Journal of Articles in Support of the Null Hypothesis
Vol. 16, No. 2
 1539–8714
www.jasnh.com

Response-Order Effects for Self-report Questionnaires: Exploring the role of Overclaiming Accuracy and Bias

Sean P. Mackinnon
Mengyao Wang
Dalhousie University

Primacy effects refer to the tendency for participants to choose questionnaire response options that are closer to the beginning of a list. We sought to replicate this effect using measures of personality and well-being. We also explored accuracy and bias on the Overclaiming Questionnaire (OCQ) as moderators. Participants were undergraduates (N = 774; 79.2% female; 73.1% Caucasian). We used a two-group, between-subjects design which manipulated the presentation order for response options on a 5-point Likert scale. The two conditions were ascending (*Strongly Disagree* first) and descending (*Strongly Agree* first). We did not find support for OCQ accuracy/bias as a moderator. Because effect sizes were very small, primacy effects may be of little practical importance in this context. Open Data/Methods: https://osf.io/aec25/

Keywords: overclaiming, response-order effects, accuracy, bias, Likert

Author Note:
We would like to thank Samantha Firth for her research assistance. Samantha Firth's research assistantship was supported by a research grant from the Social Sciences and Humanities Research Council.

Even minor modifications to features of questionnaires (e.g., the order of response options to close-ended questions) can yield significant differences in participants' responses. Primacy effects are when respondents tend to choose the options at the beginning of a response list when categorical items are presented visually (Krosnick & Presser, 2010). Studies on primacy effects and potential moderators could enable researchers to better understand response-order-related bias in their data. In our current study, we used an online self-report questionnaire on personality and perfectionism, and conducted a between-subject experiment to examine whether primacy effects exist by comparing two conditions where responses were presented in ascending order (*Strongly Disagree* to *Strongly Agree*) or descending order (*Strongly Agree* to *Strongly Disagree*). Furthermore, we used the overclaiming technique to measure accuracy and and bias as general response patterns. In the present study, we explored whether response tendencies using the overclaiming technique moderate the magnitude of primacy effects.

Primacy effects

Krosnick and Alwin (1987) provided evidence of primacy effects by altering response order in a personality traits questionnaire using a between-subjects design. They also proposed an interaction between condition and cognitive sophistication, such that participants with low education level and verbal ability showed stronger primacy effects. The idea of "satisficing" is one theory that might explain primacy effects (Krosnick, 1991). Satisficing theory suggests that participants minimize the cognitive effort when completing a task, rather than thoroughly evaluating all the response options and providing an optimal answer; thus, respondents tend to choose the first merely acceptable option that was visually presented to them.

Primacy effects in self-report questionnaires have been replicated in various domains, such as self-rated health (Garbarski, Schaeffer, & Dykeme, 2015) and political opinion (Malhotra, 2008). Mackinnon and Firth (2018) found primacy effects using self-reported drinking questionnaires, finding that the response option *Strongly Agree* on a Likert scale was chosen more frequently when it was presented first in response list, compared to when presented last. Malhotra (2008) found that lower education level predicted larger magnitude of primacy effects. Moreover, he suggested the amount of time spent on completing the survey interacts with education level on primacy effects; that is, low education respondents with less time spent on the survey are most prone to primacy effects.

Overclaiming technique

Paulhus, Harms, Bruce, and Lysy (2003) originally developed the Overclaiming Questionnaire (OCQ) as a measure for self-enhancement. This technique requires participants to indicate their level of familiarity with items on a scale from 0 (never heard of it) to 6 (very familiar). Most items are common knowledge in modern Western society, while some items are "foils" that do not actually exist. Overclaiming refers to the tendency to claim familiarity to foils. Signal detection analysis (Macmillan & Creelman, 2005) is employed to sort responses into one of the four categories: hits (real items that are correctly rated as familiar), false alarms (foils that are incorrectly rated as familiar), misses (real items

that are incorrectly rated as unfamiliar), and correct rejections (foils that are correctly rated as unfamiliar). Indices of accuracy and bias can be calculated based on the proportions of hits and false alarms.

Accuracy refers to participants' ability to differentiate foils from real items and is calculated as the difference score of the hit and false alarm rates. Paulhus and Harms (2004) found a strong positive association between accuracy and cognitive ability, specifically, between IQ and crystalized intelligence. Paulhus and Dubois (2014) further explored the potential of the OCQ as an alternative scholastic assessment to multiple choice questions and found it had good convergent validity, predictive validity, and reliability. Pesta and Pozanski (2009) found that overclaiming accuracy is positively related to grades for MBA students. Ziegler, Kemper and Rammstedt (2014) found that a modified vocabulary version of the OCQ was positively associated with general and verbal knowledge. Ludeke and Makransky (2015) also found that accuracy is positively correlated to intelligence and knowledge. Moreover, they indicated a negative correlation between accuracy and careless indicators, suggesting that attentive respondents show higher accuracy compared to careless respondents.

Bias refers to "Yes" rate, that is, one's tendency to claim items as "familiar" versus "unfamiliar" to both real items and foils. Paulhus et al. (2003) suggested that bias is correlated with measures of self-enhancement and socially desirable responding. However, Ludeke and Makransky (2015) and Kam, Risavy, and Perunovic (2014) disagree, and argue that the OCQ bias index has weak validity when predicting self-enhancement or socially desirable responding. Ludeke and Makransky (2015) pointed out that bias index is consistently related to the measures of careless responding (e.g., survey errors). Based on the evidence from past studies, OCQ accuracy and bias could be conceptualized as indicators of cognitive ability and careless responding, respectively.

Satisficing Theory and Overclaiming

In the present study, OCQ accuracy is a proxy measurement for cognitive ability, based on the above literature review. According to "satisficing" theory and prior research (Ludeke & Makransky, 2015; Malhotra, 2008) respondents high in cognitive ability will likely be less prone to primacy effects. Participants who are high in OCQ accuracy could be considered as attentive respondents, with high cognitive ability and attention to detail. On the other hand, to minimize the cognitive effort on completing questionnaires, careless respondents (i.e., high in OCQ bias) may tend to employ "satisficing" strategy more frequently, and thus show a larger primacy effect. As a result, their responses might fail to reflect accurate information, and thus might introduce additional measurement error.

Current Study

We used a between-subjects design to explore the moderation effect of OCQ indices on the magnitude of primacy effects using questionnaire on personality and perfectionism. We predicted that respondents would show primacy effects using well-being and perfectionism questionnaires and that OCQ accuracy indices moderate the magnitude of primacy effects. Therefore, our hypotheses are as follow:

H1: Participants will choose *Strongly Agree* more frequently on closed-ended

responses to personality and well-being questionnaires when the response options are presented in descending order (from left to right: *Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree*), as opposed to being presented in ascending order (from left to right: *Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree*).
H2: There will be a significant interaction between OCQ accuracy indices and condition (i.e., ascending or descending) when predicting the frequency of each response option. Specifically, as OCQ accuracy increases, the magnitude of the primacy effects will become smaller.

We also explored whether OCQ bias moderates primacy effects, but given prior research (Ludeke & Makransky, 2015), we were less certain about the direction of this effect. Thus, we had an additional secondary research question:

RQ1: Does OCQ bias interact with condition (i.e., ascending or descending) when predicting the frequency of each category of responses?

Method

Power Analysis

We conducted the power analysis for H1, based on a prior published study (Mackinnon & Firth, 2018). In this study, they observed an effect size of $W = .036$ and an intraclass correlation (ICC) of 0.06. Bland (2004) proposed a design effect (Deff) formula to correct the reduced statistical power due to the effect of clustered data, as follows:

$$D_{eff} = 1 + (m - 1) \times ICC$$

where m is the number of observations in a cluster. In our study, $m = 82$ (i.e., 82 items for each participant, except for OCQ).

Thus, we estimated the sample size for clustered data by first computing a required sample size for an unclustered design, and then multiplying that sample size by the design effect to get our total sample size. Assuming an effect size of $W = .036$ (Mackinnon & Firth, 2018), 80% power, alpha = .05, and a design effect of 5.86 we require a minimum sample size of 663 participants to achieve 80% power.[1]

Participants

Of the 931 participants who opened the survey link, 774 completed the survey, and were included in the study (i.e., 83.1% of participants who opened the survey and read the consent form clicked through to complete study tasks and consented to share their data. We

1 We thought to include including random effects for both participant *and* item only after reading Barr, Levy, Scheepers, and Tily (2013). This was, unfortunately, after data were collected. Thus, this power analysis may not perfectly represent the analysis described below. Nonetheless, this power analysis was the a-priori basis for the sample size we collected. We exceeded this number by 111 participants, as data collection was more rapid than expected.

recruited 774 participants online through online ads, flyers, and the Participant Pool of a Canadian university. There were no inclusion/exclusion criteria for the study. Participants were primarily young (M_{age} = 21.14, SD_{age} = 5.97), female (79.2%), and Caucasian (73.1%).

Measures

Demographics. The demographic questionnaire included three items to obtain participants' age, sex, and ethnicity, respectively. Age was measured in years. Sex was reported in three categories, i.e., male/female/other (please specify). Ethnicity was reported using open-ended text (e.g., Asian, Caucasian/White, First Nations, etc.) and was converted to categories post-hoc.

Over-claiming questionnaire (OCQ; moderator variable). We used a 90-item version of the OCQ (Paulhus et al., 2003). The items are from six domains of knowledge (i.e., historical names and events; physical sciences; twentieth century names; books and poems; authors and characters; and social science and law). Fifteen items are presented in each category. Twelve out of every 15 are real items and the remaining three items are foils that do not actually exist. Participants were asked to rate their familiarity with each item using a 7-point Likert scale ranging from 0 (Never Heard of It) to 6 (Very Familiar).

To score the OCQ responses, we first dichotomized each response into unfamiliar (0 in original rating) and any level of familiar (1 to 6 in original rating) and then applied signal detection analysis to the data following the procedure suggested by Paulhus and Harms (2004). For each respondent, we calculated hit rate by dividing the number of real items that are correctly rated as familiar by the total number of real items, and false-alarm rate by dividing the number of foils that are falsely rated as familiar by the total number of foils. To avoid the computational problems when calculating accuracy and bias caused by extreme values in the hit or false alarm rates (i.e., 0 or 1), we employed a correction approach (Stanislaw & Todorov, 1999) by adding 0.5 to the number of hits and false alarms and adding 1 to total number of real items and foils, before calculating hit and false-alarm rates.

The accuracy index d' and the bias index c derived from signal detection theory (SDT) (Macmillan & Creelman, 2005) are examined. The formulas are as follow:

$$d' = \Phi^{-1}(hitrate) - \Phi^{-1}(falsealarmrate)$$
$$c = (\Phi^{-1}(hitrate) - \Phi^{-1}(falsealarmrate))/2$$

We used the PROBIT command in SPSS for the Φ^{-1} function.

Questionnaire Measures (target measures). We modified and standardized the original anchors of all questionnaires to be 5-point Likert scales: from 1 (*Strongly Disagree*) to 5 (*Strongly Agree*) in ascending condition and from 1 (*Strongly Agree*) to 5 (*Strongly Disagree*) in descending condition. This allows for a unified data analysis scheme that can include all questionnaire items in a single analysis. All questionnaires except for perfectionism referred to the measurement period of "the past year."

Depression and anxiety. Depression and anxiety were measured using three subscales (i.e. 7-item stress subscale, 7-item depression subscale, and 7-item anxiety

subscale) of the Depression Anxiety Stress Scale-21 (DASS-21). A sample item includes: "I couldn't seem to experience any positive feeling at all." (Lovibond & Lovibond, 1995). The unmodified measure used a 4-point scale (0 = Did not apply to me at all to 3 = Applied to me very much or most of the time). Research supports internal consistency of the DASS-21 (αs > .80), though some research suggests the highly intercorrelated subscales might have a bifactor structure (Osman et al., 2012). In the present study, Cronbach's alpha was .83 for stress, .90 for depression, and .82 for anxiety.

Satisfaction with life. The 5-item Satisfaction with Life Scale measures overall life satisfaction ("I was satisfied with my life;" Diener et al., 1985). The unmodified measure used 7-point scales from 1 (*Strongly Disagree*) to 7 (*Strongly Agree*). Pavot, Diener, Colvin, and Sandvik (1991) supported the internal consistency (α = .83) and criterion validity (*rs* > .50 when compared to peer ratings) of this unmodified measure. In the present study, Cronbach's alpha was .84.

Positive and negative affect. The 20-item Positive and Negative Affect Scale (PANAS; Watson et al., 1988) is comprised of two subscales, positive affect (e.g., "Active") and negative affect (e.g., "Afraid"). In the unmodified measure, each item is rated on a 5-point scale of 1 (not at all) to 5 (very much). Watson et al. (1988) reported that the one-year PANAS version has strong test-retest reliability (*rs* ≥ .60) and a clear two-factor structure. In the present study, Cronbach's alpha was .89 for positive affect and .89 for negative affect.

Flourishing. The Flourishing Scale (Diener et al., 2010) assesses general eudemonic well-being of participants. This scale consists of 8 items and the unmodified measure used a 7- point scale from 1 (*Strongly Disagree*) to 7 (*Strongly Agree*). A sample item is: "I lead a purposeful and meaningful life". Diener et al. (2010) reported that the original flourishing measure has good internal consistency (α = .87) and convergent validity with other well-being measures (rs from .28–.62). Cronbach's alpha was .89 in the present study.

Perfectionism. We measured perfectionism using six short-form subscales developed by Cox, Enns, and Clara (2002): The 5-item self-oriented perfectionism subscale ("One of my goals is to be perfect in everything I do;" Hewitt & Flett, 1991), the 5-item socially prescribed perfectionism subscale ("The better I do, the better I am expected to do;" Hewitt & Flett, 1991), the 5-item other oriented perfectionism subscale ("Everything that others do must be of top-notch quality;" Hewitt & Flett, 1991), the 4-item personal standards subscale ("I set higher goals than most people.;" Frost et al., 1990), the 5-item concern over mistakes subscale ("If I fail at work/school, I am a failure as a person;" Frost et al., 1990), and the 4-item doubts about actions subscale ("Even when I do something very carefully, I often feel that it is not quite right;" Frost et al., 1990). Participants responded to perfectionism items using the timeframe "over the past several years." The items in the first three subscales (i.e. self-oriented perfectionism subscale, socially prescribed perfectionism subscale, and other oriented perfectionism subscale) originally used 7-point scales from 1 (Strongly Disagree) to 7 (Strongly Agree). Cox et al. (2002) supported the factor structure of these short forms. Cronbach's alpha was .87 for self-oriented perfectionism, .78 for socially prescribed perfectionism,.75 for other-oriented perfectionism, .85 for personal standards, .86 for concern over mistakes, and .85 for doubts about actions in the present study.

Procedure

This study was reviewed by a research ethics board at our institution. Data were collected between February 20, 2018 and January 21, 2019. We conducted a between-

subjects experimental design with two conditions. The order of response options differed in the two conditions. In the ascending condition, the close-ended response options for each item in the questionnaires were presented in ascending order (i.e., from left *Strongly Disagree* to right *Strongly Agree*). In the descending condition, each set of response options were presented in descending order (i.e., from left *Strongly Agree* to *Strongly Disagree*). Items were presented in grid format, with multiple items per page, each questionnaire on a separate page (for an example, see: https://osf.io/aec25/). No questionnaires had reversed items.

Participants accessed our online Opinio survey (Object Planet, 2018) by signing up in the university participant pool or by contacting the researchers via email. Participants first completed a baseline demographic questionnaire along with the Overclaiming Questionnaire (OCQ 90; Paulhus et al., 2003). The OCQ responses order was always presented in ascending order for all participants. Then participants were then randomly assigned into either ascending condition or descending condition. Participants in both conditions sequentially completed the Depression Anxiety Stress Scale-21 (Lovibond & Lovibond, 1995), the 5-item Satisfaction with Life Scale (Diener et al., 1985), the 20-item Positive and Negative Affect Scale (PANAS; Watson et al., 1988), the Multidimensional Perfectionism Scale Short Form (Hewitt & Flett, 1991), the Revised Frost Multidimensional Perfectionism Scale (Frost et al., 1990) and the Flourishing Scale (Diener et al., 2010), in that order. The study took participants a median of 12 minutes (IQR 7 min) to complete. Response time per question was not measured. After completing the whole survey, participants could receive course credit if eligible, and/or enter a draw to win $100 gift card. Only one gift card was given out as a prize.

Data Analysis

Data were cleaned and restructured in SPSS 25 software. Data were analyzed and plotted in R. First, descriptive statistics (i.e., counts and proportions) for each of the five Likert scale options were reported, and raw data were visualized using stacked bar plots of proportions. Next, five dummy-coded outcome variables were created from the 1-5 Likert scale data: (SA) *Strongly Agree* = 1, else = 0, (A) *Agree* = 1, else = 0, (N) *Neutral* = 1, else = 0, (D) Disagree = 1, else = 0, (SD) *Strongly Disagree* = 1, else = 0.

Data were analyzed using generalized linear mixed models with a binomial family and a logit link using the glmm function of the lme4 package in R.[2] Random intercepts were specified for participant ID and item, consistent with recommendations from Barr et al. (2013). Random slopes were not included, as our design was between-subjects, so condition and OCQ accuracy/bias scores did not vary across participant or item.

OCQ Accuracy and bias were included in separate models to avoid collinearity, as these two predictors are derived from the same questionnaire and are highly correlated, $r = .44$, 95% CI = [−.49, -.38]. Prior to analysis, condition was deviation coded as −0.5 (Descending) and 0.5 (Ascending) and OCQ accuracy and bias were mean centered. Interaction terms were calculated using the multiplicative product of condition and accuracy or bias. Centering in this manner improves interpretation of coefficients in the

2 A-priori, we had planned to use ordinal regression. However, these data did not meet the proportional odds assumption, as will be clear from data visualizations in the results. We also considered multinomial logistic regression as an alternative; however, given the absence of an obvious reference category for our hypotheses, we instead chose the method described above.

presence of an interaction term. In the presence of an interaction effect, the coefficient for condition should be interpreted as the effect of condition on the outcome, when accuracy/bias = 0 (i.e., the mean, since the data are mean centered). The coefficients for accuracy/bias should be interpreted as the effect of accuracy/bias on the outcome, collapsing across both conditions (i.e., like a classical main effect). Thus, we ran a total of 10 generalized linear mixed models (five models for each of the five Likert scale options, with separate models for OCQ accuracy and OCQ bias scores). Results of these analyses were visualized using model-predicted probability plots. The equation for each model broadly follows this form:

$$Y = b_0 + S0_s + I0_i + b_1Condi + b_2OCQ + b_3Cond*OCQ + e$$

Where b_0 = intercept, $S0_s$ = random intercept for subject, $I0_i$ = random intercept for item, b_1 = slope for condition, b_2 = slope for OCQ accuracy or bias, and b_3 = slope for the interaction effect, and e = error. Two effect sizes are reported. The marginal R^2 values refer to the proportion of variance explained by the fixed effects only, and the conditional R^2 refers to the variance explained by both fixed and random effects (Nakagawa, Johnson, & Schielzeth, 2017).

Results

Descriptive statistics are presented in Table 1. Overall, participants were most likely to select "Agree" to most items, and *Strongly Disagree* was the least likely to be endorsed. Nonetheless, there was substantial variation in responding to questions, and little missing data (0.4%). Raw data were plotted using bar plots of proportions grouped by condition and questionnaire (Figure 1). Proportions in Figure 2 are calculated from totals within condition to improve interpretation in the presence of unequal sample sizes across conditions (e.g., 4076 *Strongly Disagrees* in the ascending condition / 32,566 total observations in the ascending condition = 13%). Inspection of this plot suggests that participants are not endorsing "Strongly Agree" more frequently in the descending condition, as hypothesized. There appears to be a slight trend towards endorsing *Strongly Disagree* more often in the ascending condition and "Agree" in the descending condition, but effects are extremely small. However, there appears to be variation across questionnaires, as the effect of condition

Table 1. *Descriptive Statistics*

	Count	%
Strongly Disagree	7,403	11.7%
Disagree	14,005	22.1%
Neutral	12,278	19.3%
Agree	19,827	31.3%
Strongly Agree	9,696	15.3%
Missing	239	0.4%
Total	63,468	100%

*Note. Number of observations is derived from 774 participants * 82 items = 63,468.*

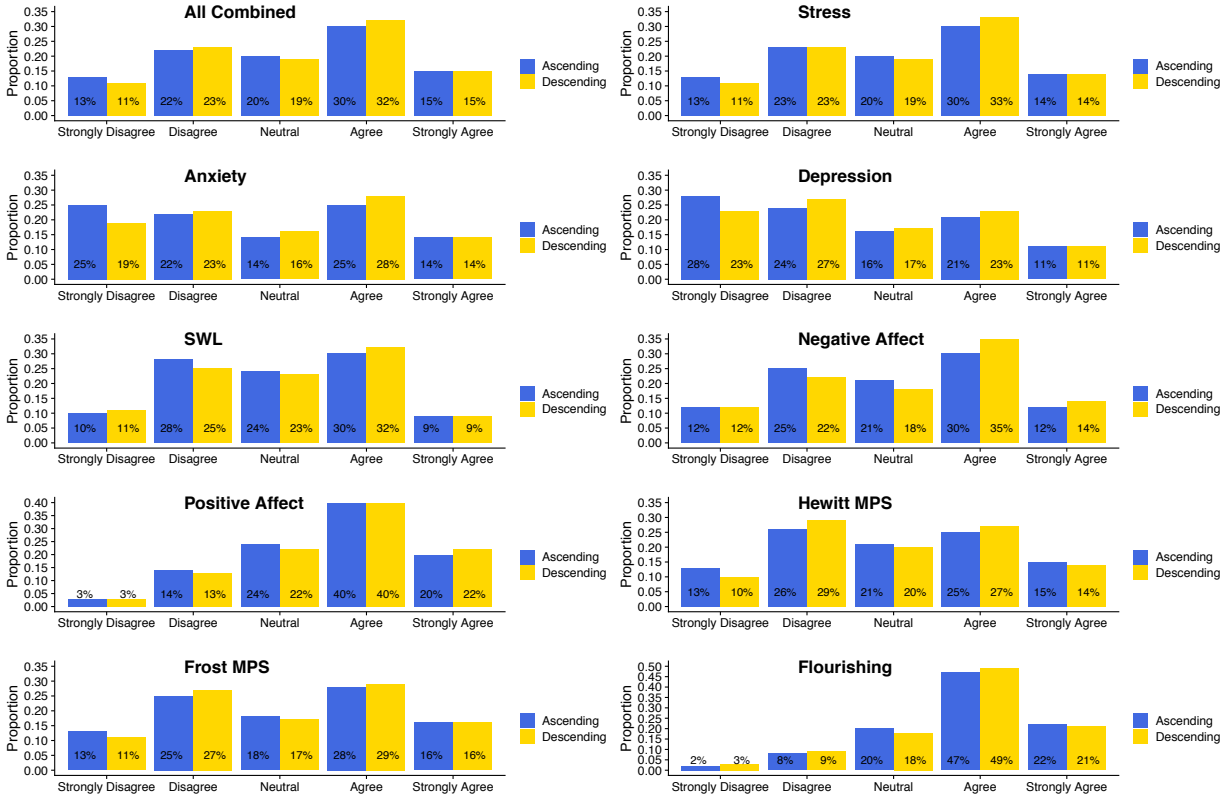

Figure 1. *Bar plots of raw data showing the levels of agreement, split by questionnaire type and experimental condition.*

appears strongest in the DASS-21. This variation across questionnaire supports our choice to include a random variance component for item in subsequent analyses.

Coefficients for generalized linear mixed models are presented in Tables 2 and 3. The ICCs for participant range from 0.09 to .28 and the ICCs for item range from .02 to 20, suggesting that there is generally some consistency across participant and items. This supports our choice to incorporate random components for both participant and item. Consistent with the bar plots in Figure 1, participants in the ascending condition were ~1.3 times less likely to select *Strongly Disagree*, and ~1.1 times more likely to select Agree. Broadly speaking, this fails to support H1. OCQ accuracy and bias had small main effects when predicting some outcomes. These effects were not predicted a-priori but, are of some interest as exploratory analyses. Participants were more likely to select *Strongly Disagree* as OCQ accuracy increased (OR = 1.23) and as OCQ bias decreased (OR = 0.81). Participants were also slightly less likely to select "Neutral" as OCQ accuracy increased (OR = 0.89). All the multiplicative interaction terms were non-significant, failing to support H2. That is, the effects observed for condition did not vary as OCQ accuracy or bias changed. Broadly, the marginal R^2 values were very small (< .01) while the conditional R^2 values tended to be much larger (from .13 to .46). This shows that experimental condition and the OCQ had little predictive power; most of the variance predicted can be accounted for by the random effects for item and participant. Model-predicted probability plots for each of the ten models are presented in Figure 2. These visualizations confirm that the effect sizes are quite small. Moreover, these visualizations show that high levels of OCQ accuracy/bias were comparatively rare; thus, the model may not generalize well to people with high levels of accuracy or bias.

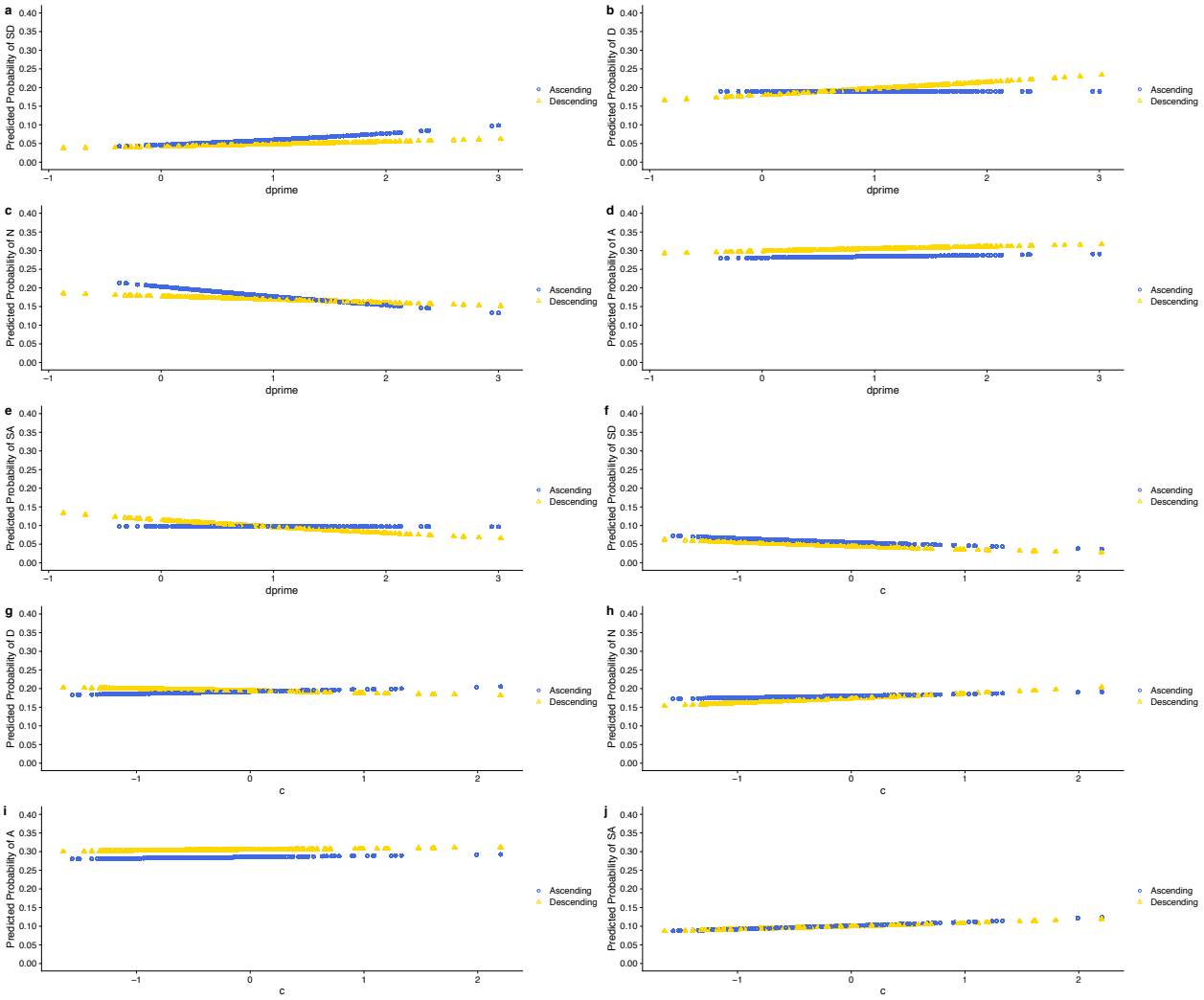

Figure 2. *Model-predicted probability plots from the logistic regression analyses. The y-axis is the model-predicted probability of selecting a given response (SD = Strongly Disagree; D = Disagree; N = Neutral; A = Agree; SA = Strongly Agree). The x-axis refers to either accuracy (d') or bias (c). The yellow dots refer to the descending condition and the blue dots refer to the ascending condition.*

Despite a few statistically significant findings, the lower bound of the confidence intervals for the odds ratios are nearly 1.0 and the *p*-values were between .01 to .05 in all cases. Given the number of tests and relatively large number of observations, this gives low confidence in the reproducibility of the statistically significant effects we did find. Indeed, a slightly more stringent *p*-value criterion of .01 would reduce all findings to non-significance. Broadly then, results fail to support our hypotheses.

Discussion

Satisficing theory (Krosnick, 1991; Krosnick and Presser, 2010) proposes that participants will choose the first acceptable answer presented to them in a list. A previous study with similar methodology (Mackinnon and Firth, 2018) found that participants were more likely to choose "Strongly Agree" when it is presented as the first response option (i.e.,

Table 2. *Generalized linear mixed models (logistic) with d' and condition predicting Likert scale options*

	Predictor	OR	95% CI	p	R_m^2 / R_c^2	ICC_{ID}	ICC_{ITEM}
SD	Condition	1.26	1.04 – 1.53	**0.017**	0.005 / 0.462	0.26	0.20
	Accuracy	1.23	1.04 – 1.45	**0.016**			
	Interaction	1.14	0.82 – 1.59	0.438			
D	Condition	0.96	0.88 – 1.06	0.437	0.001 / 0.176	0.09	0.09
	Accuracy	1.06	0.97 – 1.15	0.184			
	Interaction	0.90	0.76 – 1.05	0.185			
N	Condition	1.07	0.97 – 1.18	0.203	0.002 / 0.130	0.11	0.02
	Accuracy	0.89	0.82 – 0.97	**0.010**			
	Interaction	0.90	0.75 – 1.07	0.223			
A	Condition	0.91	0.83 – 0.99	**0.028**	0.001 / 0.145	0.08	0.07
	Accuracy	1.02	0.95 – 1.10	0.548			
	Interaction	0.99	0.85 – 1.15	0.874			
SA	Condition	1.01	0.85 – 1.21	0.886	0.001 / 0.357	0.27	0.08
	Accuracy	0.90	0.77 – 1.05	0.194			
	Interaction	1.22	0.89 – 1.66	0.216			

Note. $N_{observations}$ *= 63,229.* $N_{participants}$ *= 774. OR = Odds ratio. 95% CI = 95% Confidence interval.* R_m^2 *= marginal* R^2*.* R_c^2 *= conditional* R^2*.* ICC_{ID} *= Intraclass correlation for participant ID.* ICC_{ITEM} *= intraclass correlation for questionnaire item. SD = Strongly Disagree; D = Disagree; N = Neutral; A = Agree; SA = Strongly Agree. Condition was coded as −0.5 (descending) and 0.5 (ascending). Accuracy was mean centered prior to analysis. Interaction is the multiplicative product of condition and Accuracy.*

descending order). We did not replicate this effect, failing to support H1. Instead, results showed a very small tendency for participants to choose *Strongly Disagree* more often when it was presented first. This is still in line with satisficing theory, but it is not clear why results for *Agree* are different from Mackinnon and Firth (2018). This discrepancy could be due to using a different set of items – in the present study, we examined perfectionism and well-being vs. personality, alcohol use and motives in Mackinnon and Firth (2018). However, a recent study may shed light on the discrepancy. Terentev and Maloshonok (2019) examined primacy effects in a large sample (N = 22,910) of online students responding to a post-course survey. They found evidence of primacy effects when questions were presented in a vertical, item-by-item layout (i.e., top-to-bottom for response options and one item per page) with more participants endorsing "I am completely new to this subject area" when it was presented first (35% vs. 30%). However, when items were presented in a grid format (i.e., left-to-right with multiple items on a single page, as in the present study; see https://osf.io/aec25/), they found that *Strongly Disagree* was less likely to be selected when it was presented first. This too does not match our findings – nor Mackinnon and Firth (2018) who also primarily used a grid format – but it does highlight that the primacy effect may be less consistent when questions are presented in a grid format relative to an item-by-item format.

Accuracy and bias on the OCQ did not moderate the effect of experimental condition, failing to support H2 and H3. Since the OCQ was measured only once per participant, analyses may have been underpowered. Moreover, the university student sample may have introduced selection bias – indeed, Figure 2 shows that few participants

Table 3. *Generalized linear mixed models (logistic) with c and condition predicting Likert scale options*

	Predictor	OR	95% CI	p	R_m^2 / R_c^2	ICC_{ID}	ICC_{ITEM}
SD	Condition	1.25	1.03 – 1.51	**0.022**	0.005 / 0.462	0.26 $_{ID}$	0.20 $_{item}$
	Bias	0.81	0.70 – 0.95	**0.011**			
	Interaction	1.04	0.75 – 1.42	0.829			
D	Condition	0.96	0.88 – 1.06	0.432	0.000 / 0.176	0.09 $_{ID}$	0.09 $_{item}$
	Bias	1.00	0.93 – 1.08	0.983			
	Interaction	1.08	0.92 – 1.26	0.353			
N	Condition	1.07	0.97 – 1.19	0.179	0.001 / 0.130	0.11 $_{ID}$	0.02 $_{item}$
	Bias	1.06	0.98 – 1.16	0.155			
	Interaction	0.95	0.80 – 1.12	0.522			
A	Condition	0.91	0.83 – 0.99	**0.030**	0.001 / 0.145	0.08 $_{ID}$	0.07 $_{item}$
	Bias	1.01	0.95 – 1.09	0.687			
	Interaction	1.00	0.87 – 1.16	0.970			
SA	Condition	1.02	0.85 – 1.21	0.853	0.001 / 0.357	0.28 $_{ID}$	0.08 $_{item}$
	Bias	1.10	0.95 – 1.28	0.187			
	Interaction	1.01	0.75 – 1.36	0.940			

Note. $N_{observations}$ *= 63,229.* $N_{participants}$ *= 774. OR = Odds ratio. 95% CI = 95% Confidence interval.* R_m^2 *= marginal* R^2*.* R_c^2 *= conditional* R^2*. ICCID = Intraclass correlation for participant ID.* ICC_{ITEM} *= intraclass correlation for questionnaire item. SD = Strongly Disagree; D = Disagree; N = Neutral; A = Agree; SA = Strongly Agree. Condition was coded as −0.5 (descending) and 0.5 (ascending). Bias was mean centered prior to analysis. Interaction is the multiplicative product of condition and Bias.*

had high levels of accuracy or bias. Nonetheless, the non-significant results leave us in an inconclusive position; we do not know whether accuracy/bias plays a moderating role in this process.

Though we did not have hypotheses for main effects of accuracy and bias, these exploratory findings are worth discussing briefly. Specifically, there were main effects of accuracy and bias on the probability of choosing *Strongly Disagree*. That is, participants who were higher in accuracy or lower in bias were more likely to choose *Strongly Disagree*. This might suggest that the participants in the "ascending" condition are responding more accurately. That is, the slight increase in SD responses in the descending condition might reflect more accurate or truthful responses. This is an interesting consideration, given that it is presently unknown which participants' data are more valid (i.e., the ascending vs. descending conditions). Nonetheless, it is an exploratory finding pending future replication.

This study has numerous limitations. Young, Western university student samples have well-known limits to external validity. Because we used a grid format for question presentation, results may not generalize to item-by-item stimulus presentation. Our use of a between-subjects design did not allow for random slopes and may have reduced statistical power. Further, our choice of items was arbitrary, so results may not generalize to other questionnaire items. Moreover, the OCQ is not a true measure of cognitive ability; indeed, what it measures may be a mix of social desirability concerns, cognitive ability, and careless responding (Ludeke and Makransky, 2015; Palhus & Harms, 2004). Indeed, some

researchers have found that OCQ accuracy is so weakly related to intelligence as to be better considered an entirely different construct (Hülür, Wilhelm & Schipolowski, 2011).[3] Thus, it is likely that the failure to support hypotheses is also due to problems with measurement; to the extent that the overclaiming questionnaire is a poor proxy for cognitive ability, our ability to detect a signal from noise is greatly impaired. Future research may wish to explore a within-subjects design with item-by-item stimulus presentation and a better measure of cognitive ability (e.g., IQ).

Broadly speaking, we did not find strong support for primacy effects and satisficing theory in this experiment. If such an effect exists in this context, it is very small and potentially of little practical importance. Nonetheless, we present some preliminary exploratory evidence that the OCQ may add value to this literature. If future research shows it can reliably predict responding on questionnaires, it may be useful as a way to determine which participants are providing the most accurate/truthful data. Such an advance would be of considerable benefit and is worth exploring further.

References

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255–278. doi:10.1016/j.jml.2012.11.001

Bland, M. (2004, April 6). Sample size in guidelines trials. Retrieved from https://www-users.york.ac.uk/~mb55/clust/bupa.htm

Cox, B. J., Enns, M. W., & Clara, I. P. (2002). The multidimensional structure of perfectionism in clinically distressed and college student samples. *Psychological Assessment, 14*, 365–373. doi:10.1037/1040-3590.14.3.365

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment, 49*, 71–75. doi:10.1207/s15327752jpa4901_13

Diener, E., Wirtz, D., Tov, W., Kim-Prieto, C., Choi, D., Oishi, S., & Biswas-Diener, R. (2010). New well-being measures: Short scales to assess flourishing and positive and negative feelings. *Social Indicators Research, 97*, 143–156. doi:10.1007/s11205-009-9493-y

Frost, R. O., Marten, P., Lahart, C., & Rosenblate, R. (1990). The dimensions of perfectionism. *Cognitive Therapy and Research, 14*, 449–468.

Garbarski, D., Schaeffer, N., & Dykema, J. (2015). The effects of response option order and question order on self-rated health. *Quality of Life Research, 24*, 1443-1453. doi:10.1007/s11136-014-0861-y

Hewitt, P. L., & Flett, G. L. (1991). Perfectionism in the self and social contexts: Conceptualization, assessment, and association with psychopathology. *Journal of Personality and Social Psychology, 60*, 456–470. doi:10.1037/0022-3514.60.3.456

Hülür, G., Wilhelm, O., & Schipolowski, S. (2011). Prediction of self-reported knowledge with over-claiming, fluid and crystallized intelligence and typical intellectual engagement. *Learning and Individual Differences, 21*, 742–746. doi:10.1016/j.lindif.2011.09.006

Hülür, G., Wilhelm, O., & Schipolowski, S. (2014). "Prediction of self-reported knowledge with over-claiming, fluid and crystallized intelligence and typical intellectual engagement": Corrigendum. *Learning and Individual Differences, 35*, 153–154. doi:10.1016/j.lindif.2014.07.014

Kam, C., Risavy, S. D., & Perunovic, W. Q. E. (2015). Using over-claiming technique to probe social desirability ratings of personality items: A validity examination. *Personality and Individual Differences, 74*, 177–181. doi:10.1016/j.paid.2014.10.017

3 However, note also the erratum for this article (Hülür, Wilhelm & Schipolowski, 2014) where the correlations between overclaiming accuracy and crystalized intelligence are larger than originally reported due to a database misalignment issue.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*, 213–236. doi:10.1002/acp.2350050305

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. The *Public Opinion Quarterly, 51, 201–219. doi:10.1086/269029*

Krosnick, J. A., & Presser, S. (2010). Questionnaire design. In J. D. Wright & P. V. Marsden (Eds.), *Handbook of survey research* (2nd ed.). West Yorkshire, England: Emerald Group.

Lovibond, P. F., & Lovibond, S. H. (1995). The structure of negative emotional states: Comparison of the depression anxiety stress scales (DASS) with the beck depression and anxiety inventories. *Behaviour Research and Therapy, 33*, 335–343. doi:10.1016/0005-7967(94)00075-U

Ludeke, S. G., & Makransky, G. (2016). Does the over-claiming questionnaire measure overclaiming? Absent convergent validity in a large community sample. *Psychological Assessment, 28*, 765–774. doi:10.1037/pas0000211

Mackinnon, S. P., & Firth, S. (2018). The effect of question structure on self-reported drinking: Ascending versus descending order effects. *Journal of Research in Personality, 73*, 21–26. doi:10.1016/j.jrp.2017.10.004

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). New York: Cambridge.

Malhotra, N. (2008). Completion time and response order effects in web surveys. *Public Opinion Quarterly, 72, 914–934*. doi:10.1093/poq/nfn050

Nakagawa, S., Johnson, P. C., & Schielzeth, H. (2017). The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface, 14*, 1–11. doi:10.1098/rsif.2017.0213

Object Planet. (2018). Opinion survey software [computer software] Retrieved from http://www.objectplanet.com/opinio/

Osman, A., Wong, J. L., Bagge, C. L., Freedenthal, S., Gutierrez, P. M., & Lozano, G. (2012). The depression anxiety stress Scales—21 (DASS-21): Further examination of dimensions, scale reliability, and correlates. *Journal of Clinical Psychology, 68*, 1322–1338. doi:10.1002/jclp.21908

Paulhus, D. L., & Dubois, P. J. (2014). Application of the overclaiming technique to scholastic assessment. *Educational and Psychological Measurement, 74*, 975–990. doi:10.1177/0013164414536184

Paulhus, D. L., & Harms, P. D. (2004). Measuring cognitive ability with the overclaiming technique. *Intelligence, 32*, 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*, 890–904. doi:10.1037/0022-3514.84.4.890

Pavot, W., Diener, E., Colvin, C. R., & Sandvik, E. (1991). Further validation of the satisfaction with life scale: Evidence for the cross-method convergence of well-being measures. *Journal of Personality Assessment, 57*, 149–161. doi:10.1207/s15327752jpa5701_17

Pesta, B. J., & Poznanski, P. J. (2009). The inspection time and over-claiming tasks as predictors of MBA student performance. *Personality and Individual Differences, 46*, 236–240. doi:10.1016/j.paid.2008.10.005

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers, 31*, 137–149. doi:10.3758/BF03207704

Terentev, E., & Maloshonok, N. (2019). The impact of response options ordering on respondents' answers to rating questions: results of two experiments. International *Journal of Social Research Methodology, 22*, 179–198. doi:10.1080/13645579.2018.1510660

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology, 54*, 1063–1070.

Ziegler, M., Kemper, C., & Rammstedt, B. (2013). The Vocabulary and Overclaiming Test (VOC-T). *Journal of Individual Differences, 34*, 32–40. doi:10.1027/1614-0001/a000093

Received: 7.15.2019
Revised: 10.30.2019
Accepted: 10.31.2019