R., & Greenglass, E. (2014). The inter-item standard deviation (ISD): An index that discriminates between conscientious and random responders. *Personality and Individual Differences*.

Marjanovic, Z., Struthers, C. W., Cribbie, R., & Greenglass, E. R. (2014). The Conscientious Responders Scale: A new tool for discriminating between conscientious and random responders. *SAGE Open, 4*(3), 1-10.

Meade, A. W., & Craig, S. B. (2012). Identifying Careless Responses in Survey Data. *Psychological Methods, 17*(3), 437–455.

Nichols, D. S., Greene, R. L., & Schmolck, P. (1989). Criteria for assessing inconsistent patterns of item endorsement on the MMPI: Rationale, development, and empirical trials. *Journal of Clinical Psychology, 45*, 239-250.

Niessen, A. S., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting careless respondents in web-based questionnaires: Which method to use? *Journal of Research in Personality, 63*, 1-11.

Nugter, M. A., & Buwalda, V. J. (2012). Achtergronden en gebruiksmogelijkheden van ROM in de ggz [Background and possible use of ROM in mental health care]. *Tijdschrift voor Psychiatrie, 54*, 111-120.

Osborne, J. W., & Blanchard, M. R. (2011). Random responding from participants is a threat to the validity of social science research results. *Frontiers in Psychology, 1*, 1-7.

Powers, D. M. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies, 2*(1), 37-63.

Raîche, G., Walls, T. A., Magis, D., Riopel, M., & Blais, J.-G. (2013). Non-graphical solutions for Cattell’s scree test. *Methodology, 9*, 23–29.

Swets, J. A. (1973). The relative operating characteristic in psychology: A technique for isolating effects of response bias finds wide use in the study of perception and cognition. *Science, 182*, 990-1000.

Tellegen, A. (1988). The analysis of consistency in personality assessment. *Journal of Personality, 56*, 621-663.

Tjur, T. (2009). Coefficients of determination in logistic regression models — A new proposal: The coefficient of discrimination. *The American Statistician, 63*, 366-372.

Trauer, T. (Ed.). (2010). *Outcome measurement in mental health: Theory and practice.* Cambridge, UK: Cambridge University Press.

van Kampen, D., de Beurs, E., & Andrea, H. (2008). A short form of the Dimensional Assessment of Personality Pathology-Basic Questionnaire (DAPP-BQ): The DAPP-SF. *Psychiatry Research, 160*, 115-128.

van Noorden, M. S., van Fenema, E. M., van der Wee, N. J., van Rood, Y. R., Carlier, I. V., Zitman, F. G., & Giltay, E. J. (2012). Predicting outcomes of mood, anxiety and somatoform disorders: The Leiden routine outcome monitoring study. *Journal of Affective Disorders, 142*, 122–131.

Watson, D., Weber, K., Assenheimer, J. S., Clark, L. A., Strauss, M. E., & McCormick, R. A. (1995). Testing a tripartite model: I. Evaluating the convergent and discriminant validity of anxiety and depression symptom scales. *Journal of Abnormal Psychology, 104*, 3-14.

Wing, J. K., Beevor, A. S., Curtis, R. H., Park, S. B., Hadden, S., & Burns, A. (1998). Health of the Nation Outcome Scales (HoNOS): Research and development. *British Journal of Psychiatry, 172*, 11-18.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer, 3*(1), 32-35.

Zijlstra, W. P., van der Ark, L. A., & Sijtsma, K. (2007). Outlier detection in test and questionnaire data. *Multivariate Behavioral Research, 42*, 531-555.

Zijlstra, W. P., van der Ark, L. A., & Sijtsma, K. (2011). Outliers in questionnaire data: Can they be detected and should they be removed? *Journal of Educational and Behavioral Statistics, 36*, 186–212.

Appendix

Table A1. Average 'index bias' per Factor Type, separately for conscientious responders, careless responders and both

Index	Responders	Ran25	Ran50	Ran25sec	Lstr25	Lstr50	Mix
lz	Conscientious	-0.087	-0.145	-0.071	-0.050	-0.070	-0.088
	Careless	0.666	1.132	0.539	0.382	0.544	0.670
	Both	-0.002	-0.002	-0.003	-0.001	-0.002	-0.001
Gu	Conscientious	0.372	1.365	0.448	0.262	0.751	0.471
	Careless	671.896	1172.434	433.546	363.105	530.844	636.291
	Both	78.547	137.405	50.794	42.514	62.211	74.430
nGu	Conscientious	0.002	0.004	0.002	0.002	0.003	0.002
	Careless	0.102	0.154	0.083	0.062	0.088	0.098
	Both	0.013	0.021	0.012	0.009	0.013	0.014
ISD	Conscientious	0	0	0	0	0	0
	Careless	0.249	0.396	0.221	0.112	0.136	0.223
	Both	0.029	0.046	0.026	0.013	0.016	0.026
Ma	Conscientious	-0.578	-0.902	-0.434	-0.316	-0.387	-0.558
	Careless	4.588	7.329	3.545	2.503	3.226	4.445
	Both	0.000	0.000	0.000	0.000	0.000	0.000
EO	Conscientious	0	0	0	0	0	0
	Careless	0.197	0.457	0.170	0.071	0.079	0.195
	Both	0.023	0.053	0.020	0.008	0.009	0.023
O	Conscientious	0.002	0.003	0.002	0.053	0.169	0.017
	Careless	49.646	99.186	47.021	20.445	35.978	50.880
	Both	5.796	11.584	5.483	2.405	4.274	5.944
Syn	Conscientious	0	0	0	0	0	0
	Careless	0.226	0.340	0.151	0.163	0.223	0.221
	Both	0.026	0.040	0.018	0.019	0.026	0.026
Ant	Conscientious	0	0	0	0	0	0
	Careless	0.154	0.264	0	0.136	0.224	0.155
	Both	0.018	0.031	0	0.016	0.026	0.018
Lmax	Conscientious	0	0	0	0	0	0
	Careless	-1.077	-1.753	-0.385	7.361	20.419	4.906
	Both	-0.125	-0.204	-0.045	0.859	2.382	0.573
Lmean	Conscientious	0	0	0	0	0	0
	Careless	-0.074	-0.119	-0.049	0.405	0.950	0.223
	Both	-0.009	-0.014	-0.006	0.047	0.111	0.026

Table A2. Average sensitivity and specificity per index and condition

Index	Base	Ran25		Ran50		Ran25sec		Lstr25		Lstr50		Mix	
	rate	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.
lz	.05	98.9	99.1	100.0	100.0	97.7	98.0	84.5	89.2	88.4	93.0	92.3	95.7
	.1	98.9	99.0	100.0	100.0	97.5	97.6	84.7	89.0	87.7	92.6	92.1	95.6
	.2	98.8	98.9	100.0	100.0	96.7	96.9	83.9	88.5	86.5	91.8	91.1	95.4
Gu	.05	97.9	98.0	100.0	100.0	91.3	91.4	76.8	82.9	80.3	87.6	87.3	91.6
	.1	98.0	98.0	100.0	100.0	91.3	91.3	76.9	82.7	80.0	87.5	87.1	91.5
	.2	97.9	98.0	100.0	100.0	91.1	91.1	76.8	82.6	79.0	87.2	86.9	91.6
nGu	.05	94.2	93.5	99.3	98.7	92.7	90.7	78.8	83.2	81.9	88.0	87.4	90.7
	.1	94.2	93.7	99.2	98.7	92.4	91.0	78.9	83.3	81.5	87.8	87.3	90.7
	.2	93.9	93.7	99.1	98.8	92.0	91.2	78.6	83.3	80.9	87.5	87.2	90.7
ISD	.05	92.4	91.7	99.6	99.5	85.7	86.7	68.7	66.0	68.3	69.3	80.2	82.7
	.1	92.6	91.8	99.6	99.5	85.8	86.5	68.9	65.7	68.0	69.0	80.2	82.5
	.2	92.6	91.6	99.6	99.5	86.1	86.3	69.1	65.5	68.2	68.7	80.3	82.6
EO	.05	78.3	84.5	93.5	95.0	77.2	82.8	59.9	71.8	60.1	72.7	69.8	82.7
	.1	78.2	84.7	93.5	95.1	77.0	83.1	60.0	71.9	60.2	72.9	69.8	82.8
	.2	78.1	84.8	93.4	95.1	76.6	83.2	59.5	72.0	60.2	73.1	69.9	82.8
Ma	.05	97.9	97.6	100.0	100.0	95.2	94.7	74.3	81.8	77.1	85.7	87.0	91.9
	.1	98.0	97.6	100.0	100.0	94.2	93.6	73.8	81.0	75.4	84.3	86.4	91.5
	.2	97.8	97.4	100.0	100.0	92.0	91.2	72.7	79.6	72.2	81.5	85.7	91.0
O	.05	66.2	69.8	94.9	78.7	61.9	69.3	49.0	63.6	54.7	67.0	63.9	70.0
	.1	66.5	69.9	94.3	79.3	61.8	69.3	48.9	63.3	54.3	66.7	63.6	70.0
	.2	66.2	69.9	93.3	80.1	61.8	69.4	48.9	63.1	53.6	65.6	63.4	69.9
Syn	.05	68.8	69.6	79.5	77.0	63.0	64.4	62.6	65.4	68.3	69.5	67.8	69.1
	.1	68.6	69.4	79.0	76.9	62.9	64.3	62.5	65.3	68.3	69.4	68.1	69.2
	.2	68.6	69.5	79.0	76.9	63.0	64.3	62.6	65.4	68.3	69.4	67.9	69.2
Ant	.05	57.1	60.6	65.8	64.4	51.6	50.4	54.7	60.1	60.1	63.6	56.8	60.9
	.1	57.0	60.6	66.0	64.5	50.7	50.8	54.6	60.1	59.9	63.5	56.8	60.8
	.2	57.0	60.6	65.8	64.5	50.6	50.7	54.4	60.1	59.9	63.5	56.9	60.8
Lmax	.05	65.2	49.0	80.8	49.1	69.7	36.1	97.7	84.9	100.0	96.5	44.1	82.6
	.1	65.3	49.1	80.7	49.1	71.6	34.4	91.4	88.6	100.0	97.1	43.8	83.8
	.2	65.1	49.1	80.7	49.1	72.0	34.2	91.0	88.7	100.0	97.2	43.7	84.1
Lmean	.05	71.6	48.6	79.7	62.0	67.5	43.3	96.9	86.0	99.5	96.4	43.5	78.5
	.1	71.7	48.8	79.5	62.1	67.6	43.3	96.1	86.6	99.2	96.7	43.6	78.7
	.2	71.6	48.8	79.5	62.1	67.6	43.3	95.1	87.3	99.0	96.9	43.5	78.9