The assessment and impact of careless responding in routine outcome monitoring within mental Health care
Conijn, Judith M.; Franz, Gunhild; Emons, Wilco H. M.; De Beurs, Edwin; Carlier, Ingrid V. E.

Published in:
Multivariate Behavioral Research

Document version:
Publisher's PDF, also known as Version of record

DOI:
10.1080/00273171.2018.1563520

Publication date:
2019

Citation for published version (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright, please contact us providing details, and we will remove access to the work immediately and investigate your claim.
The Assessment and Impact of Careless Responding in Routine Outcome Monitoring within Mental Health Care

Judith M. Conijn, Gunhild Franz, Wilco H. M. Emons, Edwin de Beurs & Ingrid V. E. Carlier

To cite this article: Judith M. Conijn, Gunhild Franz, Wilco H. M. Emons, Edwin de Beurs & Ingrid V. E. Carlier (2019) The Assessment and Impact of Careless Responding in Routine Outcome Monitoring within Mental Health Care, Multivariate Behavioral Research, 54:4, 593-611, DOI: 10.1080/00273171.2018.1563520

To link to this article: https://doi.org/10.1080/00273171.2018.1563520

© 2019 The Author(s). Published with license by Taylor & Francis Group, LLC

Published online: 19 Apr 2019.

Submit your article to this journal

Article views: 775

View related articles

View Crossmark data
The Assessment and Impact of Careless Responding in Routine Outcome Monitoring within Mental Health Care

Judith M. Conijn, Gunhild Franz, Wilco H. M. Emons, Edwin de Beurs, and Ingrid V. E. Carlier

Research Institute of Child Development and Education, University of Amsterdam, Amsterdam, the Netherlands; Institute of Psychology, Leiden University, Leiden, the Netherlands; Tilburg School of Social and Behavioral Sciences, Tilburg University, Tilburg, the Netherlands; Department of Psychiatry, Leiden University Medical Centre, Leiden, the Netherlands

ABSTRACT

Careless responding by mental health patients on self-report assessments is rarely investigated in routine care despite the potential for serious consequences such as faulty clinical decisions. We investigated validity indices most appropriate for detecting careless responding in routine outcome monitoring (ROM) in mental health care. First, we reviewed indices proposed in previous research for their suitability in ROM. Next, we evaluated six selected indices using data of the Brief Symptom Inventory and the Mood and Anxiety Symptom Questionnaire from 3,483 outpatients. Simulations showed that for typical ROM scales the $L_{\text{max}}$ index, Mahalanobis distance, and inter-item standard deviation may be too strongly confounded with the latent trait value to compare careless responding across patients with different symptom severity. Application of two different classification methods to the validity indices did not converge in similar prevalence estimates of careless responding. Finally, results suggest that careless responding does not have a substantial biasing effect on scale-score statistics. We recommend the $p_{\text{z}}$ person-fit index to screen for random careless responding in large ROM data sets. However, additional research should further investigate methods for detecting repetitive responding in typical ROM data and assess whether there are specific circumstances in which simpler validity statistics or direct screening methods perform similarly as the $p_{\text{z}}$ index.

KEYWORDS

Careless responding; mental health care; person-fit indices; routine outcome monitoring; validity indices

Introduction

There is growing interest in routinely monitoring clients’ progress during psychiatric or psychological treatment, termed routine outcome monitoring (ROM; de Beurs et al., 2011). ROM has been implemented in many countries and in many different mental health care settings (Carlier & van Eeden, 2017; Trauer, 2010). It can provide feedback to clinicians and to clients themselves on how clients fare during treatment. Such feedback may improve clinical judgment and may allow for better decision-making regarding the clinical course of the therapy. Furthermore, ROM data in aggregated form (e.g., aggregated over groups of clients, departments, or institutions) may provide managers and policy makers with salient information about the effectiveness of health care that is offered (Barendregt, 2015; Trauer, 2010). ROM data are collected using clinician-rated scales and self-report instruments that assess clients’ psychological problems and symptoms of psychopathology. Often a combination of generic and disorder-specific scales is used. Some mental health care institutions use a battery of instruments at intake and during (e.g., every 6 weeks or 3 months) and after treatment, whereas others use fewer scales at more frequent intervals (de Beurs et al., 2011).

When using self-report instruments, the underlying assumption is that respondents answer the questions as truthfully and accurately as possible. Regarding ROM, Boswell, Kraus, Miller, and Lambert (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 occasionally may be compromised. For example, some patients may...
intentionally over-report their symptoms, whereas others may underreport their symptoms (Ben-Porath, 2013). These content-based invalid response styles and their effects have been studied extensively (Ingram & Ternes, 2016).

In this study, we focus on careless responding (e.g., Meade & Craig, 2012), also called “insufficient effort responding” (Huang, Liu, & Bowling, 2015). Careless responding refers to answering items without sufficiently considering item content or item instructions. As careless respondents are not purposefully manipulating their answers, careless responding is a non-content-based response style (McGrath, Mitchell, Kim, & Hough, 2010). Recently, several studies have shown that noncontent-based response styles may be prevalent in mental health care patients because the cognitive problems of psychopathology may induce patients to satisfice (i.e., to provide nonoptimal responses) to questionnaires (Conijn, van der Ark, & Spinhoven, 2017; Gervais et al., 2017; Keeley, Webb, Peterson, Roussin, & Flanagan, 2016). Especially on longer instruments, extensive assessment batteries, or repeated assessments, mental health care patients may become tired and lose concentration or motivation. Research suggests that this may induce two different types of careless responding, including random careless responding and repetitive careless responding (e.g., Kam & Chan, 2018; Meade & Craig, 2012). Random careless responders tend to respond in an unsystematic manner, while repetitive careless responders tend to repeat the same answer regardless of item content. Respondents may show varying degrees of careless responding, ranging from completely random or repetitive item-score patterns to “nonoptimal” response patterns with only some degree of unsystematic or repetitive responses (Krosnick, 1991).

The consequences of careless responding (and invalid response styles in general) in clinical practice are described in Armistead-Jehle and Green’s (2016) model and include first order (e.g., faulty diagnosis decisions), second order (e.g., poor treatment outcomes), and third order effects (e.g., societal financial costs). Although no studies have assessed the possible negative effects of careless responding in ROM on these levels, ample studies in various other settings (e.g., employee job satisfaction, educational intervention study) have assessed the effects of careless responding on research results. Most studies concluded that careless responding can substantially bias results, such as estimated group differences in mean scores, bivariate correlations, and reliability coefficients (e.g., Burchett et al., 2016; Holtzman & Donnellan, 2017; Huang, Liu et al., 2015; Kam & Meyer; 2015; McGonagle, Huang, & Walsh, 2016; Osborne & Blanchard, 2011). Other studies found only small (or no) effects of removing careless respondents on the factor structure of questionnaire data, criterion test validity, bivariate correlations or Cronbach’s alpha (Conijn, Emons, & Sijtsma, 2014; Meijer, 1997; Widhiarso & Sumintono, 2016; Zijlstra, van der Ark, & Sijtsma, 2011). The varying effects of carelessness are likely to be a result of the different psychometric aspects evaluated, the type of careless response strategy, the prevalence of aberrant responding, the study design (simulated vs. real data), and the method used to detect careless respondents. The varying effects also imply that the effect of careless responding needs to be investigated in specific settings.

**Methods for detecting careless responding**

Some popular clinical test batteries, such as the Personality Assessment Inventory (PAI; Morey, 1991) and the Minnesota Multiphasic Personality Inventory-2 (MMPI-2; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989), include questionnaire-specific scales to detect random responding (e.g., the Variable Response Inconsistency scale; Handel, Ben-Porath, Tellegen, & Archer, 2010). However, ROM systems rarely include validity checks, and therefore a general (instead of a questionnaire-specific) approach to detect careless responding is required. Most approaches to detect careless responding originate from survey methodology and 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 to detect careless responses. For example, researchers can ask participants to estimate how reliable their answers are (Meade & Craig, 2012). Researchers can also deliberately include items that every conscientious responder should or should not endorse (Fervaha & Remington, 2013). Direct screening methods have been shown to be effective (e.g., Huang, Bowling, Liu, & Li, 2015; Kam & Chan, 2018; Meade & Craig, 2012), but they increase test length and may lead respondents to feel that their answers are not trusted or taken seriously. The latter is especially problematic in routine mental health care practice, where direct screening methods may negatively influence patient’s treatment motivation.
Alternatives to direct screening methods are statistical post hoc indices that detect careless responding by quantifying how unusual or inconsistent a response pattern is. The performance of these validity indices has been investigated by means of simulation studies and experimental research (e.g., Huang, Curran, Keeney, Poposki, & DeShon, 2012; Meade & Craig, 2012; Niessen, Meijer, & Tendeiro, 2016). However, most research on these indices has been done in the context of personality research (Emons, 2008; Huang et al., 2012; Meade & Craig, 2012; Niessen et al., 2016). The indices found suitable for personality data, such as long string indices, item response theory (IRT)-based person-fit indices (Meijer, Niessen, & Tendeiro, 2016) or the Mahalanobis distance, may not necessarily be suitable for typical ROM data. Psychopathology measures are characterized by skewed item-score data, items worded in the same direction, and subscale data that do not fit unidimensional IRT models well (Reise & Waller, 2009; Thomas, 2011).

To assess the suitability of different validity checks in ROM, two important issues also need to be addressed. First, a method for categorizing respondents as careless versus noncareless should be chosen. A categorization method is required to inform clinicians at what value of a validity index, the patients’ test score is likely to be invalid, and to inform data analysts of aggregated ROM data at which cutoff value a response pattern should be excluded from the analyses. Currently it is unclear how to best determine cutoff values for post hoc indices (DeSimone et al., 2015; Niessen et al., 2016), but two different methods have been suggested. First, IRT-based Monte Carlo simulations can generate null distributions—the distribution of the index in a clean data set—for a validity index, and cutoff values can be computed based on these distributions and a desired Type I error rate (e.g., Emons, 2008; Seo & Weiss, 2013; Sinharay, 2017). Although most often used, the drawback of this method is that its suitability may depend on the fit of the estimated IRT model to the data. Second, Tukey’s fences method, also known as the boxplot method, can be used to identify respondents with outlying values on the validity indices (Zijlstra, van der Ark, & Sijtsma, 2007). This method is relatively easy because the cutoff value solely depends on the observed distribution of the statistic and does not require the use of IRT. To our knowledge, no study has assessed whether these two methods for categorizing respondents provide consistent results.

Second, it should be investigated to what extent the null distributions of different validity indices are confounded by the respondent’s trait level (Emons, 2008). A strong relation between trait level and a validity index’s null distribution is problematic for two reasons. First, in that case, the probability of incorrectly being categorized as careless depends on the respondent’s level of symptom severity. Consequently, excluding careless respondents based on a common cutoff may lead to selection bias by excluding respondents with most severe psychopathology (Roivainen, Veijola, & Miettunen, 2016). Second, appropriate cutoff values (e.g., the 95th percentile derived from a simulated null distribution) may vary across different patient samples depending on the symptom severity in the sample. Consequently, one cannot meaningfully compare prevalence estimates based on the same cutoff value between subgroups with different average symptom severity levels; variation in prevalence may be due to differences in symptom severity instead of differences in careless responding. For individual decision-making, a similar problem exists: cutoff values may be more or less suitable for a patient, depending on his or her latent trait level.

**Study aim**

To summarize, recent research has concluded that screening for careless responding should be common practice and that the consequences of careless responding are underestimated (e.g., Curran, 2016; Godinho, Kushnir, & Cunningham, 2016; McGonagle et al., 2016; Osborne & Blanchard, 2011). Screening for careless responding in ROM can improve diagnosis and treatment decisions of individual patients, and improve decision-making by managers and policy makers based on aggregated ROM data.

In the first part of this article, we use previous research results to discuss the suitability of various post hoc validity indices for detecting careless responding in typical ROM data. In the second part of this article, using six selected validity indices, we address three research questions using simulated data and empirical ROM data of 3543 psychiatric outpatients. Addressing these questions is fundamental for selecting appropriate validity checks for use in ROM and for assessing the necessity of such checks in the analyses of aggregated ROM data.

1. To what extent are the validity index values in clean data confounded with the latent trait level measured by the scale?
2. What is the effect of different categorization methods on prevalence estimates?
3. To what extent does careless responding affect means and standard deviations of scale scores in aggregated ROM data?

**Post hoc validity indices in ROM**

Below, we review eight post hoc validity indices that have been most often applied in recent studies (e.g., Godinho et al., 2016; Kam & Meyer, 2015; Marjanovic, Holden, Struthers, Cribbie, & Greenglass, 2015; Meade & Craig, 2012; Niessen et al., 2016), and discuss their suitability given typical ROM data properties (for a more general discussion of validity indices, see Curran, 2016). The typical ROM data properties include the following. Most of the ROMs are characterized by multiple short subscales assessing different dimensions of psychopathology and skewed item-score data (i.e., skewed in the direction of stronger symptoms) (e.g., Reise & Waller, 2009). Examples of frequently used ROM scales and test batteries include the Brief Symptom Inventory (BSI; Derogatis & Melisaratos, 1983) including 54 items and 9 subscales, the Outcome Questionnaire 45 (OQ-45; Lambert et al., 2004) including 45 items and 3 subscales, and the Treatment Outcome Package (TOP; Boswell, Kraus, Castonguay, & Youn, 2015) including 93 items and 11 subscales. Also, even though the ROM instruments measure different dimensions, the traits measured by various subscales are often strongly related (e.g., Caspi et al., 2014; Urbán et al., 2014), and items are often worded in the same direction (i.e., indicative of psychopathology). The subscale data designed to assess a single dimension are commonly not well-described by unidimensional factor models or IRT models (e.g., Thomas, 2011). Finally, although not previously investigated, we expect ROM data to mainly include mild careless responding (i.e., instead of fully random or repetitive patterns) and a relatively low prevalence of careless respondents (<10%), as patients have self interest in accurately completing the ROM.

**Inconsistency indices**

Indices to detect inconsistent or random careless responding are the statistical synonyms and antonyms indices, the even-odd consistency (EO) index, and the inter-item standard deviation (ISD) index. To compute the statistical synonyms and antonyms indices (e.g., Johnson, 2005; Meade & Craig, 2012), pairs of items that are strongly correlated (r = |.60| is a commonly used cutoff) are selected. Next, the within-person correlation between the synonymous pairs and the antonymous pairs is calculated, resulting in two indices. For computing the EO index, all unidimensional subscales are divided into an even and odd part (i.e., the even-numbered items and the odd-numbered items), and for every even- and odd-scale the person’s score is calculated (e.g., Johnson, 2005). Then, the correlation between the even- and odd-averages is calculated and corrected for test-length using the Spearman–Brown prophecy formula. Both the statistical synonyms and antonyms indices and the EO index are unlikely to be suitable for application in ROM. The statistical synonym/antonym indices require a large number of item pairs to render the indices reliable, as the number of item pairs is the sample size for computing the correlation coefficient (Meade & Craig, 2012; Niessen et al., 2016). ROM scales usually include no more than 100 items, resulting in a maximum number of 50 pairs. Similarly, the EO index is only reliable when data of a large number of subscales is available, as the number of subscales constitutes the sample size (Niessen et al., 2016; Meade & Craig, 2012). Another drawback of the EO index for ROM data is that it can detect fully random response patterns but performs poorly in detecting partial careless responding (Meade & Craig, 2012).

The ISD index is the standard deviation of a respondent’s item scores on a unidimensional subscale (Marjanovic, Holden, Struthers, Cribbie, & Greenglass, 2015). Despite the simplicity of the index, Marjanovic et al. (2015) found that this measure could distinguish completely random response patterns from nonrandom response patterns on a 60-item personality inventory, and Franz (2016) found that the index could detect response patterns with at least 25% random item scores on a test battery of 261 items, with high sensitivity and specificity. However, other studies suggested that the ISD index should be used with caution and its value may be strongly confounded with the substantive trait measured (Curran, 2016; Golay, Fagot, & Lecerf, 2013).

**Long string index**

Repetitive careless responding can be detected by determining the length of strings of identical answers (Curran, 2016; DeSimone et al., 2015; Johnson, 2005). The $L_{max}$ index is the maximum length of a string of consecutive identical answers. Some researchers have computed $L_{max}$ as the maximum length of a string on single web or questionnaire page (Meade & Craig, 2012); others have computed its value across the
complete questionnaire (Kam & Chan, 2018; Niessen et al., 2016). $L_{\text{max}}$ is most appropriate for multidimensional questionnaires with a mixture of differently scored items (i.e., scored either positively or negatively), and many response categories (DeSimone et al., 2015). Such scale characteristics reduce the possibility that a conscientious respondent produces long strings. For ROM scales, long string indices may have limited value as many of such scales assess highly related dimensions and items are often scored in a single direction.

**Multivariate outlier index**

The Mahanalobis distance (MD) is a multivariate outlier statistic (e.g., Johnson & Wichern, 2008) that quantifies the distance between a respondent’s item-score pattern and the mean item scores in the remaining data, while taking into account inter-item correlations. Several (simulation) studies have recommended this index for detecting random careless responding (e.g., Meade & Craig, 2012; Maniaci & Rogge, 2014; DeSimone et al., 2015). A potential disadvantage of this index is that it may also detect persons with extreme trait values, whereas an extreme trait value does not necessarily imply careless responding.

**Person-fit indices**

$L_{p}^{z}$ index

The parametric $L_{p}^{z}$ person-fit index for polytomous items (Dragogow, Levine, & Williams, 1985) is the standardized log-likelihood of a response pattern given the estimated unidimensional IRT model. The $L_{p}^{z}$ index mainly detects random careless responding but also has the potential to pick up other types of carelessness if these lead to inconsistencies with the IRT model (Emons, 2008). Disadvantages of the $L_{p}^{z}$ index are that large samples are required to accurately estimate IRT models for polytomous data. Also, IRT item parameters may be biased due to the misfit commonly observed between psychopathology data and unidimensional IRT models (Thomas, 2011). On the other hand, simulation studies suggest that sensitivity and specificity of $L_{p}^{z}$ is not affected much by parameter bias or violations of unidimensionality (Conijn et al., 2014, 2015). Therefore, the $L_{p}^{z}$ index may be appropriate for application to ROM data, despite some possible misfit between the estimated IRT model and the observed data.

$G_{n}^{p}$ and $G_{n}^{p}$ indices

The nonparametric Guttman statistic, $G_{n}^{p}$, is the sum of the number of Guttman errors in a response pattern (Emons, 2008; Guttman, 1950). The $G_{n}^{p}$ index weights the number of errors (i.e., $G_{n}^{p}$) by the maximum number of errors for the total score of the corresponding response pattern (Emons, 2008). For items with a polytomous response format, a Guttman error occurs when a respondent endorses a less popular item category (i.e., indicative of high symptom severity) without endorsing a more popular one (i.e., indicative of low symptom severity). Item popularity is based on the average category endorsement rates in the sample. $G_{n}^{p}$ and $G_{n}^{p}$ are highly related to $L_{p}^{z}$ (e.g., Niessen et al., 2016) and are also mostly sensitive to careless responding that leads to response inconsistencies, such as random responding. Comparing $G_{n}^{p}$ and $G_{n}^{p}$, $G_{n}^{p}$ has the advantage of being less confounded with the respondents’ trait value but $G_{n}^{p}$ has the advantage of higher detection rates for random careless responding (Emons, 2008; Franz, 2016).

**Comparison studies**

Two recent studies conducted a comprehensive comparison of post hoc validity indices, including also person-fit indices. In the first study, Franz (2016) used item parameters of three different psychopathology scales (resulting in a total of 261 items) to generate data using IRT models, and compared detection rates of 11 statistics. Results showed that for detecting severe random careless responding (e.g., response patterns including 50% random responses), $G_{n}^{p}$, $G_{n}^{p}$, $L_{p}^{z}$, MD, and ISD all had very high sensitivity and specificity values. For detecting less severe random careless responding (e.g., 25% random responses), $L_{p}^{z}$ performed best, MD, $G_{n}^{p}$, and $G_{n}^{p}$ performed quite similarly, while ISD performed somewhat more poorly. The EO and the statistical synonyms and antonyms indices performed poorly in all conditions. $L_{\text{max}}$ performed best in detecting repetitive responding, with high sensitivity and specificity for detecting severe repetitive responding. In the second study, Niessen et al. (2016) compared the performance of $L_{\text{max}}$, $G_{n}^{p}$, $L_{p}^{z}$, EO, and MD using data of a 100-item big five inventory using both an experiment (respondents were instructed to respond carelessly) and a simulation study (partial random response patterns were inserted in real data). Results were largely consistent with those of Franz (2016): EO was outperformed by all indices; the MD index was outperformed by $G_{n}^{p}$ and $L_{p}^{z}$ particularly when the prevalence of careless responding was low.
(5–10%) and data included only mild types of carelessness. A comparison of $L_p$ and $G_p$ showed that detection rates were either comparable (in the experimental study) or somewhat higher for $L_p$ (simulation study). In the experimental study, the $L_{max}$ index had detection rates that were similar to those of $L_p$ and $G_p$.

Given the expected low prevalence and mild types of carelessness in ROM data, these two studies suggest that person-fit indices may have best performance for detecting random careless responding in ROM data, and that the EO and statistical synonyms and antonyms indices are least useful. Furthermore, the two studies suggest that the $L_{max}$ index may also perform well in detecting repetitive careless responding. However, both studies are not representative for typical ROM data, for example, Franz (2016) used a very large amount of items and Niessen et al. (2016) used data of a personality inventory.

**Method**

**Participants and procedure**

We used secondary data from ROM baseline measurements from outpatients referred to the Department of Psychiatry of the Leiden University Medical Center (LUMC) and to the Regional Mental Health Care Provider GGZ Rivierduinen for treatment of mood, anxiety, and/or somatoform disorders. Patients with insufficient command of the Dutch language, who were illiterate, or who suffered from serious cognitive impairments were excluded from the ROM procedure. Because ROM is considered an integral part of the treatment process, no written informed consent of patients is required, and anonymized ROM data can be used for scientific purposes (LUMC Ethics approval).

The standard ROM baseline assessment included a structured diagnostic interview (Mini-International Neuropsychiatric Interview Plus; Sheehan et al., 1998) and several observer-rated and self-report instruments (de Beurs et al., 2011). Additionally, demographic data were collected, such as housing situation, employment status, and nationality. The data were collected using touch-screen computers, showing one item at a time on a screen, and items of different subscales of a single questionnaire were shown in mixed order. The ROM baseline session, including the MINI-plus interview, took 1–2 h. For more details on ROM, see de Beurs et al. (2011).

In the current study, we included the anonymized data of 3,543 psychiatric outpatients, who had their baseline ROM measurement between 2005 and 2009 and who completed the BSI (Derogatis & Melisaratos, 1983) and the Mood and Anxiety Symptoms Questionnaire (MASQ; Watson et al., 1995). We excluded 60 patients (1.6%) who had missing item score values, so our final sample consisted of 3,483 patients (64.6% woman), with ages ranging from 17 to 91 years ($M_{age}=39.0, SD=12.7$). Of these patients, 7.6% completed primary education, 29.4% lower secondary education, 34.6% higher secondary education, and 17.7% higher professional or university education. Diagnosis and ethnicity information was available for a subsample ($n=2,026$). In this sample, 9.6% of the patients were of nonDutch ethnic origin, and the percentages in different diagnostic categories were as follows: pure depression or dysthymia (31.6%); pure anxiety disorder (10.2%); pure somatoform disorder (3.5%); comorbid anxiety, depression, or somatoform disorder (37.3%); and other disorders (17.5%).

**Measurement instruments**

From the many ROM self-report instruments used in the assessment, we selected the BSI and the MASQ because they were completed by a large part of the sample and because they are often used in ROM procedures (Carlier & van Eeden, 2017). The BSI is a shortened version of the Symptom Checklist (SCL-90) and consists of 53 items. Of the 53 items, four items are used as independent clinical indicators. The remaining 49 items are assigned to one of nine subscales consisting of four to seven items each: somatization, obsessive–compulsive, interpersonal sensitivity, depression, anxiety, hostility, paranoid ideation, psychoticism, and phobic anxiety (de Beurs & Zitman, 2006; Derogatis & Melisaratos, 1983). Respondents indicate how much a problem bothered or distressed them during the past week on a five-point Likert scale ranging from 0 (not at all) to 4 (extremely). The total score on the BSI reflects the degree of general psychopathology (de Beurs, den Hollander-Gijsman, Helmich, & Zitman, 2007). Some research found support for the theoretical nine-factor structure (de Beurs & Zitman, 2006; Derogatis and Melisaratos, 1983) but more recent studies found support for a bi-factor model suggesting a strong general dimension underlying all items (Thomas, 2012; Urbán et al., 2014). In the current study, the Cronbachs’ alpha value for the psychoticism subscale equaled .71 but was at least .80 for the other subscales. Correlations between subscale total scores ranged from .41 to .85.
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). In the present study, we used the Dutch adaptation of the MASQ (de Beurs et al., 2007), which consists of 90 items asking about symptoms of depression and anxiety using a five-point Likert scale ranging from 1 (not at all) to 5 (extremely). Of these 90 items, 14 items are used as independent clinical indicators. The other 76 items are assigned to one of three main subscales: Anhedonic Depression (AD; 22 items), Anxious Arousal (AA; 17 items), and General Distress (GD; 37 items). GD is further divided into three specific subscales: General Distress Depression (GDD; 12 items), General Distress Anxiety (GDA; 11 items), and General Distress Mixed (GDM; 14 items). Fifteen of the MASQ items, nearly all of which belong to the AD subscale, describe positive feelings and are reverse scored.

Research results on the factor structure of the MASQ generally support the three main scales, but there is limited support for the three GD subscales (e.g., de Beurs et al., 2007; Watson et al., 1995). Another study showed neither support for the three-factor nor the five-factor solution and suggested that alternative measures should be used for measuring anxiety and depression (Boschen & Oei, 2006). Research shows sufficient evidence for the reliability and validity of the Dutch MASQ (de Beurs et al., 2007). In our study, Cronbach’s alpha values for the subscale scores ranged from .86 to .95. Correlations between subscale total scores ranged from .48 to .78.

**Validity indices**

We applied validity indices separately to the BSI and the MASQ. We used all indices reviewed in the “Post hoc validity indices in ROM” section, apart from the statistical synonyms and antonyms indices and the OE index. These indices were excluded because (1) the number of subscales and synonym/antonym item pairs in the MASQ and BSI were too low for a reliable detection of (partial) careless response patterns, and (2) simulation studies showed poor performance compared to the other validity indices (see “Post hoc validity indices in ROM” for more elaborate explanations).

**Computation**

$L_{\text{max}}$ was computed as the maximum length of a string of consecutive identical answers across the complete questionnaire. We could not compute $L_{\text{max}}$ for separate webpages (Meade & Craig, 2012) because only one item was shown on a screen. $L_{\text{max}}$ was based on the original item-scores; the other indices were computed after recoding the negatively worded MASQ items. The validity indices requiring unidimensional data (ISD, $E_p$, $G_p^p$, and $G_p^n$) were first computed for each subscale and next averaged into an overall multiscale validity index. For the person-fit indices, this approach corresponds to computing multiscale person-fit statistics (e.g., Conijn et al., 2014; Niessen et al., 2016). Although the MD index does not assume unidimensionality, it is computationally intensive when applied to many items and therefore the same multiscale approach was taken (Meade & Craig, 2012; Zijlstra et al., 2011). In calculating the ISD, $E_p$, $G_p^p$, $G_p^n$, and MD indices, we did not include the four BSI items and the 13 MASQ items that did not belong to a subscale (see “Measures” section). Consistent with previous studies, the $E_p$ index was computed with respect to the graded response IRT model (GRM) (e.g., Wanders, Wardenaar, Penninx, Meijer, & de Jonge, 2015; Wardenaar et al., 2015). Apart from the $E_p$ index, all validity indices could be interpreted such that higher positive values indicate more careless responding. We therefore multiplied $E_p$ by $-1$ so that a higher value was indicative of more careless responding.

**Simulated null distribution for the indices**

We simulated null distributions for the validity indices (i.e., their distribution in clean data without careless responding) for two purposes: assess how the indices relate to the total scale score in clean data (i.e., RQ1) and to derive the cutoff values for categorizing respondents as careless (i.e., RQ2, see the “Classification methods” section). To simulate null distributions, we generated BSI and MASQ subscale data based on various unidimensional and multidimensional GRMs that were estimated using the real BSI and MASQ data. For items with five ordered categories, as is the case for all BSI and MASQ items, the unidimensional GRM describes each item by four category-threshold parameters ($\beta_1$ to $\beta_4$), indicating the popularity of a response category, plus one discrimination parameter ($\chi$). The GRM estimates a single latent trait value for each respondent (e.g., depression symptom severity on the BSI depression subscale). The multidimensional GRM (e.g., Reckase, 2009) extends the unidimensional GRM by positing multiple latent traits and multiple corresponding discrimination parameters.
To obtain a null distribution for the $L_{max}$ index, we needed to produce data for all items, not only those belonging to a subscale. The BSI and MASQ included four and 14 items, respectively, that did not belong to any subscale. Therefore, we assessed with which subscale total score these “single” items correlated the highest and then attributed each item to the appropriate subscale. These correlations ranged from .27 to .69. One of the single MASQ items (Item 32) did not correlate substantially with any subscale ($r<.10$) and was therefore excluded.

The first step of our null-distribution simulations was to assess dimensionality for each of these “extended” subscales. We used the nfactors R package (Raiche, 2010) to assess dimensionality of the subscale data by means of three different approaches: the number of eigenvalues >1, parallel analysis, and a visual inspection of the scree plot. When methods provided a different number of factors, we chose the larger number to minimize the risk of missing important factors. Results of the three different approaches showed that the BSI scales could be well-described by a single factor. The MASQ subscales were dominated by either two dimensions (AD or AA) or three dimensions (GDD, GDA, and GDM), see Supplementary data, Appendix A.

The second step was to estimate either a unidimensional GRM or a multidimensional exploratory GRM for each of the subscales using the mirt R package (Chalmers, 2012) and results of the dimensionality analyses. The Metropolis–Hastings Robbins–Monro algorithm was used for parameter estimation because it has been found to perform well for both unidimensional and multidimensional exploratory IRT models (Cai, 2010). Supplementary data, Appendix A provides model and item-fit indices for the estimated (M)IRT models.

In the last step, we used the estimated (M)IRT models to generate 200 replicated clean data sets (i.e., without careless responses) of $N=5,000$ each. For every simulated participant, trait values per subscale were drawn from a multivariate normal distribution using the covariances between the trait estimates derived from the ROM data (Supplementary data, Appendix B provides results on a comparison between the real BSI and MASQ data and the simulated data). Finally, we computed the validity index values for each simulated data set, resulting in 200 replicated null distributions for each index. Note that we used multidimensional models to produce MASQ subscale data but we applied validity indices designed for unidimensional data (i.e., the person-fit indices and the ISD index) to the subscale data. We followed this procedure to obtain representative results, as model-data misfit can be expected in practice.

**Classification methods**

The first method used the null distribution of the validity indices (e.g., Seo & Weiss, 2013; Sinharay, 2017) and nominal Type I error levels of .01 and .05 (e.g., Meijer et al., 2016; Wanders et al., 2017). Specifically, for each of the 200 simulated data sets of $n=5,000$, we computed the 99th and 95th percentile values of each validity index and averaged these values into two single cutoff values for each index. We classified respondents with a validity index value larger than the cutoff value as being careless.

The second classification method was based on Tukey’s fences (Tukey, 1977, pp. 43–44; also known as the boxplot method) and followed Zijlstra et al. (2007). Classification of respondents was based on percentile values in the observed distribution of the validity index values: the 25th percentile value (i.e., Q1), the 75th percentile value (i.e., Q3), and the interquartile range (IQR; computed as $Q3 − Q1$). We classified respondents with a validity index value larger than Q3 + 1.5 IQR as careless respondents.

**Main analyses**

To address our first goal, we used the first simulated clean data set ($n=5000$), to graphically assess the relation between the total scores on the MASQ and BSI and the validity index values. To assess the extent that a Type I error depends on the person’s latent trait value we plotted the cutoff values derived from the three categorization methods. Next, we applied the validity indices to the real BSI and MASQ data, inspected the distribution of the indices and the relations between indices. To address our second goal, we compared the estimated prevalence of careless responding (i.e., number of respondents classified as careless divided by the total number of respondents) in the BSI and MASQ data based on the two different methods of classification. To address our third goal, we assessed the effect of removing respondents classified as careless from the real ROM data on the means and standard deviations of the BSI and MASQ (sub)scale scores. The (sub)scale scores were computed as the average of the subscales’ item scores (ranging from 1 to 5). We chose to use this method to compute (sub)scale scores because it allowed for a straightforward comparison of the bias across scales with a different number of items.
Results

Relation with total score

Figures 1 and 2 show the relation between the validity index values and the total BSI and MASQ score, respectively, in simulated data without careless respondents. Cutoff values based on the simulated 95th and 99th percentile and on Tukey’s fences are represented with horizontal lines. The figures show that $L_{\text{max}}$ values depend (strongly) on the BSI and the MASQ scale scores, respectively. The Type I errors mainly concern respondents with low trait values. The
shape of the distribution suggests that appropriate cut-off values depend on the samples’ symptom severity. Lower values are needed in samples with higher symptom severity and higher values are needed in samples with low symptom severity. For the ISD index, the distribution also depends strongly on the BSI and MASQ scale scores. Type I errors mainly concern respondents with moderate trait values. Furthermore, the distribution suggests that in samples with more extreme (either high or low) symptom severity lower cutoff values are needed compared to samples with moderate symptom severity. For the MD index, the shape of the distribution differs across the BSI and the MASQ data. The bivariate distribution of

Figure 2. Scatterplots indicating the relation between the total score and validity indices in simulated clean MASQ data.
between the total MASQ score and the Gp. The main differences between the real data and simulated data. Considering that values but Type I errors in the BSI data are rather weak. For the ISD index but the relationships are weaker. For ISD, MD is not problematic for the MASQ but for the BSI, the relation is strongly positive and Type I errors do not occur for respondents with high trait values. The ISD index has distributions similar to that of the ISD index but the relationships are weaker. For ISD, MD, and Gp, Type I errors do not occur for a particular trait values. For ISD, Gp, Type I errors in the MASQ data occur mainly for persons with low trait values but Type I errors in the BSI data are rather constant across scale values.

To summarize, the distribution of ISD, MD and Lmax values were most strongly confounded with the trait values, particularly for the BSI. The bivariate distribution of the Gp, Gnp and Zp indices with the scales scores seem more favorable for making fair classification decisions based on common cutoff values. However, of these three person-fit indices, only for the Lmax index, the probability of a Type I error was (rather) constant across scales scores for both the BSI and MASQ.

**Descriptive statistics**

**Prevalence and classifications**

**Table 1.** Spearman’s rho correlation between validity indices and with total score in real BSI and MASQ data.

<table>
<thead>
<tr>
<th>BSI (Total)</th>
<th>BSI (Gp)</th>
<th>BSI (Gnp)</th>
<th>BSI (MD)</th>
<th>MD</th>
<th>Lmax</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lp</td>
<td>1</td>
<td>.03</td>
<td>-.03</td>
<td>.03</td>
<td>-.37</td>
</tr>
<tr>
<td>Gp</td>
<td>.81</td>
<td>1</td>
<td>.82</td>
<td>.81</td>
<td>.47</td>
</tr>
<tr>
<td>Gnp</td>
<td>.84</td>
<td>.78</td>
<td>1</td>
<td>.56</td>
<td>.21</td>
</tr>
<tr>
<td>ISD</td>
<td>.61</td>
<td>.88</td>
<td>.64</td>
<td>1</td>
<td>.80</td>
</tr>
<tr>
<td>MD</td>
<td>.73</td>
<td>.93</td>
<td>.75</td>
<td>.86</td>
<td>.78</td>
</tr>
<tr>
<td>Lmax</td>
<td>-.21</td>
<td>-.45</td>
<td>-.21</td>
<td>-.56</td>
<td>.41</td>
</tr>
</tbody>
</table>

**Table 2.** Prevalence estimates for careless responding using the three different cutoff values.

<table>
<thead>
<tr>
<th>BSI (Total)</th>
<th>BSI (Gp)</th>
<th>BSI (Gnp)</th>
<th>BSI (MD)</th>
<th>MD</th>
<th>Lmax</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lp</td>
<td>1</td>
<td>.03</td>
<td>-.03</td>
<td>.03</td>
<td>-.37</td>
</tr>
<tr>
<td>Gp</td>
<td>.81</td>
<td>1</td>
<td>.82</td>
<td>.81</td>
<td>.47</td>
</tr>
<tr>
<td>Gnp</td>
<td>.84</td>
<td>.78</td>
<td>1</td>
<td>.56</td>
<td>.21</td>
</tr>
<tr>
<td>ISD</td>
<td>.61</td>
<td>.88</td>
<td>.64</td>
<td>1</td>
<td>.80</td>
</tr>
<tr>
<td>MD</td>
<td>.73</td>
<td>.93</td>
<td>.75</td>
<td>.86</td>
<td>.78</td>
</tr>
<tr>
<td>Lmax</td>
<td>-.21</td>
<td>-.45</td>
<td>-.21</td>
<td>-.56</td>
<td>.41</td>
</tr>
</tbody>
</table>

Prevalence estimates based on the simulated percentiles values suggested the presence of repetitive careless responding in the BSI data, but the MASQ prevalence estimates did not exceed the nominal Type I error rates. The negatively worded items in the MASQ may have triggered respondents to be more attentive and prevented them from producing long strings of identical item scores.

**The prevalence of careless responding**

We applied the indices to the real BSI and MASQ data. Table 1 shows the rank order correlations. The ISD, MD, Gp, Gnp and Zp indices were highly interrelated (.53–.93). The shape of the relations between the validity indices and the total scores was similar to those in the simulated clean data (Figures 1 and 2). The main differences between the real data and simulated clean data were a stronger positive linear relation between the total MASQ score and the Gp, Gnp and MD indices in the real data, and a weaker linear negative relation between the total score and Lmax index in both the BSI and MASQ real data. Considering that
compared to the BSI. Possible explanations are that the negatively worded items in the MASQ facilitate picking up response inconsistencies and that the longer MASQ test length increases power to detect careless responding.

Table 2. Estimated prevalence of careless responding using three different cutoff values.

<table>
<thead>
<tr>
<th>Classification method</th>
<th>Brief Symptom Inventory (BSI)</th>
<th>Mood and Anxiety Symptom Questionnaire (MASQ)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( l_p ) ( G_p ) ( G_p )</td>
<td>( l_p ) ( G_p ) ( G_p )</td>
</tr>
<tr>
<td>Simulated distribution</td>
<td>95th percentile</td>
<td>.125 .121 .100 .088</td>
</tr>
<tr>
<td></td>
<td>99th percentile</td>
<td>.060 .058 .042 .020</td>
</tr>
<tr>
<td>Tukey’s fences</td>
<td>95th percentile</td>
<td>.026 .016 .023 .076</td>
</tr>
</tbody>
</table>

Figure 4. Venn diagrams of the respondents classified as careless using Tukey’s fences and by at least one of three methods (\( l_p \), \( G_p \), and \( G_p \)) on the BSI and MASQ, respectively.

Agreement between indices
Although \( l_p \), \( G_p \) and \( G_p \) were highly correlated (Spearman’s rho \( \geq .78 \)) and some prevalence estimates were quite similar across these indices (e.g., for Turkey’s fences), the individuals that were classified as
Effect of carelessness on (sub)scale score means and standard deviations

To assess the effects of careless responding on scale score means and standard deviations, we used \( G^p \) and \( L_{\text{max}} \) to exclude careless respondents. First, because the combination of these indices resulted in the highest detection rates. Second, because these indices were most strongly related to the total score. This way, we could estimate the “maximal effect” of excluding careless respondents on the scale mean scores and standard deviations.

Tables 3 and 4 show the effect of excluding respondents classified as careless on the (sub)scale means and standard deviations. Results are shown separately for each index but also for excluding respondents that were flagged by either \( L_{\text{max}} \) or \( G^p \). For the (sub)scale score means, the largest effect was found for the AA subscale score when we used the \( G^p \) index and the simulated 95th percentile cutoff value. The difference in the mean score was 0.11, which corresponds to 15% of the score standard deviation in the full sample. For the (sub)scale score standard deviations, we found the largest difference for the BSI when we excluded respondents based on \( G^p \) and \( L_{\text{max}} \) using the simulated cutoff values—the standard deviation equaled 0.70 for the full sample and 0.63 for the clean sample.

Table 3. The differences in (sub)scale mean score values when comparing the cleaned empirical dataset with full empirical dataset \( (M_{\text{clean}} - M_{\text{clean + careless}}) \).

<table>
<thead>
<tr>
<th>Validity index</th>
<th>BSI total</th>
<th>AD</th>
<th>AA</th>
<th>GDD</th>
<th>GDA</th>
<th>GDM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full dataset – mean of the (sub)scale scores (^a)</td>
<td>2.19</td>
<td>3.43</td>
<td>1.88</td>
<td>2.58</td>
<td>2.32</td>
<td>2.79</td>
</tr>
<tr>
<td>95th percentile</td>
<td>(G^p)</td>
<td>(-0.08)</td>
<td>(-0.04)</td>
<td>(-0.11)</td>
<td>(-0.09)</td>
<td>(-0.11)</td>
</tr>
<tr>
<td>(L_{\text{max}})</td>
<td>(0.07)</td>
<td>(0.02)</td>
<td>(0.01)</td>
<td>(0.03)</td>
<td>(0.02)</td>
<td>(0.03)</td>
</tr>
<tr>
<td>(G^p) and (L_{\text{max}})</td>
<td>(-0.02)</td>
<td>(-0.02)</td>
<td>(-0.10)</td>
<td>(-0.06)</td>
<td>(-0.08)</td>
<td>(-0.06)</td>
</tr>
<tr>
<td>99th percentile</td>
<td>(G^p)</td>
<td>(-0.04)</td>
<td>(-0.01)</td>
<td>(-0.07)</td>
<td>(-0.05)</td>
<td>(-0.06)</td>
</tr>
<tr>
<td>(L_{\text{max}})</td>
<td>(0.02)</td>
<td>(0.00)</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>(G^p) and (L_{\text{max}})</td>
<td>(-0.02)</td>
<td>(-0.01)</td>
<td>(-0.06)</td>
<td>(-0.04)</td>
<td>(-0.05)</td>
<td>(-0.04)</td>
</tr>
<tr>
<td>Tukey’s fences</td>
<td>(G^p)</td>
<td>(-0.08)</td>
<td>(-0.00)</td>
<td>(-0.02)</td>
<td>(-0.02)</td>
<td>(-0.02)</td>
</tr>
<tr>
<td>(L_{\text{max}})</td>
<td>(0.07)</td>
<td>(0.05)</td>
<td>(0.03)</td>
<td>(0.06)</td>
<td>(0.08)</td>
<td>(0.06)</td>
</tr>
<tr>
<td>(G^p) and (L_{\text{max}})</td>
<td>(0.05)</td>
<td>(0.05)</td>
<td>(0.01)</td>
<td>(0.07)</td>
<td>(0.04)</td>
<td>(0.07)</td>
</tr>
</tbody>
</table>

*The (sub)scale scores are computed as the sum of the item scores (ranging from 1 to 5) divided by the number of items.

Effect of carelessness: post hoc analyses

When applying validity indices to the real data, not all careless respondents are excluded (i.e., due to a lack of power) and some respondents are excluded who are not responding carelessly (i.e., the Type I error rate). Therefore, we also assessed the effect of excluding careless response patterns from simulated data sets. To this end, we modified the 200 clean data sets of \( N = 5,000 \) (i.e., the data that was used to establish the null distribution of the validity indices) by replacing some of the clean response patterns to represent random careless respondents. We simulated a strong degree of carelessness in the data: a prevalence of 20% careless respondents with each 50% random item score.

Table 4. The differences in (sub)scale SD values when comparing the cleaned empirical dataset with full empirical dataset \( (SD_{\text{clean}} - SD_{\text{clean + careless}}) \).

<table>
<thead>
<tr>
<th>Validity index</th>
<th>BSI total</th>
<th>AD</th>
<th>AA</th>
<th>GDD</th>
<th>GDA</th>
<th>GDM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full dataset – SD of the (sub)scale scores (^a)</td>
<td>(0.70)</td>
<td>(0.75)</td>
<td>(0.73)</td>
<td>(0.99)</td>
<td>(0.76)</td>
<td>(0.85)</td>
</tr>
<tr>
<td>95th percentile</td>
<td>(G^p)</td>
<td>(-0.03)</td>
<td>(-0.02)</td>
<td>(-0.04)</td>
<td>(-0.00)</td>
<td>(-0.02)</td>
</tr>
<tr>
<td>(L_{\text{max}})</td>
<td>(-0.05)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(-0.01)</td>
<td>(-0.01)</td>
<td>(-0.01)</td>
</tr>
<tr>
<td>(G^p) and (L_{\text{max}})</td>
<td>(-0.07)</td>
<td>(0.02)</td>
<td>(-0.04)</td>
<td>(-0.01)</td>
<td>(-0.02)</td>
<td>(-0.01)</td>
</tr>
<tr>
<td>99th percentile</td>
<td>(G^p)</td>
<td>(-0.03)</td>
<td>(-0.01)</td>
<td>(-0.03)</td>
<td>(-0.00)</td>
<td>(-0.01)</td>
</tr>
<tr>
<td>(L_{\text{max}})</td>
<td>(-0.05)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
</tr>
<tr>
<td>(G^p) and (L_{\text{max}})</td>
<td>(-0.07)</td>
<td>(0.01)</td>
<td>(-0.03)</td>
<td>(0.00)</td>
<td>(-0.02)</td>
<td>(0.00)</td>
</tr>
<tr>
<td>Tukey’s fences</td>
<td>(G^p)</td>
<td>(-0.03)</td>
<td>(-0.00)</td>
<td>(-0.01)</td>
<td>(-0.00)</td>
<td>(-0.01)</td>
</tr>
<tr>
<td>(L_{\text{max}})</td>
<td>(-0.05)</td>
<td>(-0.02)</td>
<td>(-0.02)</td>
<td>(-0.03)</td>
<td>(-0.04)</td>
<td>(-0.05)</td>
</tr>
<tr>
<td>(G^p) and (L_{\text{max}})</td>
<td>(-0.04)</td>
<td>(-0.02)</td>
<td>(-0.03)</td>
<td>(-0.03)</td>
<td>(-0.04)</td>
<td>(-0.05)</td>
</tr>
</tbody>
</table>

*The (sub)scale scores are computed as the sum of the item scores (ranging from 1 to 5) divided by the number of items.

Note. The largest mean differences for each (sub)scale are indicated in bold. AD = anhedonic depression, AA = anxious arousal, GDD = general distress depression, GDA = general distress anxiety, GDM = general distress mixed. The percentage of respondents excluded from the data for the BSI by both \( G^p \) and \( L_{\text{max}} \) was 21% (95th percentile), 8% (99th percentile), and 9% (Tukey’s fences). For the MASQ these percentages were 22% (95th percentile), 12% (99th percentile), and 14% (Tukey’s fences). The (sub)scale scores are computed as the sum of the item scores (ranging from 1 to 5) divided by the number of items.
responding for a given respondent, items were randomly selected from the questionnaire, and the IRT-based item scores were replaced by random item scores. For half of the careless respondents, the random item scores came from a uniform distribution; for the other half, the scores came from a normal distribution (Meade & Craig, 2012). Finally, we calculated the difference between the means and standard deviations of the (sub)scale total scores before and after inserting the careless responses for every simulated response pattern. These differences were averaged into mean difference values (Table 5).

The differences in mean scale scores before and after including all of the simulated random careless responses in the clean data sets was small for the mean BSI total score (−0.08) and for the mean MASQ subscales scores (0.05 to −0.11). The differences in standard deviation values equaled at most 0.09 (GDD scale). For this scale, the standard deviation equaled, on average, 1.32 in the clean samples and 1.41 in the careless condition. So, even in data where 20% of the respondents had 50% random item scores, the effects of careless responding were not large.

### Discussion

Screening for careless responding in ROM is complicated by the fact that direct validity checks may give the patient the impression that his or her responses are not taken seriously. Implementing indirect post hoc validity checks can mitigate this problem, and may be especially feasible when ROM data are collected by means of computerized tests (Wanders et al., 2017). Our study investigated which post hoc validity checks can best be used to detect careless responding in ROM data. Moreover, we assessed the necessity of implementing these checks in the analyses of aggregated ROM data.

Our review of different validity indices and their performance in previous comparison studies (e.g., Franz, 2016; Meade & Craig, 2012; Niessen et al., 2016) suggests that the person-fit indices $\ell_{p}$ and $G^p$ are likely to have highest sensitivity and specificity for detecting random careless responding in ROM data. Person-fit indices take more (useful) information into account compared to, for example, the MD and ISD index. They use multiple parameter estimates for each item to compute the likelihood of a particular response pattern ($\ell_{p}$) or the number of response inconsistencies ($G^p$ and $G^c_n$) for each person. The ISD index does not take any item properties into account, and the MD index only takes the mean item scores and item correlation matrix into account. However, another important conclusion from the review is that additional (simulation) studies are needed to compare detection rates between person-fit statistics and simpler validity indices. Such studies should take into account the specific properties of ROM data (e.g., different violations of model assumptions, short scale length and different types of carelessness) and assess which index is most powerful when a parametric IRT model does not fit the data well.

Our simulations suggest that particularly the null distributions of the $L_{\text{max}}$, MD and ISD indices may be substantially confounded with the total scale score; in particular, when scales are short, item score distributions skewed, and all items worded in the same direction (i.e., scales such as the BSI). This is problematic in practical applications where a common cutoff value is applied to validity-index data of respondents with a large range in latent trait values. These cutoff values are more or less suitable depending on the trait value, and Type I errors may only occur for persons with specific trait values. The $\ell_{p}$ person-fit index had the most favorable null distribution for using a common cutoff in subgroups of patients with different symptom severity. The null distribution of the $G^c_n$ index should be further investigated as previous research found that the $G^p_n$ null distribution was unrelated to the scale score (Emons, 2008), while we found relatively high Type I errors for low trait values on the MASQ. Regarding the $L_{\text{max}}$ index, post hoc analyses of the real BSI data showed that long strings on the BSI were mostly strings of 0 values. Such strings may be a sign of careless responding, but they may also be due to a floor effect or purposeful under-reporting of symptoms. Future applications of the $L_{\text{max}}$ index to psychopathology scales should therefore compute different cutoff values for different response options (i.e., a separate cutoff value for the 0-score and for the 4-score) to take into account the skewed nature of the data (Johnson, 2005).

<table>
<thead>
<tr>
<th>(Sub)scalea</th>
<th>BSI total</th>
<th>AD</th>
<th>AA</th>
<th>GDD</th>
<th>GDA</th>
<th>GDM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_{\text{clean}}$</td>
<td>2.17</td>
<td>3.53</td>
<td>1.87</td>
<td>2.62</td>
<td>2.41</td>
<td>2.81</td>
</tr>
<tr>
<td>$SD_{\text{clean}}$</td>
<td>0.59</td>
<td>0.95</td>
<td>0.74</td>
<td>1.32</td>
<td>1.01</td>
<td>1.08</td>
</tr>
<tr>
<td>$M_{\text{clean}}$−$M_{\text{clean+ careless}}$</td>
<td>−0.08</td>
<td>−0.05</td>
<td>−0.11</td>
<td>−0.04</td>
<td>−0.06</td>
<td>−0.02</td>
</tr>
<tr>
<td>$SD_{\text{clean}}$−$SD_{\text{clean+ careless}}$</td>
<td>−0.02</td>
<td>−0.06</td>
<td>−0.01</td>
<td>−0.09</td>
<td>−0.06</td>
<td>−0.07</td>
</tr>
</tbody>
</table>

Note: $M_{\text{clean}}$ and $SD_{\text{clean}}$ are the average mean and standard deviation values computed in the 200 clean samples of $N = 500$ each.

The (sub)scale scores are computed as the sum of the item scores (ranging from 1 to 5) divided by the number of items.

---

Table 5. The average differences in (sub)scale M and SD values when comparing the simulated clean samples with the samples including 20% careless respondents (clean sample–full sample).

---

\[ \begin{align*}
L_{\text{max}} & : 1.27 \\
G^p & : 0.08 \\
G^c_n & : 0.11 \\
\end{align*} \]
A comparison of two different methods (IRT-based simulations and Tukey’s fences) for categorizing respondents as careless resulted in a wide range of prevalence estimates in real BSI and MASQ data. We found particularly high rates of respondents classified as careless (0.14–0.20) based on person-fit indices using the IRT-based 95th percentile cutoff value. These results are consistent with results of Wanders et al. (2017) in another ROM patient sample: application of the \( \ell_p \) statistic combined with a 95th percentile cutoff value to self-report depression data resulted in a classification rate of 0.24. Although use of the 95th percentile is common practice in person-fit research (Meijer et al., 2016), for using person-fit indices to screen for carelessness in ROM, we recommend the use of more conservative cutoff values (i.e., such as the 99th simulated percentile or Tukey’s fences value). First, because the nominal Type I error rate of .05 is relatively large compared to the expected low prevalence of carelessness in ROM data. When the prevalence is low and the test liberal, the majority of identified respondents may be false positives. Second, because there is always some degree of misfit between the IRT-based simulated data (and resulting cutoff values) and the empirical psychopathology data (such misfit was also present in our analysis; see Appendix B). This misfit may further increase Type I error rates.

Third, the high classification rates in our study and previous studies suggest that the 95th percentile cutoff also detects (very) mild forms of carelessness; such mild carelessness may only result in negligible test score bias. For practical use in ROM, it is important that only those response patterns are detected for which the carelessness leads to biased test scores. Based on the prevalence estimates in this study, we expect that the conservative categorization methods have enough power to detect such cases.

Finally, the results suggest that the effects of careless responding in ROM data on scale score means and standard deviations are not large. There are two explanations. First, many patients are motivated (at least to some extent) to produce valid data because the results may affect treatment decisions. Consequently, severe degrees of careless responding seldom occur, and most respondents classified as careless in this study showed only modest carelessness. Second, as shown by the post hoc simulated data results, even if there is a substantial degree of random responding in a sample, this still has small effects on scale score means and standard deviations. If careless responding in ROM data is similar to pure random responding, part of the random errors are canceled out on the (sub)scale score level. Despite these plausible explanations, more research should be done to verify and to expand our findings. For example, is the degree of bias also negligible when using direct screening methods or response time to exclude respondents? What is the degree of bias resulting from repetitive careless responding?

The literature review and results in this article suggest that a relatively complicated person-fit index (\( \ell_p \)) may be most useful for detecting random carelessness in ROM, as the more simple indices may be confounded with the trait level and have lower power to detect partial careless responding. This is a notable outcome as recent reviews on methods to detect careless responding did not include \( \ell_p \) (e.g., DeSimone et al., 2015; Curran, 2016). For larger mental-health care institutions, ROM systems are practically always computerized. So, despite its relative complexity, the \( \ell_p \) index seems feasible for application in ROM, even for providing real-time information on careless responding (Wanders et al., 2017). In such systems, the cutoff values for categorizing patients as careless should be established separately for different ROM measures, and be preferably based on data of large and representative samples. When validity indices are used that are not strongly confounded by the latent trait value (\( \ell_p \) and possibly also \( G_{\ell}^c \)), these cutoff values are equally applicable to groups of respondents with different degrees of symptom severity.

**Strengths and limitations**

The main strength of this study is that we reviewed and investigated indices for careless responding within the specific context of ROM, in a large and representative ROM sample. Consequently, our results provide recommendations regarding the implementation of validity checks in that context. Next to that, our results are also useful for other researchers collecting data with psychopathology scales, and provide suggestions on how to further investigate the performance of validity indices for detecting carelessness in psychopathology data.

However this study had limitations, which lead to recommendations for further research. First, we could not identify a validity index suitable for detecting repetitive careless responding in ROM data. Second, we did not compare the performance of post hoc validity checks to the performance of direct screening methods. Future research may focus on developing direct screening methods for ROM that are effective but do not lead respondents to feel that their answers...
are not trusted. Third, our study focused on careless responding but other invalid response styles such as consciously under-reporting or over-reporting symptoms (Godinho et al., 2016) are likely to be present in ROM data. These response styles may have affected our prevalence estimates, especially those based on $L_{max}$. The prevalence estimates based on the person-fit indices are less likely affected by purposeful under-or over-reporting because such respondents adopt another trait value and respond consistently with respect to that trait value (Zickar & Drasgow, 1996). An important topic for future research is to find ways to divide aberrant response patterns into a careless group and over-and under-reporting group. Response times could be used for this purpose, as careless respondents are likely faster than respondents purposefully over- and underreporting symptoms. Alternative measures could be self-reported motivation and fatigue during the completion of self-report measures. Additional advantages of including external measures such as response time in future research would be to validate the response-pattern-based validity indices and to potentially identify a subgroup of careless respondents not identified by the statistical validity indices.

Conclusion

Our study suggests that various post hoc validity indices that have been recommended for survey and personality research should be used with caution in ROM and psychopathology data. Some of these indices lack power for detecting partial carelessness, and using other indices Type I errors may only occur for respondents with specific latent trait values. Among the investigated post hoc indices in this study, person-fit index $L^P_z$, combined with conservative cutoff values, seems most appropriate for detecting random carelessness in large ROM data sets. However, future studies are needed that (1) compare the performance of $L^P_z$ with of simpler (direct) screening methods given typical ROM data properties and direct screening methods, and (2) develop more appropriate methods for detecting repetitive responding. Finally, we found that the effect of careless responding on results of aggregated data analysis may not be very large, which suggests that screening for careless responding may be most useful for improving individual decision-making.

Conflict of interest disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was not supported by a Grant.

Role of the funders/sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgments: We gratefully acknowledge the mental health care provider GGZ Rivierduinen and the patients participating in this study. The ideas and opinions expressed herein are those of the authors alone, and endorsement by the authors’ institutions is not intended and should not be inferred.

ORCID

Wilco H. M. Emons http://orcid.org/0000-0001-7576-0037

References


