Reviewer Comments:

Comment 1: I think this is a valuable research question, but the way the study was carried out might have compromised the ability to test one of the hypotheses. Also, results went opposite the hypothesized direction for one of the hypotheses and that finding was not addressed. The snowball sampling method appears to not have resulted in a strongly differentiated "twice-removed" sample. Maybe if this is tested again the authors could ask the target to identify people who know them well and people who know them less well. The strongest predictor of skipping among the readability indices gave a result that was opposite the other indices, and opposite the hypothesized effect direction. Why was that? Is there an important difference in how the indices calculate readability maybe? The interaction depicted in Figure 2 was not significant, so the commonly accepted rule is it shouldn't be inspected or interpreted. If there is a reason to make an exception to that rule, it should be explained in the figure caption.

Comment 2: Overall, I thought this was an interesting study that aims to shed more light on why items are left incomplete. However, the fact that so few items had non-responses does detract from the contribution the study could make. I have a few comments below. A related issue for big five personality informant ratings is that the informant ratings tend to have higher correlations (and less discriminant validity) across the big five than self ratings. I'm not clear on the rationale for the once and twice removed raters. It is possible that the second rater knows the participant as well as the first or even better as is evidenced by the chi-square analyses in the beginning of the Result section. I think a better design would be to have many other/informant raters for each participant and then have the participants and the informants rate how well the other/informant knows the participant using some type of behaviorally-anchored rating scale. The fact that 22% of the data were screened out is concerning. It is also not clear why the full 460 once-removed matched dataset was not used in the analyses. In terms of readability, I think there are other aspects of this that are not captured in the readability indices. Word choice, writing quality, grammar, syntax, clarity, etc. are also important. It is a little concerning that so few prompts had a non-response (1.5%). This suggest that the general problem that the study aims to address (i.e., determining what factors lead to non-response) is not a significant one since 98.5% of items have a response. It's also not clear what is meant by an item that was responded to. Are these items that had a 100% response rate across all subjects or were the t-tests conducted within each subject? I wonder if some type of logistic regression might have been a better analysis option. The DV in this case would be whether the item had a response and the IVs would be the readability and social desirability scores. Although if each participant received one of four different forms, then the analysis would have to conducted on each of the four subsamples. I agree with the future research suggestion that larger databases with more missing data could be of interest. Perhaps vendors like Hogan or others might have this data available.

SUBMISSION TYPE

Poster

TITLE

Predicting Selective Non-response via Item Characteristics

SHORTENED TITLE

Survey Non-response

ABSTRACT

This paper documents item non-response and its association with social desirability. Results from self as well as other-ratings exhibit similar patterns of missingness that become more common as items become increasingly saturated with undesirable content.

PRESS PARAGRAPH

Although missing data is commonly observed, the *reason* for it, or type of missing data is an important consideration. This paper explores patterns of survey non-response (“skipping” questions) by item characteristics including item desirability and indices of readability, as well as the occurrence of missingness across different rating sources (i.e., from the target, and two informant raters, one of which was not selected by the target him or herself). The results suggest that item “skips” are one strategy used (by all raters) to avoid responding to undesirable items, and that this effect is as strong or stronger than that of item readability.

WORD COUNT

3,000

Researchers have long grappled with the issue of missing data, which result from a very wide range of sources (e.g., fatigue, inattention, intentional “skipping” of questions or subscales, administration or data entry errors). Although missing data is commonplace with survey data, it is also problematic - particularly (we contend) in studies that rely on multiple informants to provide information about a target (Blozis et al., 2013; Phares, 1992). Of particular interest here is a basic question: Are there non-random, predictable, and volitional reasons why these informants *do not* respond to particular item prompts? The current paper aims to explore patterns of missing data by item characteristics including item desirability and indices of readability, as well missingness across different rating sources (i.e., from the target, and two informant raters, one of which was not selected by the target him or herself). Uncovering patterns in missingness would suggest that the data are not missing completely at random, but rather point toward a systematic bias – one that would call for our attention in the area of personality assessment.

**Missingness Mechanisms: Reasons for Missing Data**

Within the survey response literature, there are three widely agreed upon classifications of missing data. First, missing data can be completely at random (MCAR), which means there are not systematic differences between observed and missing data. Stated another way, “missingness” is entirely independent of the observed and the missing data. Data can also be missing at random (MAR), in which missingness is independent of the missing data, but it is dependent on the observed data. Blozis et al. (2013) provide a clear example of how this might occur. If studying father reports of marital satisfaction and some are missing, they could be “missing at random,” for fathers with low levels of self-esteem who are less likely to report their marital satisfaction. Thus, those missing data depend not on someone’s marital satisfaction, but rather their self-esteem. Data can also be missing not at random (MNAR). Here, missingness is tied to the missing data even after accounting for dependencies of the missingness on the observed data.

**The Use of Informant Ratings in Personality Assessment**

The historical focus on the detection and mitigation of method effects (of which we consider MNAR and MAR to be specific manifestations) in psychological assessment grew with Campbell and Fiske’s (1959) MTMM approach. Others have continued to broaden the scope of investigation since (e.g., Kammeyer-Mueller et al., 2010; Podsakoff et al., 2012; Podsakoff et al., 2003). Within these frameworks, informant raters are commonly considered unique sources of information, and they have been used to bypass the widely acknowledged measurement deficiencies in self-reports (e.g., Berry et al., 2012; Borkenau, & Ostendorf, 1989; Connelly & Ones, 2010; Connolly et al., 2007; John & Robins, 1993; Kammeyer-Mueller et al., 2010; Kolar et al., 1996; Leising et al., 2015; Letzring et al., 2006; Mount et al., 1994; Oh et al., 2011; Paunonen, & Kam, 2014). Podsakoff et al. (2003), in fact, recommend obtaining measures of independent and dependent variables from different sources to prevent any one individual rater from introducing associations due to intra-person factors such as consistency motifs or mood states. In line with this recommendation, Chang et al. (2012) applied meta-analysis to informant ratings, treating these sources as distinct methods of assessment. Conway and Lance (2010), however, argue quite strongly that researchers (and reviewers) have often misconstrued rating *source* as a measurement *method*. They further contend that informants/rating sources may, in fact, be prone to the same biases as self-raters – we also evaluate this possibility in the current study.

**Informants versus Selves – Divergent or Comparable Response Orientations**

Theory suggests that people have an unconscious desire to present themselves in ways they would like others to perceive them (Hogan & Blickle, 2018). Not unexpectedly, similar motives, when considered from the perspective of fellow members of one’s “in group” (social identity theory; Tajfel & Turner, 1979), would compel us to view those in our inner circles in a positive light (e.g., Murray et al., 1996). This theory runs counter to elements in the personality and common method variance literatures that argue the use of informant ratings is a viable strategy to mitigate measurement artifacts. For example - self-ratings of job performance or organizational citizenship behaviors have received criticism for being methodologically flawed. Subsequently, many investigations and meta-analyses have investigated the comparative equivalence of self- and informant-ratings of these constructs, with the implication being that self-informant agreement represents evidence that self-ratings are acceptable (see, for example, the meta-analyses of Carpenter et al., 2014 and Heidemeier & Moser, 2009). In these studies, rating alignment is taken as a comfort. This implication of course depends on the quality of the comparative standard.

Alternatively, evidence of equivalence *could* indicate that the rating sources are merely susceptible to similar flaws. In fact, others *have* acknowledged this potential limitation in using informant ratings. Our interest in the current study is whether self and informant non-response can be predicted by item content. Researchers have suggested methods by which we might tease apart ratings of targets’ actual characteristics and perceivers’ evaluation of the positivity or negativity of item/characteristic (e.g., Leising et al., 2015; Peabody, 1967). Within other applications, researchers have documented that respondents provide socially desirable ratings of both significant others (e.g., Brown & Han, 2012), and children (e.g., Lench et al., 2006). This tendency to rate others that we like more favorably has in some contexts been referred to as the “letter of recommendation problem” (Klonsky et al., 2002).

**Current Study and Hypotheses**

Alicke and Sedikides (2011) proposed self-enhancement and self-protection motives that describe healthy conceptualizations of the self. Self-enhancement refers to “a tendency to claim greater standings on a characteristic, or more credit, than is objectively warranted” (p. 2). Self-protection refers to “tactics that are adopted to avoid falling below a desired standard” (p. 2). These motives suggest that a target’s view of themselves may be reasonably expected to be positively biased. However, Alicke and Sedikides (2011) also proposed that these motives likely extend to others in whom one is invested (e.g., family). This theoretical foundation regarding the way in which people, whether consciously or unconsciously, view themselves also therefore frames how we might view others in whom we have some personal investment.

The current study extends this perspective to examine potential reasons for missing data. Theory would suggest that both the target and informant ratings are likely to “skip” over particularly undesirable questions with greater frequency than those that are qualitatively more desirable. We explore this question taking a variety of readability indices into account as well given the research on missing data suggests that there is an association between missing data and reading level (e.g., see Fongwa et al., 2010). We made the following predictions:

*H1: Item social desirability is negatively associated with non-response.*

*H2: Item readability is negatively associated with non-response.*

*H3: Item social desirability will be more strongly related to non-response than readability.*

*H4: Those that are more distant from a target (twice removed) will be less influenced by item desirability.*

**Method**

**Participants and Procedure**

A total of 888 undergraduate students from 3 US universities were recruited via an online research participation website in exchange for course credit. Upon completion of personality and demographic questions, participants were asked to provide the email address of someone else who could make ratings of their personality. An invitation was automatically generated to the person identified, and this “once-removed” rater was then asked to rate the personality of the original participant using the same item set. Once more, this rater was asked to provide the email address of a second person to make ratings of the original participant (we refer to this source as “twice-removed”). From our 888 original student participants, we realized 460 once-removed raters and 131 twice-removed raters for an experiment-wide totalof 1,479 respondents. These decreasing sample sizes were expected based on the nature of the nomination process (email invitation) and lack of researcher control regarding opportunities for response rate enhancement. Conservative screening of the data resulted in 306 raters matched across self, once-removed, and twice-removed sources (all *n*’s = 102)[[1]](#footnote-1). Upon survey completion, participants were thanked and compensated via course credit (original student) or the opportunity to take part in a drawing for one of three gift cards (once- and twice-removed raters).

**Materials**

*Item Social Desirability*. Ratings of the social desirability content of 2,413 personality indicators from the International Personality Item Pool (http://www.ipip.org) were made along a 9-point scale ranging from: (-4) *extremely undesirable* to (+4) *extremely desirable* (see, for example, Edwards, 1957). Our graduate student raters (*n* = 6) were provided with definitions of both ends of the social desirability continuum prior to making ratings. Inter-rater consistency estimates using a 2-way random model was .96.

*Readability.* Four readability indices were used here: 1)Easy Listening Formula (ELF; Fang 1966). This simple index divides the number of 2+ syllable words by the number of sentences, where a score below 12 indicates a statement that is easy to listen to and by extension, read to oneself; 2) The Navy’s Adaptation of Gunning’s Fog Index - FOG Navy Readability Index (Kincaid et al., 1975) takes into account number of words, sentences, and syllables. The index estimates the years of formal education a person needs to understand the text on the first reading; 3) the Flesch-Kincaid Grade Level (Flesch & Kincaid, 1975) was designed to indicate how difficult a passage in English is to understand. The formula uses both average word length and sentence length to determine reading level, and does not have an upper bound; and 4) the New Dale-Chall Formula builds on the Flesch reading ease formula, and is used for both children above 4th grade and adults (Chall & Dale, 1995). The formula uses a count of “hard” words and sentence length to compute the US grade level of a piece of text. Adjusted scores of < 4.9 are rated at 4th grade level or below. Adjusted scores of 10+ are rated at the college reading level.

*Personality measure(s).* Four stratified random samplings of the 2,413 IPIP items were conducted. The strata were defined via the 9-point rating scale: stratum #1 = social desirability values below -3.01, stratum #2 = -3 to -2.01, stratum #3 = -2 to -1.01, stratum #4 = -1 to -.01, stratum #5 = 0 to 1, stratum #6 = 1.01 to 2, stratum #7 = 2.01 to 3, and stratum #8 = above 3.01 (e.g., *extremely desirable*). Ten items were randomly sampled four times from within each social desirability-defined stratum, and the very large number of original items permitted sampling without replacement. Thus, participants completed only one of four completely different versions of the inventory (all 320 administered items were completely unique). All participants made typical personality ratings on a 5-point Likert-type agreement scale ranging from *strongly disagree* to *strongly agree* to 80 items. These items were presented in fully randomized orderings. In addition to the five substantive response categories, a *not applicable* option was also provided. Here, the goal was to predict non-response, and a not applicable reponse was considered equivalent to choosing to “skip” a question.

**Results**

As a check on possible differences in familiarity with the target, all once- and twice-removed raters were asked both how well they knew as well as how frequently they interacted with the target. There were slight differences on responses to these questions across groups (“well-known” (3) = 10.03, *p* = .018; “frequently” (3) = 16.91, *p* = .001), but the differences were not sufficiently meaningful to justify inclusion of these variables into our analyses (87% of twice-removed and 95% of once-removed indicated they knew the target either *very well* or *extremely well*, although 14% of twice-removed respondents indicated they only interacted with the target once a month or less frequently compared to 2% of once-removed respondents).

A total of 24,114 (98.5%) prompts received a response; 366 (1.5%) prompts resulted in a non-response (including missing or not applicable). A series of t-tests explored the question of whether responding or not (the IV in this case) differed based on item characteristics (e.g., desirability of the item, readability). Figure 1 illustrates differences for all item characteristics, which are presented as z-scores for comparative purposes. First, item desirability was significantly higher for items that were responded to (*M* = 4.86, *SD* = 2.24) than those that were skipped (*M* = 4.11, *SD* = 2.10), *t*(377.71) = 6.73, *p* < .001, *d* = .33. In addition to item desirability, a number of readability indices were explored, the first of which was the Easy Listening Formula (ELF; Fang 1966). Here, the number of 2+ syllable words in a prompt is taken into account. The results suggest that ELF was significantly lower for items that were responded to (*M* = 2.03, *SD* = 1.43) than those that were skipped (*M* = 2.40, *SD* = 1.66), *t*(373.29) = -4.15, *p* < .001, *d* = .25. The Navy’s Adaptation of Gunning’s Fog readability index, which explores 3-syllable words, was also explored. The results suggest that Fog-NRI readability was significantly lower for items that were responded to (*M* = -1.24, *SD* = 0.30) than those that were skipped (*M* = -1.16, *SD* = 0.42), *t*(370.85) = -3.67, *p* < .001, *d* = .26.

The Flesch-Kincaid Readability Score (Kincaid et al., 1975) was also explored. The results suggest that Flesch-Kincaid readability was significantly easier for items that were responded to (*M* = 4.91, *SD* = 4.85) than those that were skipped (*M* = 5.46, *SD* = 4.16), *t*(380.26) = -2.48, *p* = .013, *d* = .11. Lastly, a second index of comprehension difficulty, the Dale-Chall Readability Score (Chall & Dale, 1995) was also explored. The results suggest that readability was significantly harder for items that were responded to (*M* = 42.87, *SD* = 17.94) than those that were skipped (*M* = 37.84, *SD* = 17.84), *t*(376.28) = -5.35, *p* < .001, *d* = .28. Directly comparing the qualitative item characteristics with ratings of socially desirable content, we applied logistic regression, treating missed and non-missed responses as binary outcomes. In this comprehensive approach, the strongest unstandardized predictors were the Dale-Chall (*b* = -.03, *z* = -6.63, *p <* .001) and item social desirability (*b* = -.15, *z* = -6.23, *p <* .001). The Fog-NRI was the only non-significant predictor of missingness (*b* = .37, *z* = 1.91, *p* =.060)

The second question of interest was whether or not item desirability influenced non-response the same or differently across rater type. Using only summaries of missingness within social desirability stratum, as well as an indication of source (self, once-removed, or twice-removed), the overall R^2 was .89 (*F*(5,18) = 28.18, *p* < .001).There were strong main effects for social desirability stratum (*F*(1,22) = 6.77, *p* = .016) as well as rater type (DeltaF(2,20) = 42.49, *p* < .001). The interaction effect (rater type x social desirability stratum) was non-significant (DeltaF(2,18) = 2.57, *p* = .100). See Figure 2. These results are in contrast with the expectation that those closer to a target would be more influenced by item desirability (“miss” a larger number of items).

**Discussion**

Self-report measures of psychological constructs are commonly criticized for their vulnerability to response biases (e.g., Spector, 2006). These perspectives regarding self-evaluations are likely not surprising to the reader who is familiar with the issue of social desirability in personality assessment (e.g., Asendorpf & Ostendorf, 1988; Dilchert et al., 2006; Edwards, 1957; Paunonen & LeBel, 2012) or common method variance across measurement domains (e.g., Kammeyer-Mueller et al., 2010; Podsakoff et al., 2003). Many researchers propose that by collecting informant ratings, they circumvent some of the psychometric concerns traditionally associated uniquely with self-ratings (see, for example, Chang et al., 2012; Kolar et al.,1996; Podsakoff et al., 2012; Robins & John, 1997). The current study explored whether the desirability of items could explain the choice to answer a question about a target 1) more strongly than the readability of the item prompts, and 2) for informant raters.

In sum, the results suggest social desirability does predict non-response, in this case, perhaps as much or more so than the readability of the items and generally across ratings sources, supporting H1-H3. These results are in contrast with the expectation that those closer to a target would be more influenced by item desirability (“miss” a larger number of items; H4). However, as noted above, all raters were quite familiar with the targets. Theory would actually suggest our obtained pattern of findings given how well informants (both once and twice removed) knew the targets they were rating. Interestingly, the pattern was nearly identical for targets and twice removed raters (moderate negative slope), but although the trend was in the same direction, this pattern was less dramatic for once removed raters. The reason for this difference is unclear, and somewhat contrary to theory would suggest - in that those close to a target would be quite similarly biased and by extension, “miss” undesirable questions with similar frequency.

**Limitations and Future Directions**

We note a number of limitations and subsequent directions for future work in this area. First, our project explored personality only. We note that missing data happens much more broadly, and for more reasons than were feasible to explore here. In addition, future study would benefit from consideration of a larger number of potential reasons for non-response, and in other contexts. The data here were also relatively complete (did not contain a large percentage of missing data). However, databases often contain a much larger percent of missing information. It would be of value to conduct a larger analysis across studies to document the percentage of missing data along to look for additional patterns of systematic non-response.

**References**

Alicke, M. D., & Sedikides, C. (2011). Historical overview and conceptual framework. In M. D. Alicke & C. Sedikides (Eds.), *Handbook of Self-Enhancement and Self-Protection.* (pp.1-19). New York: Guilford Press.

Asendorpf, J., & Ostendorf, F. (1998). Is self-enhancement healthy? Conceptual, psychometric, and empirical analysis. *Journal of Personality and Social Psychology, 74,* 955-966. https://doi.org/10.1037/0022-3514.74.4.955

Berry, C. M., Carpenter, N. C., & Barratt, C. L. (2012). Do other-reports of counterproductive work behavior provide an incremental contribution over self-reports? A meta-analytic comparison. *Journal of Applied Psychology, 97,* 613-636. https://doi.org/10.1037/a0026739

Blozis, S. A., Ge, X., Xu, S., Natsuaki, M. N., Shaw, D. S., Neiderhiser, J. Scaramella, L., Leve, L., & Reiss, D. (2013). Sensitivity analysis of multiple informant models when data are not missing at random. *Structural Equation Modeling, 20*(2), 283–298. https://doi.org/10.1080/10705511.2013.769393

Borkenau, P., & Ostendorf, F. (1989). Descriptive consistency and social desirability in self and peer reports. *European Journal of Personality, 3,* 31-45. https://doi.org/10.1002/per.2410030105

Brown, J. D., & Han, A. (2012). My better half: Partner enhancement as self-enhancement. *Social Psychological and Personality Science, 3*(4), 479-486. https://doi.org/10.1177/1948550611427607

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105. https://doi.org/10.1037/h0046016

Carpenter, N. C., Berry, C. M., & Houston, L. (2014). A meta-analytic comparison of self-reported and other-reported organizational citizenship behavior. *Journal of Organizational Behavior, 35,* 547-574. https://doi.org/10.1002/job.1909

Chall, J. S., & Dale, E. (1995). Readability revisited: The new Dale-Chall readability formula. Brookline Books.

Chang, L., Connelly, B. S., & Geeza, A. A. (2012). Separating method factors and higher order traits of the Big Five: A meta-analytic multitrait-multimethod approach. *Journal of Personality and Social Psychology, 102*, 408-426. https://doi.org/10.1037/a0025559

Connolly, J. J., Kavanagh, E. J., & Viswesvaran, C. (2007). The convergent validity between self and other observer ratings of personality: A meta-analytic review. *International Journal of Selection and Assessment, 15,* 110-117. https://doi-org/10.1111/j.1468-2389.2007.00371.x

Connelly, B. S., & Ones, D. S. (2010). Perspective on personality: Meta-analytic integration of observers’ accuracy and predictive validity. *Psychological Bulletin, 136,* 1092–1122. https://doi.org/10.1037/a0021212

Conway, J. M., & Lance, C. E. (2010). What reviewers should expect from authors regarding common method bias in organizational research. *Journal of Business in Psychology, 25,* 325-334. https://doi.org/10.1007/s10869-010-9181-6

Dilchert, S., Ones, D. S., Viswesvaran, C., & Deller, J. (2006). Response distortion in personality measurement: Born to deceive, yet capable of providing valid self-assessments? *Psychology Science, 48,* 209-225.

Edwards, A. L. (1957). *The social desirability variable in personality assessment and research.* New York: Dryden.

Fang, I. E. (1966). The easy listening formula. *Journal of Broadcasting & Electronic Media, 11*(1), 63–68.

Heidemeier, H., & Moser, K. (2009). Self-other agreement in job performance ratings: A meta-analytic test of a process model. *Journal of Applied Psychology, 9,* 353-370. https://doi.org/[10.1037/0021-9010.94.2.353](https://doi-org.ezproxy.lib.uwstout.edu/10.1037/0021-9010.94.2.353)

Hogan, R., & Blickle, G. (2018). Socioanalytic theory: Basic concepts, supporting evidence and practical implications. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *The SAGE handbook of personality and individual differences: The science of personality and individual differences* (p. 110–129). Sage Reference. https://doi.org/10.4135/9781526451163.n5

John, O. P., & Robins, R. W., (1993). Determinants of interjudge agreement on personality traits: The big five domains, observability, evaluativeness, and the unique perspective of the self. *Journal of Personality, 61,* 521-551. https://doi.org/10.1111/j.1467-6494.1993.tb00781.x

Kammeyer-Mueller, J., Steel, P. D. G., & Rubenstein, A. (2010). The other side of method bias: The perils of distinct source research designs. *Multivariate Behavioral Research, 45,* 294-321. https://doi.org/10.1080/00273171003680278

Kincaid, J.P., Fishburne, R.P., Rogers, R.L., & Chissom, B.S. (1975). Derivation of new readability formulas (automated readability index, fog count, and flesch reading ease formula) for Navy enlisted personnel. Research Branch Report 8–75. Chief of Naval Technical Training: Naval Air Station Memphis.

Klonsky, E. D., Oltmanns, T. F., & Turkheimer, E. (2002). Informant reports of personality disorder: Relation to self-reports and future research directions. *Clinical Psychology: Science and Practice, 9,* 300-311.

Kolar, D. W., Funder, D. C., & Colvin, C. R. (1996). Comparing the accuracy of personality judgments by the self and knowledgeable others. *Journal of Personality, 64,* 311-337. https://doi.org/10.1111/j.1467-6494.1996.tb00513.x

Leising, D., Scherbaum, S., Locke, K. D., & Zimmermann, J. (2015). A model of ‘‘substance’’ and ‘‘evaluation’’ in person judgments. *Journal of Research in Personality, 57,* 61-71. https://doi.org/10.1016/j.jrp.2015.04.002

Lench, H. C., Quas, J. A., & Edelstein, R. S. (2006). My child is better than average: The extension and restriction of unrealistic optimism. *Journal of Applied Social Psychology, 36*(12), 2963-2979. https://doi.org/10.1111/j.0021-9029.2006.00137.x

Letzring, T. D., Wells, S. M., & Funder, D. C. (2006). Information quantity and quality affect the realistic accuracy of personality judgment. *Journal of Personality and Social Psychology, 91,* 111-123. https://doi.org/10.1037/0022-3514.91.1.111

Fongwa, M. N., Setodji, C. M., Paz, S. H., Morales, L. S., Steers, W. N. & Hays, R. D. (2010). Readability and missing data rates in CAHPS 2.0 medicare survey in African American and White medicare respondents. Health Outcomes Research in Medicine, 1(1), e39-e49. https://doi.org/10.1016/j.ehrm.2010.03.001

Mount, M. K., Barrick, M. R., & Strauss, J. P. (1994). Validity of observer ratings of the Big Five personality factors. *Journal of Applied Psychology, 79,* 272-280. https://doi.org/10.1037/a0038436

Murray, S. L., Holmes, J. G., & Griffin, D. W. (1996). The benefits of positive illusions: Idealization and the construction of satisfaction in close relationships. *Journal of Personality and Social Psychology, 70,* 79-98.

Oh, I-S, Wang, G., & Mount, M. K. (2011). Validity of observer ratings of the five-factor model of personality traits: A meta-analysis. *Journal of Applied Psychology, 96,* 762-773. https://doi.org/10.1037/a0021832

Peabody, D. (1967). Trait inferences: Evaluative and descriptive aspects. *Journal of Personality and Social Psychology, 7*(4, Pt.2)*,* 1-18. https://doi.org/10.1037/h0025230

Phares, V. (1992). Where’s Poppa? The relative lack of attention to the role of fathers in child and adolescent psychopathology. *American Psychologist. 1992; 47:*656–664. https://doi.org/10.1037//0003-066x.47.5.656

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology, 63,* 539-569. https://doi.org/10.1146/annurev-psych-120710-100452

Podsakoff, P. M., MacKenzie, S. B., Lee, J-Y., & Podsadkoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88,* 879-903. https://doi.org/10.1037/0021-9010.88.5.879

R Core Team (2019). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

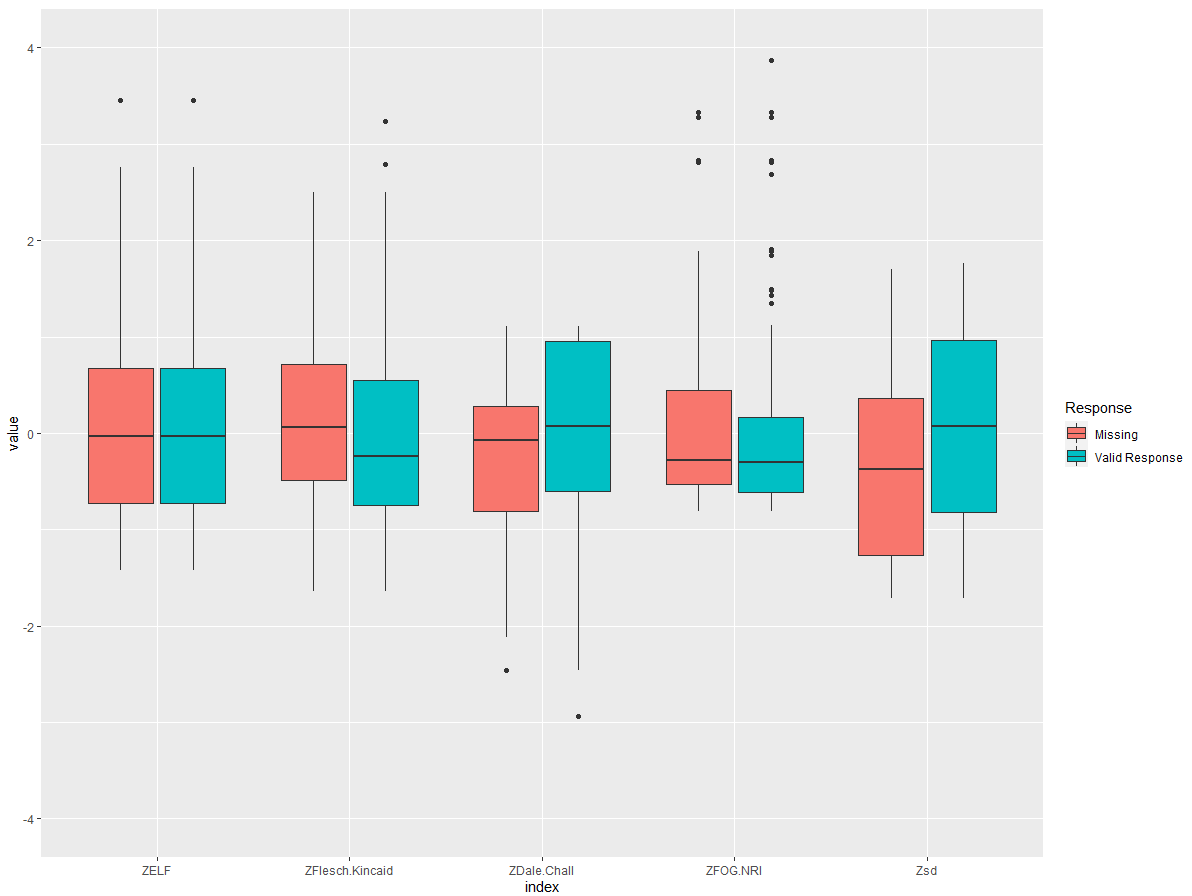
Robins, R. W., & John, O. P. (1997). The quest for self-insight: Theory and research on accuracy and bias in self-perception. In R. Hogan, J. A. Johnson, & S. R. Briggs (Eds.), *Handbook of Personality Psychology* (pp. 649-679). San Diego, CA: Academic Press.

Spector, P. E. (2006). Method variance in organizational research: Truth or urban legend? *Organizational Research Methods, 9,* 221-232. https://doi.org/10.1177/1094428105284955

Tajfel, H., & Turner, J. C. (1979). An integrative theory of inter-group conflict. In W. G. Austin & S. Worchel (Eds.), *The social psychology of inter-group relations* (pp. 33–47). Monterey, CA: Brooks/Cole.

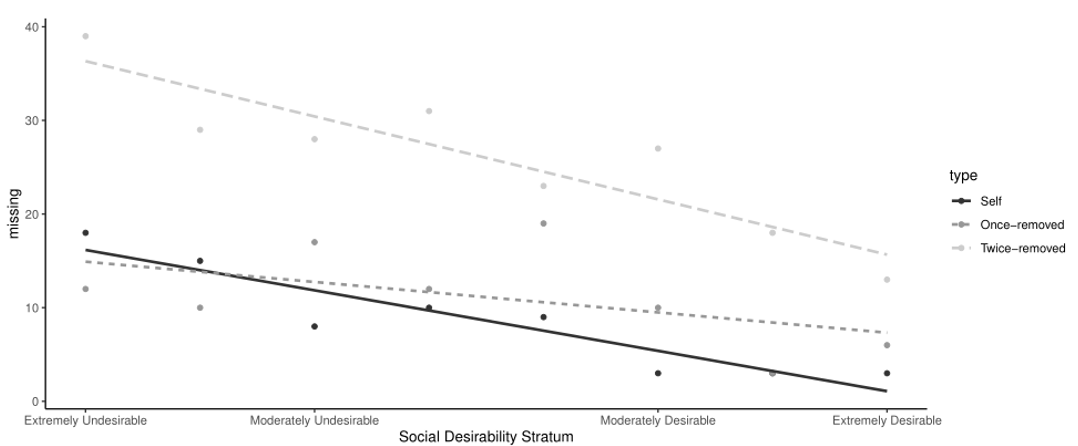
**Figure 1**

*Comparison of Standardized Values for Valid and Missing (NA) Responses across DVs*



**Figure 2**

*Smoothed (Regression-predicted) Missing Value Estimates Across Social Desirability Strata*

****

1. 1 There were two screening variables embedded within once- and twice-removed administrations, asking “who is supposed to be rated” prior to item administration and “who was rated” upon completion of the survey. Several individuals did not answer at least one of these questions with the correct source (the original self-rater) and these cases were excluded from analysis. [↑](#footnote-ref-1)