Crossing a River to get some Water? An Empirical Comparison of Classic and Contemporary Approaches to Item Social Desirability Evaluation

John T. Kulas[1], Emily J. Johnson[2], Renata García Prieto Palacios Roji[3], & Julia Wefferling[3]

[1] eRg
[2] St. Cloud State University
[3] Montclair State University

Traditional approaches to the assessment of socially desirable content within Psychological inventory indicators have been implicated as being too broadly focused. Correspondingly, an alternative method has been proposed whereby the target of rating is shifted from the item *stem* to the item's *response option* (Kuncel & Tellegen, 2009). The current study examines whether the added complexity of the more contemporary procedure is accompanied with an incrementally meaningful amount of unique information regarding the magnitude and valence of socially desirable content within Psychological inventories. Toward this pursuit, the historically traditional and more recently advocated methodologies were empirically compared and contrasted. Our interest was in collecting estimates of: 1) similarity (and uniqueness) of information, 2) inter-rater consistency (when making evaluations), and 3) cognitive difficulty of the rating processes. Results suggest that although the contemporary approach captures some unique information, this is in fact only incrementally informative in predictably particular instances. Specifically, the more cognitively taxing contemporary procedure may be best leveraged with indicators first implicated as "moderately desirable" via application of the traditional (Edwards, 1953, 1957b) approach. A more complementary application of the two approaches should benefit both researchers and item judges.

Yet to do 3/11/23: 1) graph relating Edwards to K/T (maybe look at residuals instead of subjective ratings), 2) response latencies (proxy for task difficulty), 3) inter-rater agreement (also proxy for difficulty of task)

Keywords: Social desirability, response bias, personality assessment, content validation

It may perhaps be adaptive human nature to possess an overly positive evaluation of oneself (Alicke & Sedikides, 2009, 2011; Sedikides & Alicke, 2012; Taylor & Brown, 1988). However, different contexts are also known to either prime (Birkeland et al., 2006; Donovan et al., 2014; Morgeson et al., 2007) or potentially suppress such positive bias in self-evaluation - such as, for example, when accuracy is deemed important (e.g., Dauenheimer et al., 2002). In particular, individuals may feel compelled to present themselves in a favorable manner (possibly inconsistent with their own true character) in situations that pose high-stakes consequences, such as a job interview (e.g., Barrick et al., 2009; Levashina & Campion, 2006; Weiss & Feldman, 2006) or attempting to attract a potential mate (e.g., Dimoulas et al., 1998). When applied to the domain of Psychological assessment, these proclivities are generally contextualized as acts consistent with a *socially desirable* response orientation, and reflect an individual's endorsement of characteristics that are culturally valued or desired rather than what may be objectively true of the person him or herself (Kuncel & Tellegen, 2009; Ziegler, 2011).

Procedurally, these response tendencies within Psychological assessment contexts have been most commonly examined via experimental priming (for example, instructions to "fake" or respond honestly, Birkeland et al., 2006), identification of populations assumed to have divergent response motives (for example, comparisons of job applicant versus non-applicant samples, Viswesvaran & Ones, 1999), or assessment of individual differences in likelihood of responding in a socially desirable manner (for example, Li & Bagger, 2006). Less

Correspondence concerning this article should be addressed to John T. Kulas, 250 Dickson Hall; Montclair State University; Montclair, NJ, 07043. E-mail: jtkulas@ergreports.com

common in contemporary investigations of social desirability are protocols that directly measure and evaluate the saturation of socially desirable (or undesirable) content within inventory indicators themselves.

These indicator saturation investigations *did* enjoy a brief flurry of attention in the mid 20^{th} Century (see, for example, Edwards, 1953, 1957b, 1957a), although this interest dimmed without the direct advocacy of its originating proponent and researcher, Allen Edwards. Recently, there has been a little movement toward revisiting these direct item evaluations (Cui et al., 2022; Leising et al., 2021), as well as a contemporary recommendation aimed at the *method* used to collect the evaluations (aka "ratings," e.g., Kuncel & Tellegen, 2009). The current paper empirically contrasts the traditional (aka "Edwardian") with the more recently advocated contemporary methodology. Our intent was to investigate possible redundancies in information conveyed across the approaches, as well as to seek out indicators of task complexity when judges are asked to provide such ratings.

The Role of Social Desirability in Psychological Assessment

Two contemporary methodologies have been most commonly applied in the evaluation of social desirability's impact on Psychological assessment scores, and both generally support the conclusion that social desirability should not be considered overly problematic (e.g., it is a "red herring," Ones et al., 1996). The first popular contemporary methodology involves assessing individual differences in socially desirable response tendencies via questionnaire administration. These differences in social desirable tendencies can then be leveraged to partial out social desirability effects via covariate specification - for example in the context of assessment validation. Common measures used in this application are the Balanced Inventory of Desirable Responding (BIDR), or the Marlowe-Crowne Social Desirability Scale (e.g., see Crowne & Marlowe, 1960; Li & Bagger, 2006; Paulhus, 1988).

The second set of popular contemporary methodologies employs either experimental instructions to "fake" responses or comparisons of job applicant versus non-applicant respondents (e.g., Birkeland et al., 2006; Viswesvaran & Ones, 1999). Patterns of response are then investigated under conditions thought to be susceptible to socially desirable responding (e.g., fake experimental conditions or applicant respondent samples) versus conditions purported to be lacking socially desirable influence (e.g., control or honest response honest experimental conditions, and non-applicant respondents).

The meta-analyses of Birkeland et al. (2006), Ones et al. (1996), and Viswesvaran and Ones (1999) summarize findings across studies leveraging each of these common approaches. Ones et al. (1996), for example, investigated individual differences in socially desirable responding tendencies as assessed via individual difference measures such as the BIDR, and used this information to construct semipartial correlations between Big 5 scales and work-relevant criteria (e.g., training performance, counterproductive behaviors, job performance). Using this statistical methodology, Ones et al. (1996) noted little effect of socially desirable response tendencies on criterion-related validities (the semi-partial correlations were similar in magnitude to the uncorrected coefficients).

Viswesvaran and Ones (1999) applied a similar meta-analytic lens to *experimental* investigations involving instructions to "fake good" or "fake bad", finding that Big 5 scales tended to exhibit similar levels of fakability. This study confirmed that respondents can indeed intentionally distort their responses (e.g., respond in a socially desirable manner) if instructed to do so. Regarding non-laboratory investigations where context is assumed to prime a socially desirable response orientation, Birkeland et al. (2006) similarly documented elevated Big 5 scale scores with applicant respondents relative to non-applicant respondents, but also noted that the pattern of rating elevation differed across the type of position the applicant was seeking. Note here that all methodologies encompassed by these meta-analyses are characterized by an individual difference orientation (e.g., it is differences across respondent proclivity to enhance - either driven by context or trait - within which the social desirability influence is manifest).

An Elemental Focus Alternative. Alternative to the above-noted approaches to exploring social desirability's role in Psychological assessment, there exists a much smaller subset of researchers who have focused on the assessment elements themselves (e.g., the *item*, see, for example, Edwards, 1957b). This approach appears to be more popular in non-work assessment domains than compared to the business or Industrial and Organizational assessment literatures (see, for example, Leising et al., 2012, 2015).

For roughly 60 years, the standard investigation of item-level saturation with socially desirable content had been applied in a fairly consistent manner, with little substantive deviation from the procedure first advocated by Edwards. Edwards (1953) simply asked judges to rate the content of personality items along a social desirability continuum (wherein, for example, the personality item, "I hate people" would likely be deemed less desirable than an item such as, "I regularly give money to charities in need"). Edwards specifically asked his judges to provide ratings ranging from extremely undesirable to extremely desirable along a 9-point scale, and subsequently went on to further demonstrate that the more socially desirable an item is, the more likely someone will endorse

having that characteristic (Edwards, 1953, 1957b) [1].

A Procedural Revisitation. Kuncel and Tellegen (2009) revisited the procedure, proposing that traditional measurement approaches such as Edwards' are perhaps overly simplistic if assessment specialists aim to truly understand the impact of social desirability on assessment responses. Specifically, Kuncel and Tellegen (2009) noted that previous investigations had largely ignored the potentially differential attraction to item *response options*, as opposed to (or in addition to), the level of agreement with the item stem itself. This perspective challenged the previously implicit assumption that social desirability manifests itself in a linear fashion across response options, whereby "agreement" with more (or less) of a characteristic is consistently associated with greater levels of social desirability (or *un*desirability).

As rightly noted by Kuncel and Tellegen (2009), the implicit assumption is not necessarily valid, as there are plausible characteristics with a *most* desirable standing location that is not located at either extreme (consider, for example, "being quiet" - it is likely most socially desirable to be moderate along the trait continuum for this characteristic). As an explicit alternative to the implicit assumption, Kuncel and Tellegen (2009) proposed that at least four patterns of item social desirability may commonly exist across scaled inventory response options: linear, non-linear, non-linear monotonic, and weakly non-linear monotonic. ← **check this before submitting (monotonic or non-monotonic)**

These possibilities acknowledge that the two-dimensional functional progression between an x-axis "location of response" (e.g., low, moderate, or high on the trait) and a y-axis "how desirable the location is" could be linear, logarithmic/exponential, "U"-shaped (or invertedly "U"-shaped), or perhaps even flat in several regions. The authors even suggested that *most* trait items may be best characterized by nonmonotonic or weakly monotonic relationships with social desirability and that a strictly linear relationship would be dependent on highly valued items or strongly incentivized contexts (for example, applying for a desired job).

To test their premise, Kuncel and Tellegen (2009) constructed an alternative rating system. This approach asks individual judges to rate items on *how desirable the trait to be at five different levels* of the characteristic: extremely high (top 1%), above average (top 30%), average, below average (bottom 30%), or extremely low (bottom 1%; see Figure 1, which has been reproduced from the original Kuncel and Tellegen (2009) publication). Note here that these five categories parallel the common 5-point rating system frequently encountered in self-report inventories. Applying this rating procedure, Kuncel and Tellegen (2009) found support for their premise that not all items demonstrate linear associations with social desirability and that non-monotonic relationships do exist across graded response continua.

Kuncel and Tellegen (2009)'s second study was designed to approximate real-world contexts. Here, participants were asked to act as if though they were in a pre-employment assessment situation and to explain their rationale when an extreme response was *not* chosen on the assessment. This design was intended to provide insight regarding the lack of linear manifestations of social desirability. Kuncel and Tellegen (2009) found that, across administrations, over 60% of participants did in fact choose the most extreme response options. Some of the participants who opted out of endorsing the extreme responses, however, noted that the extreme response might be poorly perceived by an evaluator (i.e., too inaccurate, bragging, too good). Taken collectively, these investigations supported the notion that trait characteristics do not necessarily manifest only strictly linear associations with the concept of social desirability.

Not Easily Upset

How desirable is it to be:

1. Extremely High in this characteristic (top 1%)

Very Undesirable	Undesirable	Neutral	Desirable	Very Desirable

2. Above Average in this characteristic (top 30%)

Very Undesirable	Undesirable	Neutral	Desirable	Very Desirable

3. Average in this characteristic

Very Undesirable	Undesirable	Neutral	Desirable	Very Desirable

4. Below Average in this characteristic (bottom 30%)

Very Undesirable	Undesirable	Neutral	Desirable	Very Desirable

5. Extremely Low in this characteristic (bottom 1%)

Very Undesirable	Undesirable	Neutral	Desirable	Very Desirable

Figure 1. Kuncel and Tellegen (2009) method for determining socially desirable saturation at the item response level.

Although there is both theoretical and empirical support for Kuncel and Tellegen (2009)'s procedure, it also quite substantially more time- and (we propose) effort-intensive than is the traditional item-rating approach (Edwards, 1953, 1957b). As technically specified, the traditional Edwards procedure requires one evaluation per item (albeit that evaluation is made across nine gradiated social desirability strata). The contemporary "Kuncel and Tellegen" procedure requires (in the case of 5-point Likert-type indicators) five evaluations across five levels of desirability per item. In addition to the greater *number* of evaluations required in the contemporary approach, we propose that the contemporary approach is also

[1] This is a very robust finding that has been replicated many times. The implications of this finding are also far-reaching, and constitute one of the reasons an exploration of the contemporary viability of Edwards' approach is deemed important. However, the focus of the current exploration is fully *procedural*, pointed directly at the *method* used to collect item social desirability ratings rather than the broader implications of attraction toward the socially desirable within Psychological assessment.

likely more cognitively demanding due to shifting objects of reference (the referent of appraisal shifts across ratings).

Given the greater time and (we argue) resource commitments required of the contemporary approach relative to the traditional, we aim to gauge to what extent these two approaches in fact capture similar versus unique pieces of information. The goal of the present investigation is therefore to directly compare these two methodologies with an "additional information" orientation - that is, is the new approach truly unique, or rather does it at least with some indicators convey somewhat redundant information as the classic, cognitively easier and less time-intensive approach?

> *Research Question 1*: Do the contemporary and traditional rating procedures capture redundant or unique information regarding social desirability saturation?

> *Research Question 2*: Is the contemporary procedure more cognitively taxing than the traditional procedure? ← **Reword after finanlize analyses - didn't collect response latencies from the Edwards form**

Study 1

Methods

Participants

Seventy-six undergraduate students made ratings of *either* item social desirability (n = 14, Edwards, 1957b), or levels of desirability associated with different trait levels (n = 62, e.g., Kuncel & Tellegen, 2009).

Materials

The IPIP-NEO is a 300-item personality measure intended to assess the Big Five personality dimensions: Conscientiousness, Agreeableness, Extraversion, Openness to Experience, and Neuroticism (Johnson, 2005). For the purposes of the current investigation, we did not collect typical responses to these 300 indicators, but were rather interested in the nature of the items themselves (or, alternatively, the evaluative content associated with differential standing along the construct implied by the item response options).

Procedure

All ratings were made via paper and pencil in an experimental laboratory. The Edwards (1957b) ratings were made along Edwards' originally specified 9-point scale ranging from Extremely Undesirable to Extremely Desirable. Because we investigated a fairly large instrument, we constructed 2 counterbalanced "Edwards" forms as an effort to limit potential fatigue effects across the rating process. The Kuncel and Tellegen (2009) ratings were collected from 60 different item stems across 10 different counterbalancings. Each rater (regardless of task; item stem or response option rating) was therefore asked to perform 300 total ratings (either 1 evaluation per 300 items or 5 evaluations per 60 items).

Results

All analyses were performed in R version 4.1.1 (R Core Team, 2021). We leveraged three different approaches comparing findings across the two item rating procedures. All three approaches focused on associations between the average item rating (aka "Edwards" value) and 2-dimensional ("Kuncel & Tellegen") plotted function. These functions imply an x-axis continuum although there exist only 5 actual x-axis categories. The 5 categories reflect progression across "Kuncel & Tellegen" frames of reference (ranging from someone who is "Extremely High in the characteristic" to someone who is "Extremely Low in the characteristic", see Figure 1). The height of the function at each of these 5 categories is determined by the average desirability rating.

Approach #1: Functional Slope. For the first investigative approach, we probed for associations between "Edwards"' item ratings and regression slope of "Kuncel & Tellegen" function. If the two procedures result in similar output, we would expect greater incidences of linearity with extremely desirable (and undesirable) items. Here, 300 individual regressions were fit using the five different rated trait locations as predictors (e.g., Kuncel and Tellegen (2009)'s "bottom 1%", "bottom 30%", "Average", "top 30%", and "top 1%" - these were treated as a scaled continuum [values of 1, 2, 3, 4, and 5]) and averaged response desirability rating as the criterion. Figure 2 helps demonstrate our approach here.

for exposition - regressions were fit to the plotted space [response category on the x-axis and average rating on the y-axis]). Within each of the 300 regressions, the expectation was that slope magnitude and valence would parallel the classic Edwards ratings of the same items. For example, the expectation was that Figure 2's "Excel in what I do" and "Believe that others have good intentions" would realize negative slope estimates, "Enjoy wild flights of fancy" would exhibit a flat slope, and "Get irritated easily" would return a moderately positive slope.

Redundant with previous paragraph → In order to capture the extremity of function across Kuncel and Tellegen

(2009) values, several regressions[2] were fit using the average (across respondents) rating as a predictor (e.g., Kuncel and Tellegen (2009)'s "Lower 1%", "Lower 30%", "Median", "Upper 30%", and "Upper 1%" were treated as a scaled continuum) and average Edwards' desirability rating as the criterion. Slopes were retained for each function, with the expectation that slope magnitude and valence would parallel the classic Edwards ratings.

Across all 300 items, the relationship was strong ($R^2 = .67$, $p < .05$), suggesting some level of similarity across procedures. Figure 2 demonstrates these relationships for 25 randomly selected items within five categorized Edwards arrays. Note that the functions (even if somewhat non-monotonic - see, for example, "Seldom Daydream" in Figure 2) tend to exhibit greater slope with Edwards' highly desirable or undesirable items, and are "flatter" with Edwards' moderate items.

Next, functional slopes of the 300 items (e.g., the Figure 2 plots for all 300 items) were rated along dimensions of "on the whole, this looks like a straight line" with possible ratings ranging from (1 = not at all, to 5 = definitely), and how much the "line rises and falls" from (1 = not at all, to 5 = a lot). These estimates were added to the first approach (defining each Edwards/Kuncel and Tellegen convergence with ratings of both functional linearity and monotonicity) and the result is presented in Figure 4. Careful inspection of the Figure 4 plot again highlights the location of non-monotonic and nonlinear Kuncel and Tellegen functions - predominantly at moderate (around neutral) Edwards rating locations.

Our third approach leveraged hierarchical polynomial regressions, with our index of interest being the change in R^2 at the second step, when the polynomial terms were specified. Most items ($n = 203$) had a very low change in R^2 (see Table 1 for a summary of these results and a specification of our definition of "low" or "very low"). Beyond our subjective categorization, F-tests indicated that 92.31% of item ratings were *not* significantly predicted when specifying the quadratic term, and only 7.69% were.

Discussion

Across approaches, results tended to converge on similar conclusions. The preponderance of our results first suggest that, in general (at least with our focal 300-item measure), linear relationships with social desirability (across response options) may be commonplace and in fact fairly well represent the plurarity of assessment item functions. Additionally, similar information seems to be available through both the traditional and contemporary measurement approaches. Although non-monotonic functions do exist, they are predominantly associated with moderately rated Edwards items. Cursory review of our randomly sampled 25 items (Figure 2) would agree with this conclusion.[3] It is quite plausible that "U" or "inverted U" shaped functions, when they occur, are reflective of some ambiguous or contextually primed desirability, and that this ambiguity or contextual moderation results in "middle ground" evaluation via the Edwards method.

Undoubtedly, the Kuncel and Tellegen (2009) procedure conveys information not contained in the classic Edwards (1957b) approach. The purpose of this investigation, however, was to document overlap between the two procedures. While it is clear nonmonotonic functions do exist for some indicators across scaled "trait levels," the vast majority of such circumstances are located within a range what the Edwards (1957b) procedure labels as merely moderately desirable or undesirable. There is surely additional information contained within these items, but the current investigation suggests that perhaps the more cognitively taxing and time-intensive procedure should be retained only for the items first identified by the cognitively-easier and less time-consuming Edwards (1957b) method. Our recommendation is to therefore retain both procedures, utilizing the cognitively easier and less time-consuming procedure as an initial evaluation and following-up with moderately desirable items to probe for more complex relationships.

Results

THIS SECTION IS ALSO NOT IN THE PAPER, BUT DID NOT WANT TO DELETE EVERYTHING WITHOUT PERMISSION

The plurality of findings do support similar infomration being conveyed through both approaches. Figure 1 captures *some* of this, as functional slopes (even if somewhat non-monotonic - see, for example, "Seldom Daydream" in Figure 1) tend to be more extreme with Edwards' highly desirable or undesirable items, and more flat with Edwards' moderate items.[4] Table 1 presents the frequency with which researcher-implicated functional shapes (linear [positive], linear [negative], nonmonotonic [U], and nonmonotonic [inverted U]) were noted within "Edwardian" strata, demonstrating that although the nonmonotonic functions do exist, they are predominantly associated with moderate Edwards items (e.g., yes these functions do occur but perhaps the ambivalence is also associated with aggregate moderation). Figure 2 presents the relationship between: 1) the functional slope relating an item response's rated level of desirability

[2] Polynomial regressions were also conducted to capture functional form but are not presented as a central index - these are available in online supplementary material.

[3] A full list of all 300 item functions are available in this paper's online resources.

[4] These items were randomly sampled from within Edwards-rating strata. Visuals of all items' functional forms are available from the authors upon request.

and the "location" of the rating, and 2) Edwards' item stem rating (on the y-axis). This strong relationship (R^2 = .67, p < .05) suggests some level of similarity across procedures.

Study Two

> put in some references regarding response latencies

One of our observations across the two item rating procedures has been that the Kuncel and Tellegen (2009) approach appears to be more cognitively taxing for "raters" than does the Edwards (1953). Study Two therefore collected ratings via computer, with response latencies recorded as well as direct self-reported estimates of difficulty. The same indicators were retained for Study Two.

Procedure

The Study One items were administered via Qualtrics (2014). Participants were administered directions for each condition: 1) Edwards rating, 2) Kuncel & Tellegen rating, and 3) modified Kuncel & Tellegen rating. This third condition was included as a possible compromise to the *anticipated* cognitive difficulty associated with the Kuncel & Tellegen procedure. Here, we ask participants to focus on only one of the Kuncel & Tellegen "levels" (e.g., across items, the rating category is held constant). We coded items for length and presence of a "negative" qualifier (e.g., I would never lie). Anecdotally, the presence of a negative element within an item stem contributed to difficulty for raters.

In addition to actual ratings, we also recorded latencies, which we interpret as indirect assessments of difficulty.

Because of the tedium and cognitive complexity of the requested task, we computed consecutive non-differentiating responses as well as intra-individual response variability estimates (see, for example, Dunn et al., 2018; Marjanovic et al., 2015) via the *careless* R package (Yentes & Wilhelm, 2021) in R version 4.1.1 (R Core Team, 2021).

Table to add: ICC's for each condition

Focus on response latencies (Edwards ratings vs. Kuncel & Tellegen vs. revised Kuncel & Tellegen (split by rated category). Indices here are: 1) response latencies, 2) inter-rater agreement/reliability [5]

Results

```
## Warning: Removed 46 rows containing missing values (geom_point).
```

The intercept was -3.08 <- just testing.

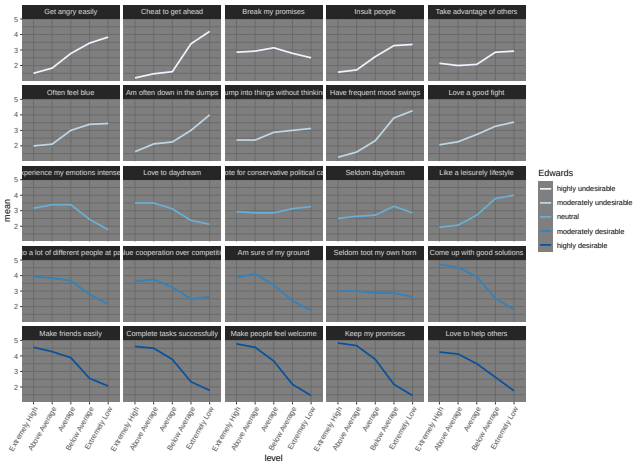

Figure 2. Kuncel & Tellegen (2009) patterns across Edwards (1953) scale values.

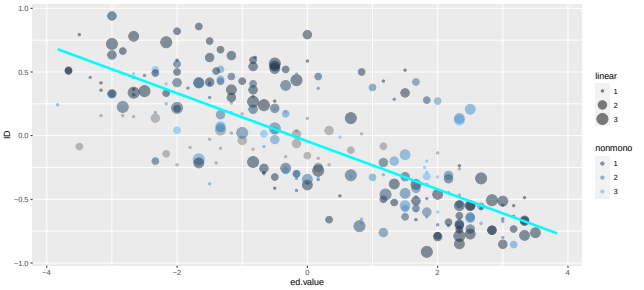

Figure 3. Response Category and Mean Rating slope across Edwards' scale values (individual scatterpoints represent items).

Discussion

Certainly the Kuncel and Tellegen (2009) procedure conveys information not contained in the classic Edwards (1957b) approach. The purpose of this investigation, however, was to document overlap between the two procedures. Clearly nonmonotonic functions do exist for some indicators across scaled "trait levels". However, the vast majority of such circumstances are located within what the Edwards (1957b) procedure labels as "moderately desirable". There is certainly additional information contained within these items, but the current investigation suggests that perhaps the more cognitively taxing and time-intensive procedure be retained for only those items first identified by the cognitively-easier and less time consuming Edwards (1957b) method as "moderately desirable".

[5] NOTE TO SELVES - WHEN WE DO THE SECOND PART (RESPONSE LATENCIES), MAKE SURE TO CODE NEGATIVELY-WORDED STEMS (THESE SEEM ESPECIALLY DIFFICULT TO RATE USING THE KUNCEL & TELLEGEN PROCEDURE)

Limitations

Our task was likely too long - in retrospect a shorter measure should have been pursued.

> CUT FROM STUDY 1: First, we fit linear regressions to all "Kuncel & Tellegen" functions (as explained below), extracting slope coefficients, then regressed Edwards' ratings on these slopes across items. Secondly, we collected visual estimates of the monotonicity and linearity of these functions, and used these estimates to help inform ranges of Edwards values along which nonlinear item functions tend to be more prominent (e.g., how many "inverted U-shaped" functions were noted in items characterized by Edwards' system as *extremely undesirable*, *undesirable*, *average*, *desirable*, and *extremely desirable*). Lastly, we conducted 300 hierarchical polynomial regressions to capture "U" or "inverted-U" functional forms, and tallied how many items were incrementally characterized by a quadratic ("above and beyond" the linear) equation.

References

Alicke, M. D., & Sedikides, C. (2009). Self-enhancement and self-protection: What they are and what they do. *European Review of Social Psychology*, *20*(1), 1–48.

Alicke, M. D., & Sedikides, C. (2011). *Handbook of self-enhancement and self-protection*. Guilford Press.

Aust, F., & Barth, M. (2022). *Papaja: Prepare american psychological association journal articles with r markdown*. https://github.com/crsh/papaja

Barrick, M. R., Shaffer, J. A., & DeGrassi, S. W. (2009). What you see may not be what you get: Relationships among self-presentation tactics and ratings of interview and job performance. *Journal of Applied Psychology*, *94*(6), 1394.

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment*, *14*(4), 317–335.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*(4), 349.

Cui, T., Kam, C. C. S., Cheng, E. H., & Ho, M. Y. (2022). Distinguishing between trait desirability and item desirability in predicting item scores: Is informant evaluation of personality free from social desirability? *Personality and Individual Differences*, *196*, 111708.

Dauenheimer, D. G., Stahlberg, D., Spreemann, S., & Sedikides, C. (2002). Self-enhancement, self-verification, or self-assessment? The intricate role of trait modifiability in the self-evaluation process. *Revue Internationale de Psychologie Sociale*.

Dimoulas, E., Wender, S., Keenan, J. P., Gallup, G., & Goulet, N. (1998). Patterns of deception in human mating strategies. *Journal of Psychology and the Behavioral Sciences*, *12*, 39–42.

Donovan, J. J., Dwight, S. A., & Schneider, D. (2014). The impact of applicant faking on selection measures, hiring decisions, and employee performance. *Journal of Business and Psychology*, *29*(3), 479–493.

Dunn, A. M., Heggestad, E. D., Shanock, L. R., & Theilgard, N. (2018). Intra-individual response variability as an indicator of insufficient effort responding: Comparison to other indicators and relationships with individual differences. *Journal of Business and Psychology*, *33*(1), 105–121.

Edwards, A. L. (1953). The relationship between the judged desirability of a trait and the probability that the trait will be endorsed. *Journal of Applied Psychology*, *37*(2), 90–93.

Edwards, A. L. (1957a). Social desirability and probability of endorsement of items in the interpersonal check list. *The Journal of Abnormal and Social Psychology*, *55*(3), 394–396.

Edwards, A. L. (1957b). *The social desirability variable in personality assessment and research.*

Kuncel, N. R., & Tellegen, A. (2009). A conceptual and empirical reexamination of the measurement of the social desirability of items: Implications for detecting desirable response style and scale development. *Personnel Psychology*, *62*(2), 201–228.

Leising, D., Ostrovski, O., & Borkenau, P. (2012). Vocabulary for describing disliked persons is more differentiated than vocabulary for describing liked persons. *Journal of Research in Personality*, *46*(4), 393–396.

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*(1), 61–71.

Leising, D., Vogel, D., Waller, V., & Zimmermann, J. (2021). Correlations between person-descriptive items are predictable from the product of their mid-point-centered social desirability values. *European Journal of Personality*, *35*(5), 667–689.

Levashina, J., & Campion, M. A. (2006). A model of faking likelihood in the employment interview. *International Journal of Selection and Assessment*, *14*(4), 299–316.

Li, A., & Bagger, J. (2006). Using the BIDR to distinguish the effects of impression management and self-deception on the criterion validity of personality measures: A meta-analysis. *International Journal of Selection and Assessment*, *14*(2), 131–141.

Marjanovic, Z., Holden, R., Struthers, W., Cribbie, R., & Greenglass, E. (2015). The inter-item standard deviation (ISD): An index that discriminates between conscientious and random responders. *Personality and Individual Differences*, *84*, 79–83.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology*, *60*(3), 683–729.

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.

Paulhus, D. L. (1988). Balanced inventory of desirable responding (BIDR). *Acceptance and Commitment Therapy. Measures Package*, *41*, 79586–79587.

Qualtrics, L. L. C. (2014). Qualtrics [software]. *Utah, USA: Qualtrics*.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Sedikides, C., & Alicke, M. D. (2012). *Self-enhancement and self-protection motives*. Oxford handbook of motivation, ed. R. Ryan. Oxford University Press.[rWvH].

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin*, *103*(2), 193.

Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement*, *59*(2), 197–210.

Weiss, B., & Feldman, R. S. (2006). Looking good and lying to do it: Deception as an impression management strategy in job interviews. *Journal of Applied Social Psychology*, *36*(4), 1070–1086.

Yentes, R., & Wilhelm, F. (2021). *Careless: Procedures for computing indices of careless responding*. https://github.com/ryentes/careless/

Ziegler, M. (2011). Applicant faking: A look into the black box. *The Industrial and Organizational Psychologist*, *49*(1), 29–36.