Does the Over-Claiming Questionnaire measure over-claiming?

Absent convergent validity in a large community sample

Ludeke, Steven; Makransky, Guido

Published in:
Psychological Assessment

DOI:
10.1037/pas0000211

Publication date:
2016

Document version
Peer reviewed version

Document license
Unspecified

Citation for pulished version (APA):
Ludeke, S., & Makransky, G. (2016). Does the Over-Claiming Questionnaire measure over-claiming? Absent convergent validity in a large community sample. Psychological Assessment, 28(6), 765-774. DOI: 10.1037/pas0000211

Does the Over-Claiming Questionnaire measure over-claiming? Absent convergent validity in a large community sample

Steven G. Ludeke*a

Guido Makransky b

a Department of Political Science, University of Southern Denmark

b Department of Psychology, University of Southern Denmark

*Corresponding author:

E-mail: stevenludeke@gmail.com

Phone: +45 2031 5064

Note: This manuscript is the last version submitted prior to acceptance, and does not reflect any changes introduced in the course of the publication process.

Abstract: The Over-Claiming Questionnaire (OCQ) aims to provide a practical and cost-effective method of assessing individual differences in the tendency to misrepresent oneself in self-reports. OCQ bias measures have strong theoretical appeal but limited empirical demonstrations of validity. Using a sample of 704 adult community members, we found minimal support for the OCQ as an assessment of misrepresentation. We assessed misrepresentation by comparing self-reports of personality and cognitive ability against other criterion indicators of these trait levels (peer reports of personality and performance on a cognitive ability measure). OCQ bias measures bore no relationship with either of these self-criterion discrepancy measures, and were also unassociated with self-deceptive enhancement scores. One OCQ index bore a modest relationship to narcissism. OCQ bias measures were instead consistently and sometimes even highly related to measures of careless responding. However, statistically controlling for careless responding only minimally improved the convergent validity of OCQ bias indices.

Keywords: Over-Claiming Questionnaire; socially desirable responding; personality assessment; careless responding; narcissism

The utility of personality traits as predictors of a range of important life outcomes highlights the need for accurate and cost-effective assessment of these traits (Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007). Self-reports of personality provide obvious benefits with respect to practical and cost-effective administration, but carry equally obvious potential costs with respect to accuracy. Recognition of this danger has motivated a massive literature aimed at identifying and measuring the tendency to misrepresent oneself in self-reports.

Much of this literature focuses on *socially desirable responding* (SDR), in which individuals represent themselves as possessing traits viewed as desirable by society at large (Paulhus, 2002). Other research highlights how individuals are particularly prone to overstate their levels of traits that they themselves see as desirable (Ludeke, Weisberg, & DeYoung, 2013). These literatures can claim some success at identifying individuals most prone to misrepresent themselves in personality self-reports, as well as the traits that a particular person is most likely to misrepresent. However, both approaches have severely limited utility for the purposes of *improving* the predictions that can be made using personality self-reports. This is because misrepresentation in self-reports appears to largely be an exaggeration of true differences in personality: those providing socially desirable responses are often, in fact, in possession of greater-than-average levels of desirable traits, and the person who particularly values a given trait is not only likely to exaggerate that particular trait but also to have higher than average levels of that trait (Ludeke et al., 2013; Ones, Viswesvaran, & Reiss, 1996).

A promising alternative has been offered by Paulhus and colleagues (Paulhus, Harms, Bruce, & Lysy, 2003; Paulhus & Harms, 2004). Called the *Over-Claiming Technique* (OCT), this approach asks survey respondents to indicate their level of familiarity with a large number of items drawn from a given content domain (e.g. academic facts, or famous musicians), where

some items are real while others are bogus items (also termed “foils”). The tendency to claim familiarity with all items (especially foils) is interpreted as a tendency to exaggerate, whereas the tendency to accurately distinguish real items from foils is indicative of true knowledge.

This approach has great intuitive appeal. Claiming familiarity with nonexistent entities seems to be a face-valid indicator of misrepresentation. Further, it is potentially one which (unlike SDR measures) is not also elevated among those who truly have the characteristic being misrepresented (here, knowledge). Empirical demonstrations of the approach’s validity and utility have typically focused on the association between such “overclaiming” and scores on other measures that have some success at identifying those prone to misrepresentation. In particular, this has involved measures of narcissism (Gebauer, Sedikides, Verplanken, & Maio, 2012; Paulhus et al., 2003; Paulhus & Harms, 2004; Paulhus & Williams, 2002) and measures of socially desirable responding (Bing, Kluemper, & Davison, 2011; Paulhus et al., 2003; Tonković, Galić, & Jerneić, 2011). Within this latter category, it is primarily measures of “egoistic” and not “moralistic” responding which relate to overclaiming, indicating that overclaiming is best characterized as an exaggeration of one’s agentic and not communal traits (Bing et al., 2011; Paulhus et al., 2003; Tonković et al., 2011).

Studies such as these are important, but as demonstrations of the success of the OCT they are clearly incomplete. This is because measures of narcissism and socially desirable responding capture not only the tendency to overstate one’s positive characteristics but also to actually have elevated levels of such characteristics: for example, the high opinion narcissists have of their abilities is partially supported by their performance on objective performance tests (Paulhus & Williams, 2002). For this reason, Paulhus and colleagues (2003) suggest that demonstrations

using discrepancies between self-reports and other indicators of trait levels such as peer reports or ability measures represent the "gold standard" for identifying misrepresentation.

Studies comparing overclaiming as identified by the OCT with self-criterion discrepancy measures have been rare and inconsistent. We are aware of only two: in the first, Paulhus and colleagues (2003, study 1) report that in a group of 137 undergraduates, overclaiming as identified using the OCT was associated with claiming higher levels of ability than were justified by performance on an IQ test, and higher levels of agentic personality characteristics than were justified by reports from classmates who had spent several hours working with the person being rated. In the second, Musch and colleagues (2012) created self-criterion discrepancy measures based on the gap between high school academic performance and self-reported knowledge and self-reported ability in a sample of 109 Germans; these discrepancy measures were significantly associated with a measure of egoistic SDR, but they were not associated with overclaiming as assessed by the OCT. Absent further support for the OCT from self-criterion discrepancy measures, its validity as an assessment of misrepresentation must thus be considered tentative.

A recent study further challenged the utility of the OCT as a measure of misrepresentation: in a large sample (N=670), Barber and colleagues (2013) found that claiming familiarity with academic foils was strongly correlated with the tendency to respond carelessly during the survey. To highlight the plausibility and the importance of this result, we must consider the nature of OCT instruments and their response scales.

Applying the Over-Claiming Technique: The Over-Claiming Questionnaire

The Over-Claiming Questionnaire (OCQ) has appeared in many versions, differing in the content domains assessed (e.g. famous musicians, popular clothing stores, philosophical figures

and concepts), in the frequency of foils, in the number of items per scale, and in the presence or absence of instructions alerting participants to the presence of foils in the OCQ. Perhaps the most common format uses three foils and twelve real items in each content domain. Items are grouped with others in their domain, with a general descriptor ("Jazz artists") preceding the items. Participants are instructed to rate their familiarity with each item, with responses ranging from "never heard of it" to "very familiar." For the purposes of illustration, we will dichotomize the participants' responses, dividing between "never heard of it" and any level of familiarity (represented by the remaining choices). Depending on the type of item (real or foil) and the participant's response (familiar or unfamiliar), there are a total of four kinds of responses, as illustrated in Table 1. We illustrate these responses with answers provided by one respondent, who accurately recognized Beethoven but was unfamiliar with the composer Stravinsky, and who claimed familiarity with the made-up "Ochberg" but not with another foil, "Rentzig."

Interpreting data from the OCQ is complicated by the fact that each of these four categories can be interpreted in multiple ways. Depending on the context in which the OCQ is used, some of these interpretations are less plausible than others: for example, just as "faking bad" is relatively uncommon outside of very specific assessment contexts such as criminal or health settings, under-claiming (reflecting a desire to appear uninformed) is probably rare in the research and classroom settings in which the OCQ is typically used. Accordingly, even though under-claiming is a conceptually possible influence on both the False and True Negative rate, True Negatives are probably best interpreted as indicating the absence of overclaiming tendencies, and False Negatives are probably best interpreted as indicating both the absence of knowledge and of overclaiming tendencies.

An important recognition by Paulhus and colleagues (2003) was that previous research which used only False Positives as indicators of overclaiming ignored how True Positive responses could be indicative not only of knowledge but also of overclaiming. Importantly, this dual interpretation of True Positives is likely relevant for most applications of the OCQ.

We have not seen any previous discussion of the multiple interpretations of False Positives, but the above-mentioned study by Barber and colleagues (2013) highlights the potential role of careless responding in creating False Positives. This problem might be particularly acute in low-stakes research contexts, where the participants feel little incentive to pay attention. Conceptually, it is easy to see how careless responding could substantially affect the rate of False Positive responses. The OCQ typically allows one response option that is appropriate for foils ("Never heard of it," scored as a 1), and four or more response options which indicate ever-increasing levels of familiarity with the item. When using a five-point response scale, completely inattentive responding might be expected to produce an average of 3 (the scale mid-point) for bogus items. In contrast, participants responding attentively and honestly (i.e. without carelessness or overclaiming) will score an average of 1 for bogus items, or perhaps modestly higher if the version of the OCQ employed contains several "attractive distractors." Because of this, it may be that even an aggressive but attentive over-claimer will obtain a lower score on the False Positive rate than will a careless responder. Meta-analyses of faking behavior highlight the relatively conservative nature of faking, showing that "fake good" instructions typically elevate scores by about half a standard deviation (Viswesvaran & Ones, 1999; a comparable elevation on OCQ foils was reported by Bing et al., 2011); if the magnitude of the gap between "fake good" and "honesty" conditions roughly parallels the gap between

aggressive over-claimers and honest responders, then the highest False Positive scores may very well belong to those not paying attention rather than those deliberately exaggerating.

Reconsidering the Over-Claiming Technique

The literature reviewed above highlights the need for further study of the validity of "overclaiming" scores obtained using the OCT. The few studies which have examined the question using the "gold standard" of self-criterion discrepancies have produced inconsistent results. Further, there are conceptual and empirical reasons to question whether the OCQ bias indices primarily represent overclaiming or whether they instead capture careless responding. In the present study, we make secondary use of data from a large, well-studied sample (the Eugene-Springfield Community Sample: Goldberg et al., 2006; Goldberg, 1999) in order to test the performance of the OCQ as a predictor of bias. Results from the OCQ in this sample have not previously been published, but previous research on the sample has demonstrated the validity of multiple measures of self-criterion discrepancies (Ludeke, Reifen Tagar, & DeYoung, 2015; Ludeke et al., 2013), suggesting that it represents a useful opportunity to test the merits of the OCQ.

We also evaluate several possible improvements to the OCQ. First, because research has demonstrated a possible influence of careless responding on the OCQ bias indices, we will explore whether the predictive power of the OCQ indices can be improved by controlling for careless responding. Second, we will evaluate whether the OCQ indices can be improved by differently weighting the hit rate and false alarm rate when computing bias indices. Finally, we will consider the impact of differences in the status which attaches to knowledge in different domains. If knowledge on some topics is seen as particularly status-enhancing, overclaiming

behavior on these “high-prestige” topics may connect more clearly (when compared to overclaiming behavior on less prestigious topics) to other indicators of self-enhancement.

Method

Sample

The present study relies on the Eugene-Springfield Community Sample, which is predominantly middle-aged (Mean = 49.87, *SD* = 12.24), female (56.6%), and Caucasian (98.6%), with a median of two years of post-secondary education. The participants were drawn from a list of homeowners in the Eugene-Springfield area of Oregon, and completed surveys by mail between 1994 and 2008 in exchange for money. All participants who completed more than 90% of the Over-Claiming Questionnaire (OCQ) items (N = 704) were included in the present study.

Measures

Over-Claiming Questionnaire. In 2003, participants completed a survey which included a version of the OCQ. The 105 items in this version were evenly distributed among seven categories: musical artists from five different genres (Jazz, Country, Classical, Rap/Hip-Hop, and Rock), clothing designers, and clothing stores. Each category included fifteen items, including three foils per scale. We excluded from analysis one item from the Rock scale which included a typo (“Paint the Earth” was presented as “Paint of the Earth”).

Participants were instructed “Please rate your familiarity with each item by filling in the appropriate number from 1 to 5.” Scores of 1 were indicated for “never heard of it,” a 3 was indicated as “somewhat familiar,” and a 5 was indicated for “very familiar.” Participants were then provided with two examples:

"1. If you're asked about POLITICIANS and the item said "Bill Clinton," you would probably mark '4' or '5' to indicate that you are familiar with him. 2. If the category was FAMOUS ATHLETES and the item said "Fred Gruneburg," you would probably mark '1' because you have never heard of him or a '2' because he sounds vaguely familiar."

The items were then presented, with the items of each category presented together under a heading for that category (e.g. "Clothing Stores").

Big Five personality. Big Five personality was assessed using a 44-item Big Five Inventory (BFI: John & Srivastava, 1999) and Saucier's (1994) 40-item Mini-Markers (MM). Participants completed these surveys in 1998 and were asked to distribute additional copies designed for peer ratings to any three people who knew them "very well." As discussed in previous work on this sample (DeYoung, 2006), peer reports were provided primarily by friends, relatives, spouses, or coworkers, and the average inter-rater correlations were highly comparable to that reported in meta-analyses on the topic. Responses were provided on five-point Likert scales, and scores for Big Five traits were obtained by taking the mean of all items for each trait from both the BFI and MM, yielding alphas between .84 and .94. Self-report data was available for 598 members of the subsample, and two or more peer reports were available for 573 members of the subsample.

Intelligence. 569 members of the subsample completed Cattell's 16 Personality Factor Questionnaire (Conn & Rieke, 1994) in 1996. The scale that Cattell labelled "Factor B" is a fifteen item performance test of cognitive ability using knowledge and reasoning problems with multiple choice answers. Previous research (e.g. Abel & Brown, 1998) has found this measure

exhibits large correlations with performance on other established ability measures such as the Wechsler Adult Intelligence Scale.

Socially desirable responding. In 1998, 635 members of the subsample completed the Balanced Inventory of Desirable Responding (Paulhus, 1991). Of interest for the present study is the Self-Deceptive Enhancement scale from this instrument, which assesses egoistic overconfidence and claims to superiority; the modest reliability for this 20-item scale (alpha = .68) is representative of studies using the instrument (Li & Bagger, 2007).

Narcissism. 626 members of the subsample completed the 40-item Narcissistic Personality Inventory (NPI: Raskin & Terry, 1988) in 2006, using a five point response scale. High scorers on the NPI, which uses items such as "I would make a great leader" to assess narcissism, are particularly prone to overstate their positive acts and qualities (Gosling, John, Craik, & Robins, 1998). The alpha reliability was .91.

Cognitive ability self-ratings. Participant's self-ratings of their intelligence were obtained using three items assessed in the same survey as the OCQ; 701 members of the subsample completed these items. Participants read the following instructions:

> *"Compared to other people of your same age and sex, how well do you think you perform in each of the following skills? About what percentage of people do you do BETTER THAN in each skill"*

Responses were provided on a nine-point scale running from "0 – 10%" to "80 – 90%." A list of skills are then presented and defined; three of these – verbal, mathematical, and spatial ability – are included as part of general cognitive ability in current conceptions. We took the average of these three items, which formed a reliable (alpha = .72) composite.

Music collection. In the same 2003 assessment that contained the OCQ, participants were asked "What kinds of music CDs, tapes, or records do you own?" and asked to provide their rough estimate for each of 14 genres, with response options for "0," "1-5," "6-10," "11-20," and "21+." For each of the five genres included in the OCQ, we predicted the number of albums reported as owned for that genre using the total number of albums reported as owned across the remaining 13 genres, and saved the residual. This residual thus represents their self-reported ownership of music in that genre after controlling for any tendency to overclaim general ownership of music.

Discrepancy measures

Ability enhancement. Our indicator of ability enhancement was the discrepancy between self-reported cognitive ability and performance on the ability measure. We used the ability measure to predict self-reports of ability, saving the residual as a discrepancy score. High scores thus indicate the tendency to claim higher levels of ability than was justified by performance on the measure.

Openness enhancement. Because the items included in the present OCQ are likely to involve egoistic and not moralistic bias (Gebauer, Sedikides, Verplanken, & Maio, 2012), OCQ bias is expected to predict overstated levels of traits such as Openness, and not traits such as Agreeableness. Supplementary analyses further supported the use of Openness, as it was the only Big Five trait for which self and peer reports were associated with accuracy on the OCQ (more open individuals were more knowledgeable), and the magnitude of the associations between accuracy and Openness were at least twice as large as with any other Big Five trait. Because true knowledge was associated with elevated levels of Openness (and no other trait), we suggest the

tendency to overclaim knowledge is most likely to associate with the tendency to overstate one's Openness. To identify differences in the tendency to overstate Openness levels, we used the mean of the available peer reports of the participant's personality to predict self-reports, saving the residual as a discrepancy score. High scores thus indicate the tendency to claim higher levels of the trait than was justified by the peer reports.

Indicators of careless responding

Survey errors. In the survey containing the OCQ we identified three instances in which participants could be scored as having committed an error. The first two of these assessed discrepancy between a participants reported usage of e-mail and of the internet other than e-mail, and the number of hours per week participants reported using each. Participants who indicated they did not use a technology and then reported using it for an hour or more per week were scored as having committed a survey error.

A third possible error could occur when participants were instructed to rank order nine skills, with each skill ranked from 1 to 9, and only one skill assigned any given number. Participants who failed to follow instructions – for example, by using the same number for more than one skill – were scored as having committed a survey error. Participant scores on survey errors were computed by summing across these different possible errors.

Mahalanobis distance. Recent research on careless responses in surveys (Meade & Craig, 2012) has identified Mahalanobis distance as an indicator of careless responding. Following Meade & Craig (2012), we used 80 personality items assessed in the same survey as the OCQ to predict study-assigned participant ID number, saving the Mahalanobis distance for each participant. High scores represent individuals whose responses on individual questions are not

necessarily abnormal when each question is considered separately, but where the combination of answers across multiple questions is unusual. That is, high scores indicate one is a multivariate outlier. In principle, such outlying scores may be obtained by truly unusual people responding honestly and attentively, but the substantial correlations reported between Mahalanobis distance and various indicators of careless responding reported by Meade & Craig (2012) highlights its usefulness for the present purposes.

Person-fit residual. We computed person-fit residuals for each participant's scores on each of the seven categories of the OCQ. Person-fit residuals are computed within the framework of Item Response Theory (IRT; Embretson & Reise, 2000), and make use of the differential difficulty among items. For example, whereas some musical artists are nearly universally recognized as fairly or highly familiar (e.g. Beethoven, Louis Armstrong, and Garth Brooks), many artists on the OCQ are much less recognized (e.g. Copland, Thelonious Monk, and Shelly West). This differential difficulty allows us to model the relative ease of different items, and to assess the extent to which a given participant's responses conform to or deviate from that model. For example, a participant with a low level of knowledge about classical artists may be expected to recognize Beethoven, but be less likely to recognize Tchaikovsky and less likely still to be familiar with Copland. Person fit residuals represent the extent to which a given participant's responses deviate from the model; for example, an individual who rates Beethoven as completely unfamiliar while claiming great familiarity with Copland and Tchaikovsky would have a high person-fit residual.

A fundamental assumption of IRT is that the scale is unidimensional. Therefore, we obtained seven person-fit residuals for each participant (one for each of the seven categories of the OCQ). There appears to be generalizable differences in the tendency to produce responses

which poorly fit the model across categories, as correlations between the seven person-fit residuals are universally positive though largely modest (ranging from .05 to .34, with an average of .15). The alpha reliability of a composite of all seven scales was a modest .46; however, for the sake of power and parsimony we computed average person-fit residuals across the categories.

Because our interest in this measure was as a measure of careless responding which may (or may not) correlate with a participant's False Positive score, we computed it using only the real items from each measure; person-fit residuals and False Positive scores thus have no items in common, increasing the interest of any observed overlap.

Person-fit residuals do not exist for participants who respond with either complete familiarity (5) or no familiarity (1) for all of the real items of a given scale. Thus, for example, the 68% of the sample who indicated no familiarity with rap artists do not have a person-fit residual for this genre.[1] (For other OCQ categories the portion of participants indicating no familiarity with a category are considerably smaller: Rock: 22%; Classical: .4%; Country: .6%; Jazz: .1%; Stores: .1%; Designers: 4%.) In addition, 3% of the sample indicated perfect familiarity (selecting the highest response option) with all 12 of the real clothing store names; no person-fit residual could be computed for these participants for this scale, for the same reason.

In the present context, our person-fit residual likely represents the best indicator we have of careless responding on the OCQ, both because it is based on OCQ items rather than other

[1] This extremely low recognition rate likely reflects both the demographic characteristics of the sample and the artists included in the assessment. Evidence of the effects of the sample include that even platinum-selling artists like Master P and Mase were unrecognized by 90 and 96% of the sample, respectively. However, nearly all of the artists responsible for the best-selling hip hop albums released before the OCQ assessment in 2003 (including Eminem, 2Pac, Beastie Boys, Fugees, & Dr. Dre) were omitted from the OCQ. These omissions likely impaired the functioning of the OCQ by limiting its power to detect differences in participants' familiarity with this genre, consistent with Mathers' advice about the inclusion of Dr. Dre (Young, Bradford, Mathers, 1999).

parts of the survey and because it is a composite of seven different person-fit residuals. This latter aspect presents a challenge to other interpretations of the person-fit residuals: for example, though some individuals might be expected to exhibit eclectic tastes within a given domain (e.g. becoming deeply familiar with Copland while largely ignoring Beethoven and Tchaikovsky), it is less likely that they would exhibit this trend consistently across the seven diverse content domains assessed in the OCQ. Truly high scores on the average person-fit residual are thus unlikely to represent honest reporting of unique tastes.

Scoring the OCQ

A brief explanation of the scoring procedure for the OCQ is necessary at this point. Paulhus and Harms (2004) offers two scoring procedures based on Signal Detection Analysis (Macmillan & Creelman, 2004). Both procedures are based off of the hit rate (the proportion of real items with which respondents claim familiarity) and the false alarm rate (the proportion of foils with which respondents claim familiarity). The first procedure is the "common-sense" approach: here, accuracy is represented as a difference score (hit rate minus false alarm rate) and bias is represented as the "yes rate" (hit rate plus false alarm rate). The second procedure represents more traditional SDA indices, using *d prime* as an accuracy index and '*c*' as the bias index. To obtain these, one first standardizes the hit rate and false alarm rate, and then performs the same calculations performed to obtain the common-sense indices. Of note, we follow standard practices in weighting the false alarm rate and the hit rate equally when computing both bias indices, even though the latter is based off of four times as many items as the former. We are unaware of any justification for this weighting, and evaluate its implications below.

The hit rate and false alarm rate are computed using dichotomous scoring for each item: it requires that one has either claimed familiarity or not, rather than allowing for indications of

varying degrees of familiarity. In order to use all of the available data, we followed established procedures in computing a hit and a false alarm rate at each cut-off point on the five-point response scale used in the present study. Thus, we calculated and averaged four accuracy scores and four bias scores, for both the common-sense and the traditional SDA indices. Because both of these bias indices will to a considerable degree reflect both true knowledge and overclaiming (due to the use of the hit rate in its computation), it is appropriate to use as a predictor of overclaiming only when used in a multiple regression alongside its corresponding accuracy index. Perhaps because of this, many authors appear to prefer using the False Positive rate as a measure of overclaiming. For this reason, we present results using it as well. One notable strength of the False Positive rate is its lack of ambiguity as an indicator of overclaiming. Two weaknesses include its relatively constrained power (many participants will claim no familiarity with any bogus item, and no differentiation can be made among such participants), and a more limited representation of the overclaiming construct: claiming familiarity with a truly unfamiliar item may be somewhat different, psychologically, than claiming extra familiarity with a somewhat familiar item.

OCQ scale characteristics. We obtained the alpha reliabilities of the OCQ measures by computing accuracy and bias measures for each of the seven content scales assessed in this version. For both the traditional and the common-sense versions of these measures, the bias indices (.78 and .75, respectively) were slightly more reliable than the accuracy indices (.73 and .67, respectively).

Results

Preliminary results. To examine the validity of scores on our three indicators of careless responding, we tested their correlations with each other and with the mean peer report of

Conscientiousness. All results (presented in Table 2) were significant in the expected direction, though universally modest in magnitude. Individuals rated by their peers as Conscientiousness were less likely to respond idiosyncratically (i.e. have high person-fit residuals; $r = -.09$, $p < .05$), to commit survey errors ($r = -.07$, $p < .05$, one-tailed), and to have a high Mahalanobis distance score ($r = -.08$, $p < .05$, $p < .05$, one-tailed). High person-fit residuals indicated the likelihood of committing survey errors and of high Mahalanobis distance scores (rs = .18 and .15, respectively, both $p < .001$), and survey errors and Mahalanobis distance were positively correlated ($r = .11$, $p < .01$). Finally, each of our three careless responding indicators was correlated with cognitive ability: the correlation with survey errors ($r = -.41$, $p < .001$) is particularly remarkable given the abbreviated nature of our survey errors indicator and the fact that cognitive ability was assessed at a different time. Our careless responding indicators might thus be characterized as measuring a combination of one's ability and willingness to provide effortful responses to surveys.

The various measures used in this study as validity criteria also correlated as expected. All four measures of agentic enhancement (Self-Deceptive Enhancement, Narcissism, and exaggeration of Openness and of cognitive ability) were positively correlated with each other. Table 2 also indicates some challenges to interpreting any association with an indicator of exaggeration similar to the limitations to classic SDR measures: for example, Narcissism correlates positively not only with overclaiming Openness ($r = .29$, $p < .001$) and intelligence ($r = .30$, $p < .001$), but also with having both traits (rs = .15 and .14, $p < .001$ and $p < .01$, respectively, for peer reported Openness and performance on the cognitive ability measure). Therefore, the most unambiguous support for any potential measure of misrepresentation would

come from an association with direct measures of exaggeration, as associations with Narcissism and classic SDR measures are somewhat ambiguous to interpret.

Prediction of validity criteria. Table 3 presents the relationships between measures of accuracy and of overclaiming (aka bias) derived from the OCQ. The False Positive rate provides a reason for initial concern about the performance of the OCQ. Most significantly, the False Positive rate was completely unassociated with any of the indicators of exaggeration.[2] Instead, the False Positive rate appears to substantially reflect careless responding, as indicated by the positive correlation with all three measures. Those with high False Positive scores also scored poorly on the cognitive ability measure (r -.20, $p < .001$).

This pattern of results is largely the same across the other bias indicators. The common-sense bias index, employed simultaneously with common-sense accuracy in a regression to predict the criteria of interest, differs only in exhibiting a significant positive relationship with number of albums owned within a given genre (average beta across five genres = .24, $p < .01$). Of interest, the accuracy index performs as expected, exhibiting a positive relationship with intelligence, peer-reported Openness, and ownership of albums in a given genre. Notably, accuracy was significantly negatively associated with each of the careless responding indicators.

This pattern is largely preserved in the traditional indices. D prime exhibits the same pattern of associations as does common-sense accuracy, and common-sense bias and c differ only in that the latter is modestly associated with Narcissism (beta = .11, $p < .05$) and is not associated with Survey Errors (beta = .07, $p > .05$).

[2] As described above, only the Self-Deceptive Enhancement scale from the BIDR (and not the Impression Management scale) is expected to associate with overclaiming on the present OCQ, as it is expected to predict egoistic and not moralistic bias. Supplementary analyses confirmed that high scorers on Impression Management are not prone to overclaim on this OCQ.

Evaluating potential explanations for lack of convergent validity for OCQ Overclaiming measures.

We consider three possible explanations for the lack of convergent validity of OCQ Overclaiming measures: careless responding, ineffective weighting of items, and ineffective selection of categories.

Careless responding. To the extent that the limited relationship between OCQ overclaiming measures and other indicators of bias is due to careless responding influencing the former, controlling for the level of careless responding might be expected to improve the association between OCQ overclaiming measures and the other indicators of bias. To test this, we added the careless responding indicators (singly, and then in a separate analysis all at once) alongside the OCQ indexes in multiple regressions, predicting each of the bias criteria in turn.

In a few instances, the expected effect was observed. For example, including either survey errors or person-fit residuals alongside the traditional measures lead to a now-significant modest relationship between bias and overclaiming Openness (with betas = .10 and .13, respectively, both $p < .05$), and including survey errors alongside the common-sense measures lead to a significant association between bias and narcissism (beta = .10, $p < .05$). Because these associations were those closest to significance when analyzed without indicators of careless responding included it is sensible that they were the ones most likely to become significant. However, these improved associations are somewhat inconsistent, as neither is significant when all three careless responding indicators are included simultaneously. Additionally of note, there is a significant association between narcissism and common-sense *accuracy* (beta = .13, $p < .03$) when person-fit residuals are included as a predictor; such a finding is not altogether unexpected,

as results from previous research and from Table 2 demonstrate that narcissism is elevated among those with knowledge.

An advantage to including the measures of careless responding as a predictor is that it allows one to analyze responses for all participants, rather than dropping those deemed to be careless responders from the analyses. Dropping participants is challenging in the present context, as there is no clear cut-point to differentiate the careless from the careful. For our least continuous measure of careless responding (Survey Errors), 74% of our sample was scored as completely attentive, so we (somewhat arbitrarily) chose that as our cut-off point on each measure. (Similar results were observed in supplementary analyses using other similarly arbitrary values as thresholds for identifying careless versus careful responders.) Among the 74% of the sample scoring lowest on a composite measure of careless responding (which averaged the three careless responding indicators together), results resembled those from the full sample presented in Table 3. Similar results were also found when using person-fit residuals to differentiate careless versus careful responders. When using Survey Errors to distinguish between careless and careful responders, we found that traditional bias was significantly associated with overclaiming Openness (beta = .12, $p < .04$) among careful responders; when using Maholanobis distance, common-sense bias was significantly associated with Narcissism (beta = .15, $p < .05$).

Rebalanced bias scales. We recalculated the common-sense bias scale and c to reflect the fact that real items make up 80% of the OCQ whereas bogus items make up only 20%; in the absence of any reason to weight them equally in the construction of these scales, we weighted the Hit rate four times more than the False Alarm rate. No results changed with respect to becoming statistically significant or insignificant, though modest changes in the predicted

directions were observed for all associations; for example, the associations between common-sense accuracy and both intelligence and peer-rated Openness were modestly improved when a rebalanced common-sense bias measure was used rather than the original common-sense bias measure (betas = 0.49 and 0.35, respectively, when used alongside the rebalanced measure, and betas = 0.41 and 0.31, respectively, when used alongside the original measure).

Comparison of different OCQ scales. Many studies using the OCQ rely on scales concerning knowledge that is more academic than the present study. One possible explanation for the present findings is that individuals were not motivated to express any tendencies they had to overclaim given the low stakes nature of the assessment procedure (confidential and by mail) and the low prestige of the items used in this version of the OCQ. Unfortunately, we do not have any data on how participants viewed the prestige attached to knowledge of the various categories assessed in the present version of the OCQ, so a rigorous comparison of the effects of scale prestige on overclaiming is not possible at present. Accordingly, the following results were obtained using sufficient "researcher degrees of freedom" (RDF: Simmons, Nelson, & Simonsohn, 2011) such that conventional hypothesis testing is inappropriate; we present them only as an overview of some potentially interesting trends in the data.

In the absence of a rigorous means of classifying the content domains used in this OCQ, the authors used their judgment in determining that the sample likely views Classical and Jazz as "high prestige" scales whereas Rap and Rock were "low prestige" scales. Consistent with our assessment, the rate at which participants claimed at least some familiarity with bogus items from Rap and Rock (3%) was lower than both the average across all scales (6%) and for Classical and Jazz (7%) in particular. For comparison, participants claimed at least some familiarity with 40% of real items across all scales.

Grouping these four genres to create high and low prestige scales, we found some evidence in favor of the hypothesis that scale prestige predicts performance on OCQ measures. Using the common-sense version of the accuracy and bias scales, accuracy on the high prestige scales was markedly associated with cognitive ability and peer ratings of Openness (betas = .41 and .35, respectively; both $p < .001$, uncorrected for RDF), whereas this was less pronounced for the low prestige scales (betas = .22 and .16, respectively; $p < .001$ and $p < .01$, respectively, uncorrected for RDF). Bias on the high prestige scales was a significant predictor of Openness exaggeration (beta = .13, $p < .03$, uncorrected for RDF), but this was not the case for low prestige scales (beta = .05, $p > .35$, uncorrected for RDF). However, results with other bias criteria were insignificant for both high and low prestige versions.

Discussion

The present study sought to provide a thorough evaluation of the validity of the OCQ using an established sample which has been successfully used in previous studies of misrepresentation in self-reports. Several themes emerged from the results presented above.

First, OCQ accuracy measures functioned quite well in the present sample, exhibiting the expected positive associations with relevant criteria such as performance on an ability measure and peer reported Openness. Further, scores on OCQ accuracy measures are depressed among those not paying attention to the assessment. Future research using the OCQ may consider whether the accuracy measure's true power to predict criteria is adequately demonstrated when it is used alongside bias measures which overweight the importance of False Positives. More broadly speaking, these results are consistent with the use of the OCQ and related measures as predictors of knowledge – as they have been used, for example, in academic contexts (Paulhus & Dubois, 2014).

Second, OCQ bias measures performed very poorly, exhibiting no association with either discrepancy measure or with self-deceptive enhancement. Only the traditional bias measure exhibited a relationship with narcissism, and this relationship was quite modest. This latter finding highlights some important limitations to the generalizability of the present study: although the literature on the OCQ is decidedly unclear on whether OCQ bias measures predict discrepancy measures, the associations with narcissism and egoistic self-enhancement are better established (Bing et al., 2011; Gebauer et al., 2012; Paulhus et al., 2003; Paulhus & Harms, 2004; Paulhus & Williams, 2002; Tonković et al., 2011; though see Barber et al., 2013). It is thus important to consider limitations of the present study as a potential explanation for the observed results.

One possibility concerns the setting of the assessment: whereas many previous studies of the OCQ were conducted in classrooms and research laboratories, participants completed the present study wherever they chose, and submitted their answers confidentially. The other study which reported no association between the OCQ bias and discrepancy measures does not provide details of the assessment procedure (Musch et al., 2012); the previous study which first reported the large correlations between OCQ False Positives and careless responding was conducted in a similarly informal and confidential setting (Barber et al., 2013). However, participants in the present sample clearly exhibit meaningful differences in their tendencies to misrepresent personal characteristics in a positive light, as discrepancy measures for personality and ability can be meaningfully predicted in the present sample not only by narcissism and SDR measures but also by ratings of trait desirability, education, sociopolitical attitudes, and religion (Ludeke & Carey, 2015; Ludeke et al., 2015, 2013; Ludeke, 2014). As with the OCQ, these characteristics were typically assessed several years after the assessments used to create the discrepancy

measures, so we see no reason why one should expect only the OCQ bias measures to be uniquely invalidated in the present assessment context. This same reasoning presents a stiff challenge to any attempt to explain the present results as a consequence of sample characteristics: the present sample exhibits predictable discrepancies in self-reports, and we see no reason why its demographic characteristics should eliminate the expected association between those discrepancies and OCQ bias scores.

There are also hurdles to attributing the poor performance of the OCQ bias measures to the particular version of the OCQ used. OCQ bias is, in the present sample, a reasonably coherent measure. It exhibited a greater degree of internal consistency than did OCQ accuracy, and as with OCQ accuracy, differences in OCQ bias were readily predictable. The challenge is simply that OCQ bias was most clearly associated with the "wrong" things – specifically, with measures of careless responding rather than of egoistic overclaiming or of self-criterion discrepancies.

At the same time, the content assessed in the present study's version of the OCQ likely accounts for at least some of the observed results. Most versions of the OCQ use content that is relatively more prestigious – for example, knowing famous philosophical doctrines (used in many OCQ studies) probably strikes many as more impressive than knowing the names of large clothing retailers (used in present study). Our exploratory analyses were consistent with the idea that overclaiming knowledge on more prestigious scales was associated with the discrepancy measure for Openness (though not intelligence), suggesting that future users of the OCT should carefully consider how the content areas they include in their assessment may improve or impair their study.

Consistent with the suggestion that the present study's version of the OCQ assesses domains which are generally lower in prestige, the present sample claimed knowledge with bogus items at a rate (6%) much lower than the 25% reported by Paulhus and colleagues (2003). This gap may in part reflect the large mean-level differences in narcissism between the college students from that study and the middle-aged adults used in the present sample (Foster, Campbell, & Twenge, 2003).[3] To the extent that this is the case, expectations for associations between OCQ bias and discrepancy measures need not be fundamentally altered, as such age trends suggest only that the rate and not the meaning of overclaiming behavior differs between the two samples. However, to the extent that the present's study's decreased rate of overclaiming reflects a failure of the present study's OCQ domains to inspire egoistic overclaiming, a decreased association of OCQ bias with both egoistic measures (such as narcissism and self-deceptive enhancement) and discrepancy measures is to be expected. Because of this, our failure to find support for the OCQ bias measures is not as daunting as it might otherwise be; future study looking to establish the validity of OCQ scores very may well succeed where we failed if a more appropriate version of the OCQ is used.

The role of careless responding. Although the survey containing the OCQ did not contain any purpose-built measures to identify careless responding, we were able to create three distinct indicators of careless responding, each of which used different pieces of information. The observed negative relationship between each of these measures and the OCQ accuracy indices further supports their use, as it indicates participants scoring high on the careless responding indices were unable to effectively discriminate between real and bogus items on the OCQ. Their

[3] Because the Narcissistic Personality Inventory (NPI) traditionally uses a dichotomous response format and our sample used a five-point Likert scale, it is not possible to meaningfully compare the mean NPI score in our sample against most of the literature. However, a modest negative correlation between age and NPI in the present sample (r = -.11, $p < .01$) suggests age trends are still relevant for the present version of the NPI.

associations with the OCQ bias indices are thus of particular interest – particularly the correlations with the False Positive rate.

Understandably, the careless responding measure most associated with the bias measures (person-fit residuals) was the careless responding measure obtained from the same measure. Importantly, in the case of the correlation with the false positive rate, these two measures use no common information: person-fit is based on true items, whereas false positive is based exclusively on bogus items. But a high correlation is nonetheless expected because the careless responding would be occurring (in two different ways) on the exact same instrument; to the extent that individuals do not exhibit the same level of carelessness across the entire survey (but instead exhibit some variation across the components), the correlation between the two different indicators should be highest when they are both based on the same part of the survey, the OCQ.

The associations between the careless responding indicators and the False Positive rate are comparable to those reported by Barber and colleagues (2013), which, like our study, was conducted confidentially and outside of the standard classroom and research laboratory environment. Importantly, Barber and colleagues (2013) used a more conventional version of the OCQ, suggesting that the present results cannot be attributed solely to the low prestige nature of the content assessed in our OCQ. The present study replicates and extends these findings, showing that careless responding modestly impairs the predictive power of the OCQ bias indices: incorporating them into a multiple regression leads to some improvements in the ability of the OCQ bias indices to predict validity criteria such as narcissism and the personality discrepancy measure, as does simply eliminating the portion of the sample scoring highest on some careless responding measures. That these improvements were present even when using our post-hoc measures of careless responding is a point of hope for the OCQ, as the use of longer,

purpose-built measures may well yield greater improvements. However, this hope pertains primarily to the validity and not the utility of OCQ scores. This is because, as noted above, the discrepancy measures in this sample are predicted much more robustly by a range of constructs other than the OCQ bias measures. Given the size and quality of the sample, statistical significance is a very minimal standard for the OCQ to surpass. To be justified in future applications, the OCQ needs to be a much more robust predictor of discrepancy measures than we have observed it to be.

The associations between OCQ bias and careless responding in the present study further highlight the importance of assessing content domains in the OCQ which are likely to be over-claimed at an elevated rate. It also suggests the OCQ may function best where it is needed most: in high stakes contexts, where participants have incentives to both pay attention and to over-claim. This is because (as discussed above) the mechanics of OCQ scoring allow the False Positive rate to be substantially influenced by careless responding. This is particularly true when overclaiming behavior is minimal, as this entails that the average scores of attentive responders is very far from the scale midpoint, whereas careless responders likely tend towards average scores of the scale midpoint and thus make up a significant portion of those with the highest False Positive scores. In contrast, when using an OCQ version and an assessment context in which the average False Positive rate is closer to the scale midpoint, the scores of careless responders will be less extreme relative to the scores of careful responders, and thus potentially less likely to impair the predictive power of OCQ bias indices.

Other results of interest. The focus of the present paper has been on attempting to evaluate the validity of scores on the OCQ bias indices. Momentarily suspending this question, it is worth considering what might be inferred from other results presented here. When computed

only using items from a particular musical genre, both the common-sense and the traditional OCQ bias measures exhibited a significant positive association with album ownership in that genre (after controlling for the participant's general stated levels of album ownership). This might be interpreted as a flaw in the OCQ bias index: album ownership might be most expected to relate to true knowledge within a given musical domain, and indeed it is positively associated with OCQ accuracy indices. This suggests a possible unfortunate commonality between OCQ bias indices and other attempts to measure misrepresentation in self-reports, in which those who are identified as engaging in misrepresentation are also those with truly elevated levels of the traits in question (Ones et al., 1996).

However, an association between OCQ bias and album ownership might also be a psychologically meaningful result. For example, those who own many classical albums may be more reluctant to admit lack of knowledge with an ostensible classical artist than would those who own fewer classical albums, inspiring them to claim more familiarity than is accurate. Of interest, the False Positive rate is less markedly associated with album ownership than are the other OCQ bias indices. One possible interpretation of this pattern of results is that those owning many albums in an area are not particularly willing to claim familiarity with names that are truly unfamiliar, but they are willing to inflate their claims of familiarity with names they truly recognize. If correct, this would highlight the importance of using the broad overclaiming indices recommended by Paulhus and colleagues (2003, 2004) rather than the False Positive rate alone, as overclaiming on truly unrecognized foils may represent a somewhat psychologically distinct process from overclaiming on partially recognized real items. Future research on the OCT may consider whether it is appropriate to collapse both kinds of overclaiming into a single index, or

whether there is a need and a possibility to separately assess and evaluate each manner of overclaiming.

Conclusions

The conceptual appeal behind the OCT as a measure of the tendency to misrepresent oneself in self reports is obvious and substantial. This conceptual appeal is, at present, not paralleled by a supportive empirical record. Overclaiming on the OCQ was associated primarily with careless responding and not with tendencies towards misrepresentation or overconfidence, even though accurate responding on the OCQ exhibited the expected associations with its criterion measures. Future work further establishing the validity of scores on the OCQ bias measures and exploring the best practices for OCT assessment is clearly needed.

References

Abel, M. H., & Brown, L. K. (1998). Validity of the 16PF reasoning ability scale. *Psychological Reports, 83*, 904–906.

Barber, L. K., Barnes, C. M., & Carlson, K. D. (2013). Random and Systematic Error Effects of Insomnia on Survey Behavior. *Organizational Research Methods, 16*(4), 616–649. doi:10.1177/1094428113493120

Bing, M., Kluemper, D., & Davison, H. K. (2011). Overclaiming as a measure of faking. *Organizational Behavior and Human Decision Processes*, *116*, 148–162. doi:10.1016/j.obhdp.2011.05.006

Conn, S. R., & Rieke, M. L. (1994). *The 16PF 5th edition technical manual*. Champagne, IL.

DeYoung, C. G. (2006). Higher-order factors of the Big Five in a multi-informant sample. *Journal of Personality and Social Psychology*, *91*(6), 1138–51. doi:10.1037/0022-3514.91.6.1138

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Psychology Press.

Foster, J. D., Campbell, W. K., & Twenge, J. M. (2003). Individual differences in narcissism: Inflated self-views across the lifespan and around the world. *Journal of Research in Personality*, *37*(6), 469–486. doi:10.1016/S0092-6566(03)00026-6

Gebauer, J. E., Sedikides, C., Verplanken, B., & Maio, G. R. (2012). Communal narcissism. *Journal of Personality and Social Psychology*, *103*(5), 854–78. doi:10.1037/a0029629

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. J. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7, pp. 7–28). Tilburg, The Netherlands: Tilburg Univesity Press.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, *40*(1), 84–96. doi:10.1016/j.jrp.2005.08.007

Gosling, S. D., John, O. P., Craik, K. H., & Robins, R. W. (1998). Do people know how they behave? Self-reported act frequencies compared with on-line codings by observers. *Journal of Personality and Social Psychology*, *74*(5), 1337–1349. doi:10.1037//0022-3514.74.5.1337

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (pp. 102–138). New York, NY: Guilford Press.

Li, A., & Bagger, J. (2007). The balanced inventory of desirable responding (BIDR): A reliability generalization study. *Educational and Psychological Measurement, 67*(3), 525–544. doi:10.1177/0013164406292087

Ludeke, S. G. (2014). Truth and fiction in the association between Openness and education: The role of biased responding. *Learning and Individual Differences*, *35*, 137–141. doi:10.1016/j.lindif.2014.07.008

Ludeke, S. G., & Carey, B. (2015). Two mechanisms of biased responding account for the association between religiousness and misrepresentation in Big Five self-reports. *Journal of Research in Personality*, *57*(August), 43–47.

Ludeke, S. G., Reifen Tagar, M., & DeYoung, C. G. (2015). Not as different as we want to be: Attitudinally consistent trait desirability leads to exaggerated associations between personality and sociopolitical attitudes. *Political Psychology*. doi:10.1111/pops.12221

Ludeke, S. G., Weisberg, Y. J., & DeYoung, C. G. (2013). Idiographically desirable responding: Individual differences in perceived trait desirability predict overclaiming. *European Journal of Personality*, *27*(6), 580–592. doi:10.1002/per

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. Psychology Press.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–55. doi:10.1037/a0028085

Musch, J., Ostapczuk, M., & Klaiber, Y. (2012). Validating an inventory for the assessment of egoistic bias and moralistic bias as two separable components of social desirability. *Journal of Personality Assessment*, *94*(6), 620–9. doi:10.1080/00223891.2012.672505

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology*, *81*(6), 660–679. doi:10.1037//0021-9010.81.6.660

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measurement of Personality and Social Psychological Attitudes* (pp. 17–59). New York, NY: Academic Press.

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 49–69). Mahwah, NJ: Erlbaum.

Paulhus, D. L., & Dubois, P. J. (2014). Application of the Overclaiming Technique to Scholastic Assessment. *Educational and Psychological Measurement*, *74*(6), 975–990. doi:10.1177/0013164414536184

Paulhus, D. L., & Harms, P. D. (2004). Measuring cognitive ability with the overclaiming technique. *Intelligence*, *32*(3), 297–314. doi:10.1016/j.intell.2004.02.001

Paulhus, D. L., Harms, P. D., Bruce, M. N., & Lysy, D. C. (2003). The over-claiming technique: Measuring self-enhancement independent of ability. *Journal of Personality and Social Psychology*, *84*(4), 890–904. doi:10.1037/0022-3514.84.4.890

Paulhus, D. L., & Williams, K. M. (2002). The Dark Triad of personality: Narcissism, Machiavellianism, and psychopathy. *Journal of Research in Personality*, *36*(6), 556–563. doi:10.1016/S0092-6566(02)00505-6

Raskin, R., & Terry, H. (1988). A principal-components analysis of the Narcissistic Personality Inventory and further evidence of its construct validity. *Journal of Personality and Social Psychology*, *54*(5), 890–902.

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science*, *2*(4), 313–345. doi:10.1111/j.1745-6916.2007.00047.x

Saucier, G. (1994). Mini-markers: a brief version of Goldberg's unipolar big-five markers. *Journal of Personality Assessment*, *63*(3), 506–16. doi:10.1207/s15327752jpa6303_8

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–66. doi:10.1177/0956797611417632

Tonković, M., Galić, Z., & Jerneić, Ž. (2011). The construct validity of over-claiming as a measure of egoistic enhancement. *Review of Psychology*, *18*(1), 13–21.

Viswesvaran, C., & Ones, D. S. (1999). Meta-Analyses of Fakability Estimates: Implications for Personality Measurement. *Educational and Psychological Measurement*, *59*(2), 197–210. doi:10.1177/00131649921969802

Table 1. Categorization of responses to the Over-Claiming Questionnaire, with illustrative responses from one participant answering "Classical Artists" items

		Response	
		Familiar	Unfamiliar
Item type	Real	True positive *Beethoven*	False Negative *Stravinsky*
	Foil	False Positive *Ochberg*	True Negative *Rentzig*

Table 2. Correlations between careless responding and validity criteria

		1	2	3	4	5	6	7	8	9
1	Conscientiousness (Peer rated)									
2	Survey Errors	-.07^								
3	Maholanobis Distance	- .08^	.11**							
4	Person-fit residuals	-.09*	.18***	.15***						
5	Cognitive ability	.04	-.41***	-.15***	-.28***					
6	Openness (Peer rated)	.14***	-.13***	-.01	-.06	.30***				
7	Cognitive ability exaggeration	.03	-.06	-.05	.05	-	.13**			
8	Openness exaggeration	-.23***	-.01	.02	-.04	.13**	-	.20***		
9	Self-Deceptive Enhancement	.15***	.02	-.07	.04	.03	-.04	.14***	.12**	
10	Narcissism	-.02	-.07	.03	.07	.14**	.15***	.30***	.29***	.29***

Note. Correlations of particular interest in grey cells. Values are marked with "-" if one variable was used to compute the other variable. ^ = $p < .05$ (one-tailed); * = $p < .05$ (two-tailed); ** = $p < .01$ (two-tailed); *** = $p < .001$ (two-tailed).

Table 3. Associations with accuracy and bias indicators

			Common-sense indices[b]		Traditional indices[b]	
		False Positive[a]	Accuracy	Bias	d'	c
Accuracy criteria	Cognitive ability	**-.20**	**.41**	**-.19**	**.32**	**-.14**
	Peer rated Openness	-.08	**.31**	-.08	**.26**	-.01
	Number of albums	.09	**.31**	**.24**	**.28**	**.22**
Bias criteria	Cognitive ability exaggeration	.01	.03	.03	.00	.03
	Openness exaggeration	.04	.05	.07	.02	.09
	Self-Deceptive Enhancement	-.02	.08	-.01	.05	.01
	Narcissism	.04	.09	.09	.05	**.11**
Careless responding indicators	Person-fit residuals	**.47**	**-.57**	**.56**	**-.46**	**.45**
	Survey errors	**.13**	**-.26**	**.12**	**-.19**	.07
	Mahalanobis	**.15**	**-.18**	**.18**	**-.13**	**.15**

Note. a = these are correlation coefficients. b = these are standardized regression coefficients, where both indices were simultaneously used as predictors. Bolded values significant at $p < .05$.