Behavior Research Methods (2020) 52:2489–2505
https://doi.org/10.3758/s13428-020-01401-8

A little garbage in, lots of garbage out: Assessing the impact of careless responding in personality survey data

Víctor B. Arias[1] · L. E. Garrido[2] · C. Jenaro[1] · A. Martínez-Molina[3] · B. Arias[4]

Published online: 27 May 2020

Abstract
In self-report surveys, it is common that some individuals do not pay enough attention and effort to give valid responses. Our aim was to investigate the extent to which careless and insufficient effort responding contributes to the biasing of data. We performed analyses of dimensionality, internal structure, and data reliability of four personality scales (extroversion, conscientiousness, stability, and dispositional optimism) in two independent samples. In order to identify careless/insufficient effort (C/IE) respondents, we used a factor mixture model (FMM) designed to detect inconsistencies of response to items with different semantic polarity. The FMM identified between 4.4% and 10% of C/IE cases, depending on the scale and the sample examined. In the complete samples, all the theoretical models obtained an unacceptable fit, forcing the rejection of the starting hypothesis and making additional wording factors necessary. In the clean samples, all the theoretical models fitted satisfactorily, and the wording factors practically disappeared. Trait estimates in the clean samples were between 4.5% and 11.8% more accurate than in the complete samples. These results show that a limited amount of C/IE data can lead to a drastic deterioration in the fit of the theoretical model, produce large amounts of spurious variance, raise serious doubts about the dimensionality and internal structure of the data, and reduce the reliability with which the trait scores of all surveyed are estimated. Identifying and filtering C/IE responses is necessary to ensure the validity of research results.

Keywords Careless responding · Insufficient effort responding · Data cleaning · Invalid response · Factor mixture modelling

Introduction

Self-report survey data is one of the most widely used information sources in psychology research. However, that some surveyed do not pay enough attention and give enough effort to provide thoughtful and accurate responses is common (Huang, Curran, Keeney, Poposki, & DeShon, 2012; Meade & Craig, 2012), thus producing invalid response vectors with the potential to alter the properties of the entire dataset (Maniaci & Rogge, 2014). If this goes unattended, the effect of careless data can have very undesirable consequences on the interpretation and replication of research results (Curran, 2016).

Imagine the following situation: A researcher constructs a scale to measure a theoretically one-dimensional construct. This scale is balanced and consists of five direct items, and five reverse-keyed items. The researcher collects data from an online sample—say, Mechanical Turk or a similar platform—and performs the usual factorial analyses. Unfortunately, the analyses result in a very poor fit of the one-dimensional model, and favor a two-dimensional model (one relating to direct items and another to reverse-keyed items). Now let's suppose that, in reality, (a) the misfit of the one-dimensional model was caused by a set of spurious correlations resulting from the inconsistent responses of a small percentage of careless respondents, and therefore, (b) the one-dimensional model is fundamentally correct despite its poor statistical fit. This situation is plausible given that even 5% of careless respondents can cause spurious relationships between variables that would not otherwise be correlated (Huang, Liu, & Bowling, 2015), and a low percentage of unexpected response patterns can prevent the one-dimensional model from achieving an acceptable fit (Reise, Kim, Mansolf, & Widaman, 2016). If the

✉ Víctor B. Arias
vbarias@usal.es

[1] Faculty of Psychology, University of Salamanca, Avda. de la Merced, 109-131, 37005 Salamanca, Spain

[2] Faculty of Psychology, Pontificia Universidad Católica Madre y Maestra, Santiago De Los Caballeros, Dominican Republic

[3] Faculty of Psychology, Autonomous University of Madrid, Madrid, Spain

[4] Faculty of Education, University of Valladolid, Valladolid, Spain

researcher is not aware of this possibility and does not correct for the adverse effect of careless responding, he may make wrong decisions. At best, faced with the impossibility of explaining his results, he will abandon his research convinced that no reputable journal will publish a scale with such poor fit indices. At worst, the researcher will reject the one-dimensional hypothesis (type I error), and will modify his theoretical model to adapt it to a well-fitted but deeply erroneous multidimensional psychometric model. If the researcher manages to publish these spurious results as meaningful, we have a good example of what Huang et al. (2015) called "insidious confound in survey data".

Several studies revealed that even a low proportion of careless respondents can produce substantial alterations in the correlation between variables, statistical power, data dimensionality, and the size of effects (Baumgartner & Steenkamp, 2001; Woods, 2006; Huang et al., 2012; Rammstedt, Kemper, & Borg, 2013; Maniaci & Rogge, 2014; Wood, Harms, Lowman, & DeSimone, 2017; DeSimone & Harms, 2018; DeSimone et al., 2018). The purpose of this study was to extend previous research on the effect of careless responses on data properties, deepening aspects related to validity, reliability, and interpretation of research results on the measurement of psychological variables. To this end, we investigated the impact of careless/insufficient effort responding on the structure and dimensionality of survey data, the interpretability of measurement models, and the reliability of trait estimates. To detect careless respondents, we used a hybrid between factorial analysis and latent class analysis (factor mixture model). We performed the analyses on four personality scales that were applied to two independent samples under two different testing conditions.

Definition and characteristics of careless/insufficient effort responding

Initially, *random responding* was the most commonly used term to refer to response patterns resulting from inattention or neglect (e.g., Eden & Leviatan, 1975). However, humans are not naturally capable of generating random numbers (Neuringer, 1986), and inattentive responses tend to exhibit different degrees of systematicity, even if people are instructed to respond randomly (Huang et al., 2012). Subsequently, the phenomenon acquired other names, such as content-independent responding, content-nonresponsivity, inconsistent responding, and careless responding (Meade & Craig, 2012). Given the objectives of our study, we have used the term careless/insufficient effort (C/IE) (Curran, 2016) to refer to response vectors resulting from lack of attention or effort, where the individual responds without sufficient attention to the content and semantic polarity of the items. Thus, although a C/IE response does not necessarily imply a deliberate attempt at manipulation, it is not related to either the content of the items or the trait or state to be measured.

One of the most prevalent forms of C/IE responding is straightlining (SL), where the person provides similar responses regardless of the content and direction of the item (DeSimone, DeSimone, Harms, & Wood, 2018). The intensity of SL can vary, from individuals who give exactly the same response to all items to less obvious patterns, where long strings of invariant response are not observed but responses are concentrated on the positive or negative side of the response scale, regardless of the direction of the item (Dunn, Heggestad, Shanock, & Theilgard, 2018). Another form of C/IE response is random responding (RR), where the person does not attend to the content of the item, but intentionally uses all response categories to appear to respond thoughtfully (DeSimone & Harms, 2018). The SL modality has been observed to have the most pronounced impact on data properties (DeSimone et al., 2018).

Adverse effects of careless/insufficient effort responding

The prevalence of C/IE cases varies widely depending on the study, the sample, and the method for detecting them, with some consensus around 8–12% (Curran, 2016). However, even a low proportion of C/IE vectors can produce significant alterations in data quality, causing spurious relationships between non-correlated variables (Huang et al., 2015), inflation/deflation of internal consistency and one-dimensionality (Wood et al., 2017), unacceptable fit in one-dimensional models and the appearance of method factors in balanced scales (Kam & Meyer, 2015), biasing of experimental manipulation and meaningful relations between variables effects (Maniaci & Rogge, 2014), and alterations in the factor structure of the data (Johnson, 2005).

The studies cited agree that the presence of a relatively low proportion of C/IE data may bias the results towards substantial inflation or deflation of effects. C/IE responses are a common phenomenon in survey data (Curran, 2016) that can give rise to two potentially serious undesirable consequences. First, much of the research in psychology is based on inspecting the fit of the theoretical model to the data, and a favorable fit may support a theory, while a poor fit will lead us to question or even reject the theoretical model (Lai & Green, 2016). The problem occurs when the misfit comes not from errors in the theory or mis-specifications of the statistical model, but from the presence of a limited amount of very poor-quality data. If a researcher is not aware of the potential effect of C/IE data, he may fall into the error of rejecting a correct hypothesis or even accepting spurious results as meaningful. Second, some studies have shown that C/IE data can both mask meaningful effects and produce fictitious effects (Huang et al., 2015; Maniaci & Rogge, 2014). This can lead to serious problems

for replicating results, since different studies may find or deny the same effect only because their samples contain different proportions of C/IE data (Curran, 2016).

The present study

Most of the research on the impact of C/IE vectors on data quality has focused on (a) estimates of reliability based on internal consistency (Cronbach alpha), (b) the magnitude of Pearson correlations between items, (c) the dimensionality of data according to the size of eigenvalues in exploratory factor analysis or principal component analysis (PCA), and (c) statistical tests on the relationships between observable variables. The objective of this study is to deepen and extend research on the impact of C/IE respondents on key aspects of measurement validity related to the size of the systematic error variance and its effect on the fit of the theoretical model, the interpretation of the results, and the impact of C/IE responses on the reliability of trait estimates.

Fit and interpretability of psychometric models Many of the constructs investigated in psychology are represented as bipolar dimensions (e.g., extroversion-introversion). In practice, this bipolarity is transferred to the test by means of positive and negative items on more or less balanced scales. However, balanced scales usually produce wording factors in factor analysis (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003; Podsakoff, MacKenzie, & Podsakoff, 2012; Weijters et al., 2013), often leading to poor fit of the data to the theoretical model, and discrepancies between authors about the dimensionality of the construct and the nature—spurious or substantive—of the additional factors. Without a clear understanding of the effects of C/IE responding, unraveling these discrepancies may require much time and research effort (Curran, 2016). In research on psychological measurement, it is common for us to evaluate the validity of our substantive hypotheses by fitting the psychometric model to the data (Markus & Borsboom, 2013). The usual question is "how well does our model fit the data obtained from this sample (population)?" This question assumes that all respondents belong to a homogeneous population for which there is only one correct model (Reise et al., 2016). However, if there are a certain number of C/IE respondents in the sample, the question ceases to be useful, since our theoretical model will hardly be able to explain responses that have little or nothing to do with the variable we are trying to measure. In this case, a more useful question might be "what proportion of people have given answers that can be explained by our model?" It is therefore important to thoroughly investigate to what extent the presence of C/IE responses contributes to the misfit of the theoretical model, the appearance of spurious factors, and the distortion of the actual structure of the data.

In this study we have focused on assessing the impact of C/IE responses on the structure of personality scales composed of positive and negative items, with the following research questions:

Question 1
To what extent do C/IE responses affect the fit and interpretability of the theoretical model?
Question 2
To what extent do C/IE responses contribute to the emergence and size of spurious factors?

Data reliability In this study, we investigated the impact of C/IE responses on reliability understood as the opposite of the error with which the subjects' actual scores are estimated in the assessed trait (Hambleton, Swaminathan, & Rogers, 1991; Thissen, 2000). From this perspective, reliability is measurement precision, that is, the difference between the observed score and the value of the trait/state predicted by a well-fitted measurement model. The greater the spread between the estimated scores and the values predicted by the model, the lower the accuracy of the estimates and the less useful the information we can obtain from the test scores. Thus, we can calculate the overall estimation accuracy by averaging the residuals between the estimated and predicted scores (Embretson & Reise, 2013). The presence of C/IE responses can alter the magnitude of the measurement error in two ways. First, a C/IE vector is by definition invalid, so the model will have more difficulty estimating the latent score associated with that vector. This will result in estimates with a lot of error, thus altering our perception of the overall reliability of the data (given that these biased estimates could act as outliers in the calculation of the mean of measurement errors). Second, as noted above, one of the usual effects of the C/IE response is the appearance of spurious factors. Fitting a one-dimensional model under a strong violation of conditional independence implies that the probabilities of the response pattern are reproduced inappropriately (Embretson & Reise, 2013), producing, in turn, problems in the estimation of model parameters and latent scores (Zenisky, Hambleton, & Sireci, 2001). This implies that the C/IE vectors might not only lead to inaccurate estimates of the C/IE subjects themselves, but also alter the precision with which the latent scores of the thoughtful respondents are estimated. Given the importance of knowing the exact measurement error when evaluating and making decisions about test scores (Thissen, 2000), our goal is to know to what extent the presence of C/IE patterns can bias the accuracy with which the test estimates the latent trait in all test subjects. To this end, we will attempt to answer the following research question:

Question 3
To what extent does the presence of C/IE responses affect the estimation accuracy of individual scores of the trait?

In an attempt to answer the three research questions, we have analyzed the impact of C/IE responses under two testing conditions. To detect C/IE respondents, we designed a factor mixture model based on explicit predictions about the properties of the data in thoughtful and careless samples.

Method

Participants

We used two samples recruited through Prolific Academic, an online tool specialized in data collection for social and behavioral science research (cf. Palan & Schitter, 2018). Responses were completely anonymous, and all participants gave expressed consent for their responses to be used in research.

Sample 1 consisted of 725 participants (61% male) aged 18 to 75 years (M = 34.7, Mdn = 32, SD = 11.7). All participants were U.S. citizens, and English was their native language. Regarding the maximum educational level reached, 4.4% reported no formal qualifications, 16.2% finished secondary school, 35.8% had an undergraduate degree, 24.5% completed college/A levels, 23.7% graduate degree, and 6.3% doctorate degree. Each participant was compensated $1.50 USD (US dollars).

Sample 2 consisted of 405 participants (52% male) aged 18–72 years (M = 34.2, Mdn = 31, SD = 12.7). All participants were U.S. nationals and had English as their first language. Regarding the maximum educational level reached, 3.4% had no formal qualifications, 15% completed secondary school, 37.5% undergraduate degree, 29.8% completed college/A levels, 10.6% graduate degree, and 3.2% doctorate degree. Each participant was compensated $3.00 USD.

Variables and instruments

Extroversion, emotional stability, and conscientiousness The instrument was composed of 18 pairs of adjectives (36 items) from the 100 unipolar markers of the Big Five (Goldberg, 1992): six pairs relating to extroversion, six pairs to conscientiousness, and six pairs to emotional stability. Each item evaluates an aspect of the trait in positive (e.g., "Bold") or negative (e.g., "Timid") polarity. We used the instructions suggested by Goldberg (1992), asking participants to indicate how accurately each adjective described their general character on a five-point scale (very inaccurate, moderately inaccurate, neither accurate nor inaccurate, moderately accurate, and very accurate). We used Goldberg's markers for three reasons: first, that the items are semantic antonyms facilitates the task of evaluating the degree of incoherence of responses in suspicious C/IE cases; second, we needed simple and easily understood items to reduce misresponses due to low verbal ability (Johnson, 2005; Krosnick, 1999) and item ambiguity (Podsakoff et al., 2003), as much as possible, thus focusing the analysis on misresponses due to carelessness, inattention, and low effort; third, we needed a balanced scale with positive and negative items, given that straightliner responders are undetectable unless we have items with different semantic polarity (Reise et al., 2016).

Dispositional optimism We used the revised version of the Life Orientation Test Revised (LOT-R; Scheier et al., 1994). The LOT-R contains six items designed to measure generalized outcome expectancies (e.g., "In uncertain times, I usually expect the best"). Three items are reverse-keyed (e.g., "I rarely count on good things happening to me"). The respondent must indicate to what extent they agree with the item on a five-point scale (from strongly disagree to strongly agree). Theoretically, the construct measured by the LOT-R is one-dimensional and bipolar (optimism-pessimism). However, several studies have argued that pessimism and optimism are separable traits based on the best fit of the two-dimensional model and differences in correlations with external criteria (Creed, Patton, & Bartum, 2002; Herzberg, Glaesmer, & Hoyer, 2006; Marshall, Wortman, Kusulas, Hervig, & Vickers, 1992). However, it is not clear how a person can be simultaneously optimistic and pessimistic generalized across situations, so some studies have suggested that the multidimensionality of the LOT-R is due to method variance related to the wording of items (Kam & Meyer, 2015; Maydeu-Olivares & Coffman, 2006).

Tendency to blind acquiescence This instrument consisted of a subset of seven items from the Greenleaf scale (Greenleaf, 1992). From a broad set of indicators, Greenleaf selected those with inter-item correlations closest to zero. This scale is therefore not intended to measure anything, but was designed to quantify phenomena related to extreme responding. In our case, we used it to investigate the extent to which the C/IE subsample showed a generalized (dis)acquiescent response pattern. The selected items were: “I am a homebody”, “Advertising insults my intelligence”, “Investing in the stock market is too risky for most families”, “I like to feel attractive to members of the opposite sex”, “My days seem to follow a definite routine—eating meals at the same time each day, etc.”, “A college education is very important for success in today's world”, and “I will probably have more money to spend next year than I have now”. The instructions and response scale were the same as those used for the LOT-R.

Attention To evaluate the respondent's attention, items containing an explicit instruction with only one possible valid

answer are frequently used, e.g., “Please select moderately inaccurate for this item” (Huang et al., 2012). Those who give any response other than the one indicated are probably not paying attention to the content of the item. The mechanism underlying the response to this type of items is not yet clear, so a failure of the check must be interpreted with caution (Curran & Hauser, 2019). In our study, we included the following item: "For this statement, please do not check any response option" to avoid random hits. We tried to ensure that the attention check was placed between items of similar length, so that it would be difficult to identify the attention check with a quick glance.

Data collection procedure

All raw data used in the analysis are available on https://osf.io/n6krb . We collected the data in two phases. In the first phase, sample 1 responded to the 36 Big Five markers, the seven items on the Greenleaf scale, and the attention check. We divided the markers into two blocks of 18 items, so that no pair of antonyms would be shown simultaneously in the same block. Each participant responded first to a block of markers, then to the Greenleaf items (on a separate page) and finally to the second block of markers (third page). Half of the participants received block 1 first and the other half block 2. Within each block, each participant received the items in random order. The objective of completely randomizing the presentation of items and blocks was to avoid, as much as possible, spurious inflations/deflations of correlations caused by the mere physical proximity or remoteness of the items (Weijters, Geuens, & Schillewaert, 2009). The attention check was placed between two of the Greenleaf items of similar length, since placing it between the markers would have been too obvious.

In the second phase, sample 2 received the LOT-R within a longer testing session than the one described in phase 1. The complete battery consisted of 150 items. The LOT-R items were presented in blocks, the three positive items first followed by the three negative ones, in the final part of the battery (from item 136 to item 143). The seven Greenleaf items were presented immediately afterwards. The attention check was embedded between the LOT-R items. All subjects received the items in the same order and position within the battery.

Data analysis

Method for screening, cut-offs, and validity checks

For this study we designed a factor mixture model (FMM) based on predictions about C/IE responses on one-dimensional scales with positive and negative items. As discussed in the introduction, one of the effects of C/IE responses is the alteration of the correlation matrix between items and the appearance of additional factors not expected by the theoretical model. Next we will explain the characteristics and specification of the FMM, as well as the underlying rationale. The MPlus code used to estimate the models can be found at https://osf.io/n6krb.

An FMM is a hybrid model that combines latent class/profile analysis (LCA/LPA) and factorial analysis (cf. Clark et al., 2013; Lubke & Muthén, 2005). The LCA is a useful method for statistically identifying internally homogeneous groups from continuous or categorical multivariate data. The LCA uses probabilistic models of belonging to unobservable subgroups, unlike other methods based on the detection of clusters by means of theoretical or arbitrary distance measurements (Hagenaars & McCutcheon, 2002). Classes are categorical variables, so they have zero variance and do not allow intra-class variability (i.e., individual differences are completely explained by class membership). However, this assumption can sometimes be overly restrictive. For example, suppose we are investigating the distribution of a psychological disorder in the general population: we might expect the existence of classes (e.g., affected and unaffected), but also intra-class individual differences (e.g., differences in severity between those affected). To allow for this intra-class variability, the FMM uses a hybrid model of categorical (classes) and continuous (factors) variables. Thus, once individuals have been classified, the FMM allows for individual intra-class differences by estimating a factorial model for each class (Clark et al., 2013).

Our FMM model hypothesizes the existence of two classes. Class 1 is the majority and contains thoughtful respondents, whose responses are congruent with the content and direction of the items (we therefore assume that the application of the test has been done under normal conditions, where most of those examined respond thoughtfully). Class 2 is in the minority and groups the C/IE respondents, whose responses have little or no relation to the direction and content of the items. The thoughtful class and the C/IE class are therefore qualitatively distinct. However, each of these classes has variance: the subjects of the thoughtful class will surely have varying levels of the target trait. The C/IE class subjects will possibly present different degrees of carelessness, from extreme individuals with responses in an invariant straightline, to slighter cases with some incoherent responses resulting from attention fluctuations and sporadic errors in the selection of the response category (Meade & Craig, 2012). To account for this variability, we specify a factorial model for each class. Suppose we are assessing extroversion through a one-dimensional scale of six items (three positive and three negative), where the individual is asked to value his way of being in general, not in reference to specific situations or contexts. After recoding the negative items, a one-dimensional model estimated from thoughtful responses would come from a matrix

where all correlations between items are positive, regardless of the semantic direction of the items (see Fig. 1):

Now, a C/IE subject, particularly if they tend to straightline, will make consistency errors in their responses to positive and negative items (e.g., they could simultaneously affirm "I am the life of the party" and "At parties I prefer to go unnoticed"). Under these circumstances, in the matrix computed from a set of C/IE vectors, the sign of the correlations will be consistent with the semantic polarity of the items, *even though the negative items are recoded*. Consequently, this matrix will give rise to a model with negative loads on the negative items (we have called this factor "whatever" because, in theory, it has no substantive interpretation beyond being an inextricable amalgam of response styles; see Fig. 2). To detect the C/IE patterns, the model relies on response inconsistencies to positive and negative items. The model is therefore especially sensitive to straightliner response patterns, which are the ones that will produce the most inconsistent responses.

Based on these hypotheses, we specify the FMM as follows: (a) for the first class, all items have the same non-standardized positive load of 1 on the factor; (b) for the second class, all items have a non-standardized load of 1, except for inverse items, for which the load is negative (−1). The variance of the factor was freely estimated in both classes. Intercepts and correlations between residuals were set to equality between classes, in order to focus differences on the signs of factor loads. Factor means were set to zero in both classes to ensure model identification. The model assigns each subject a probability of belonging to class 1 or 2, depending on whether their response pattern is more or less compatible with one or another class. As a cut-off to flag a case as C/IE, we chose a probability greater than 50% of belonging to class 2.

There are other post hoc screening methods, such as Mahalanobis distance, psychometric antonyms/synonyms, even-odd index, inter-item standard deviation, l_z^p index, and Guttman errors (cf. Curran, 2016; DeSimone, Harms, & DeSimone, 2015; Meade & Craig, 2012; Niessen, Meijer, & Tendeiro, 2016). As far as we know, the FMM model proposed here has not been used to detect C/IE respondents.

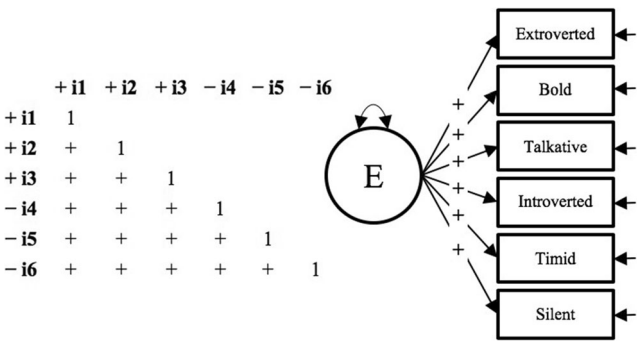

Fig. 1 Expected correlation matrix and factor loads of attentive class

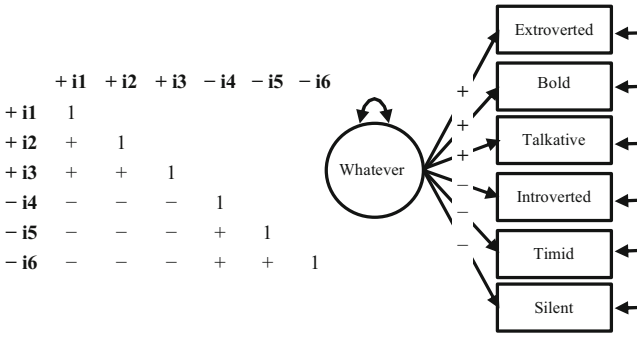

Fig. 2 Expected correlation matrix and factor loads of C/IE class

However, we have opted for FMM instead of other screening methods for four reasons:

(1) The FMM is based on explicit predictions of how data vectors from individuals who engage in C/IE responding take shape. For example, in the case of the Mahalanobis distance, it is not clear why a multivariate outlier should indicate carelessness and not an attentive but atypical response pattern in relation to the distribution of the rest of the data.

(2) The cut-off point of the FMM is purely empirical, and does not depend on a priori estimates of the number of C/IE cases present in the sample. Other screeners, such as psychometric antonyms/synonyms, even-odd index, or inter-item SD, do not have a precise mechanism to obtain the sample cut-off, so they are usually based on either universal cut-offs, whose generalizability has not yet been demonstrated, or on arbitrary cut-offs (e.g., flagging 5% of the respondents with more extreme scores in the screener, without knowing the true proportion of C/IE subjects in the sample). In the FMM, the cut-off point depends on the probability of each response vector belonging to the careless class. Thus, one could flag the vectors with a probability higher than 50%, or even higher (e.g., 75%) if a more conservative classification is desired. Although this decision can be partly arbitrary, the researcher can guide his or her decision with information such as the classification accuracy of the model (entropy values and probability distributions of each class).

(3) Most screeners are not based on modeled data, but work with raw scores. In contrast, latent variable models such as the FMM provide much more flexibility in dealing with measurement error, an advantage that often justifies the lower parsimony and greater difficulty of data analysis.

(4) There is evidence that some screeners, such as Mahalanobis distance and inter-item SD, could be confused with the substantive trait (Conijn, Franz, Emons, de Beurs, & Carlier, 2019), a situation that is clearly undesirable due to the increased probabilities of

selection bias and type I error in the identification of C/IE vectors.

Validity checks

Apart from the predictions made by the FMM model itself, external evidence is necessary to ensure that we are actually identifying C/IE respondents. We do this in the following ways:

(1) One of the assumptions of the FMM model is that the classes are qualitatively distinct, and therefore the C/IE class subjects' responses are independent of the content of the items. If this condition is not met, the probability of being assigned to the C/IE class will depend, in part, on the subjects' score on the scale. Eliminating these cases could lead to selection bias and, consequently, an artificial alteration of the assessment results (Conijn et al., 2019; Thomas & Clifford, 2017). To verify the independence between the sum score and the screener, we (a) correlated the probability assigned to each subject of belonging to class C/IE with the total score on the scale, and (b) compared the distributions of the sum scores between the complete sample and the sample of purged C/IE subjects. If the scores on the scale and on the screener are independent, then the correlation between the two will be close to zero, and there will be no substantial differences between the distributional properties of the complete and clean samples.

(2) In classifying a subject as C/IE, we are hypothesizing that they did not pay sufficient attention to item content. To verify that flagged individuals were more inattentive, we used an attention check item (see section measures). Although the performance of this type of attention check is not exact (Curran & Hauser, 2019), really careless subjects should be more likely to fail than subjects flagged as attentive.

(3) Those surveyed may vary in the intensity and pervasiveness with which they manifest C/IE responses, from individuals who respond with a consistent SL or RR pattern throughout the entire session, to individuals who respond inattentively sporadically (Meade & Craig, 2012). For the screener to be useful, it should be able to successfully identify at least the most pervasive C/IE responders across scales. To verify this, we used a set of items that does not attempt to measure anything except acquiescent/extreme response patterns (Greenleaf scale, see methods section). This set of theoretically uncorrelated items is designed to have no internal consistency (e.g., an expected Cronbach alpha of 0). However, we expect that at least the C/IE individuals with a more pervasive tendency for straightlining will contribute to an increase in the correlations between these items, given their inclination to respond systematically in the same region of the response scale. To test this, we examined the internal consistency (Cronbach's alpha) of the Greenleaf scale in the sub-samples of flagged subjects. If the screener has successfully captured the most pervasive straightliners, the consistency of the Greenleaf scale should be substantially greater in the C/IE sub-samples than in the attentive sub-samples.

Specification of the measurement models

We estimate three factorial models for each personality dimension: a one-dimensional model, a correlated traits-correlated methods minus 1 model (CT-C($M-1$)), and a restricted CT-C($M-1$) (see Fig. 3). As an example, Fig. 3a shows the one-dimensional extroversion model, consistent with the theoretical structure of the latent variable (Goldberg, 1992). We specified six correlations between the unicities of each pair of antonyms, given that semantic similarity between items usually produces residual systematic variance that is necessary to model for a correct estimation of the rest of the parameters (Saris, Satorra, & Van der Veld, 2009). Figure 3b shows a correlated traits-correlated methods minus 1 model (CT-C($M-1$); Eid, 2000). In this model, the general factor explains the common variance supposedly due to the substantive trait. The specific factor models the common variance associated with item wording, beyond and above the substantive factor. The specific factor can be understood as a method factor. Note that in this example, we are measuring a trait (extroversion) by means of two methods (positive items and negative items). When we measure the same trait with two or more methods, we expect the methods to achieve a high degree of convergence; however, a certain amount of discrepancy due to the specificities of each method is also to be expected. In the CT-C($M-1$) model, the extroversion factor contains the common variance that converges between the two methods. The method-specific factor is a residual factor that accounts for systematic covariation between negative items not explained by the trait factor. This residual factor could represent response styles, pure trait effects (e.g., "pure" introversion), or a mixture of both (Geiser, Eid, & Nussbeck, 2008). Finally, the restricted CT-C($M-1$) model represented in Fig. 3c (Geiser et al., 2008) is conceptually similar to the CT-C($M-1$) model. We are dealing with a structural equation model composed of two common factors, one measured by the positive items and the other by the negative ones. The factor of negative items is regressed into the positive factor. The latent regression (ß) is an estimator of the degree of convergence between estimates of the same trait made with different methods. The residual variance of the negative factor is conceptually similar to the residual factor of model 1b, and represents the discrepancy

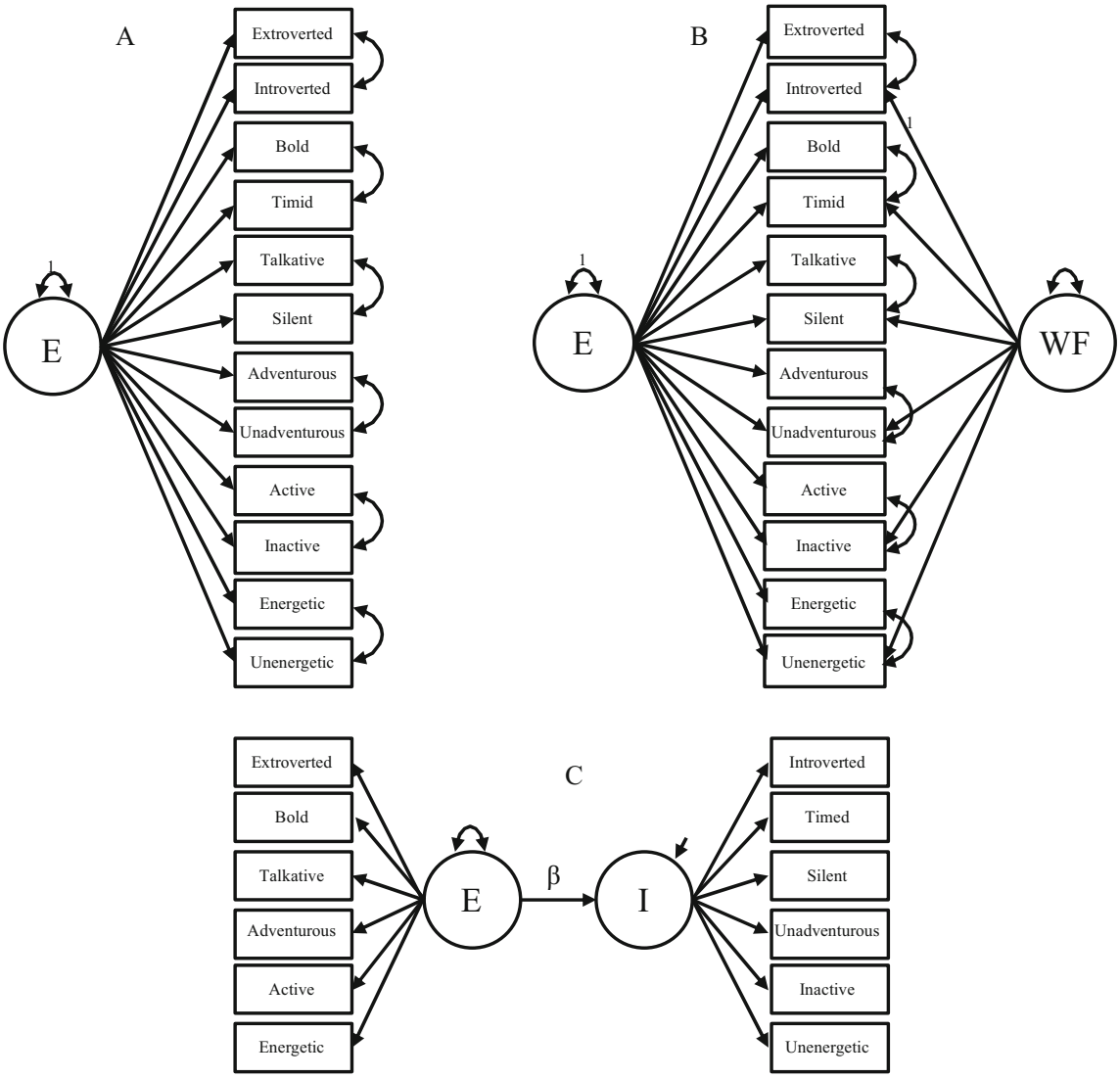

Fig. 3 **a**–**c** Conceptual representations of estimated confirmatory models. Note: E = extroversion; WF = wording factor; I = introversion. For clarity, (**c**) does not show the correlations between error terms

between the methods when estimating the same trait. If we assume that both subscales are measuring the same latent variable, a high degree of convergence between the factors is expected (i.e., a standardized ß close to 1). If ß is too low to guarantee the convergence of the factors, it could be that (a) these factors actually represent different substantive variables, or (b) an excess of method variance is deflating the empirical relationship between the factors.

Data quality indicators

In this phase we compared the results of the complete sample with those of each of the sub-samples resulting from the screening data. We used five quality indicators, which can be categorized into (a) indicators relating to the model fit (root mean square error of approximation, RMSEA and comparative fit index, CFI), and (b) indicators relating to the distribution of the common variance (explained common variance, hierarchical omega, and convergence between wording factors).

One-dimensional model fit We use two of the most common approximate fit indexes in factor analysis, RMSEA (Browne & Cudeck, 1992; Steiger & Lind, 1980, may) and the CFI (Bentler, 1990). Both indices summarize different aspects of the fit of the data to the model (Lai & Green, 2016; McNeish, An, & Hancock, 2018). Briefly, a lower RMSEA indicates greater proximity between the empirical correlation matrix and the matrix expected by the theoretical model, and a higher CFI indicates a more superior reproduction of the empirical matrix by the specified model with respect to the base model (which usually hypothesizes the absence of correlation between the observable variables). We expect that, in balanced scales, the C/IE responses (a) reduce the magnitude of the correlations between items, leading to a decrease in the CFI,

and (b) alter the matrix of correlations, making it more different from that expected by the model, leading to an increase in the RMSEA. Consequently, we expect that the model estimated on the complete sample presents a worse fit than the one estimated on the clean sample.

Explained common variance (ECV) and hierarchical omega (ωh) In a bifactor model such as CT-C(M−1) (Fig. 3b), the ECV (Brunner, Nagy, & Wilhelm, 2012) estimates the common variance explained by the substantive general factor versus that explained by the wording-specific factor, so the ECV can be interpreted as an index of one-dimensionality (Revelle & Wilt, 2013). Thus, as ECV approaches 1 (e.g., > .80), the loads of the general factor will be increasingly similar to those obtained by means of a one-dimensional model (Rodriguez, Reise, & Haviland, 2016). Conversely, low ECV values (e.g., < .70) indicate the presence of a substantial amount of multidimensionality in the data. The ωh (Zinbarg, Revelle, Yovel, & Li, 2005) is the ratio of reliable variance captured by the substantive factor once the effect of the wording factor has been partialed out. As ωh approaches 1, the trait factor will be the dominant source of variance in the responses to the items. Unlike ECV, the ωh calculation uses all sources of item variance (i.e., taking into account the error terms), so a high ωh indicates, in addition to unidimensionality, high measurement quality. Thus, ωh values higher than .80 suggest that the measure can be considered essentially one-dimensional and of sufficient quality for a correct interpretation of the scores in the general factor (Rodriguez et al., 2016).

β^2 As described in the section on measurement models, in the restricted CT-C(M−1) (Fig. 3c), β is the standardized weight of the regression of positive items factor in the negative items factor (consequently, β^2 is the variance explained by the first factor in the second). Assuming that both subscales measure the same latent variable, we expect β^2 values of 1. In practice, we will observe some degree of discrepancy due to method specificities (Geiser et al., 2008), or even possible differences between the psychological processes involved in the response to antonyms items (e.g., Kamoen, Holleman, van den Bergh, & Sanders, 2013). In any case, we estimate that a β of at least .90 (β^2 = .81 = 81% of explained variance) suggests sufficient convergence between subscales. If not, there is a problem with convergent validity which is more serious the further β^2 is from 1. The presence of C/IE vectors could affect the value of β^2: Inconsistent responses to positive and negative items can lead to a reduction in the correlations between items with respect to those expected by the one-dimensional model, and consequently to a deflation in the relation between subscales represented by β^2.

Fit and internal structure of the exploratory factorial model To evaluate the impact of C/IE responding on the internal structure of the multidimensional model, we estimated a three-factor exploratory structural equation model (ESEM; Asparouhov & Muthén, 2009) with oblique geomin rotation, using the 36 extroversion, emotional stability, and conscientiousness items. We estimated the correlations between the residuals of each pair of antonyms, as in one-dimensional models. In this case the ESEM results can be interpreted as those of a traditional exploratory factor analysis (EFA). We opted for ESEM instead of EFA given the ability of the former to easily accommodate additional specifications, such as correlated residuals. The model was estimated on both the complete sample and the clean sample.

Measurement accuracy To evaluate the impact of C/IE responding on data reliability, we calculated the standard error of measurement (SEM) associated with the estimated a posteriori score of each subject obtained from a graded response model for each scale (Samejima, 2016), first in the complete sample and then in each of the clean samples. To estimate the differences in precision, we calculated the mean of the differences in SEM between the estimates derived from the complete and clean samples. Data was analyzed with IRTPRO 4.0 (Cai, Du Toit & Thissen, 2011).

Results

The results are presented as follows: first, we review the results of FMM models and validity checks; second, we analyze the impact of C/IE responding on the fit, dimensionality, internal structure, and interpretability of one-dimensional and multidimensional models; and finally, we analyze the impact of C/IE responding on the accuracy of trait estimates.

Classification of C/IE responses and validity checks

The FMM models had no problems in achieving convergence and replicating best log-likelihood in 800 iterations. The proportion of respondents assigned to the C/IE class was 4.7% (conscientiousness), 6% (emotional stability), 7.3% (extroversion), and 10% (dispositional optimism). Entropy values were high (between .78 on the extroversion scale and .85 on the conscientiousness scale), suggesting that the model was able to classify response patterns with high precision. Consequently, we decided to retain the probability above 50% as a cut-off point for classifying cases. In sample 1, a total of 78 subjects (10.7%) were assigned to the C/IE class on at least one of the three scales (see Fig. 2). Of these, 21 (27%) were flagged on all scales, 11 (14%) on two scales, and 46 (58%) on one scale (Fig. 4).

Table 1 shows the results of the validity checks. The complete samples acquired Cronbach alpha values between .88 and .91; the elimination of C/IE vectors was a very small

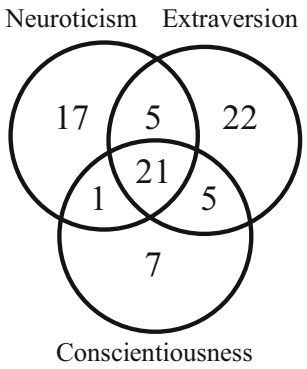

Fig. 4 Number of flagged respondents (sample 1)

improvement (between 0.01 and 0.02 points). However, all C/IE sub-samples obtained negative alpha values (from –.36 in responsibility to –.75 in optimism), indicating severely inconsistent response patterns. As expected, the Cronbach alpha of the Greenleaf scale was close to zero in the complete (0.13 and 0.04) and clean (between −0.05 and 0.04) samples. In contrast, the C/IE sub-samples obtained substantially higher values (between .38 and .57), suggesting a moderate overall tendency to respond consistently on one side of the response scale.

Regarding attention check items, in sample 1, the failure rate was 5.10% and from 2.3% to 2.7% in the clean samples. In the C/IE sub-samples, the failure rate was substantially higher (between 40% and 53%). In sample 2, the pattern of results was similar with 2.2% of failures in the clean sample and 16% in the C/IE sub-sample, although the failure rate in the C/IE sub-sample was lower than expected given the greater length of the testing session and the location of the attention check.

The means and standard deviations calculated from the sum scores were practically identical between the complete and clean samples, so the screening did not produce relevant alterations in the distribution of sum scores. Finally, the correlations between trait and the probability of being assigned to the C/IE class were zero (extroversion and optimism) or low (−0.12 in emotional stability and −0.23 in conscientiousness).

The results described in this section suggest that cases classified as C/IE (a) were extremely inconsistent in their responses, (b) were moderately consistent in a set of items designed to be uncorrelated, suggesting a generalized tendency to indiscriminate preference for one side of the response scale, (c) showed higher prevalence of attention failure than in the non-C/IE sample, and (d) did not relate to target trait scores or produce alterations in data distribution. In summary, there is reasonable evidence that the flagged responses were the result of carelessness and did not represent valid but atypical response patterns.

Finally, Table 2 shows an example of response patterns classified according to their probability of belonging to the C/IE class. Case 1 is an extreme straightliner, which is very damaging to the quality of data. On the contrary, case 2 has used all the points of the response scale, with preference for the left side. Although case 2 has given some apparently coherent answers (e.g., organized/disorganized), for the rest he either responded very inconsistently, or produced unexpected response patterns, such as declaring himself to be simultaneously very responsible and organized but having very little conscientiousness. This profile resembles a random pattern, although the tendency to prefer one side of the scale suggests some systematicity in the responses. Cases 3 and 4 are difficult to classify as SL or RR. There is a general preference for the

Table 1 Validity checks for C/IE classes

Sample	Scale	n	Cronbach α	α-G	ACI failed	Mean (SD)	rt-s
Full	Ext	725	.88	.13	5.1 %	36.2 (9.8)	.03
	Con	725	.89	.13	5.1 %	46.4 (8.4)	–.20
	Sta	725	.88	.13	5.1 %	41.5 (9.3)	–.10
	DOpt	405	.88	.03	3.7 %	18.3 (5.2)	.04
Clean	Ext	672	.90	.04	2.3 %	36.2 (10)	
	Con	691	.90	.04	2.7 %	46.8 (8.4)	
	Sta	681	.90	.03	2.6 %	41.7 (9.5)	
	DOpt	364	.91	–.10	2.2 %	18.3 (5.5)	
C/IE	Ext	53 (7.3 %)	–.57	.51	40 %	36.7 (6.4)	
	Con	34 (4.7 %)	–.36	.57	53 %	38.5 (3.5)	
	Sta	44 (6.0 %)	–.52	.56	43 %	37.8 (3.5)	
	DOpt	41 (10 %)	–.75	.41	16 %	18.0 (2.0)	

Note: Ext = extroversion; Con = conscientiousness; Sta = emotional stability; DOpt = dispositional optimism; Ext (IPIP) = extroversion factor from IPIP markers scale; α-G = Cronbach alpha from Greenleaf scale; ACI = attention check item; ETTSD = expected score standardized difference; rt-s = Pearson correlation between trait sum score and probability of belonging to the C/IE class

Table 2 Examples of non-recoded responses of cases flagged as C/IE

Item	Case 1 (p = 1)	Case 2 (p = .90)	Case 3 (p = .70)	Case 4 (p = .52)
Organized	+	+ +	+	+
Disorganized	+	− −	−	+
Responsible	+ +	+	+	+ +
Irresponsible	+	− −	+ +	− −
Conscientious	+	− −	0	+
Negligent	+ +	− −	−	− −
Practical	+	0	+	+ +
Impractical	+ +	−	+ +	+
Thorough	+ +	0	+	+
Careless	+	− −	− −	+
Hardworking	+ +	− −	+ +	+ +
Lazy	+	−	+	− −

Note: p = probability of belonging to the C/IE class; + + = very accurate; + = accurate; 0 = neither; − = inaccurate; − − = very inaccurate. Reverse items are in bold

right side of the response scale, and approximately half of their responses to antonym pairs were very inconsistent. These profiles could be compatible with attention fluctuations and errors in the selection of the response category.

Impact on the fit and structure of the one-dimensional model

Table 3 shows the results of the factorial analysis. Since the results were very similar on the three scales of the Big Five and the LOT-R, we will discuss only the results of the conscientiousness scale in detail. In the complete sample, the one-dimensional model obtained a rather poor fit (RMSEA = .108, CFI = .84). The fit improved drastically with the inclusion of the wording factor in the CT-C($M-1$) model (RMSEA = .053, CFI = .97). In this model, 73% of the common variance was explained by the substantive factor (27% by the method factor), and 76% of the reliable variance in scores was attributable to the trait (ωh = .76). An ECV value of .73 implies a substantial deterioration in the unidimensionality caused by the method factor. In the restricted CT-C($M-1$) model, the impact of the method on the common variance was even more evident: With a β value of .75, the positive factor explained only 56% of the variance of the negative factor. Since both factors supposedly measure the same trait, this result reveals a very serious problem of convergent validity.

Table 3 Model fit and variance explained by trait and method factors

		Model						
		Unidimensional		CT-C(M−1)				Restricted CT-C(M−1)
Scale	Sample	RMSEA	CFI	RMSEA	CFI	ωh	ECV	β (β^2)
Extroversion	Full	.132	.80	.081	.94	.77	.78	.75 (.56)
	Clean	.077	.95	.082	.95	.88	.95	.98 (.96)
Conscientiousness	Full	.108	.84	.053	.97	.76	.73	.75 (.56)
	Clean	.057	.96	.064	.96	.88	.94	.97 (.94)
Stability	Full	.111	.85	.048	.97	.77	.76	.73 (.53)
	Clean	.059	.96	.062	.96	.87	.94	.94 (.89)
Optimism	Full	.126	.93	.065	.99	.82	.75	.83 (.68)
	Clean	.054	.99	.083	.99	.91	.99	.98 (.96)

Note: CT-C(M−1) = correlated traits-correlated methods minus one model; RMSEA = root mean error of approximation; CFI = comparative fit index; ECV = explained common variance

In the clean sample, the one-dimensional model obtained a reasonably good fit (RMSEA = .057, CFI = .96), slightly lower than that obtained by the CT-C(M−1) model in the complete sample (RMSEA = .053, CFI = .97), and slightly higher than the CT-C(M−1) in the clean sample (RMSEA = .064, CFI = .96). The ECV was .94, indicating high one-dimensionality. ωh increased from .76 to .88, which implies that 12% of the variance that in the full sample model was in the method factor or in the error, became part of the trait factor in the clean sample, substantially improving the quality of measurement. The β value was .97, indicating an almost complete convergence between the positive and negative factors (94% of variance explained, compared to 56% in the complete sample). In summary, removing 4.7% of C/IE cases from the sample made the one-dimensional model fit better than the CT-C(M−1) model, and the method variance practically disappeared. This implies that after screening, (a) it was no longer necessary to specify a method factor to achieve a good fit, and (b) practically all of the method variance and one-dimensional model misfit were caused by the presence of only 4.7% of C/IE response vectors.

Let us consider two results relating to the discriminant validity of the factors and the size of the method variance in more detail. In the restricted CT-C(M−1) of the LOT-R, the mean of the factorial loads in the negative factor was .84, so the mean variance extracted by the factor (AVE; Fornell & Larcker, 1981) was $.84^2 = .70$. The variance shared between the positive and negative factors was $\beta^2 = .68$. Since the variance explained by the negative factor was greater than that shared with the positive factor, there is more evidence of discriminant than convergent validity. In other words, the model estimated using the entire sample is telling us that optimism and pessimism are empirically distinct variables. However, in the clean sample, the β^2 value of .96 and the AVE of the negative factor of .71 clearly tell us that optimism and pessimism are poles of the same dimension.

Secondly, Table 4 shows the factorial loads of the CT-C(M−1) for the extroversion scale (the detailed results for the rest of the scales are available from the first author, or can be calculated with the MPLUS code and raw data provided in the supplementary material). In the complete sample, the method factor acquired moderate loads that in half of the items surpassed the loads of the substantive factor, producing important validity problems in the negative items (given that those items were better explained by the method factor than by the trait factor). On the other hand, in the clean sample, the method factor loads were very low, the negative items loads in the trait factor recovered the lost variance until acquiring magnitudes similar to those of the positive items, and the method factor practically collapsed due to a lack of common variance (in fact, the loads in the method factor could be set to zero without detriment in the model fit).

Table 4 Factor loads from extroversion CT-C(M−1)

Full sample		Clean sample	
Extroversion	Method	Extroversion	Method
.75		.74	
.61		.59	
.68		.65	
.71		.70	
.68		.69	
.60		.57	
.59	.33	.69	.13
.36	.41	.52	.02
.47	.46	.60	.40
.51	.52	.69	.26
.56	.42	.73	.02
.36	.44	.53	.06

Note: all loads are standardized

Impact on the fit and structure of the multidimensional model

We estimated the ESEM models on the data from sample 1, specifying three factors measured by the 36 personality items (oblique geomin rotation). In order to form the clean sample, we followed a conservative criterion and eliminated only those respondents that were flagged in at least two of the personality scales (32 cases, 4.4% of the sample). The fit was poor for the complete sample (RMSEA = .071; CFI = .83), and substantially better for the clean sample (RMSEA = .048; CFI = .93). Table 5 shows the factorial loads of both models. For clarity, low factorial loads ($\lambda < .10$) are not shown. The complete sample model was unable to recover the theoretical structure of the data. The first factor captured the variance common to extroversion items. The second factor was clearly a wording factor, given that higher loads were concentrated exclusively on negative items. The third factor was an ill-defined factor of difficult or impossible interpretation, with inconsistent loads on five items of different scales. In the clean sample, the model was able to recover the theoretical structure reasonably well, with low cross-loads and without discriminant validity problems except in two pairs of antonyms in the emotional stability factor (stable/unstable and not envious/envious). This result suggests that a low proportion of C/IE cases (4.4%) was able to impede the recovery of the true structure of multidimensional data.

Impact on measurement accuracy

Figures 5 and 6 show the standard error of measurement of each respondent in the extroversion and dispositional

Table 5 Factorial loads from ESEM models

		Full sample (n = 725)			Clean sample (n = 693)		
		RMSEA = .071			RMSEA = .048		
		CFI = .830			CFI = .930		
Item		F1	F2	F3	F1	F2	F3
Extraverted		.93			.79	–.17	
Energetic		.53			.53		
Talkative		.85			.72		–.12
Bold		.64			.69		
Assertive		.71			.68		
Adventurous		.55		.42	.53		.16
Introverted		.84	.38		.71		
Unenergetic		.41	.60		.41	.26	
Silent		.78	.57	–.21	.66		
Timid		.75	.62		.66		
Unassertive		.74	.58		.68		
Unadventurous		.50	.51		.47		
Organized						.54	
Responsible						.75	
Conscientious						.48	
Practical						.53	
Thorough						.61	
Hardworking					.18	.62	
Disorganized			.70			.59	
Irresponsible			.84			.83	
Negligent			.79			.71	
Impractical			.72			.57	
Careless			.81			.76	
Lazy			.73		.22	.55	
Calm					–.16		.82
Relaxed							.85
At ease							.85
Not envious				.37		.15	.28
Stable				.70		.36	.49
Contented				.68	.17	.15	.41
Angry			.67			.22	.50
Tense			.55				.70
Nervous			.59		.26		.58
Envious			.60			.28	.27
Unstable			.74			.39	.48
Discontented			.61		.14	.26	.43
Factor correlations	F1		–.11	.39		.31	.42
	F2			.39			.42

Note: Loadings below .10 are not shown

optimism scales (the rest of the graphs can be obtained from the first author). The blue circles represent the estimates on the complete sample, the red circles the estimates on the clean sample, and the black circles the estimates on the C/IE cases. As expected, the C/IE cases contributed to reducing the reliability of trait estimates across the whole sample. A considerable shift is observed between the estimates on the complete and clean sample, with the estimates on the clean sample moving to zones of lower measurement error. On the extroversion scale, eliminating the C/IE cases led to an average reduction of 7.5% in measurement error for the entire sample, 11.8% on the LOT-R, 7.3% on the stability scale, and 4.5% on the conscientiousness scale. The gain in precision was strongly related to the number of C/IE response vectors eliminated (r = .97).

Discussion

The discussion is organized as follows: first, we will discuss the impact of C/IE responding on the fit and interpretation of the measurement model; second, the relationship between the C/IE data and the size and interpretation of the wording factors. Third, we will address the impact of C/IE responding on measurement accuracy.

Impact of careless responding on the fit of the measurement model

For all scales analyzed, the one-dimensional psychometric model derived from the theoretical model obtained an unacceptable fit. If we were only guided by statistical fit to make decisions, we would no doubt reject the one-dimensional hypothesis in favor of much better-fitting multidimensional solutions. However, after eliminating a relatively small number of C/IE cases (between 4.4% and 10%), the one-dimensional models fitted as well or better than multidimensional models, suggesting the retention of the more parsimonious model in all cases. Consequently, even a small number of C/IE cases may lead to a deterioration in the fit of such severity that it unequivocally suggests the incorrect rejection of the baseline hypothesis. C/IE vectors (especially straightliners) produced sets of inter-item correlations that cannot be explained by the one-dimensional model. Thus, the inability of the model to explain the C/IE patterns resulted in large specification errors and poor fit, requiring the inclusion of additional factors.

If we were in a situation where the one-dimensional model fits poorly and we were not aware that such a result is caused by C/IE responses, we would have at least two alternatives. The first is to fit a model with a wording factor, such as the CT-C($M-1$) used in this study. In this way we will surely improve the fit, but this procedure is not without problems: A wording factor often involves a severe violation of conditional independence that makes the interpretation of scores and the model itself much more complex, as well as the

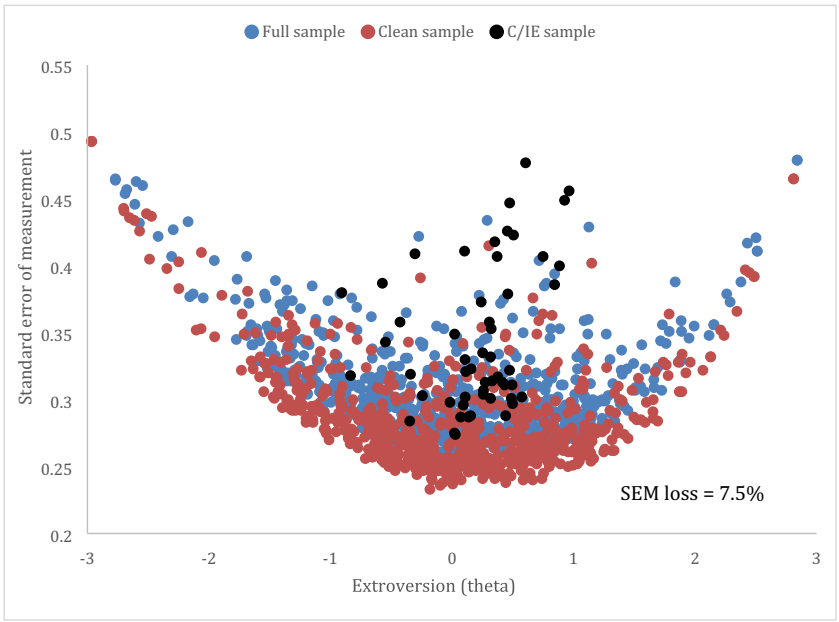

Fig. 5 Standard error of measurement (extroversion)

transfer of research results into practice (e.g., complicating the translation of the model into applied norms of correction and interpretation of test scores). The second alternative is to hypothesize that positive and negative item factors actually represent separable substantive dimensions, thus accepting a two-factor model. Let's take the case of the LOT-R as an example. Although the original model proposed dispositional optimism as a one-dimensional structure (Scheier et al., 1994), Marshall et al. (1992), in a classical study, proposed a two-dimensional model (optimism and pessimism), based on the best fit of the correlated two-factor model, and in which these factors presented different degrees of correlation with other personality variables. As we have seen, our results on the entire sample suggest retaining a two-dimensional model, supporting the proposal by Marshall and colleagues. However, the results changed drastically after cleaning the sample of C/IE cases: optimism and pessimism items reached an almost complete convergence, making the one-dimensional model the most plausible option (and supporting the original theoretical model proposed by Scheier et al.).

This leads us to suggest that, in the event of a poor fit of the theoretical model, it is necessary to investigate the reasons for the poor fit before opting for less parsimonious models. Otherwise, we could run the risk of rejecting a basically

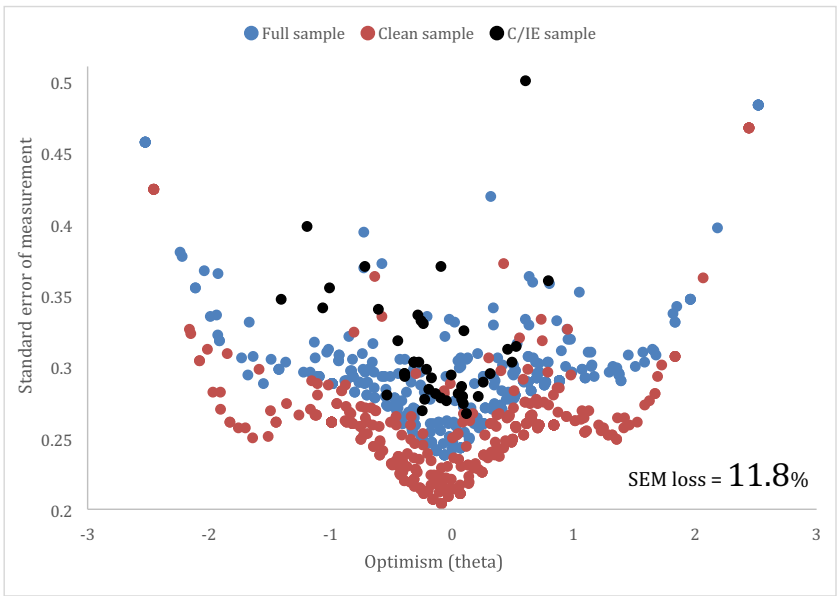

Fig. 6 Standard error of measurement (dispositional optimism)

correct model, retaining an incorrect model, and even justifying our decision by means of substantive interpretations of factors that are the product of data of dubious validity. We obtained similar results after estimating the multidimensional model. Just 4.4% of C/IE vectors can alter the internal structure of the data until it becomes unrecognizable, given the structure expected by the theoretical model. The effect of C/IE vectors in the multidimensional model was similar to that observed in the one-dimensional models, with the appearance of a strong wording factor that forces us to estimate an additional dimension if we wish to recover an interpretable structure. However, once the C/IE vectors were eliminated, the wording factor disappeared, the fit of the theoretical model improved considerably, and the internal structure of the data was clarified enough to facilitate its interpretation.

The impact of careless responding on the size and interpretation of method factors

In all the scales analyzed, the wording variance in the CT-C(M −1) models was high in the complete sample (between 22% and 27% of the common variance) but practically disappeared after eliminating the C/IE responses. This result has at least two interpretations. First, the main function of the wording factor was to accommodate a limited number of incoherent and extremely atypical response patterns given the expectations of the one-dimensional model. Second, on none of the scales did the wording factor have a reasonable substantive interpretation beyond being a useful mathematical mechanism for modeling a certain amount of systematic error. In other words, these factors were probably not measuring anything, nor reflecting any underlying trait or state. Let us take as an example the extroversion scale used in this study. The CT-C(M−1) model resulted in a wording factor that explained 22% of the common variance in the complete model, and 57% of the common variance in the subset of negative items. If we did not know that this factor was caused by 7.3% of persons with highly inconsistent responses, how should we interpret the factor? According and Geiser et al. (2008), a wording factor in a CT-C(M−1) model could be interpreted as a pure method factor, as a pure trait (i.e., "pure introversion"), or as a mixture of both. If we interpret it as a method factor, we might think that something very bad happens with reverse-keyed items, and reach the conclusion that it is advisable to dispense with them to avoid that amount of nuisance variance. If, on the other hand, we interpret the factor as "pure introversion", we'll probably be reifying a mathematical artifact and giving substantivity to a factor that is merely the product of a few incoherent responses. Of course, this does not negate the interest and need to investigate the nature of wording factors, given that method factors may contain substantive variance (Podsakoff et al., 2003), and that the tendency to insufficient effort responding may be related to enduring individual differences (Bowling et al., 2016). What we propose is that we first examine the extent a wording factor can be explained by C/IE cases, and then proceed to investigate those cases in which, after data cleansing, the wording factor retains sufficient specific variance to admit a possible substantive interpretation.

The impact of careless responding on measurement accuracy

In this study, we have investigated the impact of C/IE responses on reliability, understood as the opposite of the error with which we estimate the real scores of the subjects in the evaluated trait (Hambleton et al., 1991; Thissen, 2000). For this purpose, we estimate the standard errors of measurement for each individual before and after cleaning the sample using a graduated response model. As expected, not only was the trait estimated with the most error in the C/IE subjects, but the screening led to a substantial increase in the accuracy of the trait estimates in all subjects (with improvements of between 11% and 7%). This implies that C/IE responses may impair the reliability of estimates for thoughtful subjects to a non-ignorable degree, and that cleaning the sample may lead to an appreciable improvement in the reliability of the entire dataset.

This study has several limitations. First, we have only used personality-balanced measures. Thus, it is necessary to evaluate the impact of C/IE responding on other types of instruments, particularly clinical measures applied to community samples. Clinical measures often represent unipolar dimensions (or quasi-traits; Reise & Waller, 2009). This results in sets of items with a single semantic polarity that makes it impossible to detect C/IE respondents of the straightliner type. Thus, a relevant objective of future research is to assess the impact of C/IE responses in clinical evaluation, and to design appropriate methods for the detection of these vectors in this type of instruments. Second, the effects found come from a limited number of scales. Although, in theory, careless responding does not depend so much on the scale used as on respondent variables, it would be advisable to replicate these results in a variety of instruments, in order to warrant their generalizability. Third, the FMM model is designed to maximize its effectiveness in detecting straightliner response patterns; it would therefore be necessary to investigate the impact of other types of misresponding (e.g., random or middle responding) through appropriate methods.

Conclusions

The title of this article paraphrases an expression from computer science ("garbage in, garbage out") referring to the consequences of analyzing low-quality data. In our study, some garbage in (between 4.7% and 10% of low-quality response

vectors) has produced a lot of garbage out in the form of drastic alterations to the properties of the data. We can conclude that the presence of a small number of C/IE cases potentially has the following effects: it (a) leads to the incorrect rejection of a correct model because of a drastic deterioration in its statistical fit; (b) raises serious doubts about the dimensionality and internal structure of the data; (c) causes problems of convergent validity between sets of items that theoretically measure the same construct; (d) alters the structure of the data in multidimensional models; (e) substantially reduces the reliable variance explained by the trait factor; (f) produces large amounts of systematic error variance; (g) forces the researcher to fit models with method factors to account for this systematic variance, thereby introducing unnecessary complications in the estimation and interpretation of the measurement model and test scoring; and (h) reduces the reliability with which the model estimates the latent scores of all sample.

Although more research dedicated to improving our understanding of careless responding and refining methods to detect it is needed, researchers and practitioners have several ad hoc and post hoc tools for detecting at least the worst-quality response vectors (cf. Curran, 2016). We therefore encourage researchers to regularly incorporate response quality screening techniques into their studies, and editors and reviewers to require such assessments as part of data analysis.

References

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(3), 397-438.

Baumgartner, H., & Steenkamp, J. B. E. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 38*(2), 143-156.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238.

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*(2), 218.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods and Research, 21*, 230-258. doi: https://doi.org/10.1177/0049124192021002005

Brunner, M., Nagy, G., & Wilhelm, O. (2012). A tutorial on hierarchically structured constructs. *Journal of Personality, 80*(4), 796-846. doi: https://doi.org/10.1111/j.1467-6494.2011.00749.x

Cai, L., Du Toit, S. H. C., & Thissen, D. (2011). IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. *Chicago, IL: Scientific Software International.*

Clark, S. L., Muthén, B., Kaprio, J., D'Onofrio, B. M., Viken, R., & Rose, R. J. (2013). Models and strategies for factor mixture analysis: An example concerning the structure underlying psychological disorders. *Structural Equation Modeling: A Multidisciplinary Journal, 20*(4), 681-703.

Conijn, J. M., Franz, G., Emons, W. H., de Beurs, E., & Carlier, I. V. (2019). The Assessment and Impact of Careless Responding in Routine Outcome Monitoring within Mental Health Care. *Multivariate Behavioral Research, 54*(4), 293-611. doi:https://doi.org/10.1080/00273171.2018.1563520

Creed, P. A., Patton, W., & Bartum, D. (2002). Multidimensional properties of the LOT-R: Effects of optimism and pessimism on career and well-being related variables in adolescents. *Journal of Career Assessment, 10*, 42-61.

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

Curran, P. G., & Hauser, K. A. (2019). I'm paid biweekly, just not by leprechauns: Evaluating valid-but-incorrect response rates to attention check items. *Journal of Research in Personality, 82*, 103849.

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*(2), 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*(5), 559-577.

DeSimone, J. A., Harms, P. D., & DeSimone, A. J. (2015). Best practice recommendations for data screening. *Journal of Organizational Behavior, 36*(2), 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*(1), 105-121.

Eden, D., & Leviatan, U. (1975). Implicit leadership theory as a determinant of the factor structure underlying supervisory behavior scales. *Journal of Applied Psychology, 60*(6), 736-741.

Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika, 65*(2), 241-261.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory.*: Psychology Press.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*(1), 39-50.

Geiser, C., Eid, M., & Nussbeck, F. W. (2008). On the meaning of the latent variables in the CT-C (M-1) model: A comment on Maydeu-Olivares and Coffman (2006). *Psychological Methods, 13*(1), 49.

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment, 4*(1), 26.

Greenleaf, E. A. (1992). Measuring extreme response style. *Public Opinion Quarterly, 56*(3), 328-351.

Hagenaars, J. A., & McCutcheon, A. L. (2002). *Applied latent class analysis.*: Cambridge University Press.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Herzberg, P. Y., Glaesmer, H., & Hoyer, J. (2006). Separating optimism and pessimism: A robust psychometric analysis of the Revised Life Orientation Test (LOT-R). *Psychological Assessment, 18*(4), 433.

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*(1), 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*(3), 828.

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality, 39*(1), 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*(3), 512-541.

Kamoen, N., Holleman, B., van den Bergh, H., & Sanders, T. (2013). Positive, negative, and bipolar questions: The effect of question

polarity on ratings of text readability. *Survey Research Methods, 7*(3), 181-189.
Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology, 50*(1), 537-567.
Lai, K., & Green, S. B. (2016). The problem with having two watches: Assessment of fit when RMSEA and CFI disagree. *Multivariate Behavioral Research, 51*(2-3), 220-239.
Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods, 10*(1), 21.
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.
Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning.*: Routledge.
Marshall, G. N., Wortman, C. B., Kusulas, J. W., Hervig, L. K., & Vickers, R. R. J. (1992). Distinguishing optimism from pessimism: Relations to fundamental dimensions of mood and personality. *Journal of Personality and Social Psychology, 62*, 1067-1074.
Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods, 11*(4), 344.
McNeish, D., An, J., & Hancock, G. R. (2018). The thorny relation between measurement quality and fit index cutoffs in latent variable models. *Journal of Personality Assessment, 100*(1), 43-52.
Meade, A. W., & Craig, S. B. (2012). Identifying Careless Responses in Survey Data. *Psychological Methods, 17*(3), 437-455. doi:https://doi.org/10.1037/a0028085
Neuringer, A. (1986). Can people behave "randomly?" The role of feedback. *Journal of Experimental Psychology: General, 115*, 62-75.
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.
Palan, S., & Schitter, C. (2018). Prolific. ac-A subject pool for online experiments. *Journal of Behavioral and Experimental Finance, 17*, 22-27.
Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879.
Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology, 63*, 539-569.
Rammstedt, B., Kemper, C. J., & Borg, I. (2013). Correcting Big Five personality measurements for acquiescence: An 18-country cross-cultural study. *European Journal of Personality, 27*(1), 71-81.
Reise, S. P., Kim, D. S., Mansolf, M., & Widaman, K. F. (2016). Is the bifactor model a better model or is it just better at modeling implausible responses? Application of iteratively reweighted least squares to the Rosenberg Self-Esteem Scale. *Multivariate Behavioral Research, 51*(6), 818-838.
Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual review of clinical psychology, 5*, 27-48.
Revelle, W., & Wilt, J. (2013). The general factor of personality: A general critique. *Journal of Research in Personality, 47*(5), 493-504.
Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment, 98*(3), 223-237.
Samejima, F. (2016). Graded response models. In W. J. van der Linden (Ed.), *Handbook of Item Response Theory* (Vol. 1, pp. 123-136). London: Chapman and Hall/CRC.
Saris, W. E., Satorra, A., & Van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling: A Multidisciplinary Journal, 16*(4), 561-582.
Scheier, M. F., Carver, C. S., & Bridges, M. W. (1994). Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): A reevaluation of the Life Orientation Test. *Journal of Personality and Social Psychology, 67*, 1063-1078.
Steiger, J. H., & Lind, J. C. (1980, may). *Statistically based tests for the number of common factors.* Paper presented at the Annual Meeting of the Psychometric Society, Iowa City, IA.
Thissen, D. (2000). Reliability and measurement precision. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 159-183). Mahwah, NJ: Lawrence Erlbaum Associates.
Thomas, K. A., & Clifford, S. (2017). Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior, 77*, 184-197.
Weijters, B., Baumgartner, H., & Schillewaert, N. (2013). Reversed item bias: An integrative model. *Psychological Methods, 18*(3), 320.
Weijters, B., Geuens, M., & Schillewaert, N. (2009). The proximity effect: The role of inter-item distance on reverse-item bias. *International Journal of Research in Marketing, 26*(1), 2-12.
Wood, D., Harms, P. D., Lowman, G. H., & DeSimone, J. A. (2017). Response speed and response consistency as mutually validating indicators of data quality in online samples. *Social Psychological and Personality Science, 8*(4), 454-464.
Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment, 28*(3), 186.
Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2001). *Effects of Local Item Dependence on the Validity of IRT Item, Test, and Ability Statistics.* Amherst: University of Massachusetts.
Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's a, Revelle's b, and McDonald's wH: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*(1), 123-133.