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 inventory indicators. 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 20th 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 contrasts the traditional (aka "Edwardian") with the more recently advocated contemporary methodology. Our intent was to investigate possible redundancies in information conveyed across the two 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. Historically popular measures used in this application include, for example, 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 semipartial correlations were similar in magnitude to 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 analysis 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 psyche - within which the social desirability influence is manifest).

An Elemental Focus Alternative. Alternative to the above-noted inter-individual-oriented approaches to exploring social desirability's role in Psychological assessment, there exists a 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 contemporarily popular within non-work assessment domains than 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 methodological 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 undesir-

able 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 Edwardian item rating protocol, 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 potential for social desirability to be maximally salient at locations other than trait extremes. 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). Procedurally, Kuncel and Tellegen (2009) assessed differential attraction to item *response options*, as opposed to the Edwardian focus on the *item stem.*

As rightly noted by Kuncel and Tellegen (2009), there are plausible characteristics with a *most* desirable standing location that is not located at either extreme (consider, for example, "being quiet" - it is, without additionally provided context, likely most socially desirable to be moderate along the trait continuum for this characteristic). As an explicit alternative to the implicit Edwardian assumption of linear social desirability manifestation, 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 monotonic (rate of increase is not constant), weakly non-linear monotonic (flat regions exist), and non-monotonic (pattern reversal).

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, flat in regions, or perhaps even "U"- or inverted "U"-shaped. 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* (they deem) *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 ubiquitous 5-point rating system often retained 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 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) protocol 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) cognitive effort-intensive than is the traditional item-rating approach (Edwards, 1953, 1957b). As technically specified, the traditional

---

[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.

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 likely more cognitively demanding due to shifting objects of reference (the referent of appraisal shifts across ratings - top 1%, top 30%, etc.).

Given the greater time and 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 similar information as the classic, less cognitively taxing and more time-efficient approach?

Research Question 1: Do the contemporary and traditional rating procedures capture 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 evaluative content of the items (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 4.4.1 (R Core Team, 2024). We leveraged three different approaches comparing findings across the two item rating procedures. First, we applied simple linear regressions to all "Kuncel & Tellegen" functions (as explained below), extracting slope coefficients, and then correlated Edwards' ratings with these slope coefficients across items. Secondly we conducted 300 hierarchical 2nd degree *polynomial* regressions to each function in an attempt to capture empirical "U" or "inverted-U" functional forms, and tallied how many items were incrementally ("above and beyond" the linear) characterized by the quadratic regression term. Lastly, we categorized visual perceptions of the functions, and revisited the above noted associations within *differently categorized functional forms*.

NOTE. Old but maye revisit: used these groupings 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*).

All three approaches focused on the *functional form* of "Kuncel & Tellegen" ratings, and relied upon either regression analyses to provide an empirical estimate of the function form and/or judge categorizations to provide a subjective interpretation of the function form. These functions all reflect progression across Kuncel & Tellegen frames of reference

(ranging from someone who is "Extremely High in the characteristic (top 1%)" to someone who is "Extremely Low in the characteristic (below 1%)", again see Figure 1 for exposition. The "height" of the function at each of five rated frames of reference is determined by the *average desirability rating* at each of these points of normative consideration.

Approach #1: Functional Slope (Empirical). One reasonable manifestation of similarity across the two procedures would be stronger incidences of functional linearity with extremely desirable and undesirable items (and correspondingly weaker linear associations with moderate items). For the first investigative approach, we therefore probed for associations between "Edwards" ' item ratings and *regression slope* of "Kuncel & Tellegen" function.

We first fit 300 individual regressions retaining the five different rated trait locations as a predictor (e.g., Kuncel and Tellegen (2009)'s "bottom 1%", "bottom 30%", "Average", "top 30%", and "top 1%" - these were treated as representing an equal-interval numerical continuum [values of 1, 2, 3, 4, and 5]) and averaged (Edwards) response desirability rating as the criterion. Within each of the 300 individual 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 an item such as "Believe that others have good intentions" would realize a highly desirable Edwards rating as well as a high magnitude, negatively valenced slope estimate. "Enjoy wild flights of fancy" would exhibit a moderate Edwards rating and flat slope, and "Get irritated easily" would return an undesirable Edwards rating as well as a moderately positive slope. Across all 300 items, the relationship between Edwards rating and Kuncel & Tellegen functional slope was indeed revealed to be strong ($r = $ -.76, $R^2 = $ .58, $F_{(1,298)} = $ 412.26, $p < $ .001), suggesting a non-trivial association between procedures.

For added visual exposition, items were also categorized within arrays of values such that "Kuncel & Tellegen" functional forms could be observed within meaningful ranges of Edwardian values (e.g., Somewhat Undesirable, Extremely Desirable, etc.). Figure 2 presents 25 Kuncel & Tellegen functional forms randomly sampled from within each of 5 different Edwardian arrays[2], and the figure *rows* reflect the array strata. Note that the functions (even if somewhat non-monotonic - see, for example, "Seldom Daydream" in Figure 2) tend to exhibit "steeper slope" with Edwards' highly desirable or undesirable items, and are "flatter" with Edwards' moderate items - this first set of explorations merely confirms a general pattern such that more extreme "Edwards" items tend to have steeper "Kuncel & Tellegen" functions, but is not indicative of functional form (e.g., this exploration is not directly reflective of, for example, nonmonotonicity).

Approach #2: Functional form (Empirical). Our second approach leveraged hierarchical polynomial regressions,
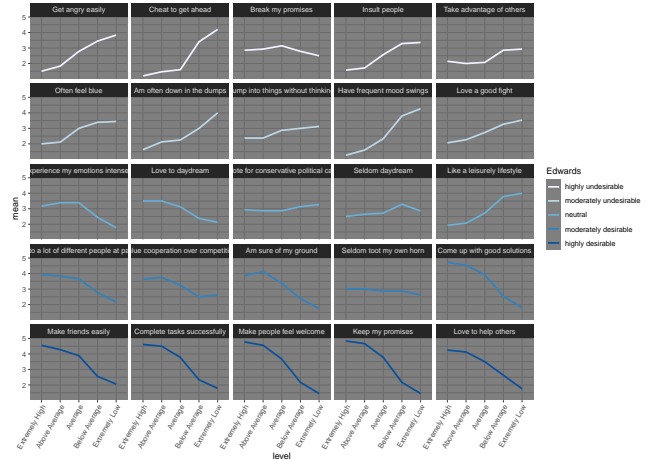

Figure 2. Kuncel & Tellegen (2009) patterns across Edwards (1953) scale values.

with the index of interest being the change in R^2 associated with nonlinear association, as operationalized as a quadratic (second degree) polynomial term. Most item functions ($n_k = $ 203) exhibited a very low change in R^2 (see Table 1 for a summary of these results and explicit articulations of our subjective classifications of "low" or "very low").[3] These results suggest that, although subjective judgements may be able to visually distinguish functional form[4], empirical estimates are not as sensitive to deviations from linearity.

Approach #3: Functional form (Visual). Next, the Figure 2 plots for all 300 items were presented to judges who performed an inductive content analysis (Miles & Huberman, 1994) – grouping item functions by *perceived functional similarity* without further instruction. There were 11 total cate-

[2]Note that these plotted functions progress along only 5 actual x-axis values. The "connected dots" do not imply continuous x-axis values but are presented to facilitate observation of linear, monotonic, or non-monotonic functional patterns (e.g., Kuncel & Tellegen, 2009).

[3]F-tests associated with the ΔR^2 indicated that 92.31% of item functions were not *significantly* improved when specifying the quadratic term ($\alpha = $ .05). This information is being presented for completeness, although it should be noted that F critical values are very large in this atypical application of regression and effect sizes are more appropriately informative.

[4]Compare with the % of "linear" and maybe "egyptian" that we subjectively categorized - 5/30/24. Also add to discussion – there is a blurry line between a function being "linear" and "non-linear". Subjectively allowances are given that can be mutually agreed upon. There comes a point, however, where one judge deems a function "linear" and another disagrees. Similarly, there comes a point where both judges agree that the function exhibits non-linearity. Indeed, humans in general may not be very good at dissociating curved and straight lines (Bales & Follansbee, 1935; Gibson, 1933; Ogilvie & Daicar, 1967; Watt et al., 1987).

Table 1
Polynomial effects, organized by magnitude

Example Item	ΔR^2	$\Delta R^2 Range$	n_k
Make friends easily	0.03	Below .05	152
Would never cheat on my taxes	0.06	.05 - .10	51
Have a vivid imagination	0.14	.10 - .15	20
Trust others	0.20	.15 - .20	15
Experience my emotions intensely	0.27	.20 - .40	33
Love large parties	0.60	.40 - .60	15
Am always busy	0.80	.60 - .80	12
Never splurge	0.82	Above .80	2

Note. ΔR^2 refers to the incremental second-order polynomial effect estimated via hierarchical regression.

Table 2
Average Edwards rating and number of functional classifications

label	category	n
m= -1.31	linear (p)	21.00
m= -1.48	non-mono (s)	46.00
m= -1.16	non-mono (hook mp)	16.00
m= -0.85	non-mono (hook ep)	22.00
m= -1.26	non-mono (j)	15.00
m= -0.45	non-mono (flat)	11.00
m= -0.34	non-mono (peaked)	46.00
m= 1.88	non-mono (hook en)	39.00
m= 0.54	non-mono (hook mn)	17.00
m= 1.79	non-mono (z)	38.00
m= 1.86	linear (n)	26.00

gories identified, and all but three item functions fit into one of these 11 categories (item functions were assigned exclusively to only one category)[5]. Figure 3 presents an exemplar function for each of the 11 categories as well as panel scatterplots and correlations executed within each of the 11 categories. Note that 10 of the categories represent "mirror" functions (e.g., the "leftmost" category is a symmetrical mirror of the "rightmost" category). The average Edwards rating for all items grouped within each category is presented in Table 2 along with the number of items classified as exhibiting the functional form.

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

Study 1 Discussion

Across approaches, results tended to converge on similar patterns. The preponderance of our results first suggest that, in general (at least with our focal 300-item measure), linear or "close-to" 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 meaningful deviations from linear functions do exist, they are predominantly associated with *moderately rated* Edwards items. Although not exhaustively confirmatory, consultation of the 25 randomly sampled item functions (Figure 2) also reflects this pattern of non-monotonicity.[6] 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.
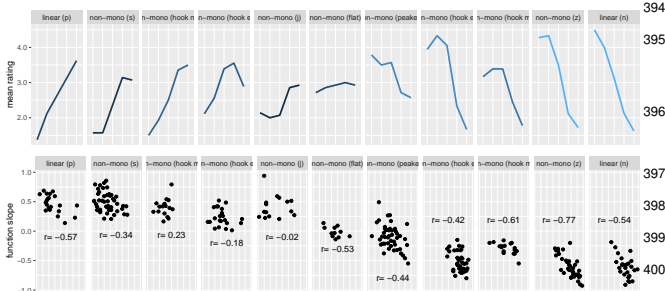
Study 2

One of our suppositions across the two item rating procedures has been that the Kuncel and Tellegen (2009) approach is more cognitively taxing for raters than is the Edwards (1953) rating task. Study Two therefore collected ratings via computer, with response latencies recorded as estimates of difficulty.

Methods

Participants

One hundred and thirty one undergraduate students from two universities participated in item ratings. After data cleaning, 87 students ratings were analyzed.

---

[5]The unclassified items were, "Am not interested in other peoples' problems", "Postpone decisions", and "Readily overcome setbacks".

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

Procedure

All ratings and response latencies (measured in seconds 's') were gathered online using Qualtrics (2014), utilizing the Kuncel and Tellegen (2009) rating scale. The ratings included fifteen of the most archetypal items from Study 1, focusing on the three functional forms: asymmetric, linear, and quasisymmetric.

During data cleaning, we screened out total survey durations less than 500 seconds and longer than 2000 seconds (n=46) as well as individuals who had total response latencies less than 20 seconds across the final 20 items (on average less than 1 second per item response; n=8). These conservative screens were applied based on observations from experiment proctors that several participants were not fully engaged in their responses. Additionally, questions were implemented to gauge carelessness and task difficulty. Computation of consecutive non-differentiating responses as well as intra-individual response variability estimates was done due to tedium and cognitive complexity of the task (see, for example, Dunn et al., 2018; Marjanovic et al., 2015) via the *careless* R package (Yentes & Wilhelm, 2023) in R version 4.4.1 (R Core Team, 2024).

Results

Response Latencies. All latencies were taken and compiled into one list so that outliers could be removed. Any latencies that were 0 were first removed, and afterwards and any time below the 10th percentile and above 90th percentile, from the remaining times, were removed. Afterwards, mean latencies were gathered for four conditions: the item (e.g., Worry about things), the level (e.g., extremely high), the type (e.g., linear), as well as if the item was negatively worded or not.

Latency data had some differences in type with quasisymmetric taking the least amount of time ($M = 2.8$ seconds) and asymmetric taking the most time ($M = 3.5$ seconds). Negatively worded items took on average 0.5s longer than non-negatively worded items. The extremely High had the largest latency time at 6.0 seconds, while the other levels ranged from 2.0s to 3.0s. However, this is likely an artifact due to extremely high levels always being the first rating after reading the item characteristic, rather than it being harder to rate.

Interrater Reliability. ICC(2,1) was used for interrater reliability ((**shroutfleiss?**)) for Study 1 and Study 2 comparisons on the Edwards (1957b) as well as Kuncel and Tellegen (2009) methodologies. For Edwards, the ICC(2,1) was run for the three functional forms, and then all ratings were combined to obtain an overall ICC. However, for the Kuncel and Tellgen method, interrater reliability was first run on the item itself, and then all interrater reliabilities were averaged to obtain the three functional form types for Study 1 and Study 2. All ICCs are shown in Table 3.

The Edwards interrater reliability was moderate for linear functional forms (ICC = 0.525), and had no interrater reliability for asymmetric (ICC = 0.001) or quasisymmetric (ICC = 0.000) items. Overall reliability, of all 15 items, was poor (ICC = 0.277). The Kuncel and Tellegen functional form interrater reliability for study 1 was moderate for linear (ICC = 0.503), and poor for asymmetric (ICC = 0.445) and quasisymmetric (ICC = 0.099). Study 2, however, had poor interrater reliability for all function forms: linear (ICC = 0.324), asymmetric (ICC = 0.303), quasisymmetric (ICC = 0.141).

Averaged ICC by functional form and condition are presented in Figure 4 via bars. In addition, the Kuncel Tellegen conditions include error terms related to the item with the lowest and highest ICCs in each functional form category. Error terms were highest for asymmetric forms in both Study 1 (error bar span = 0.41) and Study 2 (0.44), with linear in both studies, and quasisymmetric in study 2 were in the 0.30s range (study 1 linear: 0.31, study 2 linear: 0.35, study 2 quasi: 0.36). The lowest error was the quasisymmetric form in study 1 at 0.18 units.

Functional Forms. Figure 5 and 6 represent individual ICCs with their functional forms in the Kuncel and Tellegen

Table 3
Individual ICC

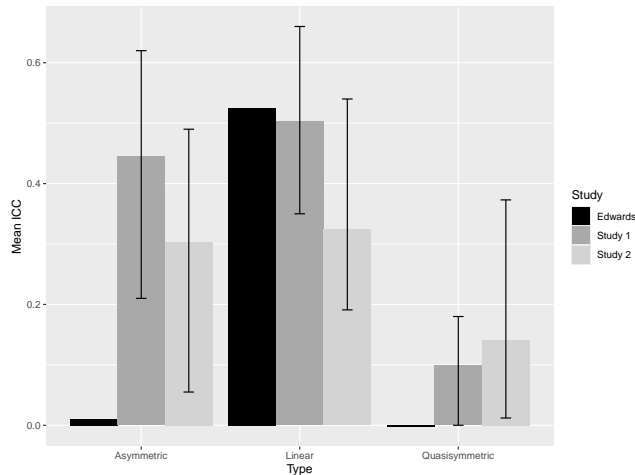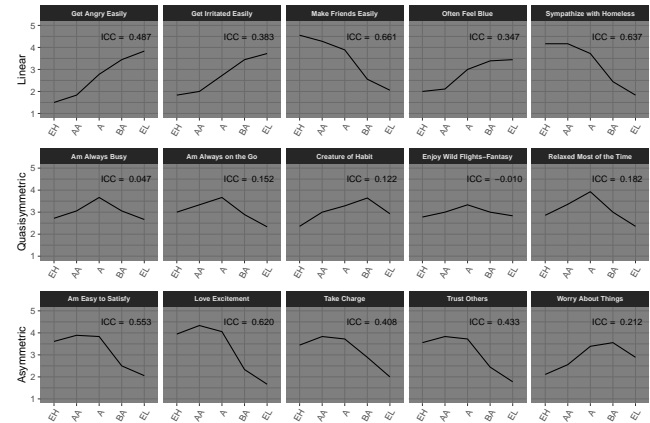|	Study 2	Study 1
Worry about things	0.06	0.21
Trust others	0.36	0.43
Take charge	0.44	0.41
Am easy to satisfy	0.17	0.55
Love excitement	0.49	0.62
Make friends easily	0.54	0.66
Get angry easily	0.22	0.49
Often feel blue	0.19	0.35
Sympathize with the homeless	0.44	0.64
Get irritated easily	0.23	0.38
Am always busy	0.05	0.05
Enjoy wild flights of fantasy	0.22	-0.01
Am always on the go	0.05	0.15
Am relaxed most of the time	0.37	0.18
Am a creature of habit	0.01	0.12
All items	0.25	0.32
Asymmetric Mean	0.30	0.44
Linear Mean	0.32	0.50
Quasi Mean	0.14	0.10

Figure 4. Mean ICC by Item Type

Figure 6. Study 1: Functional Form and ICC by Item

conditions for both study 2 and study 1 for comparison. Visually, linear functional forms stayed consistent between study 1 and study 2. Concerning asymmetric forms, study 1 had more extreme visuals in comparison to study 2, going from ep to mp (but these would still be classified as asymmetric forms). Quasisymmetric forms had two items that visually switched functional forms: "Enjoy Wild Flights of Fantasy" (to linear), and "Relaxed Most of the Time" (to asymmetric).
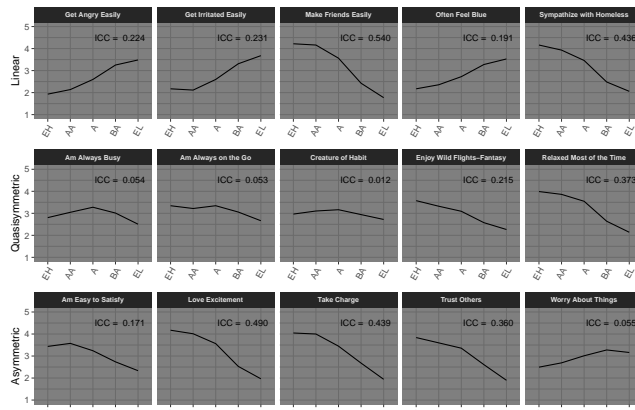

Figure 5. Study 2: Functional Form and ICC by Item

Discussion

Note. Anecdotally the functions appeared more pronounced on one side versus the other - look at Figure 3 and some "hooks" tended to be stronger (more gentle slopes with the other side)

The plurality of findings do support similar information being conveyed through both approaches. Figure 1 captures

some of this, as functional slopes (even if somewhat nonmonotonic - 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 and the "location" of the rating, and 2) Edwards' item stem rating (on the y-axis). This strong relationship (r = -.76) suggests some level of similarity across procedures.

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.

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".

Kuncel and Tellegen (2009) proposed 4 functional types - visual inspection of our empirical functions suggest that additional information lies in the *location* of the function as well. Witches hat has different implications than egyptian or captain hook. Captain hook and witches hat are both "non-monotonic", but differ in "high point" location, either at neutral or moderately desirable locations. Captain hook can likely be treated as effectively linear whereas witches hat should not. Our functional slope index captures these deviations and should perhaps be considered in future investigations (intercept too? If it's witches hat then intercept *most likely* neutral-ish although not necessarily so).

Limitations

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

> **Redundant with previous paragraph** → In order to capture the extremity of function across Kuncel and Tellegen (2009) values, several regressions 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.

Our analytical methodology is atypical - the focus of Kuncel and Tellegen (2009) was visual functions. Our approach therefore also focused on these functions - extracting regression coefficients as an (imperfect) empirical index. Follow-up analyses retained slope coefficients as meaningful representations of the Kuncel and Tellegen (2009) functions. Some of our descriptive correlations are essentially correlations of "correlations", and distal indices such as these are

not ideal (for example, for anticipated replication). We chose to retain and present these analyses as descriptive, but acknowledge that alternative methodologies should be pursued in future investigations.

> Commentary at the end of the references: rated the visual functions 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 3. Careful inspection of the Figure 3 plot again highlights the location of non-monotonic and nonlinear Kuncel and Tellegen functions - predominantly at moderate (around neutral) Edwards rating locations. [NOTE - LOOKS LIKE WE MAY NEED TO REDO THESE ESTIMATES; THERE ARE 96 "NOT AT ALL LOOKS LIKE A STRAIGHT LINE" and 130 "NOT AT ALL RISES AND FALLS"; ALTERNATIVELY GET MORE CREATIVE WITH RESIDUALS ANALYSES (ON DESKTOP)]

> NOTE. Maybe do something with residuals to replace current subjective rating graph - lm object is named, "try" 8/12/23

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. (2023). *Papaja: Prepare american psychological association journal articles with r markdown.* https://github.com/crsh/papaja

Bales, J. F., & Follansbee, G. L. (1935). The after-effect of the perception of curved lines. *Journal of Experimental Psychology*, *18*(4), 499.

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. *In-*

ternational 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.*

Gibson, J. J. (1933). Adaptation, after-effect and contrast in the perception of curved lines. *Journal of Experimental Psychology*, *16*(1), 1–31.

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.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Sage.

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.

Ogilvie, J., & Daicar, E. (1967). The perception of curvature. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *21*(6), 521.

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. (2024). *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.

Watt, R., Ward, R., & Casco, C. (1987). The detection of deviation from straightness in lines. *Vision Research*, *27*(9),

1659–1678.

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. (2023). *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.