The Leadership Quarterly xxx (xxxx) xxxx

Contents lists available at ScienceDirect

The Leadership Quarterly

journal homepage: www.elsevier.com/locate/leaqua

Careless responding in questionnaire measures: Detection, impact, and remedies☆

Philippe Goldammer[a,*], Hubert Annen[a], Peter Lucas Stöckli[a], Klaus Jonas[b]

[a] Department of Military Psychology and Pedagogics, Military Academy at ETH, Zurich, Birmensdorf, Switzerland
[b] Department of Psychology, University of Zurich, Zurich, Switzerland

ARTICLE INFO

Keywords:
Careless responding
Screening techniques
Factor structure
Interrater agreement and reliability

ABSTRACT

Undetected carelessly given responses in survey data diminish the credibility of study findings. We therefore examined two pressing issues: the accuracy of popular screening indices, and the impact of careless responding on the psychometric properties of constructs. In an experiment in Study 1, induced response sets were used to examine the effectiveness of seven indices in detecting careless responding. Response time per item, personal reliability, psychometric synonyms, psychometric antonyms, and Mahalanobis distance were effective. However, the longstring and intra-individual response variability were ineffective. In Study 2, the effects of careless responding were examined under normal study conditions. In this sample, 33% of the participants were identified as careless responders. Careless responding inflated item variances, biased item means towards the scale midpoint, increased residual variances of construct indicators, and reduced the within-group agreement on consensus-based constructs. To enhance the credibility of findings, therefore, screenings for careless responding should be applied regularly.

Introduction

Employees, customers, subordinates, and other stakeholders are commonly surveyed to assess the quality of leadership and/or the climate (e.g., work satisfaction, organizational commitment, organizational citizenship behavior [OCB]) in units and organizations. The use of these sources demands that the responses are given carefully and in line with the survey instructions. However, it is rare that all respondents will do so. It is more likely that a certain percentage of the sample will respond rather carelessly (Kam & Meyer, 2015; McGonagle, Huang, & Walsh, 2016). This type of response behavior has been labeled careless responding (Meade & Craig, 2012), insufficient effort responding (Huang, Curran, Keeney, Poposki, & DeShon, 2012), and random responding (Credé, 2010). In contrast to more purposeful or attentive response styles (e.g., positive and negative impression management), careless responding describes responding patterns in which participants are unmotivated to respond accurately and do not pay attention to the item contents and the survey instructions (Huang et al., 2012; McGrath, Mitchell, Kim, & Hough, 2010; Meade & Craig, 2012).

The negative consequences of undetected careless responding seem to be profound. Simulation studies have suggested that a small proportion of 10% (Woods, 2006) or even only 5% (Credé, 2010) of careless responders in a sample may be enough to alter the results and lead to different conclusions regarding the hypotheses. In addition, previous research has suggested that careless responding can have biasing effects on observed item correlations (Credé, 2010; McGonagle et al., 2016; McGrath et al., 2010), reliability estimates (Huang et al., 2012; Maniaci & Rogge, 2014), and factor loadings (Kam & Meyer, 2015; Meade & Craig, 2012) and may also distort the construct dimensionality (Huang et al., 2012; Kam & Meyer, 2015; Woods, 2006). In practice, careless responding could in turn increase the risk of implementing ineffective management strategies or making the wrong personnel decisions.

Despite the relevance of this topic, it is far from routine to screen survey data for careless responding in organizational and leadership research (Huang, Liu, & Bowling, 2015). This lack of screening might be because up to now there are still many unanswered questions concerning one of the most basic issues—the detection accuracy of careless response indices. In addition, previous studies have primarily examined the effects of careless responding in the context of personality inventories, which, of course, raises the question of whether similar effects occur in the context of leadership scales and related climate measures.

Thus, the present paper has two major aims: first, to examine the

☆ This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

* Corresponding author at: Militärakademie an der ETH Zürich, Kaserne Reppischtal, CH-8903 Birmensdorf, Switzerland.
E-mail address: philippe.goldammer@milak.ethz.ch (P. Goldammer).

https://doi.org/10.1016/j.leaqua.2020.101384
Received 26 February 2018; Received in revised form 6 December 2019; Accepted 15 January 2020

accuracy of seven commonly applied indices in detecting careless responding participants, and second, to examine to what extent careless responding affects item-level (i.e., means, [co]variances) and construct-level measures (e.g., measurement model fit, factor loadings, residuals) and the ratings of consensus-based constructs, such as group leader ratings (i.e., the aggregated group rating, but also the interrater reliability and interrater agreement of this group rating). By addressing these two issues, the present paper may help to make careless response screenings more popular in the field of organizational and leadership research. In addition, it may also provide researchers with guidance on dealing with careless responding in their data.

Factors influencing careless responding

A good start for handling careless responding is being prepared for the fact that a certain percentage of participants in the sample will have responded rather carelessly. It generally seems reasonable to expect a rate of at least 10% to 15% in surveys (Curran, 2016; Huang et al., 2012; Huang et al., 2015; Meade & Craig, 2012). However, the actual rate of careless respondents occurring in a specific study sample depends on characteristics of the respondents and characteristics of the questionnaire and study administration (Edwards, 2019).

On the side of the respondents, an obvious aspect is the respondents' motivation (Schwarz, 1999). If respondents are motivated—intrinsically (e.g., interest in the survey or its results) or extrinsically (e.g., by incentives) (Huang et al., 2012)—lower rates of careless responding can be expected. Moreover, recent studies have demonstrated the importance of the respondent's personality. For instance, conscientiousness, agreeableness, and emotional stability were each negatively related to careless responding (Bowling et al., 2016; Grau, Ebbeler, & Banse, 2019).

In addition, characteristics of the questionnaire and study administration may also impact the rate of careless responding. One example is survey length. Longer questionnaires generally require the respondents to stay attentive for a longer period, so they tend to be more prone to careless responding than shorter ones (Meade & Craig, 2012). Another example is the type of instructions. For instance, when participants were alerted that responding without effort would be detectable, the rate of careless responding was lower (Huang et al., 2012).

Detection of careless responding

Once researchers have found that careless responding might be an issue in their data, detecting survey participants with careless response patterns usually becomes the next priority. Because careless responding can have different forms, different indices have been proposed. DeSimone and Harms (2018) grouped these indices into two broad categories: direct and indirect measures.

Direct measures

Direct measures are items that are explicitly included in a survey in advance of administration. One option is to ask the participants directly for their self-reported response effort during the survey. Another option is to include instructed items, where the participants are explicitly instructed to choose a particular response option (e.g., "Please choose for the next item the response option 'completely disagree'"). A third technique for directly assessing the participants' response effort is to use infrequency or bogus items (e.g., "I have 17 fingers"; DeSimone, Harms, & DeSimone, 2015, p. 173), which can only be answered meaningfully by using one response option. Usually, participants who admit to low response effort or participants who fail to choose the correct option over several instructed or infrequency items are considered as careless responders.

However, directly assessing the respondents' effort has not been without criticism—first, because these measures can be faked (DeSimone & Harms, 2018; Edwards, 2019), and second, because the wording of these items can have an influence on the participants' responses (Edwards, 2019) (for a detailed discussion on the potential caveats of direct measures, see also Curran, 2016, pp. 13–15). A more reliable approach to assessing participants' careless responding might therefore be the use of indirect or unobtrusive measures.

Indirect measures

Indirect measures highlight irregularities in the participants' response patterns over the course of the questionnaire. They require no questionnaire modification and are not detectable for the participants (DeSimone & Harms, 2018). These indirect measures can be grouped into three subtypes (Curran, 2016; DeSimone & Harms, 2018; Edwards, 2019): measures of response invariability, response time, and response consistency.

Response invariability. Longstring and intra-individual response variability (IRV) can be subsumed under the response invariability measures (Edwards, 2019). The idea behind these measures is that careful respondents are expected to choose different response options for dissimilar items. Accordingly, little or no response variability over several (even dissimilar) items may indicate lack of effort. In the case of the longstring, the number of invariant responses (i.e., the length of the strings) over a series of consecutive items is counted (DeSimone et al., 2015). Based on this information, either the average longstring (i.e., average string length) or the maximum longstring (i.e., the longest string) is computed (Meade & Craig, 2012). For instance, if a participant has used the same response option for 20 consecutive items in a section of a questionnaire, the participant's maximum longstring is 20 (assuming that there were no longer strings of invariant responses in the questionnaire) (DeSimone et al., 2015). In the case of the IRV, simply the within-person standard deviation of responses across a specific set of consecutive items is calculated (Dunn, Heggestad, Shanock, & Theilgard, 2018). For instance, if a participant has used the response option '4' 10 times and the response option '5' 10 times over 20 consecutive items, the participant's IRV is 0.51. However, if a participant has used the response option '3' five times, the response option '4' 10 times, and the response option '5' five times over 20 consecutive items, the participant's IRV is 0.73.

Response time. Response time measures (e.g., time to complete the whole survey or parts of it) are based on the fact that a minimum amount of time is needed to read an item and choose an appropriate response option. If a participant's response time falls below the absolute minimum that is necessary to properly process these steps, the trustworthiness of these responses may be doubted (Huang et al., 2012).

Response consistency. Within-person consistency indices, such as personal reliability or psychometric antonyms/synonyms, follow the idea that participants are not expected to contradict themselves over the course of the questionnaire. These indices are obtained by calculating within-person correlations on the basis of several item pairs (a minimum of three pairs are needed) (Curran, 2016; DeSimone et al., 2015). For personal reliability, the within-person correlation can be calculated across the averages of the even- and odd-numbered items of the questionnaire scales, for instance (Jackson, 1976). In the case of psychometric antonyms and synonyms, the within-person correlation is computed across highly positively or highly negatively correlated item pairs respectively (e.g., greater than $|0.60|$ [Meade & Craig, 2012]) that were identified in the sample correlation matrix.

In contrast to the within-person consistency indices, the Mahalanobis distance reflects another type of consistency—(in)consistency with the normative response pattern of the sample. Even though the Mahalanobis distance has been typically used for searching for multivariate outliers that are not represented well in the models under study, recent studies have illustrated its use for identifying careless responders (DeSimone & Harms, 2018; Meade & Craig, 2012). Using it in the context of careless response screenings is based on the idea that a strongly deviating response pattern from the sample norm (which results in a larger Mahalanobis distance) might also indicate that the responses were given in a random or careless manner (DeSimone et al., 2015; Meade & Craig, 2012).

Remedies for careless responding

After several of these indices have been computed for each participant, the researcher has the choice between different types of remedies that allow the effect of careless responding to be contained.

Separating careless from careful respondents

The most often chosen approach is to separate careless from careful respondents. To do so, previous studies have mainly used one of the following strategies—the multiple hurdle approach or the latent class analysis approach (below, we also refer to these two approaches as overall classification approaches). In the multiple hurdle approach (Curran, 2016; DeSimone et al., 2015), for each index a cut score is defined that serves as an individual hurdle. To be considered as a careful responder, all the hurdles need to be passed. Alternatively, it is also possible to identify the assumed sample heterogeneity within the careless indices (i.e., careful vs. careless responders) through a latent class analysis (Kam & Meyer, 2015; Meade & Craig, 2012). In this case, the indices serve as indicators of the latent groups.

This grouping may then be used to examine to what extent the estimation results of interest were affected—either by directly comparing the estimates of the carefully responding group with those of the carelessly responding group or by comparing the estimates of the carefully responding group with those of the total sample (Curran, 2016; Edwards, 2019).

Modeling careless responding as control variable

Alternatively, researchers can partial out the effects of careless responding (while using the complete data set) by including it as control or moderator variable in their analyses of interest (Edwards, 2019). In the observed variable case, this idea can be implemented by including individual indices or the average of the standardized scores of the indices (Bowling et al., 2016) as a further covariate in the analysis. In the context of structural equation models (SEM), this idea might also be implemented by modeling a careless response style factor on which all the model relevant items are regressed (Huang et al., 2012; Williams & McGonagle, 2016). However, if such a response factor is modeled, the inherent limitation is that a homogenous careless response pattern has to be assumed in the data.

Treating careless responses as missing values

Another potential remedy for purifying the model estimation results from the effects of careless responding is to treat the careless responses as missing values (Edwards, 2019). These newly defined missing values could then be handled like all other missing values through maximum likelihood estimation for missing data or multiple imputation (e.g., Enders, 2010). Like the control variable approach, the missing value approach leaves the data intact (i.e., the complete data set is used) and takes into account that careless responding may vary over the course of the questionnaire (Edwards, 2019). However, to make this approach work, effective indices are needed that can be computed for individual items or isolated parts of the questionnaire.

Detection of careless responding: knowns and unknowns

Despite the increased interest in careless responding in the last years and the advances that have been made so far (e.g., DeSimone & Harms, 2018; Johnson, 2005; Meade & Craig, 2012), there are still many unanswered questions concerning one of the most basic issues—the detection accuracy of careless response indices. For instance, many of the proposed cut scores for the indices are based on rules of thumb (see Curran, 2016; DeSimone et al., 2015) that were hardly ever tested under experimental conditions (see Huang et al., 2012, for an exception). Even more worrying is that there has been no investigation so far of the classification accuracy of the two most commonly chosen overall classification approaches (i.e., multiple hurdles and latent class analysis approach). In addition, the few experimental studies that have been conducted up to now (Huang et al., 2012; Niessen, Meijer, & Tendeiro, 2016) examined the effectiveness of the indices only in the context of personality inventories, which, of course, raises the question of whether they have the same properties in the context of leadership scales and related climate measures (e.g., organizational commitment, OCB). Thus, the present paper examines the following research questions:

Research question 1. What indices are effective in detecting careless respondents in the context of leadership scales and related climate measures (e.g., organizational commitment, OCB)?

Research question 2. How effective are scale-specific careless response indices (i.e., indices that are only computed across the scale-specific items) compared to their global counterpart (i.e., indices that are computed across all questionnaire items)?

Research question 3. What is the classification accuracy of the multiple hurdle approach and the latent class analysis approach?

In addition to the uncertainties surrounding the detection accuracy of careless responding, there is also little research evidence available on the effects of careless responding within the context of leadership scales and related climate measures (e.g., organizational commitment and citizenship behavior) (McGonagle et al., 2016). Moreover, so far there has been no investigation of how consensus-based constructs, such as group leader ratings (i.e., the aggregated group rating, but also the interrater reliability and interrater agreement of this group rating) are affected. Traditionally, it has been assumed that careless responding mainly adds unsystematic variance and thereby increases the risk of making a type 2 error (Huang et al., 2012; McGrath et al., 2010). More recently, however, studies found that careless responding can add systematic variance under certain circumstances (e.g., in the case of unidirectionally keyed scales, or when scale means of the attentive respondents depart from scale midpoints) and therefore can also increase the risk of making a type 1 error (Huang et al., 2015; Kam & Meyer, 2015).

A further objective of the paper was therefore to gain insights about the nature of the effects of careless responding in the context of leadership scales and related climate measures (e.g., organizational commitment, OCB). Thus, the following three research questions were also addressed:

Research question 4. What effects does careless responding have on item-level measures (i.e., means, variances, covariances)?

Research question 5. What effects does careless responding have on construct-level measures (i.e., measurement model fit, factor loadings, intercepts, residual variances, and composite reliability)?

Research question 6. What are the effects of careless responding on consensus-based constructs (i.e., the effects on the aggregated group rating but also the effects on the interrater reliability and interrater agreement of this group rating)?

The six research questions were examined with two studies. In Study 1, an experiment was conducted in which careful and careless responding response sets were induced; this study setting allowed us to determine the accuracy of different indices and the two overall classification approaches. In addition, we wanted to obtain initial insights about the effects. Study 2 then made use of the results of Study 1 and examined the effects of careless responding under normal study conditions.

Study 1

Method

Participants and study setting

The participants were German-speaking conscripts (i.e., recruits) who were doing their military service in summer 2018 in two randomly selected training camps of the Swiss Armed Forces. The total sample consisted of 359 participants, who were mostly men (n = 357; 99.4%), but the sample also included two female participants who were doing voluntary military service. The participants were on average 20 years old (SD = 1.14). A quarter of the participants (25.1%, n = 90) had completed upper secondary school, whereas the majority (n = 250, 69.6%) had completed a certified apprenticeship. Only a minority of the participants (n = 19, 5.3%) had completed only the nine years of compulsory schooling. The participants were nested in 12 groups (i.e., platoons), of which each had on average 29.92 members (SD = 8.49). Each group was led by one group leader (i.e., platoon leader), and at the time the study took place, the participants had been led by their group leader for about 9 to 10 weeks.

In advance, the participants were told that they would take part in an experiment which would be about how to best identify careless responding in survey data and that all participants would receive 10 Swiss francs as compensation for their efforts after completion of the survey.

Experimental conditions and survey arrangement

The data was gathered group-wise (i.e., each of the 12 groups was surveyed separately). After providing a general introduction, we randomly assigned the members of each group to one of the three experimental conditions—either to the careful responding condition (n = 121), or to one of the two careless responding conditions (i.e., random responding [n = 119], opposite responding [n = 119]). Civilian instructors, who were randomly assigned to one of the three conditions (and randomly reassigned after each run), then led the three subgroups into separate labs, where they provided the participants with the instructions that were specific to each condition. The participants in the careful responding condition were instructed to answer all items on the questionnaire accurately. The participants in the careless responding condition (random) were instructed to answer randomly on 50% of the items in each scale (i.e., they could choose any response option for the specific items, regardless of whether the option applied to them or not). The participants in the careless responding condition (opposite) were instructed to give opposite responses on 50% of the items in each scale (i.e., if the answer for a specific item would have been "5," they had to answer "1" instead; "4" became "2," "3" stayed "3,", "2" became "4,"and "1" became "5").

The questionnaire was arranged in six scale-specific blocks (according to the six scales described below in the section 'Substantive measures'). The order of the six blocks within the questionnaire was randomized. Each scale-specific block was further divided into two separate survey pages. On the first page of each of these scale-specific blocks, a random selection (without replacement) of 50% of the scale items was displayed. For this selection of items, all participants were instructed to respond carefully. On the following page of the block, the items that were not selected in the previous step were displayed with the instruction that was specific to each experimental condition (i.e., for the careful responding condition "Complete the questions below exactly as they apply to you"; for the careless responding condition [random] "Complete the questions below as follows: Choose any response option, no matter whether it applies to you or not"; for the careless responding condition [opposite] "Complete the questions below as follows: Always choose the opposite of what applies to you").

Another special element of the questionnaire was the 'answer check,' which we defined for each survey page, with the result that proceeding to the next survey page was only possible if all items on a page had been answered.

The data from this experimental study is stored as Supplementary material.

Substantive measures

We included three leader behavior measures (i.e., transformational, passive-avoidant, and authentic leadership), two relational correlates of leadership (leader-member exchange, organizational commitment), and one follower effectiveness criterion (organizational citizenship behavior) in our questionnaire. These measures were selected because they represented different aspects of the integrative theoretical leadership framework (see Banks, Gooty, Ross, Williams, & Harrington, 2018) and because they could be readily adapted to the present study context.

Transformational leadership (TFL) was assessed with the items of Multifactor Leadership Questionnaire (MLQ; Bass & Avolio, 1995). In the validated German adaption of the MLQ (Felfe, 2006), TFL is measured by 24 items (e.g., "My supervisor speaks enthusiastically about what is to be achieved"; "My supervisor considers my individuality and doesn't treat me as just one of many subordinates") that are assessed on a five-point Likert scale ranging from 1 = *never* to 5 = *frequently, almost always*.

Passive-avoidant leadership (PAL) was assessed with the eight items of the two MLQ subscales management-by-exception passive and laissez-faire (Bass & Avolio, 1995) of the validated German MLQ adaption (Felfe, 2006). These items (e.g., "My supervisor fails to interfere until problems become serious"; "My supervisor avoids making decisions") were assessed on a five-point Likert scale ranging from 1 = *never* to 5 = *frequently, almost always*.

Authentic leadership (AL) was measured with the publisher's (i.e., Mindgarden) German translation of the Authentic Leadership Questionnaire (Avolio, Gardner, & Walumbwa, 2007). These 16 items (e.g., "My supervisor admits mistakes when they are made"; "My supervisor displays emotions exactly in line with feelings") were assessed on a five-point Likert scale ranging from 1 = *never* to 5 = *frequently, almost always*.

Leader-member exchange (LMX) was measured with the validated German translation (Paul & Schyns, 2014) of Liden and Maslyn's (1998) multidimensional measure. The participants assessed the 12 items (e.g., "My supervisor is the kind of person one would like to have as a friend"; "I am impressed with my supervisor's knowledge of his job") on a five-point-Likert scale ranging from 1 = *does not apply at all* to 5 = *applies completely*.

The participants' organizational commitment (OC) was measured with the validated German adaptation (COBB; Felfe & Franke, 2012) of Meyer and Allen's (1990) commitment measure. The 14 scale items (e.g., "I am proud to be part of this organization"; "I would feel kind of guilty if I left this organization now") were assessed with a five-point Likert scale ranging from 1 = *does not apply at all* to 5 = *applies completely*.

The participants' organizational citizenship behavior (OCB) was measured with the 20 items of the validated German adaptation (Staufenbiel & Hartz, 2000) based on the OCB scale proposed by Podsakoff, MacKenzie, Moorman, and Fetter (1990). Because the dimension "courtesy" could not be replicated in the German validation samples, the German version contains only four of the five subscales that Organ (1988) proposed. The participants assessed these items (e.g., "I help my colleagues when they are overloaded with an assignment"; "I actively try to prevent difficulties with my colleagues") on a five-point Likert scale ranging from 1 = *does not apply at all* to 5 = *applies completely*.

Together with three sociodemographic variables, one item for the group assignment (i.e., their group leader's last name), one item for the participants' self-reported carefulness, and one item to check the participants' compliance with the instructions, the questionnaire therefore included exactly 100 items.

Operationalization of detection approaches

We operationalized careless responding with seven indirect indices: maximum longstring, intra-individual response variability (IRV), average response time per item, personal reliability, psychometric synonym index, psychometric antonym index, and Mahalanobis distance.

For each of the seven indices we computed a global index, for which all items of the questionnaire were used, and scale-specific versions of the indices (i.e., for each of the six scales), in which only the items of the corresponding scales were used for computation. In the case of response time per item, the scale-specific version corresponded to the webpage-specific average response time per item (detailed information on how these indices were computed is provided in the Supplementary material).

Results

Compliance check

We first checked whether the participants complied with the survey instructions and therefore used the item "I always answered the questions according to the instructions," which the participants had to assess on a five-point scale (1 = *does not apply at all* to 5 = *applies completely*) at the end of the questionnaire. Almost all participants stated that they had complied or completely complied with the survey instructions ($n = 352$, 98.1%). In addition, a Wald test of parameter constraints (using Mplus Version 8.2 [Muthén & Muthén, 1998–2017] with robust maximum likelihood estimation[1]) showed that the participants in the three response conditions were equally compliant with the survey instructions, Careful: $M = 4.70$, $SD = 0.57$; Careless (random): $M = 4.82$, $SD = 0.44$; Careless (opposite): $M = 4.69$, $SD = 0.56$, $\chi^2(2) = 5.52$, $p = .06$. Based on this high and comparable level of compliance, we could proceed with the main analyses.

Accuracy of careless response indices and overall classification approaches

In the first part of the main analyses in Study 1, which was concerned with the detection accuracy of the individual indices and the overall classification approaches, we proceeded as follows: First, we determined how well the seven global careless response indices performed in detecting careless responding and how well they performed compared to each other. Next, we examined the performance of the scale-specific indices and how well they performed compared to their global counterparts. And finally, we derived cut scores for the indices and examined the accuracy of the overall classification approaches. In addition to these main analyses, we also conducted two supplementary analyses. We compared the means of global and scale-specific indices across the response conditions and examined the correlation matrix of global indices and those of their scale-specific counterparts. The results of these analyses are available as Supplementary material (i.e., Tables S1 to S4).

Accuracy of the global indices. To examine how accurately the global careless response indices performed in detecting careless responding participants (i.e., random and opposite responding) and whether some of the indices performed better than others, we plotted for every index a receiver operating characteristic (ROC) curve (see Fig. 1) and examined the corresponding area under the curve (AUC). An ROC curve illustrates how the true positive rate (i.e., sensitivity) and the corresponding false positive rate (i.e., 1-specificity) vary over the entire range of an index. Accordingly, the corresponding AUC can be interpreted as the probability that a careless response indicator yields a higher score for a randomly chosen individual who is carelessly responding than for a randomly chosen individual who is not (see Lasko, Bhagwat, Zou, & Ohno-Machado, 2005, p. 407; Streiner & Cairney, 2007, p. 125). Thus, if an index is effective in detecting careless responding participants, its curve lies above the diagonal (with an AUC significantly larger than 0.5), and if an index performs no better than chance, its curve lies close to the diagonal (with an AUC not significantly different from 0.5).

For plotting the ROC curves and for estimating the AUCs we used the nonparametric method, and for calculating the standard errors for each AUC and the differences between AUCs, we used the method proposed by DeLong, DeLong, and Clarke-Pearson (1988). All these analyses were performed in Stata (StataCorp, 2017).

The inspection of the corresponding AUCs and their confidence intervals showed (see Fig. 1) that five of the seven global careless response indices (i.e., personal reliability, Mahalanobis distance, psychometric synonyms, psychometric antonyms, and average response time per item) were effective in detecting careless responding participants. In contrast, the longstring and IRV did not perform well. The longstring performed only as well as chance; the IRV classified more participants incorrectly than correctly. Accordingly, the omnibus test for equality indicated a significant difference between the AUCs, $\chi^2(6) = 555.73$, $p < .001$. The subsequently conducted pairwise Bonferroni-corrected comparisons of the AUCs showed, for instance, that the personal reliability index and the Mahalanobis distance were equally effective and that both were more effective than all other indices (detailed results can be found in the Supplementary material, Table S5). Even though the performance of the average response time per item was only mediocre in the total sample ($\text{AUC}_{[95\%\ \text{CI}]} = 0.62_{[0.56–0.68]}$), the performance of this index seemed clearly better ($\text{AUC}_{[95\%\ \text{CI}]} = 0.85_{[0.80–0.90]}$) if only the careful and random responding participants were used as subsample (see Supplementary material, Fig. S1)—an observation that could be confirmed when the univariate logit estimates of the total sample ($0.16_{[95\%\ \text{CI} = 0.06–0.27]}$) and that of subsample ($0.90_{[95\%\ \text{CI} = 0.66–1.13]}$) were compared with seemingly unrelated estimation (SUEST; Weesie, 1999) and proved to be significantly different, $\chi^2(1) = 40.30$, $p < .001$.

Accuracy of the scale-specific indices. We proceeded similarly and relied on the same tools (i.e., ROCs and AUCs), when we examined how accurately the scale-specific indices performed in detecting careless responding participants (i.e., random and opposite responding) and how well they performed compared to the global version of the indices.

In the case of the longstring and the IRV, the Bonferroni-corrected comparisons of the AUCs showed that the scale-specific counterparts were as ineffective as the corresponding global indices. In contrast, most of the scale-specific versions of personal reliability, psychometric synonym, and the Mahalanobis distance were effective in detecting careless responding participants—however, all of them with significantly lower detection accuracy than the corresponding global index. The only index for which the scale-specific counterparts (i.e., the average response per item of the instructed webpages) turned out to be as effective as the global version was the average response time per item. More importantly, when rerunning the comparisons for this index with the subsample that only included the careful and random responding participants, the detection accuracy of some of the scale-specific indices was even better than that of the global index (detailed results can be found in the Supplementary material, Table S6).

Cut scores for careless response indices. Based on these results of the ROC

[1] In Study 1, the reported results are based on analyses in which no adjustments for clustering were applied. We decided not to apply adjustments for clustering, because in most of the analyses of Study 1 we were estimating more parameters than there were clusters (i.e., 12 groups), and using adjustments for clustering (i.e., TYPE = COMPLEX in Mplus or vce(cluster) in Stata) under such circumstances may have produced downwardly biased standard errors (Maas & Hox, 2005; McNeish & Stapleton, 2016), which in turn could have resulted in parameter tests in which the H_0 would have been rejected too often. Accordingly, the results of the cluster-robust analyses of Study 1 that we provide as a supplement (see Supplementary material, Excel file 'Cluster robust analyses of Study 1') should be interpreted with caution. Most notably however, the cluster-robust analyses of Study 1 led to only minor changes in the standard errors of most estimates, without affecting our main conclusions.

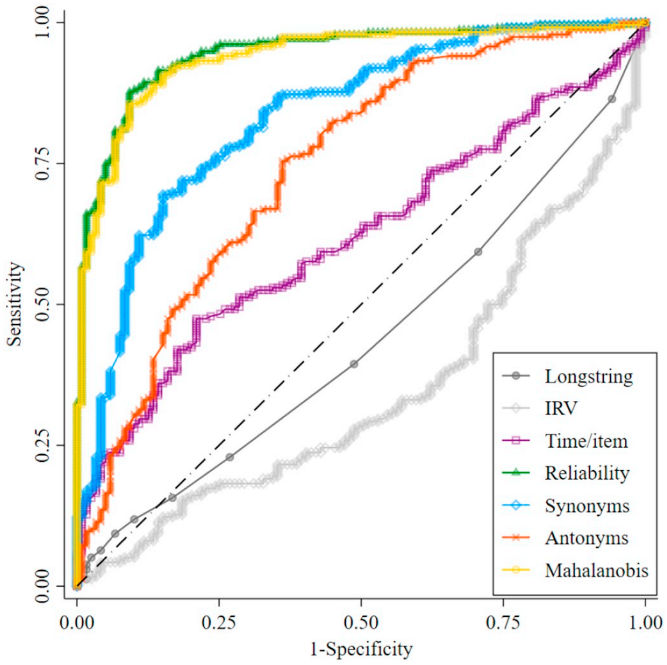

Fig. 1. Receiver operating characteristic (ROC) curves of the seven global careless response indices in the total sample (n = 355) of Study 1. For three participants, the psychometric antonyms could not be calculated, and for one participant, the psychometric synonyms could not be calculated, which was because of no variance in one of the item pair vectors. Longstring = longstring index, AUC$_{(95\%\ \text{confidence interval [CI]})}$ = 0.44$_{(0.38–0.50)}$; IRV = intra-individual response variability (reversed), AUC$_{(95\%\ \text{CI})}$ = 0.35$_{(0.29–0.41)}$; time/item = average response time per item (reversed), AUC$_{(95\%\ \text{CI})}$ = 0.62$_{(0.56–0.68)}$; reliability = personal reliability (reversed), AUC$_{(95\%\ \text{CI})}$ = 0.94$_{(0.92–0.97)}$; synonyms = psychometric synonyms (reversed), AUC$_{(95\%\ \text{CI})}$ = 0.84$_{(0.79–0.88)}$; antonyms = psychometric antonyms, AUC$_{(95\%\ \text{CI})}$ = 0.74$_{(0.69–0.80)}$; Mahalanobis = Mahalanobis distance, AUC$_{(95\%\ \text{CI})}$ = 0.94$_{(0.91–0.96)}$.

Table 1
Cut scores and levels of specificity and sensitivity for global careless response indices in Study 1.

Index	Cut score	Specificity	Sensitivity
Heuristics/existing cut scores			
Average response time per item	< 2 s per item[a]	1	0
Personal reliability	< 0.30[b]	0.98	0.62
Psychometric synonyms	< 0.22[c]	0.85	0.67
Psychometric antonyms	> −0.03[b]	0.85	0.44
Mahalanobis distance	> 117.63[c,d]	0.99	0.34
Cut scores set for 95% specificity			
Average response time per item	< 5.56 s per item	0.95	0.24 (*0.45*)
Personal reliability	< 0.42	0.95	0.75
Psychometric synonyms	< −0.03	0.95	0.34
Psychometric antonyms	> 0.36	0.95	0.16
Mahalanobis distance	> 94.81[e]	0.95	0.71
Cut scores set for 99% specificity			
Average response time per item	< 4.97 s per item	0.99	0.13 (*0.25*)
Personal reliability	< 0.26	0.99	0.57
Psychometric synonyms	< −0.30	0.99	0.15
Psychometric antonyms	> 0.55	0.99	0.07
Mahalanobis distance	> 105.03[f]	0.99	0.56

Note. If several cut scores had the same level of specificity, we reported the cut score with the largest sensitivity. Italicized values represent the sensitivities that were obtained when the subsample (i.e., careful and random responding participants; n = 240) was used; s = seconds.

[a] Based on Huang et al.'s (2012) heuristic.

[b] Based on Johnson's (2005) results.

[c] Based on DeSimone and Harms' (2018) suggestion.

[d] Represents the critical $\chi^2(94)$ value at an α level of 0.05.

[e] Represents the $\chi^2(94)$ value set for 95% specificity, which corresponds to an alpha level of 0.46.

[f] Represents the $\chi^2(94)$ value set for 99% specificity, which corresponds to an alpha level of 0.21.

analyses, we then addressed the issue of defining appropriate cut scores for the careless response indices. Naturally, the definition of cut scores will depend on the purposes of a test and thus whether the correct classification of the target cases (i.e., sensitivity) or the corrected classification of the non-target cases (i.e., specificity) is considered as more important. In our case, we considered a high level of specificity as more important than a high level of sensitivity, because the classification as careless responder may ultimately also entail the segregation of the respective participants from the rest of the sample, which in the case of falsely classified careless responders would be an unnecessary loss of power.

In our examination, we focused on the five indices that turned out to be effective in detecting careless responding (i.e., average response per item, personal reliability, psychometric synonyms, psychometric antonyms, and Mahalanobis distance) and evaluated for them the suitability of three sets of cut scores. Whereas the first set of cuts scores was based on previous research and heuristics that had been proposed, the other two sets of cut scores were empirically derived from our data with fixed levels of specificity at 95% and 99%, respectively (i.e., the cut scores were set such that only 5% or 1% of the careful respondents would be misclassified).

Table 1 shows the cut scores and the corresponding levels of specificity and sensitivity of the five effective global indices; Table S7 in the Supplementary material shows those of their scale-specific counterparts.

When comparing the derived cuts scores and levels of specificity and sensitivity of three sets, several aspects became evident. For instance, the heuristic of screening participants whose global average response time per item was faster than two seconds turned out to be too strict, because we would not have identified any of the careless responding participants if we had applied this cut score. Instead, screening participants whose global average response time per item was faster than five seconds would have been the most effective in our case. When focusing on the average response time per item on the instructed webpages, however, the heuristic of 'two seconds per item' turned out to be very effective, with specificities that ranged from 0.98 to 1 and sensitivities that ranged from 0.16 to 0.39 (in the subsample in which only the careful and random responding participants were included the sensitivities even ranged from 0.33 to 0.79).

In contrast, the proposed cut scores showed quite good performance in the case of personal reliability and Mahalanobis distance, with only a small proportion of falsely identified careful respondents and a moderate to large proportion of correctly identified careless respondents. In this context, it is also noteworthy that the proposed cut score for the Mahalanobis distance (i.e., screening the top 5% of the χ^2 distribution) was one of the existing cut scores, apart from the response time heuristic, that turned out to be effective in detecting careless responding on the scale level, with specificities that ranged from 0.98 to 1 and sensitivities that ranged from 0.13 to 0.22.

Accuracy of overall classification approaches. Based on the cut scores that we derived above, we then examined which of the two overall classification approaches (i.e., multiple hurdle approach or latent class analysis) would be more accurate in classifying the respondents into a careful and careless responding group.

The accuracy of each approach was examined under two alternative specifications. Whereas the accuracy of the multiple hurdle approach was examined for the two sets of cut scores that were derived with a specificity of 95% and 99%, the accuracy of the latent class analysis approach was examined for two alternative model specifications—for a two-class model in which local independence and equal indicator

Table 2
Accuracy of the overall classification approaches in Study 1.

Overall classification approach	Specificity	Sensitivity
Multiple hurdle		
95% specificity of each careless response index	0.79	0.96
99% specificity of each careless response index[a]	0.96	0.75
Latent class analysis		
2 classes, LI, free indicator variances across classes	0.82	0.97
2 classes, LI, equal indicator variances across classes	0.98	0.77

Note. LI = local independence.

[a] The cut scores displayed in the lower part of Table 3 were used, except for the Mahalanobis distance, for which we used 117.63 (α = 0.05) instead of the 105.03 (α = 0.21) as cut score, because we also wanted to take into account the error probability according to the alpha level.

variances across classes were specified and for a two-class model in which local independence and free indicator variances across classes were specified (detailed results of the latent class model selection are reported in the Supplementary material and the Supplementary Table S8). Table 2 shows the levels of specificity and sensitivity for each of the two alternative specifications.

It turned out that using the five indices as multiple hurdles of which each had 95% specificity resulted in nearly complete detection of all careless respondents (i.e., 96% sensitivity). However, this amount of sensitivity came at the cost of an accumulation of the individual false positive rates and eventually with an overall specificity of only 79%, which we considered as too low for our purposes. In contrast, when each of the five indices was used with a 99% specificity, the overall level of specificity remained on an acceptable level of 96% and at the same time allowed detection of 75% of the careless respondents in the sample.

The results also showed that the two-class model with freely estimated indicator variances across classes would have been too liberal in classifying participants as careless responders (i.e., a false positive rate of 18%). In contrast, the two-class model, in which the indicator variances were constrained to equality across classes, exhibited the demanded level of specificity (i.e., 0.98) and at the same time allowed detection of 77% of the careless respondents in the sample.

In sum, similar levels of overall specificity could be achieved with either approach. However, we nevertheless considered the multiple hurdle approach as more favorable than the latent class analysis approach in terms of applicability and interpretability. Whereas the classification results of the multiple hurdle approach left little room for interpretation once the 'correct' cut scores were defined, the results of the latent class analysis entailed certain ambiguities (e.g., extracting the 'correct' number of classes, or the parameter specifications within and across classes) that had to be dealt with, without guarantee that the best fitting and most interpretable latent class solution would be the best in terms of careless responder classification.

Effects of careless responding

After we had explored the accuracy of the indices and the overall classification approaches, we then turned to the second part of the main analyses in Study 1, which was concerned with the effects of careless responding on the latent construct indicators, the factor structures, and the subordinates' ratings of their group leader.

Item level. We first examined the effects of careless responding on the indicators of the latent constructs—more precisely, how careless responding affected the average inter-item covariance, the average item variance, and the average item mean. Because of the limited sample size of the three response conditions, we focused for these item level analyses on one arbitrary selected subscale of each of the six constructs. We then specified six multiple group models (i.e., one for each subscale), in which the co(variances) and means of the items were freely estimated across the three response conditions. We then separately constrained the newly introduced test parameters (i.e., averages of inter-item [co]variances and means) to equality across the three conditions. This resulted in three equality tests per subscale, making a total of 18 equality tests on the indicator level, of which each was examined at a Bonferroni-corrected alpha level. If the omnibus test passed the critical value, it was followed by pairwise Bonferroni-corrected post-hoc tests.

Table 3
Comparison of item-level measures of the transformational leadership subscale 'individual consideration' across conditions in Study 1.

Item-level measures	Response conditions			$\chi^2(2)$
	Careful	Random	Opposite	
	M (*SE*)	*M* (*SE*)	*M* (*SE*)	
Covariances	0.45^{a} (0.07)	0.11^{b} (0.06)	0.05^{b} (0.07)	22.76*
Variances	1.02^{a} (0.08)	1.27^{ab} (0.07)	1.43^{b} (0.10)	12.13*
Means	3.38^{a} (0.07)	3.17^{ab} (0.06)	3.07^{b} (0.06)	11.78*

Note. Parameter equality was tested with Wald tests. In each row, means with different superscripts are significantly different from each other (i.e., α = 0.025/54 with a critical z value of 3.31). Careful = participants in the careful responding condition; random = participants in the random responding condition; opposite = participants in the opposite responding condition. *M* = mean; *SE* = standard error of the mean estimate. Estimation results are based on robust maximum likelihood estimation for missing data (MLR) using Mplus Version 8.2.

* Larger than the critical $\chi^2(2)$ value of 11.77 (i.e., α = 0.05/18), which indicates global inequality between the means.

Table 3 shows the results of these comparisons for the transformational leadership subscale 'individual consideration.' The item level results of the other subscales are displayed in Table S9 in the Supplementary material, but they were mostly comparable to the item level results of the subscale 'individual consideration.'

All three item-level measures were affected by careless responding. Compared to the average item covariance of the careful group, the average item covariances of the careless responding groups were significantly smaller and even close to zero in the case of the opposite responding group. In addition, the average item variance of the opposite responding group turned out to be significantly larger than that of the careful responding group. The comparisons further revealed that the average item mean of the opposite responding group was significantly lower than that of the careful group.

Construct level. We then proceeded with determining the effects of careless responding on the construct level—more precisely, how the fit of the selected measurement models and the individual model parameters (i.e., factor loadings, intercepts, residuals, composite reliability) were affected. We therefore wanted to run single as well as multiple group confirmatory factor models. However, when we fitted the factor models in the two careless responding groups, the model estimation either converged not at all (mostly in opposite responding group) or resulted in solutions in which the factors had non-significant loadings on their indicators (mostly in the random responding group). In contrast, we always obtained proper solutions with factors that loaded significantly on their indicators when we examined the factor models in the careful responding group.

Thus, careless responding distorted the (co)variance structure so much that not even the most basic form of invariance (i.e., configural invariance) could be established, which in turn also meant that the prerequisite for further comparisons regarding model fit and model parameters was not met.

Group leader ratings. Finally, we examined to what extent subordinates' group leader ratings and measures of interrater reliability and interrater agreement were affected by careless responding.

Table 4
Comparison of group leader TFL ratings and the within-group agreement on the TFL ratings across response conditions in Study 1.

Group	Careful			Random			Opposite			$\chi^2(2)$
	n_{rater}	M (SE)	$r_{wg(j)}$	n_{rater}	M (SE)	$r_{wg(j)}$	n_{rater}	M (SE)	$r_{wg(j)}$	
1.	12	3.57^{a} (0.12)	0.95	12	3.23^{ab} (0.10)	0	13	3.02^{b} (0.07)	0	$16.78^{\#}$
2.	13	3.71^{a} (0.13)	0.96	13	3.28^{ab} (0.10)	0	13	2.90^{b} (0.07)	0	$34.45^{\#}$
3.	13	3.53 (0.20)	0.87	12	3.25 (0.08)	0.94	13	3.04 (0.09)	0.34	6.14
4.	13	4.15^{a} (0.12)	0.97	14	3.74^{a} (0.09)	0.82	14	3.07^{b} (0.08)	0	$69.12^{\#}$
5.	9	3.54^{ab} (0.19)	0.93	9	3.54^{a} (0.13)	0.22	8	2.93^{b} (0.06)	0	$24.54^{\#}$
6.	11	3.56^{a} (0.17)	0.90	13	3.25^{a} (0.11)	0	12	2.90^{a} (0.09)	0	$13.94^{\#}$
7.	11	3.46 (0.20)	0.92	10	3.44 (0.17)	0	11	3.02 (0.07)	0	9.11
8.	10	3.48 (0.17)	0.95	10	3.13 (0.12)	0.58	10	3.10 (0.09)	0	3.86
9.	7	3.74^{a} (0.18)	0.96	4	3.09^{ab} (0.20)	0	5	2.88^{b} (0.06)	0	$19.91^{\#}$
10.	7	2.55^{a} (0.09)	0.98	7	2.65^{ab} (0.10)	0.88	7	3.14^{b} (0.10)	0	$21.18^{\#}$
11.	7	2.49^{a} (0.22)	0.88	8	2.52^{a} (0.17)	0.80	7	3.10^{a} (0.06)	0	$15.61^{\#}$
12.	8	4.06^{a} (0.09)	0.98	7	3.34^{ab} (0.20)	0	6	3.03^{b} (0.07)	0	$87.72^{\#}$
ICC(1)		0.37, $F(11, 109) = 7.00^{***}$			0.35, $F(11, 107) = 6.44^{***}$			0.00, $F(11, 107) = 0.97$		
ICC(2)		0.86			0.84			$-0.03^{\dagger}$		
RGR, $\bar{z}_{\Delta\text{variance}}$		-3.10^{**}			-1.94			-0.47		

Note. TFL = transformational leadership. Careful = participants in the careful responding condition; random = participants in the random responding condition; opposite = participants in the opposite responding condition. M = mean; SE = standard error of the estimate. $r_{wg(j)}$ = within group agreement index for the 24 TFL items, based on a slightly skewed null distribution with a random variance of 1.34. In each row, means with different superscripts are significantly different from each other (i.e., $\alpha = 0.05/144$, with critical $\chi^2(1)$ value of 12.80). ICC = intra-class correlation coefficient; RGR = random group resampling procedure. A significant negative z-score indicates that the average within-group variance of the real groups was significantly smaller than that of the pseudo groups. Because the difference between the average within-group variance of the real groups and that of the pseudo-groups will depend on the random draw of the pseudo groups, we drew 1000 sets of pseudo-groups and averaged the z-scores (i.e., $\bar{z}_{\Delta\text{variance}}$) of the 1000 difference parameters of the within-group variances.

[†] Normally the ICCs stay within the boundaries of 0 and 1; however, if the mean square of the error variance is larger than the mean square of the between-subject variance, they may also become negative.

[#] Larger than the critical $\chi^2(2)$ value of 13.77 (i.e., $\alpha = 0.05/48$), which indicates global inequality between the means.

[**] $p < .01$.

[***] $p < .001$.

We therefore first compared the group leader ratings in the three response conditions within each of the 12 groups using the SUEST procedure in Stata. For each of the 48 equality tests that we conducted (i.e., for the mean ratings of TFL, PAL, AL, LMX in each of the 12 groups), we used a Bonferroni-corrected alpha level. If the global equality test reached the critical value, the mean differences between the conditions were further explored with Bonferroni-corrected pairwise equality tests.

Within each response condition, we then assessed the interrater reliability across the sub-group units (i.e., a third of each group) with the intra-class correlation coefficients (ICC) 1 and 2, and the interrater agreement within each sub-group unit with the $r_{wg(j)}$ (James, Demaree, & Wolf, 1984) and the random group resampling procedure (RGR; Bliese & Halverson, 2002). For the computation of these group-level indices, we used the *multilevel* (Bliese, 2016) package in R (R Core Team, 2018).

Table 4 shows results for the subordinates' transformational leadership ratings of their group leader. The results for the other leadership scales are available as Supplementary material (see Tables S10, S11, S12), but they were mostly comparable to those of the transformational leadership scale.

The comparisons of the group leader ratings across the response conditions revealed for almost all groups global inequality of the mean ratings—in most cases because the mean rating of the group members in the opposite responding condition was significantly different from that of the group members in the careful responding condition. Whereas the mean ratings of careful responding sub-group units consistently deviated from the scale midpoint (i.e., 3), those of the opposite responding sub-group units were always located close to the scale midpoint. Although none of the difference tests between the mean rating of the careful and the random sub-group units reached the Bonferroni-corrected alpha level in the case of the transformational leadership ratings, the mean ratings of the random responding group members also tended to be biased towards the scale mid-point in almost all groups.

In addition to the impact on the average ratings of the group leaders, the interrater reliability and interrater agreement were adversely affected as well. When using consensus-based constructs, such as our group leader ratings, optimally a significant proportion of the rating variance can be attributed to the group membership (indicated by high ICC 1 values [e.g., above 0.05] and a significant *F*-test), the corresponding group means can be reliably distinguished from each other (indicated by high ICC 2 values [e.g., above 0.70]), and the individual raters show high agreement with their co-raters on the target (indicated by a high $r_{wg(j)}$ value [e.g., above 0.70] and/or when the average within-group variance of the real groups is smaller than that of the randomly drawn pseudo-groups) (e.g., Woehr, Loignon, Schmidt, Loughry, & Ohland, 2015).

All of these prerequisites were met when we examined them for the careful responding sub-group units. In contrast, they were only partially met in the case of the random responding sub-group units, and none of the prerequisites were met in the case of the opposite responding sub-group units. Thus, careless responding distorted the interrater reliability and interrater agreement so much that the intended consensus-based construct (or aggregated measure) might have been no longer considered as meaningful.

Discussion

Study 1 helped to determine the detection accuracy of seven careless response indices and two overall classification approaches (research questions 1 to 3) and provided us with initial insights about the effects of careless responding in the context of leadership scales and related climate measures (i.e., organizational commitment, OCB) (research questions 4 to 6).

The analyses conducted yielded several valuable findings. First, five of the seven examined screening indices (i.e., response time per item, personal reliability, psychometric synonyms, psychometric antonyms, and Mahalanobis distance) turned out to be effective in detecting careless responding participants. However, two of the seven indices (i.e., longstring and IRV) turned out to be ineffective. Second, global careless response indices generally outperformed scale-specific ones. Third, the multiple hurdle approach, in which the cut scores for each index were set to 99% specificity, turned out to be our preferred overall classification approach: for one, because it resulted in an acceptable level of overall specificity, and for another, because of its ease of applicability. Finally, the analyses showed that the (co)variance and mean structures of the constructs were severely affected by careless responding. Compared to the careful responding condition, increased item variances, reduced inter-item covariances, and item means that tended to be biased towards scale mid-point were observed in the careless responding conditions. Because of this addition of measurement error, factor models and consensus-based constructs therefore also turned out to be no longer tenable.

In sum, the results of Study 1 therefore demonstrated the effectiveness of most of the screening procedures examined and gave us initial insights into the adverse effects of careless responding on the psychometric properties of leadership and related constructs. But even more importantly, these results provided the basis for investigating the effects of careless responding on leadership scales and related climate measures (i.e., organizational commitment, OCB) under normal study conditions.

For this purpose, in Study 2 we analyzed a dataset that we gathered between summer 2013 and summer 2015 as part of a larger research project that aimed to determine the predictive validity of the personality and intelligence tests used in cadre selection in the Swiss Armed Forces.

Study 2

Method

Participants and study setting

The sample comprised 8838 recruits who served in 32 different military training camps of the Swiss Armed Forces. The participants were predominantly male conscripts (n = 8776, 99.3%), but the sample also included 62 (0.7%) female recruits who were doing voluntary military service. A third of the recruits (32.6%, n = 2882) had completed upper secondary school; the majority (n = 5956, 67.4%) had completed the nine years of compulsory schooling and completed a certified apprenticeship. Because Switzerland is a multilingual country, with German, French, and Italian being the most widespread languages, the survey was conducted in these three languages. The majority of the participants (81.4%, n = 7198) completed the questionnaire in German, 13.8% (n = 1217) in French, and 4.8% (n = 423) in Italian. The participants were nested within 503 groups (i.e., platoons), of which each had on average 17.47 members (SD = 9.94). Each group was led by one group leader (i.e., platoon leader), and at the time the study took place, the participants had been led by their group leader for about 5 to 6 weeks.

The data collection was conducted group-wise (i.e., each of the 503 groups was surveyed separately) and guided by a civilian instructor, who was present while the recruits completed the anonymous online questionnaire. Answer checks were applied for the whole questionnaire, such that proceeding to the next survey page was only possible if all items had been answered. In case the internet connection failed, paper and pencil questionnaires were used as a backup (n = 102). The main purpose of the research project—namely, the evaluation of the cadre selection tools and the gathering of leadership ratings—remained unclear to the participants. Instead, the identification of stress sources during basic military training was emphasized as the main focus of the study. After completing the survey, all participants received a chocolate bar as a thank-you.

The data from Study 2 is also stored as Supplementary material.

Substantive measures

In the questionnaire the participants assessed the leadership of the group leader and the exchange with him/her, their own organizational commitment and citizenship behavior, their (dis)satisfaction with military basic training, and their personal motivation to lead.

To assess the leadership of the group leader, the 51 items of the validated German adaptation (Felfe, 2006) of the Multifactor Leadership Questionnaire (MLQ; Bass & Avolio, 1995) were used. These items (e.g., "My supervisor [group leader] speaks enthusiastically about what is to be achieved"; "My supervisor [group leader] considers my individuality and doesn't treat me as just one of many subordinates") were assessed on a five-point Likert scale that ranged from 1 = *never* to 5 = *frequently, almost always*.

Leader-member exchange (LMX) was assessed with the validated German translation (Schyns, 2002) of Graen and Uhl-Bien's (1995) LMX scale. The seven items (e.g., "My supervisor [group leader] knows my development possibilities") were measured on a five-point-Likert scale that ranged from 1 = *does not apply at all* to 5 = *applies completely*.

The participants' organizational commitment (OC) was measured with the validated German adaptation (COBB; Felfe & Franke, 2012) of Meyer and Allen's (1990) commitment measure. The 14 items of this scale (e.g., "I am proud to be part of this organization [Swiss Armed Forces]") were assessed with a five-point Likert scale that ranged from 1 = *does not apply at all* to 5 = *applies completely*.

Five items (e.g., "I help a comrade when he is struggling with a task") were used to assess the recruits' organizational citizenship behavior (OCB; Annen, Goldammer, & Szvircsev-Tresch, 2015), 25 items (e.g., "The limited privacy is stressful for me") to assess their satisfaction with the basic training (Brühlmann & Stgier, 2010), and six items (e.g., "I feel confident about taking on a military leadership position") to assess their motivation to lead (Swiss Armed Forces, 2012). All of these items were assessed on a 4-point Likert scale that ranged from 1 = *does not apply at all* to 4 = *applies completely*.

Together with seven sociodemographic variables and four items that asked the participants to indicate their military assignment (e.g., their group leader's last name), the questionnaire contained 119 items in its basic version. This German version was then translated into French and Italian by the Armed Forces' professional translation service. The translations were again reviewed by a team of psychologists and military officers, but the process included no back-translation.

Careless response detection and overall classification

For detecting careless respondents in the Study 2 data, we relied on four of the five indirect indices that in Study 1 turned out to be effective in detecting careless responding: average time per item, personal reliability, psychometric synonyms, and Mahalanobis distance.[2] Whereas all questionnaire items were used to calculate a global average response time per item, only the 72 items of the validated scales (i.e., MLQ, LMX, OC)[3] were used to calculate the two consistency indices (i.e., personal reliability, psychometric synonyms) and the Mahalanobis distance (detailed information on how these indices were computed is provided in the Supplementary material).

We screened respondents based on a multiple hurdle approach in which the cut score of each index was set to 99% specificity using the

[2] The psychometric antonym index was not used for screening respondents in Study 2, because only three item pairs with rather small negative correlations (i.e., close to −0.3) could be identified in the correlation matrix of the 72 items. In addition, this index could not have been calculated for 1960 participants. But the most important thing in the decision not to use an antonym index under these conditions was the rather poor detection accuracy ($\text{AUC}_{[95\%\ \text{CI}]} = 0.59_{[0.52–0.66]}$) of a similar index (i.e., 3 pairs with correlations close −0.3) in the Study 1 data, for which the cut score with a 99% specificity could not even be determined.

[3] The journal editor pointed out that the use of unvalidated scales might increase the risk of reporting effects that occurred not because the respondents were careless but because the questionnaire measures were not good. We therefore used only the 72 items of the validated scales.

cut scores derived from Study 1. Of the 8838 recruits, 761 responded with a faster rate than 4.97 s per item, 829 had a personal reliability index that was lower than 0.26, 149 had a value on the psychometric synonym index that was below −0.30, and 1740 had a Mahalanobis distance value that was larger than the critical χ^2 value of 92.80 (at an α-level of 0.05). Based on these four indices, 2941 (33.3%) of the 8838 recruits were therefore flagged as careless responders and 5897 (66.7%) of the recruits as careful responders.

Results

In the main analyses of Study 2, we then examined the effects of careless responding on the item and construct level and the group leader ratings. In addition to these analyses, we also conducted three supplementary analyses, in which we compared the means of the indices in the careful and careless responding groups, examined the correlation matrix of the indices, and examined to what extent the sample characteristics (i.e., sociodemographic variables, size of the language groups, and groups' descriptives) altered when careless responders were disregarded. The results of these analyses can be found in the Supplementary material, Tables S13 to S15.

Effects of careless responding

Item level. We examined the effects of careless responding on the item-level measures (i.e., the average inter-item covariance, average item variance, and average item mean) in two complementary ways: first, by directly comparing the estimates of the carefully responding group with those of the carelessly responding group, for which we used conventional multiple group SEM, and second, by comparing the estimates of the careful group with those of the total sample, which allowed us to examine the practical relevance of having unscreened careless responders in the data. For these nested sample comparisons, we used the generalized SEM command (gsem) in Stata, which can be used to combine and then compare estimation results of such “stacked models,” as SUEST does in the case of observed parameter estimates (Canette, 2014) (Stata and Mplus example codes for running nested sample comparisons are provided in the Supplementary material). The effects on the test parameters were then examined within each subscale,[4] and each of these comparisons was evaluated at a Bonferroni-corrected alpha level.

Table 5 shows the results of these comparisons for the transformational leadership subscale ‘individual consideration.’ The item level results of the other subscales are displayed in Table S16 in the Supplementary material and were mostly comparable to the results of the subscale ‘individual consideration.’

Compared to the careful responding group, increased average item variances and item means that were biased towards scale mid-point could be observed for the careless responding group. In contrast to results of Study 1, however, the average inter-item covariance of the careless group turned out to be larger and not smaller than that of the careful group. In addition, the nested sample comparisons revealed that the full sample was substantially affected by the inclusion of the careless respondents. Compared to the careful responding group, increased average item (co)variances and smaller average item means could be observed for the total sample.

Construct level. To determine the effects of careless responding on the construct level, we examined two aspects in Study 2: its impact on the fit of the measurement models and its impact on the individual factor model parameters. In addition to these two analyses, we also conducted a supplementary analysis in which we investigated the effect on the measurement invariance testing of the constructs across the three language groups. The results of this analysis are reported in the Supplementary material and Supplementary Table S17.

First, we examined the effects on the fit of the measurement models and therefore ran a confirmatory factor model for each subscale separately within the careful, careless, and full sample. Table 6 shows the measurement model fit results of the transformational leadership subscale ‘individual consideration,’ and Table S18 in the Supplementary material shows those of the other subscales.

Notably and in contrast to the results of Study 1, we obtained in each of the three response samples proper solutions for all examined factor models. However, no clear pattern emerged when we examined the model fit of the subscales across the response samples. When looking at the approximate fit indices, for instance, the fit of some factor models improved, whereas the fit of others was unchanged or even worse when only the careful respondents were used. Similarly, the χ^2 of the factor models did not seem to be affected systematically. Factor models either fitted according to the χ^2 in each response sample or did not fit according to the χ^2 in each response sample.

Next, we then examined the effects on the measurement model parameters and therefore ran a series of two-group models for each subscale in which we separately tested whether factor loadings, intercepts, residuals variances, and composite reliability estimates could be constrained to equality between the careful and careless group and between the careful group and the full sample.[5] Table 7 shows the results of the parameter invariance tests across the response samples for the transformational leadership subscale ‘individual consideration.’ Table S19 in the Supplementary material shows the results of the parameter invariance tests for the other subscales, which were mostly comparable to the results of the subscale ‘individual consideration.’[6]

As had already become evident in the analyses of fit of the measurement models, the factor models did not completely collapse under the presence of careless responding, as had been the case in Study 1. The parameter invariance tests further showed that not all model parameters were affected by careless responding. Whereas the factor loadings could be considered as equal across the response samples, the item intercepts and residual variances were substantially different. Compared to the careful responding sample, lower item intercepts and

[4] Initially, we wanted to examine the effects of careless responding on the higher-order structures of the MLQ scales (e.g., TFL, TAL, PAL). However, we failed to replicate the hypothesized structures in the careful and the careless responding groups because of the poor discriminant validity of the first-order factors. We therefore decided to examine the effects on the subscale level. In addition, we also decided to only report the results of the comparisons in the total sample, because comparable effects occurred when the comparisons between the careful and careless responding group were run in the language group-specific subsamples.

[5] For these analyses, we deviated from the conventional procedure of measurement invariance testing (i.e., examining the decrease of model fit using log likelihood ratio tests as parameters become increasingly constrained), in that we examined the invariance of the measurement parameters between the (sub) samples with Wald tests within the configural model. We did so because no scaling correction factors (which are necessary for calculating scaled log likelihood difference tests) could be computed for the nested sample models, because the sample covariance matrix (i.e., H1 model) in these models was, as expected, singular (i.e., freely estimated correlations between the items of careful subsample and those of the total sample approach 1). Using Wald tests was therefore the only available option for testing the invariance of the parameters between careful and total sample, and to ensure comparability, we also used Wald tests when we tested the invariance of the parameters between the careful and careless responding sample.

[6] A reviewer of an earlier version of this manuscript had pointed out that model misfit can bias parameter estimates and make the results of cross-group comparisons less trustworthy. To rule out the possibility that the misfit of our factor models (almost all our models had a significant χ^2) and not the participants' response pattern was the reason for the difference between careful and careless respondents, we reran the parameter invariance tests between these groups with respecified models that fitted the data better. However, most of the parameter differences remained significant even when they had been tested within these respecified factor models.

Table 5
Comparison of item-level measures of the transformational leadership subscale ‘individual consideration’ across (sub)samples in Study 2.

Item-level measures	Direct comparisons[a]			Nested sample comparisons[b]		
	Careful	Careless	$\chi^2(1)$	Careful	Full sample	$\chi^2(1)$
	M (*SE*)	*M* (*SE*)		*M* (*SE*)	*M* (*SE*)	
Covariances	0.44 (0.02)	0.57 (0.02)	30.87*	0.44 (0.02)	0.52 (0.02)	85.53*
Variances	0.86 (0.02)	1.30 (0.02)	346.15*	0.86 (0.02)	1.05 (0.02)	345.00*
Means	3.71 (0.02)	3.29 (0.03)	334.40*	3.71 (0.02)	3.57 (0.02)	310.53*

Note. Parameter equality was tested with Wald tests. *M* = mean; *SE* = standard error of the mean estimate.

[a] Estimation results are based on robust maximum likelihood estimation for missing data (MLR) with TYPE = COMPLEX specification using Mplus Version 8.2.

[b] Estimation results are based on maximum likelihood and cluster robust standard error estimation (i.e., vce[cluster]) using gsem in Stata 15.1.

* Larger than the critical $\chi^2(1)$ value of 12.26 (i.e., $\alpha = 0.05/108$), which indicates inequality of the estimates.

Table 6
Confirmatory factor model fit indices of the transformational leadership subscale ‘individual consideration’ across (sub)samples in Study 2.

Fit indices	Careful	Careless	Full sample
$\chi^2(2)$, *p*-value	93.42, $p < .001$	20.31, $p < .001$	91.59, $p < .001$
RMSEA (90% CI)	0.088 (0.074–0.104)	0.056 (0.036–0.079)	0.071 (0.059–0.084)
SRMR	0.016	0.013	0.015
CFI	0.995	0.994	0.998
TLI	0.985	0.982	0.995

Note. Estimation results are based on robust maximum likelihood estimation for missing data (MLR) with TYPE = COMPLEX specification using Mplus Version 8.2. RMSEA = root mean square error of approximation; 90% CI = 90% confidence interval for RMSEA; SRMR = standardized root mean square residual; CFI = comparative fit index; TLI = Tucker-Lewis index.

Table 7
Invariance tests of the parameters of the transformational leadership subscale ‘individual consideration’ across (sub)samples in Study 2.

Estimates	Direct comparisons[a]			Nested sample comparisons[b]		
	Careful	Careless	$\chi^2(1)$	Careful	Full sample	$\chi^2(1)$
Factor loadings						
λ_1	1.00	1.00	–	1.00	1.00	–
λ_2	1.77 (0.07)	1.95 (0.11)	2.98	1.77 (0.06)	1.82 (0.06)	1.58
λ_3	1.97 (0.08)	2.23 (0.12)	4.33	1.97 (0.08)	2.01 (0.07)	0.73
λ_4	1.77 (0.06)	1.91 (0.10)	2.09	1.77 (0.06)	1.80 (0.06)	0.82
Intercepts						
τ_1	4.18 (0.02)	3.89 (0.03)	187.49*	4.18 (0.02)	4.08 (0.02)	179.97*
τ_2	3.68 (0.02)	3.18 (0.04)	112.11*	3.68 (0.02)	3.51 (0.03)	203.91*
τ_3	3.60 (0.02)	3.16 (0.03)	243.97*	3.60 (0.02)	3.45 (0.02)	228.36*
τ_4	3.40 (0.02)	2.91 (0.03)	254.69*	3.40 (0.02)	3.24 (0.02)	244.19*
Residual variances						
θ_1	0.41 (0.01)	0.75 (0.03)	193.72*	0.41 (0.01)	0.52 (0.01)	173.94*
θ_2	0.55 (0.02)	0.91 (0.04)	81.75*	0.55 (0.02)	0.67 (0.02)	76.23*
θ_3	0.22 (0.01)	0.42 (0.03)	45.34*	0.22 (0.01)	0.29 (0.01)	53.08*
θ_4	0.38 (0.01)	0.65 (0.03)	78.34*	0.38 (0.01)	0.47 (0.01)	74.62*
Composite reliability						
ρ	0.82 (0.01)	0.77 (0.01)	35.09*	0.82 (0.01)	0.81 (0.01)	6.53

Note. The first indicator was used to scale the latent variable. Changing the scaling indicator did not affect the results of the direct and nested sample comparisons. Parameter equality was tested with Wald tests. Unstandardized estimates are displayed. Standard errors of the estimates are shown in parentheses. The configural model fit of the careful/careless responding two-group model was $\chi^2(4) = 112.11$, $p < .001$; RMSEA (90% CI) = 0.078 (0.066–0.091); SRMR = 0.015; CFI/TLI = 0.993/0.980. The configural model log likelihood of the careful/full sample two-group model was −71,827.52.

[a] Estimation results are based on robust maximum likelihood estimation for missing data (MLR) with TYPE = COMPLEX specification using Mplus Version 8.2.

[b] Estimation results are based on maximum likelihood and cluster robust standard error estimation (i.e., vce[cluster]) using gsem in Stata 15.1.

* Larger than the critical $\chi^2(1)$ value of 14.99 (i.e., $\alpha = 0.05/462$), which indicates inequality of the estimates.

larger residual variances were obtained for the careless and full samples. The larger residual variances in the factor models of the careless and full samples then in turn also tended to reduce the composite reliability of the measure within these samples.

Group leader ratings. Finally, we examined to what extent subordinates' group leader ratings and measures of interrater agreement and interrater reliability were affected by careless responding. We therefore inspected the changes in the aggregated group ratings, in the ICCs 1 and 2, in the within-group agreement (i.e., $r_{wg(j)}$), and in the results of the random group resampling procedure when these indices were calculated with and without careless responding group members.

Table 8 shows the results for the aggregated TFL ratings. Table S20 in the Supplementary material shows the results of other selected leadership scales (i.e., TAL, management by exception passive [MBEP], LMX), which revealed a similar pattern.

Table 8
Comparison of aggregated group leader TFL ratings across response conditions in Study 2.

	Careful group members only	All group members	$\chi^2(1)$
	M (*SE*)	*M* (*SE*)	
Aggregated group rating	3.67 (0.02)	3.59 (0.02)	102.27[a]
$r_{wg(j)}$	0.94 (0.01)	0.90 (0.01)	39.10[a]
ICC (1)	0.31, *F*(483, 5378) = 6.53***	0.31, *F*(502, 8288) = 8.91***	
ICC (2)	0.85	0.88	
RGR, $\bar{z}_{\Delta\text{variance}}$	−10.72***	−13.36***	

Note. *M* = mean; *SE* = standard error of the estimate. $\chi^2 = \chi^2$ value that is based on robust maximum likelihood estimation for missing data (MLR) in Mplus Version 8.2 with TYPE = COMPLEX. $r_{wg(j)}$ = within group agreement indices of the 24 TFL items, based on a slightly skewed null distribution with a random variance of 1.34. ICC = intra-class correlation coefficient; RGR = random group resampling procedure. A significant negative z-score indicates that the average within-group variance of the real groups was significantly smaller than that of the pseudo groups. Because the difference between the average within-group variance of the real groups and that of the pseudo-groups will depend on the random draw of the pseudo groups, we drew 1000 sets of pseudo-groups and averaged the z-scores (i.e., $\bar{z}_{\Delta\text{variance}}$) of the 1000 difference parameters of the within-group variances.

[a] Larger than the critical $\chi^2(1)$ value of 7.48 (i.e., $\alpha = 0.05/8$), which indicates inequality of the estimates.

*** $p < .001$.

In contrast to Study 1, careless responding did not distort the leadership ratings so much that the consensus-based leadership constructs could no longer be upheld, but it nevertheless adversely affected relevant aspects of the group-level measures. Whereas the ICCs and the results of the random group resampling procedure[7] did not seem to be adversely affected, the negative effects were more clearly apparent in the case of the aggregated group leader ratings and the within-group agreement on these ratings. For the positively worded scales TFL, TAL, and LMX, for instance, the aggregated group leader ratings were on average lower when the ratings of all group members (including the carelessly responding members) were used than when only the ratings of the carefully responding group members were used. For the negatively worded scale MBEP, however, the aggregated group leader ratings were on average higher when the ratings of all group members were used than when only the ratings of the carefully responding group members were used. For all examined scales, the within-group agreement on the group leader rating was on average lower when the ratings of all group members were used than when only the ratings of the carefully responding group members were used.

In addition, the adverse effects of careless responding on these group-level measures turned out to be more pronounced as the percentage of careless responders in the group increased. For instance, a higher percentage of careless responders in the group was associated with a lower TFL rating of the group leader ($r = -0.40$, $z\ -8.33$, $p < .001$) and lower agreement within the group on this rating ($r = -0.23$, $z\ -3.61$, $p < .001$). And not surprisingly, the absolute difference between the target rating obtained from all group members and that obtained from only the careful responders increased as the percentage of careless responders in the group increased ($r = 0.49$, $z = 15.50$, $p < .001$).

Discussion

Study 2 helped us to gain insights about the effects of careless responding on leadership scales and related climate measures (i.e., organizational commitment, OCB) under normal study conditions (research questions 4 to 6).

Similar to Study 1, Study 2 illustrated several adverse effects on the item and construct level and on the group leader ratings. However, the adverse effects seen in Study 2 were not as severe as those found in Study 1. Because of careless responding, the average item mean became downwardly biased in the case of positively worded scales and upwardly biased in the case of negatively worded scales, and the average item variance and covariance became inflated. On the construct level, careless responding adversely affected the factor model estimation through the addition of residual variance, which in turn then also tended to reduce the composite reliability estimate. And finally, if careless responding group members were not excluded from the analyses, the aggregated group leader rating tended to become downwardly biased in the case of positively worded scales and upwardly biased in the case of negatively worded scales, and in any case, the within-group agreement on these ratings was more likely to be reduced.

General discussion

Even though careless responding in survey data seems to be a rather regularly occurring phenomenon, it is far from routine to screen for careless responding in organizational and leadership research. This lack of screening might be because up to now there has been no such systematic examination of the effectivity of commonly applied careless response screenings and no in-depth investigation of the impact of careless responding on frequently used scales in organizational and leadership research. We aimed to fill these gaps by conducting two studies. In Study 1, we conducted an experiment to determine the accuracy of seven indirect screening indices and two overall classification approaches in detecting careless responding participants. In addition, initial insights about the effects on item- and construct-level measures and group leader ratings could be gained. The cut scores derived from Study 1 were then used in Study 2 to examine the effects of careless responding on item- and construct-level measures and group leader ratings under normal study conditions.

Consistent with previous research (e.g., DeSimone & Harms, 2018; Huang et al., 2012), the response time (i.e., average response time per item) and the consistency indices (i.e., psychometric synonyms, psychometric antonyms, personal reliability, Mahalanobis distance) turned out to be effective in detecting careless respondents. However, in contrast to previous study findings (DeSimone & Harms, 2018; Dunn et al., 2018; Huang et al., 2012; Meade & Craig, 2012), the measures of response invariability (i.e., longstring, IRV) were ineffective. Even though this finding needs further confirmation from other studies, it may provide a first hint concerning the scale-dependent effectiveness of the invariability measures, with effectiveness in the case of 'more' balanced scales (e.g., personality inventories) and ineffectiveness in the case of unidirectionally keyed scales (e.g., leadership scales).

[7] At the first sight, the larger ICC 2 values and larger z values for the RGR procedure that were obtained in the full sample tended to indicate that higher interrater reliability and agreement are achieved if all group member were used. However, the larger values of these indices in the full sample were mainly because of the fact that the average group size was larger in the full sample than in the careful responding sample.

Table 9
Summary of recommendations when dealing with careless responding in survey data.

Prevention and precaution

1. Keep the study participants motivated during the conducting of the survey by using incentives.
2. Commit the participants to the study purposes by providing personal instructions.
3. Design short questionnaires and include only items that are necessary for the study purposes.
4. Place central items at the beginning of the survey.
5. Increase planned sample size by the expected loss of participants because of careless responding.

Detection

1. Items that directly assess the participants' response effort do not have to be included in the questionnaire.
2. Use response time measures (i.e., average response time per item) and consistency indices (i.e., psychometric synonyms, psychometric antonyms, and Mahalanobis distance) for detection. If cut scores are used, only use those validated under experimental conditions.
3. Do not use invariability measures (e.g., maximum longstring or intra-individual response variability) for detection.
4. Compute personal reliability and psychometric synonyms and antonyms by using item-pairs of the whole questionnaire.
5. Psychometric synonyms and antonyms should only be used for detection if a sufficient number of item pairs (we recommend > 5 pairs) with sufficiently large correlations (i.e., at least above 0.60 for synonyms pairs, and at least below −0.40 for antonyms pairs) can be obtained.
6. Average response time per item and Mahalanobis distance are suited for local (e.g., webpage-specific) careless response detection.
7. For overall classification, use the indices as multiple hurdles with 99% specificity for each index.

Remedies

1. Running a multiple group model across response subsamples requires a larger sample size but also allows for comparisons of measurement model parameters.
2. Including careless responding as a covariate or response style factor leaves the sample intact but assumes a homogenous careless response pattern.
3. Treating carelessly given responses as missing values leaves the sample intact. However, only the average response time per item and the Mahalanobis distance can be used for detection.

When the five effective indices were then combined to classify the respondents into careful and careless responding groups, the multiple hurdle approach with 99% specificity for each index turned out to be more convincing than all variants examined: for one, because it resulted in an acceptable level of overall specificity and a high level of sensitivity, and for another, because of its ease of applicability.

In both studies, the subsequently conducted subgroup analyses then showed that careless responding substantially affected item- and construct-level measures and group leader ratings.

In line with Huang et al.'s (2015, p. 830) assumptions, careless responding inflated the item variances and biased the item means towards the scale midpoint. In the case of positively worded leadership scales (e.g., TFL, LMX), in which the means of the careful respondents mostly lay above the scale midpoint, careless responding tended to induce a downward bias. In the case of negatively worded leadership scales (e.g., MBEP), in which the mean of the careful respondents usually lay below the scale midpoint, careless responding tended to induce an upward bias. Furthermore, careless responding also affected the item covariances. Careless responding had an attenuating effect on the item covariances in Study 1, but it had an inflating effect on them in Study 2.

The effects of careless responding also became apparent when the factor structures of the leadership scales were examined. Here, careless responding mainly increased the measurement error, which led to an almost complete collapse of the expected factor structure in Study 1 and which tended to reduce the composite reliability estimates in Study 2. In contrast to previous study findings that reported an increase of model fit after careless respondents had been removed, however, (Huang et al., 2012; Woods, 2006), no clear pattern emerged when we examined the fit of the measurement models across the response samples in Study 2.

The examination of the group leader ratings then revealed effects similar to those encountered for the item means. Once careless responding group members were included in the analysis, the aggregated group leader ratings tended to be biased towards the scale midpoint. In the case of positively worded leadership scales (e.g., TFL, LMX), in which the means of the careful responding group members mostly lay above the scale midpoint, the inclusion of the careless responding group members tended to induce a downward bias. In the case of the negatively worded leadership scales (e.g., MBEP), in which the mean of the careful responding group members usually lay below the scale midpoint, the inclusion of careless responding group members tended to induce an upward bias. In addition, careless responding also tended to reduce the within-group agreement on this rating. And not surprisingly, the effects on the group leader rating and within-group agreement became more pronounced as the percentage of careless responders in the group increased.

Besides the above-mentioned findings regarding the detection and impact of careless responding, another important point to discuss are the difficulties that we encountered when we tested the factor models. Almost all the measurement models that we tested did not fit the data, even when only the careful responding participants were used in the analyses. Two conclusions may be drawn from this finding. In the more optimistic case, it may be argued that the lack of model fit simply reflects the approximate nature of the models and underlying theories. In turn, however, this conclusion would imply that future studies need to use more appropriate estimation and testing procedures, such as Bayesian SEM (BSEM) with approximate zero priors for cross-loadings and/or residual correlations (e.g., Muthén & Asparouhov, 2012) or two stage least squares (2SLS) estimation with instrumental variables (e.g., Bollen, Gates, & Fisher, 2018), which allow the obtaining of plausible estimates in the case of only approximately fitting (by using BSEM) or even locally misspecified models (by using 2SLS). In the less optimistic case, however, it may be argued that the lack of observed model fit reflects conceptual problems with the constructs (Antonakis, Bastardoz, Jacquart, & Shamir, 2016; Van Knippenberg & Sitkin, 2013), which would generally call into question the validity of the examined leadership scales and climate measures (i.e., organizational commitment, OCB). In turn, this conclusion would imply that future studies need to find and use alternative behavioral measures when studying leadership and related phenomena. Alternative measures could be questionnaires with items that more clearly map on well-defined constructs (Antonakis et al., 2016; Banks et al., 2018) or, even better, objective measures such as video materials, transcripts of speeches, and other archival sources (Antonakis et al., 2016).

Implications and recommendations for addressing careless responding

As the results illustrate, careless responding can have substantial effects on the psychometric properties of leadership scales and related climate measures (i.e., organizational commitment, OCB). If it remains undetected or is not properly addressed, careless responding can even bias the hypothesis testing or result in leadership evaluations that do not reflect the leaders' actual performance. Therefore, the following steps should be taken to minimize the impact of careless responding. A summary of these recommendations is given in Table 9.

Prevention and precaution

Probably the best strategy to minimize the impact of careless responding is to take preventive and precautionary measures. Researchers should therefore seek to keep the participants motivated and attentive during the conducting of the survey. One strategy might be to adequately reward the participants with monetary or other equivalent incentives (Edwards, 2019). Alternatively, the participants' motivation could also be increased or maintained through giving personal instructions. In such a controlled setting, the instructions could be made more salient, and personal appreciation could make the participants more committed to the study purposes. Complementary to these strategies, careless responding in the substantive measures could be also reduced by keeping the number of items at the necessary minimum (Meade & Craig, 2012), or if a lengthy survey is inevitable, by putting the most relevant items at the beginning of the survey (see Galesic & Bosnjak, 2009). In addition, researchers should plan ahead and whenever possible increase the planned sample size by the expected loss of participants because of careless responding.

Detection

For detecting careless responding in leadership ratings, we recommend the use of response time measures (e.g., average response time per item) and consistency indices (i.e., psychometric synonyms, psychometric antonyms, personal reliability, Mahalanobis distance). However, the use of invariability measures (e.g., maximum longstring or IRV) is not recommended until further research evidence on the effectiveness of these types of indices in the context of unidirectionally keyed scales is available. Further, there seems to be little need to additionally include a large number of items that directly assess the respondents' effort, because our results showed that five indirect screening indices may do a fine job in detecting careless respondents.

In the case of personal reliability and psychometric synonym and psychometric antonym indices, we recommend that only global indices be calculated, because these indices may be computed only when items of different questionnaire parts are used and because most of the global versions of the indices had a significantly better detection rate than their scale-specific counterparts in our investigation. Moreover, we recommend computing psychometric synonyms and antonyms only if a sufficient number of item pairs (e.g., we used 12 synonym and seven antonym pairs) with sufficiently large correlations can be obtained (i.e., at least above 0.60 for synonyms pairs, and at least below −0.40 for antonyms pairs), because otherwise the effectiveness of these indices will be drastically reduced. Another advantage of using more item pairs for the computation of these within-person consistency indices is that fewer respondents will have a missing value on these indices. In contrast to the within-person consistency indices, average response time per item and Mahalanobis distance can be also used on the basis of a smaller number of items (e.g., page- or scale-wise) and thus to detect local or sporadic careless responding.

If the researcher then aims to use these indices to classify the respondents into a careful and a careless responding group, we recommend using the indices as multiple hurdles with a high level of specificity for each index (e.g., 99%). However, care must be taken when using a latent class analysis for classification, as the best fitting class solution may not necessarily be the best in terms of accuracy.

Remedies

Depending on the sample size and the focus of analysis, we recommend using one of the three strategies that we outlined in the introductory section above (i.e., running a multiple group analysis across the response subsamples, running the analyses with careless responding as covariate, or running the analyses in which the careless responses are treated as missing values) to contain the effects of careless responding.

In large samples, in which the careless responding group may reach the size of an individual sample, a multiple group analysis can be conducted. This analysis strategy has the advantage that separate (measurement) model parameters can be estimated for the careful and careless responding groups, which can then be used for direct comparisons between these groups. In smaller samples, or in the case where researchers just want to partial out the effect of careless responding from the main analyses, careless responding may be included as a covariate or response style factor. However, if the sample is too small to run a multiple group analysis or to model the effects of an additional covariate or response style factor, a researcher may seek to individually identify careless responses (using response time and Mahalanobis distance) and treat them in the subsequent main analyses as missing values.

Limitations and future research directions

The generalizability of the findings (including the cut scores derived in Study 1) may be limited because of the specific samples that were used. Indeed, samples that are drawn in the basic military training of the Swiss Armed Forces are inevitably very homogenous regarding sex (predominantly male), age (mostly 20 years old), and organizational tenure (all group members started their military service at the same time). However, the homogeneity of the current samples also has its strengths, such that the misfit of factor models, for instance, may not be attributed to omitted contextual factors (see Antonakis & House, 2014).

Another limitation concerns the setting of the study and survey administration. In this case, the compulsory character of military service and the use of answer checks could have additionally increased the rate of careless responding and thus could have led to an overestimation of its adverse effects.

In addition, it needs to be kept in mind that the way careless responding affects the results is an interplay of the rate of careless responders in the sample, the most dominant careless response pattern (DeSimone, DeSimone, Harms, & Wood, 2018), and the scale features (Kam & Meyer, 2015). Thus, even though most of the effects were in line with Huang et al.'s (2015) theoretical proposition, other patterns of effects may occur in other samples.

Another point of potential concern might be the participants' compliance in Study 1.[8] Even though the compliance check suggested that the participants were following the instructions that were specific to each response condition, it may still be possible that some participants did not understand or follow the instructions. If some participants in the careful responding condition responded to the items carelessly, the empirical cut scores of 95% and 99% were set too leniently. Conversely, if some participants in the careless responding condition answered the items carefully, the cut scores were set too strictly.

Despite these potential limitations, this investigation contributes to the literature by offering valuable insights about the effectiveness of various detection and classification approaches in the context of leadership scales and related climate measures (i.e., organizational commitment, OCB) and by offering initial insights about effects of careless responding on item- and construct-level measures and consensus-based constructs such as group leader ratings. It therefore lays the ground for future studies that examine this response phenomenon in other organizational and cultural settings.

One potential avenue for future research would be to further examine the effectiveness of the different detection, classification, and remedy approaches. For instance, it should be explored to what extent the effectiveness of the invariability measures depends on the keying of scales. On the other hand, experimental or simulation studies could also investigate the relative effectiveness of the covariate and missing value approach in recovering the unbiased estimation results.

[8] A reviewer of an earlier version of this manuscript had pointed out that the actual compliance check in Study 1 was rather weak and that the results presented for the research questions 4, 5, and 6 seem to provide stronger evidence that the manipulation worked.

In addition, it would be helpful to obtain more insights about the effectiveness of different prevention strategies. For instance, are personal survey instructions equally effective as a financial reward in preventing participants from responding carelessly? And what is the combined effect of these two approaches?

Furthermore, future studies should also examine the effects of careless responding on more complex models. Because the present study only examined the effects on the psychometric properties of individual constructs, it would be interesting to see how complete SEMs and their path are affected. Similarly, it would be interesting to find out how parts of multilevel models (e.g., random and fixed effects in two-level models) are affected by careless responding.

Finally, more research is needed for a better understanding of the underlying mechanisms of careless responding. Up to now the research field is rather atheoretical. Embedding this response behavior in a theoretical framework (e.g., Ward & Meade, 2018) may help to identify further constructs that are substantially related to careless responding, which, in turn, would make it possible to design more effective prevention strategies.

Conclusion

In sum, whenever questionnaire measures are used in organizational and leadership research, the data gathered should be routinely checked for careless responding, because this response behavior can have substantial effects on the psychometric properties of constructs, which in turn can bias the hypotheses testing or may result in leadership evaluations that do not reflect the leaders' actual performance.

By investigating two important issues—the effectivity of careless response screenings in the context of leadership scales and related climate measures (i.e., organizational commitment, OCB), and the effects of careless responding on item- and construct-level measures and group leader ratings—the present paper may contribute towards making careless response screenings more popular in the field of organizational and leadership research and may serve other researchers as good starting point for their own screenings.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.leaqua.2020.101384.

References

Annen, H., Goldammer, P., & Szvircsev-Tresch, T. (2015). Longitudinal effects of OCB on cadre selection and pursuing a career as militia cadre in the Swiss Armed Forces. *Military Psychology, 27*, 9–21.

Antonakis, J., Bastardoz, N., Jacquart, P., & Shamir, B. (2016). Charisma: An ill-defined and ill-measured gift. *Annual Review of Organizational Psychology and Organizational Behavior, 3*, 293–319.

Antonakis, J., & House, R. J. (2014). Instrumental leadership: Measurement and extension of transformational–transactional leadership theory. *Leadership Quarterly, 25*, 746–771.

Avolio, B. J., Gardner, W. L., & Walumbwa, F. O. (2007). *Authentic leadership questionnaire.* Menlo Park, CA: Mind Garden.

Banks, G. C., Gooty, J., Ross, R. L., Williams, C. E., & Harrington, N. T. (2018). Construct redundancy in leader behaviors: A review and agenda for the future. *Leadership Quarterly, 29*, 236–251.

Bass, B. M., & Avolio, B. J. (1995). *MLQ Multifactor Leadership Questionnaire: Technical report.* Redwood City, CA: Mindgarden.

Bliese, P. (2016). Multilevel modeling in R (2.6): A brief introduction to R, the multilevel package and the nlme package. Retrieved from http://rsync.udc.es/CRAN/doc/contrib/Bliese_Multilevel.pdf.

Bliese, P. D., & Halverson, R. R. (2002). Using random group resampling in multilevel research: An example of the buffering effects of leadership climate. *Leadership Quarterly, 13*, 53–68.

Bollen, K. A., Gates, K. M., & Fisher, Z. (2018). Robustness conditions for MIIV-2SLS when the latent variable or measurement model is structurally misspecified. *Structural Equation Modeling, 25*, 848–859.

Bowling, N. A., Huang, J. L., Bragg, C. B., Khazon, S., Liu, M., & Blackmore, C. E. (2016). Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality. *Journal of Personality and Social Psychology, 111*, 218–229.

Brühlmann, B., & Stgier, M. (2010). *Motivation für eine freiwillige Weiterausbildung in der Schweizer Armee – Eine empirische Langzeituntersuchung möglicher Einflussfaktoren, [Motivation for a voluntary additional military education in the Swiss Armed Forces – an empirical long-term study of possible influencing factors].* Zurich, Switzerland: ETH Zurich (Unpublished Bachelor's thesis).

Canette, I. (2014, August 18). Using gsem to combine estimation results [Stata blog]. Retrieved from https://blog.stata.com/2014/08/18/using-gsem-to-combine-estimation-results/.

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.

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology, 66*, 4–19.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics, 44*, 837–845.

DeSimone, J. A., DeSimone, A. J., Harms, P. D., & Wood, D. (2018). The differential impacts of two forms of insufficient effort responding. *Applied Psychology, 67*, 309–338.

DeSimone, J. A., & Harms, P. D. (2018). Dirty data: The effects of screening respondents who provide low-quality data in survey research. *Journal of Business and Psychology, 33*, 559–577.

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

Dunn, A. M., Heggestad, E. D., Shanock, L. R., & Theilgard, N. (2018). Intra-individual response variability as an indicator of insufficient effort responding: Comparison to other indicators and relationships with individual differences. *Journal of Business and Psychology, 33*, 105–121.

Edwards, J. R. (2019). Response invalidity in empirical research: Causes, detection, and remedies. *Journal of Operations Management, 65*, 62–76.

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

Felfe, J. (2006). Validierung einer deutschen Version des "Multifactor Leadership Questionnaire" (MLQ Form 5 × Short) von Bass und Avolio (1995) [Validation of a German version of the "Multifactor Leadership Questionnaire" (MLQ Form 5 × Short) by Bass and Avolio (1995)]. *Zeitschrift für Arbeits- und Organisationspsychologie, 50*, 61–78.

Felfe, J., & Franke, F. (2012). *Commit. Verfahren zur Erfassung von Commitment gegenüber der Organisation, dem Beruf und der Beschäftigungsform [Commit. Procedure for measuring commitment to the organization, profession and form of employment].* Bern, Switzerland: Verlag Hans Huber.

Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly, 73*, 349–360.

Graen, G. B., & Uhl-Bien, M. (1995). Relationship-based approach to leadership: Development of leader-member exchange (LMX) theory of leadership over 25 years: Applying a multi-level multi-domain perspective. *Leadership Quarterly, 6*, 219–247.

Grau, I., Ebbeler, C., & Banse, R. (2019). Cultural differences in careless responding. *Journal of Cross-Cultural Psychology, 50*, 336–357.

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.

Jackson, D. N. (1976, November). *The appraisal of personal reliability. Paper presented at the meetings of the Society of Multivariate Experimental Psychology, University Park, PA.*

James, D. L., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology, 69*, 85–98.

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from Web-based personality inventories. *Journal of Research in Personality, 39*, 103–129.

Kam, C. C. S., & Meyer, J. P. (2015). How careless responding and acquiescence response bias can influence construct dimensionality: The case of job satisfaction. *Organizational Research Methods, 18*, 512–541.

Lasko, T. A., Bhagwat, J. G., Zou, K. H., & Ohno-Machado, L. (2005). The use of receiver operating characteristic curves in biomedical informatics. *Journal of Biomedical Informatics, 38*, 404–415.

Liden, R. C., & Maslyn, J. M. (1998). Multidimensionality of leader-member exchange: An empirical assessment through scale development. *Journal of Management, 24*, 43–72.

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*, 86–92.

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.

McGonagle, A. K., Huang, J. L., & Walsh, B. M. (2016). Insufficient effort survey responding: An under-appreciated problem in work and organisational health psychology research. *Applied Psychology, 65*, 287–321.

McGrath, R. E., Mitchell, M., Kim, B. H., & Hough, L. (2010). Evidence for response bias as a source of error variance in applied assessment. *Psychological Bulletin, 136*, 450–470.

McNeish, D. M., & Stapleton, L. M. (2016). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review, 28*, 295–314.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*, 437–455.

Meyer, J. P., & Allen, N. J. (1990). The measurement and antecedents of affective, continuance and normative commitment to the organization. *Journal of Occupational Psychology, 63*, 1–18.

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods, 17*, 313–335.

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.

Niessen, A. S. M., 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.

Organ, D. W. (1988). *Organizational citizenship behavior: The good soldier syndrome.* Lexington, MA: Lexington Books.

Paul, T., & Schyns, B. (2014). Deutsche Leader-Member Exchange Skala (LMX MDM) aus der Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS) [German leader-member exchange scale (LMX MDM) from the compilation of social science items and scales (ZIS)]. Retrieved from https://zis.gesis.org/skala/Paul-Schyns-Deutsche-Leader-Member-Exchange-Skala-(LMX-MDM.

Podsakoff, P. M., MacKenzie, S. B., Moorman, R. H., & Fetter, R. (1990). Transformational leader behaviors and their effects on followers' trust in leader, satisfaction, and organizational citizenship behaviors. *Leadership Quarterly, 1*, 107–142.

R Core Team (2018). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*, 93–105.

Schyns, B. (2002). Überprüfung einer deutschsprachigen Skala zum Leader-Member-Exchange-Ansatz, [Examination of a German scale for the Leader Member Exchange approach]. *Zeitschrift für Differentielle und Diagnostische Psychologie, 23*, 235–245.

StataCorp (2017). *Stata Statistical Software: Release 15.* College Station, TX: StataCorp LLC.

Staufenbiel, T., & Hartz, C. (2000). Organizational citizenship behavior: Entwicklung und erste Validierung eines Meßinstruments [Organizational citizenship behavior: Development and first validation of a measurement tool]. *Diagnostica, 46*, 73–83.

Streiner, D. L., & Cairney, J. (2007). What's under the ROC? An introduction to receiver operating characteristics curves. *Canadian Journal of Psychiatry, 52*, 121–128.

Swiss Armed Forces (2012). *Qualifikations- und Mutationswesen in der Armee [Qualifications and redeployments in the Swiss Armed Forces].* Bern, Switzerland: BBL.

Van Knippenberg, D., & Sitkin, S. B. (2013). A critical assessment of charismatic-transformational leadership research: Back to the drawing board? *Academy of Management Annals, 7*, 1–60.

Ward, M. K., & Meade, A. W. (2018). Applying social psychology to prevent careless responding during online surveys. *Applied Psychology, 67*, 231–263.

Weesie, J. (1999). Seemingly unrelated estimation and the cluster-adjusted sandwich estimator. *Stata Technical Bulletin, 52*, 34–47.

Williams, L. J., & McGonagle, A. K. (2016). Four research designs and a comprehensive analysis strategy for investigating common method variance with self-report measures using latent variables. *Journal of Business and Psychology, 31*, 339–359.

Woehr, D. J., Loignon, A. C., Schmidt, P. B., Loughry, M. L., & Ohland, M. W. (2015). Justifying aggregation with consensus-based constructs: A review and examination of cutoff values for common aggregation indices. *Organizational Research Methods, 18*, 704–737.

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment, 28*, 189–194.