UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Kay, C. S., Weston, S. J., & Condon, D. M. (2026). "To Be" or "I am Someone Who Tends To Be": Does the Wording of Personality Items Matter? *Collabra: Psychology*, *12*(1). https://doi.org/10.1525/collabra.150375

Social Psychology

"To Be" or "I Am Someone Who Tends to Be": Does the Wording of Personality Items Matter?

Cameron S. Kay[1,2,3][a], Sara J. Weston[3], David M. Condon[3]

[1] Environmental Social Sciences Department, Stanford University, Stanford, CA, USA, [2] Psychology Department, Union College, Schenectady, NY, USA, [3] Department of Psychology, University of Oregon, Eugene, OR, USA

Keywords: personality descriptors, item wording, response biases, test-retest reliability, response time

https://doi.org/10.1525/collabra.150375

Collabra: Psychology

Vol. 12, Issue 1, 2026

The fundamental unit of data collection for self- and informant-report personality research is the item. In personality structure research, items have historically taken the form of single-word trait descriptors (e.g., "Calm") while, in most personality assessment frameworks, they consist of phrases and sentences (e.g., "I see myself as someone who tends to be calm"). This Registered Report used a randomized, longitudinal design to evaluate the effects of differences in item wording on how people respond to personality measures. Specifically, we examined the effect of four item formats—trait-descriptive adjectives presented alone (e.g., "Talkative"), with the linking verb "am" (e.g., "Am talkative"), with the additional verb "tend" (e.g., "Tend to be talkative"), and with the additional indefinite pronoun "someone" (e.g., "See myself as someone who tends to be talkative")—on how people respond to an adjective-based measure of the Big Five (i.e., the Midlife Development Inventory; Lachman & Weaver, 1997). With few exceptions, the findings indicated that item format had little effect on item means, extreme responding, acquiescent responding, internal consistency, test-retest reliability, and participants' subjective experiences completing the survey. That said, participants responded to shorter item formats (e.g., adjectives presented alone) faster than longer item formats (e.g., adjectives presented with "See myself as someone who tends to be"), especially when they were accompanied by the personal pronoun "I" (e.g., "I am talkative" versus "Am talkative"). The findings from this work have implications for ongoing research related to personality structure and assessment, including the development of more detailed taxonomies of personality, the creation of more generalizable assessment models, and the harmonization of existing assessment models.

1. Introduction

A fundamental assumption in the methodology of psychological assessment relates to the equivalence of self-report ratings based on single-word descriptors (typically, adjectives) and brief phrases containing the same terms (e.g., "Calm" versus "I see myself as someone who tends to be calm"). The assumption is that seemingly trivial differences in stimuli have no effect on response patterns or the interpretations that can be made from data collected using either of these formats. Indeed, there are numerous cases where the assumption of functional equivalence is self-evident. For example, "Talkative" and "Talk a lot" are essentially identical, as the latter is merely a definition of the former.

Yet, face validity is not always adequate to support claims of equivalence. The absence of an effect due to item wording changes is much less clear for some descriptors and wording changes than it is for others. Consider being instructed to rate yourself as "active" or "warm" relative to "I tend to be active" or "I tend to be warm." By specifically invoking tendency, the latter phrasing is more clearly prompting respondents to rate trait-level psychological characteristics across situations, reducing the likelihood that they will respond based on transient psychological (or even physical) states. Similarly, the extent of the differences in phrasing should be expected to increase the likelihood of a meaningful difference. Consider, for example, "Organized" versus "I see myself as someone who tends to be organized." The latter format improves upon the simple adjective prompt by clarifying both the trait/state ambigu-

a Correspondence concerning this article should be addressed to Cameron S. Kay, Environmental Social Sciences Department, 473 Via Ortega, Stanford, CA 94305. E-mail: cameronstuartkay@gmail.com.

ity (by referencing tendency) and the lack of clarity around between-person and within-person ratings (with "...see myself as someone who..."). To be specific, the second format may prompt respondents to rate their tendency to be organized *relative to other people*, as it is akin to saying, "I am an organized type of person" rather than "I feel my life is relatively organized at this moment." Still, the difference is subtle, and the longer form requires more reading on behalf of respondents. Does it really matter?

The primary aim of this registered report was to evaluate the effect of different item wording options using both between-person and within-person analyses. The study design required to evaluate this question also allowed for the evaluation of test-retest reliability at the item level and the effect of wording differences on response biases, response times, and participants' subjective experiences completing the survey. Main effects of item-wording format were not expected on average, though variability was expected depending on the content of the item stem. In other words, we expected that the wording of items might matter for some subset of single-word descriptors and, to the extent that this was the case, we sought to identify the characteristics of descriptors that were most affected. As we discuss below, the results of these investigations are likely to be impactful for subsequent research on personality structure, assessment, and outcomes regardless of the magnitude or absence of significant main effects for item wording.

1.1. Rationale

While open-ended personality questionnaires have been around since at least the 19th century (e.g., Proust's questionnaire responses circa 1886; Kindley, 2016), the shift towards using "objective" inventories with close-ended response options occurred in the early 20th century (Kindley, 2016), driven in part by the benefits of being able to rapidly score responses in large samples and compare them quantitatively (Gibby & Zickar, 2008; Goldberg, 1971; Gough, 1960).

Adjective checklists and phrased item prompts have both been used in these objective tests from the beginning (Horsch & Davis, 1935), and each has its own set of advantages. One benefit of using adjectives over phrased items stems from the lexical rationale for using the linguistic relations among such terms as a proxy for the structure of psychological individual differences as constructs (Ashton, Lee, & Goldberg, 2004; Saucier & Goldberg, 2001). The idea that between-person differences in psychological traits are instantiated in the lexicon of person-descriptors has been explored by many researchers (e.g., Cattell, 1943; Cutler & Condon, 2022; Goldberg, 1992; Norman, 1963; Thurstone, 1934; Tupes & Christal, 1961) since the development of an exhaustive list of terms by Allport and Odbert (1936). Some researchers (e.g., Goldberg, 1999) have argued that the use of such single-word descriptors is more appropriate for research on this topic because the universe of such terms is finite, allowing for the circumscription of potential traits. By contrast, the number and scope of phrased items is effectively unbounded.

A second potential benefit of using adjectives is their relative efficiency (see Briggs, 1992). Adjectives contain, by definition, a single word, while phrased items, by definition, contain two or more words.[1] It is, therefore, easy to imagine that a participant would respond to an adjective prompt faster than they would to a phrased item simply because there is less to read. However, it is important to note that this is only a *potential* benefit. Although a shorter administration time is often mentioned to support the use of adjectives over phrased items (e.g., Briggs, 1992; Hamby & Ickes, 2015; but see Hendriks et al., 1999), there is, to our knowledge, no published research examining differences in response times for content-matched adjectives and phrased items (but see D. Wood et al., 2017).

A third potential benefit is that, unlike phrased items, adjectives can provide both a state- and trait-based assessment of a construct of interest (M. Crowe et al., 2016; M. L. Crowe et al., 2018; Edershile et al., 2019). Consider, for example, the statement "People always seem to recognize my authority" from the Narcissistic Personality Inventory (Raskin & Hall, 1979). Responses to this item would likely be stable across time, especially given the use of the term "always." A person who believes that everyone recognizes their authority at time one is presumably also going to believe that everyone recognizes their authority at time two. However, the item is not able to assess a person's present feelings of authority.[2] In contrast, responses to the adjective "Superior" from the Narcissistic Grandiosity Scale (Rosenthal et al., 2007, 2020; see also M. Crowe et al., 2016) or the adjective "Irritable" from the Narcissistic Vulnerability Scale (M. L. Crowe et al., 2018) can, depending on the instructions, assess either a stable trait or transient state. A person who is asked if they *are* superior or irritable would likely provide the same response at time one and time two, whereas a person who is asked if they *feel* superior or irritable may very well provide different responses at time one and time two. Indeed, previous work has indicated that the Narcissistic Grandiosity Scale and Narcissistic Vulnerability Scale provide theoretically-consistent associations with both transient states (e.g., the Narcissistic Vulnerability Scale is highly positively associated with momentary measures of negative affect) and stable traits (e.g., the Narcissistic Vulnerability Scale is highly positively associated with trait-based measures of neuroticism) when used with the appropriate response format (M. Crowe et al., 2016; M. L. Crowe et al., 2018; Edershile et al., 2019).

1 Phrased items contain two or more words *by definition*, but they very often contain many more. Of the 3,320 items in the IPIP (Goldberg, 1999)—which is known, in part, for having exceptionally brief items (Ashton et al., 2007; Goldberg et al., 2006)—58% contain six or more words.

2 We should note that such statements can also be written to assess transient states (e.g., "People currently recognize my authority").

Despite the potential benefits of adjectives, researchers have recently shown a clear preference for assessment via phrased items. Though dozens of adjective-based instruments continue to be used regularly (Craig, 2005), the majority of findings reported in peer-reviewed personality journals use assessments with phrased items (Zola et al., 2021), most often based on models with five (Goldberg, 1999; McCrae et al., 2005) or six (Lee & Ashton, 2004; Thalmayer et al., 2011) factors. Phrased items—ranging in length from multi-word phrases to complete sentences—are likely preferred because they can be made less abstract and ambiguous than single-word descriptors (but see Walton et al., 2021) and allow for greater contextualization, familiarity, and interpretability (Goldberg, 1999). An item like "I am overly flattering of my co-workers", for example, is more specific, contextualized, and familiar for most respondents than similar single-word descriptors like "unctuous" or "obsequious" (or even simply "flattering"). Indeed, prior work suggests that phrased items tend to yield greater interrater agreement (DeYoung, 2006) and greater test-retest reliability (Chmielewski & Watson, 2009; Watson, 2004) than adjectives. Moreover, phrased items allow researchers to evaluate the manifestation of a construct in specific contexts (Saucier, 2020; see also Kay & Saucier, 2023) and specific time intervals (Carlson et al., 2016; Condon et al., 2020). By way of illustration, Conley & Saucier (2019) showed that being quiet *at parties* is more indicative of extraversion than being quiet *at home*. The recent preference for phrased items may also reflect the increasing consensus that personality structure is reasonably well-described by the so-called "Big Few" models (Mõttus et al., 2020). Widespread adoption of these Big Few models has allowed test developers to focus more on the improvements mentioned above—in essence, improved psychometric properties (i.e., reliability, validity)—and less on the models' structural properties. In fact, several of the most widely used, phrase-based instruments were developed from structural models originally identified with data collected using trait-descriptive adjectives (Goldberg, 1999; John & Srivastava, 1999; Lee & Ashton, 2004).

At least two conversion processes from adjectives to phrased items have been documented in some detail. The Big Five Inventory (BFI; John & Srivastava, 1999) used a "conceptually derived prototype" approach that relied on consensus in the sorting of adjectives into the Big Five domains by 10 psychologists. Terms with the highest factor loadings, based on subsequent observer reports, were then used as "the item core to which elaborative, clarifying, or contextual information was added" (p. 115). Of note, this process was explicitly motivated by prior evidence suggesting that adjectives are answered less consistently than definitions of the same terms when compared with synonyms and antonyms (Goldberg & Kilkowski, 1985; John & Srivastava, 1999). Several popular measures have since been developed, at least in part, using items from the BFI, including the Ten-Item Personality Inventory (Gosling et al., 2003) and the BFI-2 (Soto & John, 2017), as well as numerous versions in other languages (Denissen et al., 2008; Fossati et al., 2011; Lang et al., 2001; see also Ziegler & Bensch, 2013).

A second approach to conversion relied on joint administration of the Big Five Factor Markers (Goldberg, 1992) and phrased items in the International Personality Item Pool (IPIP) to the Eugene-Springfield Community Sample (Goldberg, 1999). The analytic procedures involved in this approach were straightforward, though the data collection requirements were formidable. Approximately 1,250 new phrased items (Goldberg, 1999; Hendriks, 1997) were administered to approximately 800 respondents in the Eugene and Springfield communities of Oregon (Goldberg, 1999).[3] The resulting phrased-item scales were composed of those items that most highly correlated with the adjective-based scales. As with the BFI, other measures have been developed based on the IPIP version of the Big Five Factor Markers (DeYoung et al., 2007; Donnellan et al., 2006), including numerous translations (https://ipip.ori.org/newItemTranslations.htm; Goldberg et al., 2006). This same approach has also been used to create proxies of several proprietary measures for use in the public domain (Ashton et al., 2007; Goldberg, 1999; Goldberg et al., 2006; Witt et al., 2009).

For the current study, an important aspect of the conversion of adjectives to phrased IPIP items was the creation of guidelines for writing new items. These were originally set forth for creating items in Dutch (Hendriks, 1997; Hendriks et al., 1999) but were largely maintained by Goldberg and collaborators during translation and subsequent item creation (Goldberg et al., 2006). To paraphrase Hendriks and colleagues (1999, p. 310), the items were (1) written in the third-person singular for the sake of objectivity; (2) constructed with the most simple item phrasings possible, without negation, to reduce confusion among respondents; (3) constructed, whenever possible, to avoid the use of idioms and/or phrasing that was specific or exclusionary to one or more identity groups; and (4) written *de novo* rather than borrowed from existing measures or item pools. One other guideline used at the outset seems to have been maintained somewhat inconsistently: that items should not make use of trait-descriptive adjectives or type nouns. The logic was that items including such terms would provide little utility relative to existing banks of single-word descriptors and would be less consistently interpreted by participants. Nevertheless, many single-word trait descriptors are present in the IPIP items (e.g., "Am a shy person").

Despite the predominant use of phrased-item inventories in personality science recently, several research initiatives point to the need for a better understanding of the extent to which differences in phrasing alter item-level properties. These include (1) efforts to construct personality taxonomies from the "bottom-up" by beginning with a

3 Roughly 2,050 additional items have been added to the IPIP since the initial study.

detailed characterization of content at the item level (Condon et al., 2021; Mõttus et al., 2020); (2) the ongoing search for more generalizable models of personality structure (Ashton, Lee, Perugini, et al., 2004; De Raad et al., 2010, 2014; Saucier et al., 2014; Thalmayer et al., 2024; Thalmayer, Job, et al., 2020; Thalmayer, Saucier, et al., 2020; J. K. Wood et al., 2020); and (3) the continued assessment of personality in long-running and influential longitudinal panel surveys that continue to use single-word descriptors (Graham et al., 2017; Hill et al., 2011; Juster & Suzman, 1995; Ryff et al., 2019).

With respect to the development of more detailed taxonomies, a better understanding of item wording effects is needed to integrate single-word descriptors into phrased item pools. Despite the arguments made by Hendriks and colleagues (1999), there may be good reason to include these terms in the taxonomy of phrased items. This includes the need to ensure full coverage of the trait descriptor universe and to structurally locate the terms among phrased items, as in a nomological network (Cronbach & Meehl, 1955). It remains an open question whether (and, if so, how) the formatting of these terms—as single-word descriptors or otherwise—alters their characteristics as assessment stimuli.

Beyond taxonomic work, an understanding of these effects will influence best practices for collecting ratings on these descriptors. These practices are relevant for the development of more generalizable personality measures because research in this area continues to rely on ratings collected from adjectives. Multiple research teams conducting exploratory work in numerous languages have reported structurally similar models with two and three dimensions (De Raad et al., 2010, 2014; Saucier et al., 2014; Thalmayer, Job, et al., 2020; Thalmayer, Saucier, et al., 2020). As was done with the Big Five, the next step is to convert these adjective-based models to phrased item inventories, ideally using a protocol that is empirically informed with respect to the effects of phrasing.

Understanding these effects is also relevant to projects aimed at harmonizing personality assessments. Harmonization efforts typically use one of several analytic approaches to compare scores from different measures of the same construct across studies (e.g., Kern et al., 2014; Van den Berg et al., 2014; Zimmerman et al., 2020). Because measures of personality often differ across longitudinal surveys, harmonization work is a priority for integrating findings from lifespan studies of health (e.g., the National Institute on Aging, 2020; the Program on Global Aging, Health, and Policy, 2021). For example, the Health and Retirement Study (HRS; Hill et al., 2011; Juster & Suzman, 1995) and Midlife in the United States (MIDUS; Ryff et al., 2019) have both provided unique information about the role of personality in health and aging but assess personality within a single-word-descriptor framework rather than a phrased item framework (Lachman & Weaver, 1997). Of course, it is feasible to harmonize two or more frameworks by creating a "crosswalk" (Gatz et al., 2015), but a more robust approach would involve developing a standard metric for linking (Schalet et al., 2021), which could be achieved through the creation of a comprehensive taxonomy of personality. Though this work would involve substantial statistical and data collection resources, it would extend the benefits of harmonization from simple retrospective linking between two or more existing scales towards the development of new measures comprised of the best items in each. As a first step, this work requires a better understanding of the effects of phrasing *at the item level*.

Developing such an understanding is challenging, as evidenced by the scarcity of prior work on this topic (for exceptions, see Jones, 2011, and Walton et al., 2021). Though between-person analyses of the differences across item wording formats are straightforward, the within-person analyses needed to evaluate the effects of item wording on response biases and reliability are confounded by practice effects and the stability of the psychological content assessed by each item (Revelle & Condon, 2019). With respect to practice effects (i.e., recall of prior responses to similarly worded items), the difficulty lies in distinguishing consistency due to the similarity of personality attributes described in highly similar items and the motivation of respondents to appear consistent in self-presentation (Hogan & Hogan, 1998; Marcus, 2009). The issue of stability of content is not often considered in personality assessment because psychological traits, as a group, are typically defined as highly stable. Most reporting on reliability makes use of coefficients that measure internal consistency (e.g., Cronbach's alpha) rather than stability (e.g., test-retest decay) despite evidence of variability in the state/trait-ness of various constructs, even within the timespan of a single assessment (Henry et al., 2024; Lowman et al., 2018; D. Wood et al., 2018). This variability is likely pronounced among single-item trait measures (i.e., nuances; Condon et al., 2021; Mõttus et al., 2017). In the current work, we seek to address these methodological challenges with a longitudinal study design that includes random assignment.

1.2. The Current Study

The primary aim of this study was to evaluate the effects of item wording in online, self-report personality assessments. Specifically, we examined the extent to which incremental differences in item wording affect item response distributions, scale reliabilities, item response times, and participant survey-taking experiences. For our personality assessment, we used the Midlife Development Inventory (MIDI; Lachman & Weaver, 1997; J. Smith et al., 2017). The MIDI is a 31-adjective measure of the Big Five that has been frequently used in large-scale, longitudinal surveys, including the HRS (Hill et al., 2011; Juster & Suzman, 1995); the MIDUS (Ryff et al., 2019); the Survey of Midlife in Japan (Ryff et al., 2018); the English Longitudinal Study of Ageing (Steptoe et al., 2013); and the National Health and Aging Trends Study (Kasper & Freedman, 2021). The incremental wording changes considered here include a progression from using trait-descriptive adjectives (1) by themselves (e.g., "talkative"; Format 1), (2) with the linking verb "am" (e.g., "Am talkative"; Format 2), (3) with the additional verb "tend" (e.g., "Tend to be talkative"; Format 3), and (4) with the indefinite pronoun "someone" (e.g., "See myself

as someone who tends to be talkative"; Format 4). Additionally, we considered the effect of including the personal pronoun "I" with Formats 2 through 4 above. These four formats were selected to evaluate each step of the transition from single-word descriptors into full sentences that clearly frame the item as a rating of individuation and tendency (with respect to each descriptor).

We leveraged a repeated measures design—administering each adjective three times to the same participant (with different item format combinations randomized across participants)—to remove variability in responses due to sources other than the format manipulation. As a result, our analyses had more power to detect an effect than if we had used a between-persons design. Moreover, these analyses accounted for memory effects by including data on delayed (i.e., approximately five-minute) and very delayed (i.e., approximately two-week) recall, which were assessed via a memory paradigm similar to that used in the HRS (Runge et al., 2015). Accounting for these memory effects helped ensure that a lack of observed differences between item formats was not simply a consequence of participants remembering how they responded previously and choosing to respond in a consistent manner.

Turning to our specific hypotheses, we did not expect there to be significant differences between pairs of formats in terms of average responses, the incidence of extreme responding, or the incidence of acquiescent responding.[4] We also did not expect to see significant differences in the distribution of scores on the basis of including or excluding the pronoun "I". As described above, there are a multitude of reasons that item wording *could* influence responses to personality scales, but our assumption was that none of these effects would be large enough to result in meaningful differences among the formats. Importantly, this expectation of null effects does not mitigate the significance of the research question. In practice, it seems that many researchers and scale developers have proceeded on the assumption that differences in item formats do not have meaningful effects. If this assumption is inaccurate, additional research would be warranted. In contrast, the absence of evidence (i.e., null effects) would provide some support for the status quo. As such, irrespective of the outcome, we believe the present study would provide useful information for the development and refinement of personality measures.

We also evaluated internal consistency and test-retest reliability (within session and over two weeks) among the formats. Consistent with prior research (Chmielewski & Watson, 2009; DeYoung, 2006; Watson, 2004), we expected slightly higher internal consistency and test-retest reliability estimates for item wording formats that are longer. In other words, we believed reliability would be greatest for Format 4, followed by Format 3, Format 2, and, finally, Format 1. As an added benefit, this investigation allowed us to provide insight into the internal consistencies and test-retest reliabilities of the subscales of a measure of the Big Five commonly used in longitudinal studies (i.e., the MIDI).

We further compared response times as a function of item format. This information was intended to help researchers who regularly use surveys in two ways. First, it was intended to provide empirical data on whether adjective ratings require less time than phrased items to complete, allowing researchers to make better-informed decisions when it comes to choosing between adjectives and phrased items. Second, it was intended to help researchers better estimate the duration of their surveys and, by extension, calculate an appropriate amount to pay their participants when using paid participant panels. We had no a priori hypotheses with respect to differences in response times. While longer items should take longer to read, at least one source has claimed that response times for adjectives are longer than phrased items (Hendriks et al., 1999). For at least some items, the difference in length is likely to be inconsequential. Moreover, at least some of the time needed to respond to an item is constituted by the process of self-evaluation, which should contribute equally to the response durations for each of the four formats.

We also examined whether participants' subjective experiences completing the survey varied as a function of item format. Specifically, we evaluated whether participants who responded to certain formats were more likely to report that they enjoyed completing the survey and/or perceived the survey as better designed. Knowing this information can help researchers design surveys that provide a positive experience for participants. Ensuring participants have a positive experience is not only ethical (i.e., beneficent) but may, in some cases, improve the quality of the collected data (Bowling et al., 2021; see also Meade & Craig, 2012). Again, we had no a priori hypotheses for this line of inquiry.

Finally, exploratory analyses considered preliminary evidence for differences (e.g., mean levels of responses, mean response times) based on the types of devices participants used to complete the survey (i.e., mobile, desktop/laptop, or tablet).

2. Method

The data collection protocol was reviewed by Research Compliance Services at the University of Oregon (Study ID 00000190).

2.1. Participants

Participants were recruited online from Prolific, an on-demand data collection platform. The survey for the first time point (Time 1) was launched at 18:30 UTC on March 31st, 2023. The survey for the follow-up time point (Time 2) was opened 11 days later (18:30 UTC on April 11th, 2023).

4 We expected to see some variability across specific items (at a rate greater than expected by random chance) due to idiosyncratic semantic properties, but we were not exactly sure how these differences would manifest. As such, we tested these differences in an exploratory fashion.

The goal was to have participants complete the Time 2 survey, on average, two weeks after the Time 1 survey. In total, 998 participants completed the Time 1 survey, with 900 of those participants returning to complete the Time 2 survey (9.82% attrition) an average of 12.43 days later. It took participants an average of 12m30s to complete the Time 1 survey and 4m19s to complete the Time 2 survey. Participants were compensated at the U.S. federal minimum wage (US$7.25 per hour) for completing the Time 1 survey. To motivate the participants to complete the Time 2 survey, they were compensated at more than twice the U.S. federal minimum wage (US$15.00 per hour).

The targeted sample size was based on a power analysis that made use of a small pilot sample (reported in the first stage of the Registered Report). Power was estimated using our first model (i.e., the model testing the effect of item format on response means); we expected it to have the lowest power and thus represent a conservative estimate of the necessary number of participants. For simplicity, we estimated power as if this were not a repeated measures model (i.e., by taking each participant's average response and using this as their only contribution to the model). The analysis suggested that 136 participants per condition was sufficient for achieving power greater than 90% for the Time 1 analyses. To ensure adequate power (and because our budget allowed for it), we recruited nearly twice that number: 250 participants per format or 1,000 participants overall.

The targeted participants were 18- to 90-year-olds who reported residing in the United States. To avoid collecting a sample that was disproportionately young—a potential issue for data collected through Prolific (Charalambides, 2021)—we used age quotas to ensure that, at most, 20% of participants were between the ages of 18 and 23 and, at most, 20% of participants were between the ages of 24 and 29. The participants' states of residence were collected to evaluate the possibility of overrepresentation from one or more geographic regions; we did not have sufficient power to analyze the data by state or region. We also asked respondents to self-report their fluency with English on a scale from "Not at all (need translation)" to "Very well (fluent/native)". As the entire survey was administered in English, we removed participants who did not report speaking English "well" or "very well" (n = 1). We further targeted roughly equal proportions of participants whose biological sex assigned at birth was female or male; we also included an "other" response option for biological sex assigned at birth, but we did not anticipate sufficient statistical power to analyze this subsample. We also collected data on educational attainment level, country of origin (U.S. or not, distinct from residence), race/ethnicity, and household income, though we did not plan to use these variables for our primary analyses. In addition to descriptive characterization of the sample, part of the rationale for collecting these demographic variables stemmed from needing a 5-minute delay, on average, between the recall task blocks (see the Measures and Protocol sections below). We also collected device type information for secondary analyses examining (1) whether self-reported device type aligned with device type assessed via embedded capture and (2) whether item means, response times, and participant experiences differed as a function of device type. This was assessed with a single item asking participants whether they were using a mobile phone, desktop/laptop, or tablet.

In addition to the recruitment procedures described above, various exclusion procedures were used. For each participant, we calculated three indices of careless and insufficient effort responding (see Curran, 2016): (a) their average endorsement of attention check items (described in the Measures section), (b) their longest string of identical responses, and (c) their average response time per item. We excluded any participant who (a) had an average response of 4 ("slightly agree") or greater to the attention check items (n_{T1} = 8, n_{T2} = 7), (b) provided the same response to over half of the items (⩾ 21) from a given block in a row (n_{T1} = 5, n_{T2} = 0), or (c) had an average response time of under 1 second or over 30 seconds per item (n_{T1} = 9, n_{T2} = 8). Participants who were excluded based on the criteria at Time 1 were not invited back to participate at Time 2. Following recommended practices for on-demand data collection (Curran, 2016), we also include two questions at the end of the survey asking participants whether they believed their responses were accurate and, if not, why they believed their responses were inaccurate. Responses to the first question were used to flag participants for possible exclusion. Since participants do not necessarily know what qualifies as accurate data, the decision to exclude or not exclude a participant ultimately depended on their response to the second question. No participants who were not already excluded using the other indices of careless and insufficient effort responding provided a reason that would require they be excluded. We intended to exclude participants who did not provide valid Prolific IDs (n = 0), but no participants matched this criterion. We did, however, end up excluding two participants for providing a Prolific ID at Time 2 that did not match a Prolific ID provided at Time 1 (n = 2).

In the end, we had a total sample of 975 participants (48.92% female, 50.77% male). The ages of the participants ranged from 18 to 84 (M_{AGE} = 37.14, SD_{AGE} =14.51). With respect to geographic distribution, the sample included participants from 47 states and the District of Columbia (Alaska, Montana, and Wyoming were not represented). Most participants identified as white (66.67%), with the next largest reported ethnic identities being Black (10.36%), Asian (8.31%), and Hispanic (7.69%). The modal educational attainment was a four-year college degree (33.54%). This was followed by some college or university (18.05%) and only high school (17.33%). The modal household income was between US$20,000 and US$40,000 per year (24.10%), but large proportions of the sample reported household incomes between US$40,000 and US$60,000 per year (20.52%) and between US$60,000 and US$80,000 per year (17.06%).

2.2. Measures

2.2.1. Trait-Descriptive Adjectives

The trait-descriptive adjectives used to evaluate item wording effects were taken from a revised version of the *Midlife Development Inventory* (MIDI; Lachman & Weaver, 1997; J. Smith et al., 2017), which was administered in the HRS. The revised version of the scale includes the 25 original adjective items (Lachman & Weaver, 1997) designed to measure the Big Five traits—Neuroticism (4 items), Extraversion (5 items), Openness to Experience (7 items), Conscientiousness (4 items), and Agreeableness (5 items)—and 6 additional items that were added to provide more coverage of Conscientiousness ("Reckless", "Self-disciplined", "Impulsive", "Cautious", "Thorough", and "Thrifty").[5] All of these adjectives are similar to those found in other common measures of the Big Five. As a case in point, approximately 38.71% of the adjectives in the MIDI can be found in the Big Five Factor Markers (Goldberg, 1992). We, therefore, expect that any results obtained in the present study will generalize to other measures of the Big Five. The adjectives should also be fairly familiar to the average American adult. Prior work has, for example, found that the percentage of participants who can accurately define each adjective in the MIDI ranges from 75% to 97% (*M* = 90%; Median = 92%) (Condon et al., 2022), at least when excluding the relatively ambiguous terms "intelligent" (43%), "nervous" (54%), and "warm" (68%). As such, any results obtained in the present study are unlikely to be due to participants (or a subset of participants) simply not understanding what a given adjective means. Responses to the MIDI items were reverse-scored as needed to be consistent with the MIDI Big Five scoring guidelines (see Table 1).

We also administered seven items from the *Big Five Mini-Markers* (MM; Saucier, 1994). The purpose of including these items was to create balanced subsets of two positively keyed and two negatively keyed items for each trait for the acquiescent responding analyses (described in the Analytic Strategy section below). Since the positively keyed items in the MIDI tend, on average, to be more socially desirable than the negatively keyed items in the MIDI, it was possible that participants would agree with these items not because they are acquiescing but because they are engaging in socially desirable responding. Creating key-balanced pairs was intended to obviate this issue. To preserve our ability to evaluate the MIDI as it is typically administered, these items were not used for any other analyses.

Two items from the *Invalid Responding Inventory for Adjectives* (IDRIA; Kay, 2024) were also included, in part to help evaluate the extent of inattentive responding, but also to consider the effect of item wording on these items. The IDRIA includes three adjectives intended to be endorsed by no one and three adjectives intended to be endorsed by everyone. The measure was designed for use with data collection via paid electronic surveys where participants are often not motivated to expend substantial effort. The two items used here—"asleep" and "human"—were selected from among the more unusual adjectives in the IDRIA (e.g., "Carbonated", "Triangular") because they were relatively inconspicuous and, therefore, less likely to inflate average item response times.[6] The item "human" was reverse-scored so that higher scores on both items reflected greater inattentive responding.

Table 1 shows the 40 MIDI, IDRIA, and MM descriptors used in the present study (although in only one format; see the Protocol subsection for more information about the formats). The items are listed in the same order that they were administered. This ordering is consistent with that used (without the IDRIA and MM items) in ten waves of the HRS (Hill et al., 2011; Juster & Suzman, 1995), three waves of the MIDUS (Ryff et al., 2019), and several other longitudinal panel studies (e.g., Iveniuk et al., 2014; Weston et al., 2020). Although the original MIDI recommended four response options (Lachman & Weaver, 1997), six response options were used for the MIDI, IDRIA, and MM here (1 – "Very Inaccurate", 2 – "Moderately Inaccurate", 3 – "Slightly Inaccurate", 4 – "Slightly Accurate", 5 – "Moderately Accurate", and 6 – "Very Accurate"). These options were based on those used by Goldberg (1990, 1992) for data collection with personality descriptors. Each item was administered separately (one at a time), with the next item loading automatically after participants provided a response. Response times for the MIDI and IDRIA items were measured as the amount of time that passed between each response. For example, the response time for item 22 began when participants selected a response to item 21 and ended when participants selected a response to item 22.

2.2.2. Participant Experience

To assess the participants' subjective experience responding to the survey, we administered two items immediately following the first block of trait descriptive adjectives, one assessing the participants' enjoyment of the survey (i.e., "Overall, I am enjoying responding to the present sur-

5 The original MIDI contained a sixth factor (i.e., agency) that is not used here (Lachman & Weaver, 1997).

6 A Kruskal-Wallis test indicated that there were differences among the item formats for "asleep" (χ^2 (3) = 126.58, p < .001), with follow-up Holm-corrected Wilcoxon tests indicating significant mean differences among all pairings of the formats (except for the pairing of "Tend to be asleep" with "See myself as someone who tends to be asleep"). Critically, "asleep" was rated as more accurate when framed as a tendency—"Tend to be asleep" (*M* = 2.83, *SD* = 1.55) and "See myself as someone who tends to be asleep" (*M* = 2.67, *SD* = 1.38)—than when not framed as a tendency—"Asleep" (*M* = 2.25, *SD* = 1.41) and "Am asleep" (*M* = 1.59, *SD* = 1.09). We would, therefore, caution researchers against using this attention check item with tendency-framing formats in the future. That said, it is worth noting that (a) the majority of participants did rate "asleep" as inaccurate across all four formats (64.82%-90.98%) and (b) the potential for increased false positives resulting from differences in item endorsement for "asleep" across the four formats was largely mitigated by creating a composite that included responses to both "asleep" and "human".

Table 1. The trait-descriptive adjectives.

Adjective	Measure
Outgoing	MIDI – Extraversion
Helpful	MIDI – Agreeableness
Reckless	MIDI – Conscientiousness† - R
Moody	MIDI – Neuroticism - R
Organized	MIDI – Conscientiousness
Friendly	MIDI – Extraversion
Warm	MIDI – Agreeableness
Worrying	MIDI – Neuroticism - R
Responsible	MIDI – Conscientiousness
Lively	MIDI – Extraversion
Asleep	IDRIA – Infrequency
Caring	MIDI – Agreeableness
Nervous	MIDI – Neuroticism - R
Creative	MIDI – Openness to Experience
Hardworking	MIDI – Conscientiousness
Imaginative	MIDI – Openness to Experience
Softhearted	MIDI – Agreeableness
Calm	MIDI – Neuroticism
Self-disciplined	MIDI – Conscientiousness†
Intelligent	MIDI – Openness to Experience
Curious	MIDI – Openness to Experience
Active	MIDI – Extraversion
Human	IDRIA – Infrequency – R
Careless	MIDI – Conscientiousness – R
Broad-minded	MIDI – Openness to Experience
Impulsive	MIDI – Conscientiousness† - R
Sympathetic	MIDI – Agreeableness
Cautious	MIDI – Conscientiousness†
Talkative	MIDI – Extraversion
Sophisticated	MIDI – Openness to Experience
Adventurous	MIDI – Openness to Experience
Thorough	MIDI – Conscientiousness†
Thrifty	MIDI – Conscientiousness†
Quiet	MM – Extraversion‡ - R
Unsympathetic	MM – Agreeableness‡ - R
Relaxed	MM – Neuroticism‡ - R
Uncreative	MM – Openness to Experience‡ - R
Shy	MM – Extraversion‡ - R
Cold	MM – Agreeableness‡ - R
Unintellectual	MM – Openness to Experience‡ - R

Note. MIDI = Midlife Development Inventory; IDRIA = Invalid Responding Inventory for Adjectives; MM = Big Five Mini-Markers. "†" indicates an item that was added to the revised version of the MIDI (J. Smith et al., 2017). "‡" indicates items used solely for the acquiescence analyses. "R" indicates a reverse-scored item.

vey") and one assessing the participants' perceptions of the survey's quality (i.e., "Overall, I think the present survey is well designed"). The two items were prefaced with the text: "In order to improve our future surveys, we would like to learn more about your experience taking this survey. Please indicate the accuracy of the following two statements." Participants responded to the two statements using the same six-point scale used for the MIDI, IDRIA, and MM items.

2.2.3. Recall Tasks

Following the methods used to replicate the word recall task from the HRS for online administration in the Women's Health Valuation (WHV) study (Runge et al., 2015), respondents were presented with 10 English nouns and then asked to recall them with and without a delay. The words were taken directly from four sets of words administered in the HRS. Each participant was assigned, at random, one set of words and shown only the words from that set. The four sets of words are listed below.

1. BOOK, CHILD, GOLD, HOTEL, KING, MARKET, PAPER, RIVER, SKIN, TREE
2. BUTTER, COLLEGE, DOLLAR, EARTH, FLAG, HOME, MACHINE, OCEAN, SKY, WIFE
3. BLOOD, CORNER, ENGINE, GIRL, HOUSE, LETTER, ROCK, SHOES, VALLEY, WOMAN
4. BABY, CHURCH, DOCTOR, FIRE, GARDEN, PALACE, SEA, TABLE, VILLAGE, WATER

At the beginning of the task, participants were informed that they would be shown the 10 words and asked to recall them later. They were instructed to complete the task from memory without aid or writing instruments. Each of the 10 words was then presented separately on the screen for 3 seconds.

The immediate recall task was conducted by asking participants to recall as many words as they could, in any order, by typing their responses into a text box. Scores ranged from 0 to 10, with one point being assigned for each correct word. Correct responses were allowed to include misspelling deviations of up to one letter. Scoring of these responses was automated using the {vwr} package (Keuleers, 2013) in R (R Core Team, 2021). Following the immediate recall task, respondents were asked to complete a masking task, which was estimated to take approximately 2 to 3 minutes to complete. The task instructed participants to "type the names of 20 different kinds (species) of animals." None of the words in the recall task were (non-human) animals.

After completing the masking task, as well as responding to a second block of trait descriptive adjectives and providing demographic information, the participants completed the delayed recall task. The version of the word recall task given in the WHV has a delay of up to 20 minutes after the initial presentation of the words. The delay in this case was expected to be approximately 5 minutes, consistent with the method used in the HRS. In the present study, participants also completed a *very* delayed recall task as part of the Time 2 survey (approximately two weeks later).

2.3. Protocol

The protocol included data collection at two time points. Most of the data was collected at Time 1, with the Time 2 survey being administered approximately two weeks later at the discretion of participants who chose to participate. The protocols for Time 1 and Time 2 are described separately. Figure 1 provides a visual representation of the study design.

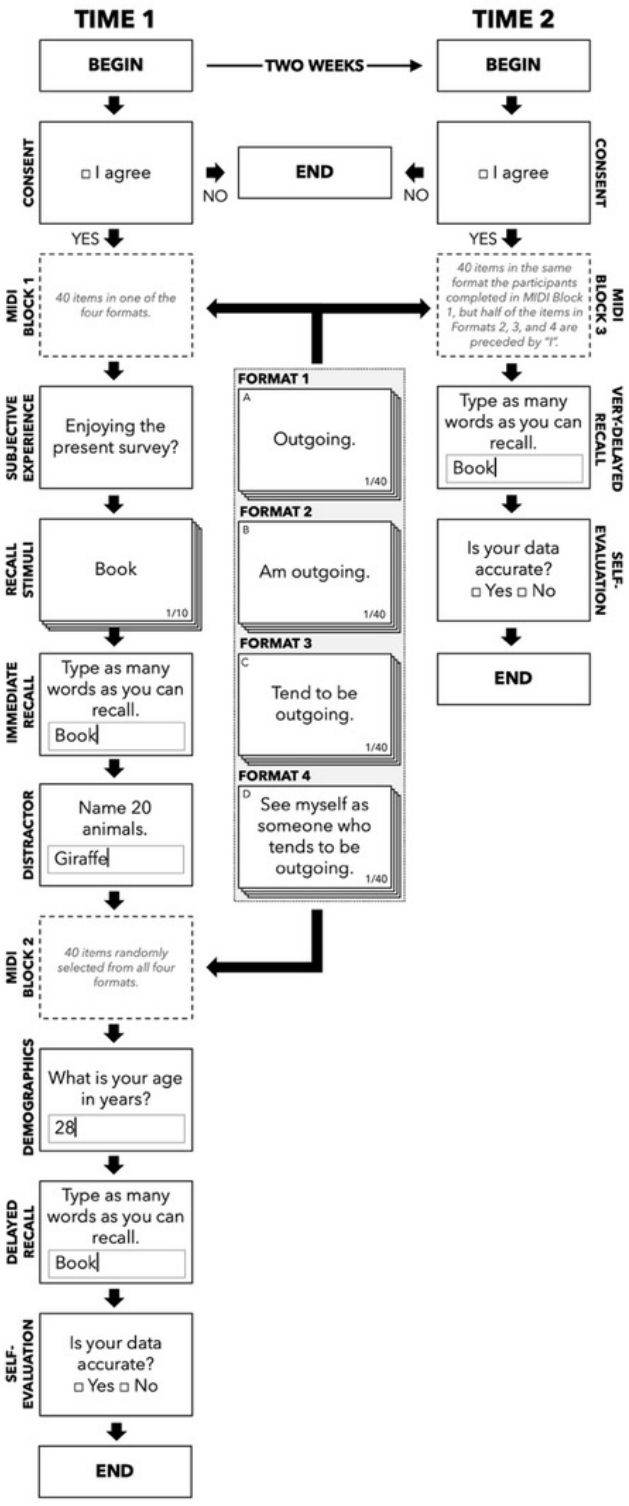

Figure 1. The study design.

2.3.1. Time 1

Participants first completed an informed consent form. They then completed MIDI Block 1. Specifically, they were presented with the simple instruction, "Please indicate how well each of the following describes you," and randomly assigned to complete the 31 items from the MIDI, the 2 items from the IDRIA, and the 7 items from the MM in one of four formats. The four formats are listed below.

- Format 1: The adjective alone (e.g., "Outgoing"). This is the original format used by the MIDI.
- Format 2: Including the first-person singular present conjugation of the linking verb "to be" (e.g., "Am outgoing").
- Format 3: Adding the verb "tend" to indicate regularity or frequency of behaving in a manner consistent with the descriptor (e.g., "Tend to be outgoing"). Besides including single-word trait descriptors, this format is consistent with the guidelines used to write IPIP items (Goldberg et al., 2006; Hendriks et al., 1999).
- Format 4: Adding the indefinite pronoun "someone" (e.g., "See myself as someone who tends to be outgoing"). This format is approximately consistent with that used by the BFI (John & Srivastava, 1999) and BFI-2 (Soto & John, 2017). It does not, however, capture the full range of verb conjugations used in the BFI ("tends to be", "can be", "is", etc.). It also does not include the first-person pronoun "I", which is included with the BFI items ("I see myself…").

For all formats, the MIDI items were presented in the original, fixed MIDI ordering. The IDRIA items ("human" and "asleep") were administered after MIDI item 10 ("lively") and MIDI item 21 ("active"), respectively. The MM items were presented after MIDI item 31 ("thrifty"). Since we wanted to closely approximate how the MIDI is typically administered, we did not want to mix the MM items in with the MIDI items. Unfortunately, this meant that most of the negatively keyed items were presented together at the end of the block.

Following MIDI Block 1, the participants responded to the two questions asking about their subjective experiences taking the survey. They were then presented with the instructions and words for the recall tasks and, subsequently, asked to complete the *immediate* recall task by typing as many words as they could remember from their assigned word group. This was followed by the 2- to 3-minute-long masking task.

The participants then completed MIDI Block 2. MIDI Block 2 included the same descriptors administered in MIDI Block 1 in the same order and with the same instructions. However, each descriptor was presented in one of the four possible formats selected at random. Using this design, an average of approximately one-quarter of the items administered to each participant in MIDI Block 2 were the same as those administered in MIDI Block 1, allowing for assessment of within-session test-retest reliability.

After MIDI Block 2, participants were asked to provide their demographic information, including age, biological sex assigned at birth, location of residence (country and state in the U.S.), self-reported English fluency, and educational attainment. They then completed the *five-minute delayed* free recall task and provided their self-evaluation of whether their data was accurate or not.

2.3.2. Time 2

At the start of Time 2, participants again completed an informed consent form. They then responded to MIDI Block 3, which included the 31 items from the original MIDI, 2 items from the IDRIA, and 7 items from the MM. The items administered to each participant in MIDI Block 3 were presented in the same format as they were in MIDI Block 1. For example, if a participant was administered items in Format 1 in MIDI Block 1, they were administered items in Format 1 in MIDI Block 3. However, each item presented in Formats 2, 3, and 4 was, at random, preceded by the first-person pronoun "I". For example, roughly half of the participants responding to items presented in Format 2 in MIDI Block 3 were shown "Am outgoing" while the other half were shown "*I* am outgoing".

Following MIDI Block 3, the participants completed the *two-week delayed* free recall task. They then self-evaluated the accuracy of their data.

2.4. Analytic Strategy

Descriptions of our statistical analyses are organized around each of the primary research questions. All data were analyzed using open-source statistical software in R (R Core Team, 2021).

2.4.1. Does Item Format Affect the Distribution of Responses?

We tested our primary research question using four linear regression models. Each analysis used data collected in all three MIDI blocks. We first tested whether item format was associated with the mean responses to the personality items. Format was represented in the model using three dummy codes, with Format 1 (adjective only) serving as the reference group. We used a multilevel model design, which allowed us to incorporate the 93 responses provided by each participant across the three blocks into a single model. The model allowed for varying intercepts across participants, items, and blocks:

Level 1:

$$Response_{ijkl} = \beta_{0jkl} + \varepsilon_{ijkl}$$

Level 2:

$$\beta_{0jkl} = \gamma_{0000} + \gamma_{0100}(F2)_j + \gamma_{0200}(F3)_j + \gamma_{0300}(F4)_j + u_{0j00} + u_{00k0} + u_{000l}$$

We used the {glmmTMB} package (Brooks et al., 2017) in R to fit this model to a long-form dataset with each row representing one observation (*i*) per participant (*j*), item (*k*), and block (*l*):

```
glmmTMB(response ~ format + (1 | id) + (1 | item) + (1 | block), data = data)
```

We used an *F*-test to determine whether format had a significant effect on the distribution of responses.

The second and third analyses tested whether item format was associated with *extreme responding* and *acquiescent responding*, respectively. For the extreme-responding analyses, responses of 1 or 6 to the MIDI items were recoded as 1 (i.e., extreme) and all other responses were recoded as 0 (i.e., not extreme). For the acquiescent-responding analyses, responses of 4, 5, or 6 to the key-balanced subsets (described in the Measures section) were recoded as 1 (i.e., yea-saying) and all other responses were recoded as 0 (i.e., nay-saying). Both models used the same equation defined above, with the difference being that the outcome variable was binary. Thus, the models were fit using binary logistic regression. For example:

```
glmmTMB(extreme ~ format + (1 | id) + (1 | item) + (1 | block), data = data, family = "binomial")
```

The fourth analysis examined the effect of including the personal pronoun "I" on the mean responses to the personality items. The presence of "I" was represented in the model using a dummy code, with the absence of "I" serving as the reference group. We tested two models. The first model used the same equation as the format analysis but with the "I" variable added as a predictor. The second model included the interaction between the "I" variable and the "format" variable:

```
glmmTMB(response ~ i * format + (1 | id) + (1 | item) + (1 | block), data = data)
```

We followed up on the four analyses described above by fitting one model for each of the trait-descriptive adjectives. These were, again, multilevel models with responses nested within-person and within-block; an *F*-test was used to determine whether format had an effect on the responses. The 31 *p*-values estimated for each follow-up analysis were adjusted using a Holm correction. For the *F*-tests that were significant after correction, we used pairwise *t*-tests to identify where the differences among the four formats lay, again using a Holm correction to adjust the six *p*-values.

2.4.2. Does Item Format Affect Reliability?

We evaluated this question in two ways. First, with respect to internal consistency, we calculated and reported both Cronbach's alpha and Omega Hierarchical estimates for all formats using data from MIDI Block 1 and MIDI Block 2 only. We calculated 95% confidence intervals for the Cronbach's alpha estimates. Differences in internal consistency were considered statistically significant if the confidence intervals for two formats did not overlap. We also visualized density distributions for all possible split halves for each of the four formats (Revelle & Condon, 2019).

Second, with respect to test-retest reliability, we fit two multilevel models: a model regressing the MIDI Block 2 responses onto the MIDI Block 1 responses and a model regressing the MIDI Block 3 responses onto the MIDI Block 1 responses. For both models, we only allowed the responses to correlate when their formats were the same (e.g., MIDI Block 1 Format 1 with MIDI Block 2 Format 1; MIDI Block 1 Format 2 with MIDI Block 3 Format 2). When standardized within-block, the slope coefficients for the first model and second model approximate five-minute and two-week test-retest reliability correlations, respectively. Multilevel modelling (and, more specifically, nesting item within participant) was further used to more appropriately adjust standard errors by taking into account dependencies among responses from the same participant.

We also tested several additional models. We examined whether the slopes in the above models were moderated by item format, allowing us to examine whether test-retest reliability differed by format. We also examined whether the slopes in the above models were moderated by the participants' performance on the memory task (standardized). When testing reliability between MIDI Block 1 and MIDI Block 2, we used the results from the five-minute delayed recall task from the Time 1 survey. When testing reliability between MIDI Block 1 and MIDI Block 3, we used the two-week delayed recall task from the Time 2 survey. We also calculated test-retest reliability for each item using correlations within session (after approximately five minutes) and between sessions (after approximately two weeks); these are reported in the supplementary material.

2.4.3. Does Item Format Affect Response Durations?

This question was evaluated using analyses that are essentially identical to those described in the "Does item format affect the distribution of responses?" section above. More specifically, we fit multilevel models predicting response times in seconds from item format using data from (1) MIDI Block 1 and MIDI Block 2 only and (2) MIDI Block 1 and MIDI Block 3 only. The second model tested whether the inclusion of the personal pronoun "I" affected response times. Plots of the pilot data suggested that response times were highly positively skewed. We, therefore, used a log-transformation to adjust the response times before analyzing the models.

2.4.4. Does Item Format Affect Participants' Subjective Experiences Completing a Survey?

To address this question, we conducted two ANOVAs. For both ANOVAs, the independent variable was the format participants were assigned in MIDI Block 1. For the first ANOVA, the dependent variable was the participants' responses to the statement asking them whether they enjoyed completing the survey. For the second ANOVA, the dependent variable was their responses to the statement asking them whether they thought the survey was well designed. As above, we used pairwise *t*-tests with a Holm correction to identify where the specific differences among the four formats lay.

2.4.5. Exploratory Analysis: Device Type

We again fit the models described above—specifically using data from MIDI Block 1 and MIDI Block 2—to test whether device type impacted participants' average responses, their response times in log-seconds, and their experiences completing the survey. We fit a model with a single categorical predictor (device), as well as a model with an interaction (device by item format).

2.5. Transparency and Openness

This project was submitted under the Registered Report format. At the time of the Stage 1 submission, we reported how we determined our anticipated sample size, the planned protocol for data collection (including all measures and demographic variables), and the analytic code to be used for testing all research questions. The pilot data and a blinded version of the code were made available to reviewers (https://bit.ly/3jqh5mA).

The code was updated upon revision and at the time of the Stage 2 submission. The following deviations from the analyses specified in our Stage 1 registered report should be noted. First, for our response distribution analyses, we switched from using the {sjPlot} package (Lüdecke, 2025) to using the {marginaleffects} package (Arel-Bundock et al., 2024). The latter better accounted for the sample size and nesting of the multilevel models. Second, for our response time analyses, we excluded twelve observations where the recorded response durations were zero. Not only was this likely a Qualtrics recording error, since it is impossible for a participant to respond to an item without some amount of delay, but taking the natural log of these values made subsequent analyses impossible (the natural log of zero is undefined). Finally, for our device type analyses, we used the Anova() function from the {car} package (Fox & Weisberg, 2019)—which evaluates effects using Chi-square tests—instead of the aov() function from the {stats} package (R Core Team, 2021)—which evaluates effects using F-tests. We encountered an issue where we were unable to extract significance values for interactions using the aov() function.

At the time of publication, all data and code were made publicly available.

3. Results

3.1. Does Item Format Affect the Distribution of Responses?

Overall, item format was associated with the average response to the personality items ($F(3, 59{,}441) = 10.89$, $p < .001$). However, no differences were apparent when considering the confidence intervals for individual pairs of formats (see Figure 2A). Moreover, the largest observed difference, which was between Format 2 and Format 3, only amounted to a mean difference of 0.06 on the six-point scale (Hedges' $g = 0.05$).

When we analyzed each item separately (correcting for multiple comparisons), item format significantly predicted the average response to 20 of the items (Table S7). However, again, very few of the pairwise *t*-tests revealed statistically significant differences (Table S8). The most consistent trend was that endorsement of adjectives tended to be lower when presented with "See myself as someone who tends to be" (Format 4) than when presented alone (Format 1), with "Am" (Format 2), or with "Tend to be" (Format 3). That said, these differences were restricted to six items for Format 1, three items for Format 2, and two items for Format 3.

Next, we tested whether item format was associated with extreme responding (i.e., selecting a 1 or a 6 on the 6-point scale). There were significant differences in the likelihood of extreme responding across the item formats ($F(3, 59{,}441) = 7.29$, $p < .001$) but, again, no differences were apparent when considering the confidence intervals for individual

pairs of formats (see Figure 2B). In this case, the largest observed difference was between Format 3 and Format 4 and was negligible (Hedges' $g = 0.04$).

Again, we tested this effect separately for each trait-descriptive adjective. We found significant differences for 12 of the items (Table S9). As with the item means, pairwise *t*-tests revealed few statistically significant differences (Table S10). The most consistent trend was that extreme responding tended to be higher when an adjective was presented with "See myself as someone who tends to be" (Format 4) than when presented alone (Format 1), with "Am" (Format 2), or with "Tend to be (Format 3). These findings were, however, restricted to four adjectives for Format 1, one adjective for Format 2, and two adjectives for Format 3.

We also tested whether item format was associated with acquiescent responding (i.e., selecting a 4, 5, or 6 on the 6-point scale). There were no differences in the likelihood of acquiescing across the item formats ($F(3, 38{,}002) = 1.96$, $p = .118$). Moreover, no differences were apparent when considering the confidence intervals for individual pairs of formats (see Figure 2C). The largest observed difference was between Format 2 and Format 4 and was negligible (Hedges' $g = 0.02$).

When tested separately for each trait-descriptive adjective, there were significant differences for 12 items (Table S11). As with the two prior outcomes, pairwise comparisons revealed few significant differences (Table S12). The most consistent trend was, again, in relation to Format 4: acquiescent responding tended to be lower when an adjective was presented with "See myself as someone who tends to be" (Format 4) than when presented alone (Format 1), with "Am" (Format 2), or with "Tend to be (Format 3). These findings were restricted to two adjectives for Format 1, two adjectives for Format 2, and one adjective for Format 3. A second trend was that acquiescent responding tended to be lower when an adjective was presented with "Tend to be" (Format 3) than when presented with "Am" (Format 2), but this was only found for three adjectives.

Finally, we used data from MIDI Block 1 and MIDI Block 3 to test whether the inclusion of the word "I" influenced the participants' responses and whether this influence was moderated by item format. We omitted the item format containing only the adjective from these analyses. After controlling for format, the presence of the word "I" ($F(1, 49{,}273) = .384$, $p = .536$) did not significantly account for variance in the responses to the personality items. The effect was negligible (Hedges' $g = 0.00$). When testing the model separately for each personality descriptor, "I" had an effect on the responses for only 3 of the 38 descriptors—responsible ($p < .001$), sympathetic ($p = .040$), and impulsive ($p = .042$)—after correcting for multiple comparisons (Table S13).

The interaction of "I" with item format ($F(2, 49{,}721) = 0.277$, $p = .758$) was also not significant. At the item level, only one of the tests of the interaction was significant (thrifty, $p = .003$) (Table S14).

3.2. Does Item Format Affect Reliability?

Cronbach's alpha and omega hierarchical estimates for each of the Big Five traits across the four formats are displayed in Table 2 and Table 3, respectively. Only one significant difference (defined by non-overlapping confidence intervals) emerged among the Cronbach's alpha estimates: the internal consistency for Openness was higher when presented with "Am" (Format 2) than when presented with "Tend to be" (Format 3). Figure 3 displays the distribution of internal consistency values for all possible split-halves for each trait and item format combination. The distributions of split-halves for all five traits were remarkably similar across the four response formats, even though the distributions differed noticeably across traits. For example, the distributions for Neuroticism were less unimodal than the distributions for Conscientiousness.

The test-retest reliability of items within Time 1 (approximately five minutes apart) was .85 (95% CI [.84, .86]). Memory (i.e., the number of items correctly recalled on the five-minute word recall task) moderated the reliability coefficient ($b = .03$, 95% CI [.02, .04]), but the effect was quite small. Item format did not significantly moderate the reliability coefficient ($F(3, 8{,}253) = .578$, $p = .629$).

The test-retest reliability of items from Time 1 to Time 2 (approximately two weeks apart) was .78 (95% CI [.77, .79]). Memory (i.e., the number of items correctly recalled on the two-week word recall task) moderated the reliability coefficient ($b = .01$, [.00, .02]). Again, this effect was weak. Item format did not significantly moderate the reliability coefficient ($F(3, 32{,}667) = 1.65$, $p = .176$). Test-retest correlations for each item are reported in the supplementary materials (Table S19), broken down by both format and time interval.

3.3. Does Item Format Affect Response Durations?

Item format was significantly associated with the time it took participants to respond to the personality items ($F(3, 73{,}111) = 453.01$, $p < .001$). As shown in Figure 4, faster responses were achieved when the adjectives were presented alone (Format 1), with "Am" (Format 2), or with "Tend to be" (Format 3) than when presented with "See myself as someone who tends to be" (Format 4). The effects were small in absolute terms but sizeable compared to the other effects observed thus far (Hedges' $g_{F1\text{-}F4} = 0.31$; Hedges' $g_{F2\text{-}F4} = 0.24$; Hedges' $g_{F1\text{-}F4} = 0.20$).

Item-specific analyses found significant differences for all of the descriptors, even after correcting for multiple comparisons (Table S21). Follow-up pairwise comparisons indicated that the differences were primarily localized to "See myself as someone who tends to be" (Format 4). For most adjectives, responses were faster when the adjectives were presented alone (Format 1), with "Am" (Format 2), or with "Tend to be" (Format 3) than when presented with "See myself as someone who tends to be" (Format 4).

We pooled data from MIDI Block 1 and MIDI Block 3 to test whether the inclusion of the personal pronoun "I" changed the time it took participants to respond to the personality items. Counterintuitively, when controlling for

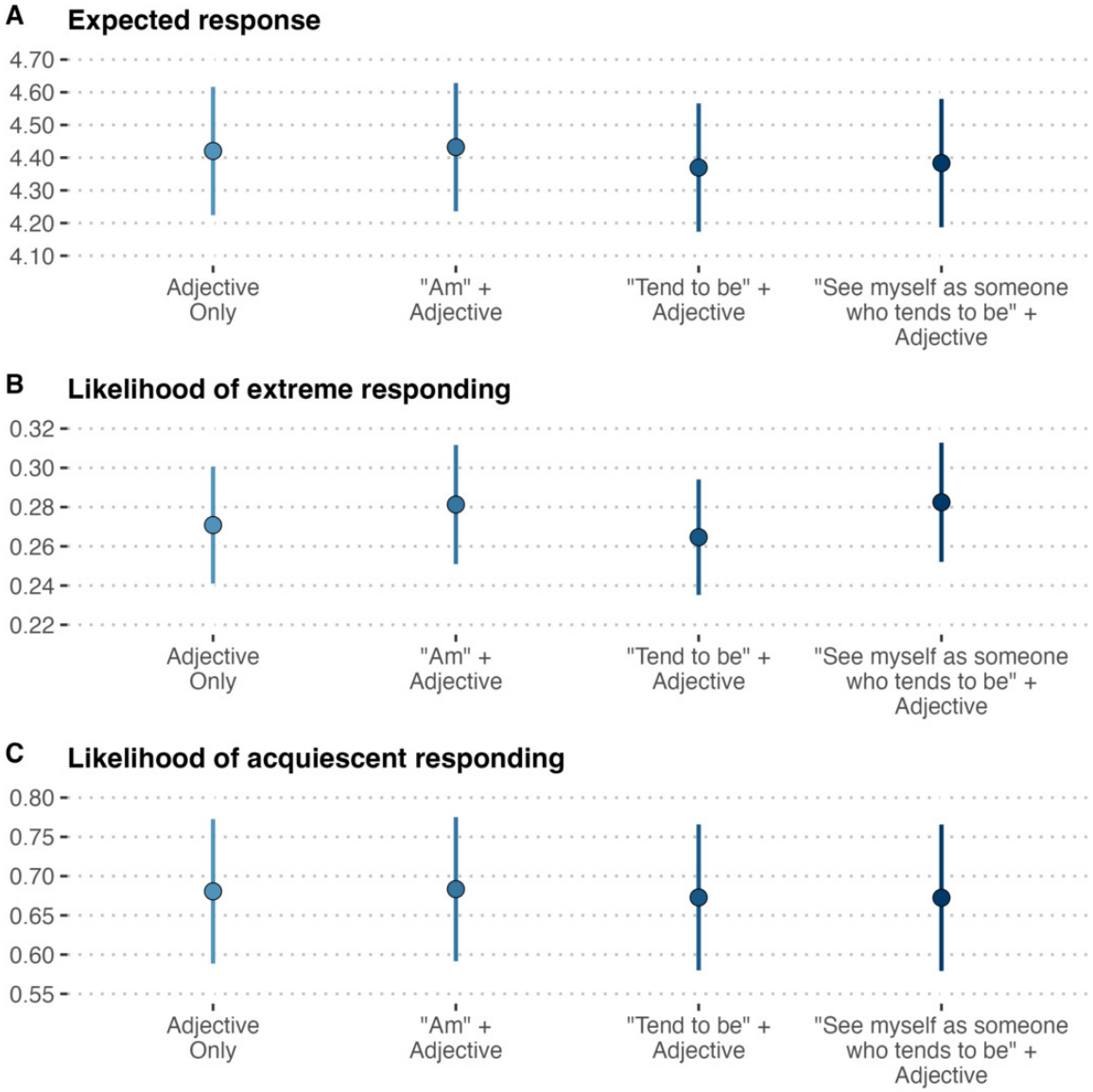

Figure 2. Effects of item format on (A) item means, (B) incidence of extreme responding, and (C) incidence of acquiescent responding.

Table 2. Cronbach's alpha estimates of internal consistency for each of the Big Five subscales.

	Format 1	Format 2	Format 3	Format 4
Extraversion (5 descriptors)	0.80 [0.77, 0.82]	0.82 [0.80, 0.85]	0.84 [0.82, 0.86]	0.81 [0.78, 0.83]
Agreeableness (5 descriptors)	0.90 [0.89, 0.91]	0.90 [0.88, 0.91]	0.90 [0.88, 0.91]	0.92 [0.91, 0.93]
Conscientiousness (10 descriptors)	0.83 [0.80, 0.85]	0.85 [0.82, 0.87]	0.80 [0.78, 0.83]	0.84 [0.81, 0.86]
Neuroticism (4 descriptors)	0.83 [0.81, 0.86]	0.86 [0.84, 0.88]	0.82 [0.79, 0.84]	0.83 [0.81, 0.86]
Openness (7 descriptors)	0.76 [0.72, 0.79]	0.68 [0.64, 0.73]	0.77 [0.73, 0.80]	0.72 [0.68, 0.76]

Note. Format 1 = Adjective Only; Format 2 = "Am" + Adjective; Format 3 = "Tend to be" + Adjective; Format 4 = "See myself as someone who tends to be" + Adjective.

item format, the inclusion of "I" led to slightly *faster* responding ($F(1, 49{,}611) = 6.46$, $p = .011$). The response duration per item with "I" (expected value of 2.03 seconds) was 1.4% faster than the response duration per item without "I" (expected value of 2.07 seconds). This difference was negligible (Hedges' $g = 0.02$).

When considering each item individually, the inclusion of "I" only had a significant effect on "outgoing". Participants took an average of 1.30 seconds longer to respond when "I" was absent from items that included this adjective (Table S61).

In a separate model, we further found that the inclusion of "I" moderated the effect of item format on timing ($F(2,$

Table 3. Omega Hierarchical estimates of internal consistency for each of the Big Five subscales.

	Format 1	Format 2	Format 3	Format 4
Extraversion (5 descriptors)	0.75	0.76	0.77	0.75
Agreeableness (5 descriptors)	0.89	0.82	0.82	0.88
Conscientiousness (10 descriptors)	0.67	0.65	0.54	0.55
Neuroticism (4 descriptors)	0.80	0.84	0.81	0.79
Openness (7 descriptors)	0.62	0.56	0.66	0.53

Note. Format 1 = Adjective Only; Format 2 = "Am" + Adjective; Format 3 = "Tend to be" + Adjective; Format 4 = "See myself as someone who tends to be" + Adjective.

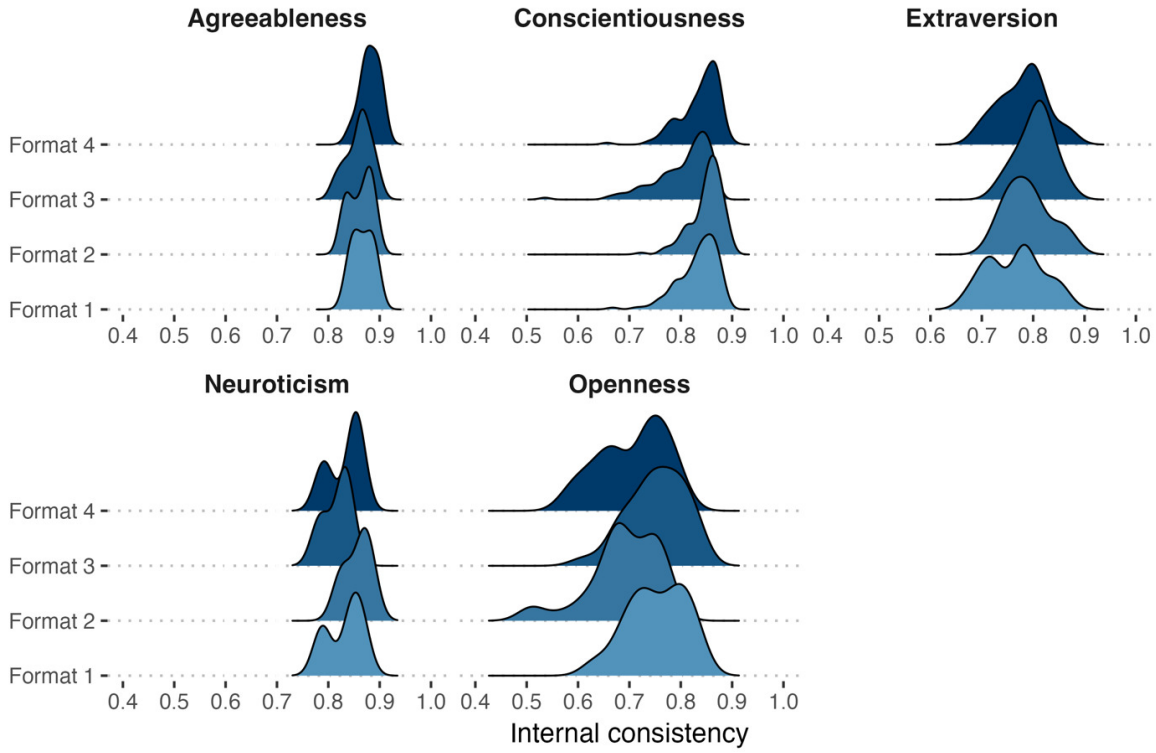

Figure 3. Distribution of internal consistency values for all possible split-halves by trait and format.

Note. Format 1 = Adjective Only; Format 2 = "Am" + Adjective; Format 3 = "Tend to be" + Adjective; Format 4 = "See myself as someone who tends to be" + Adjective.

49,609) = 7.43, $p < .001$). However, when the formats were considered individually, there did not appear to be significant differences between including and not including "I". We also did not find significant differences when considering each item individually after correcting for multiple comparisons (Table S62).

3.4. Does Item Format Affect Participants' Subjective Experiences Completing a Survey?

Item format did not significantly influence how enjoyable the participants found the survey ($F(3, 971) = 1.65$, $p = .176$) nor their belief that the survey was well-designed ($F(3, 971) = 1.26$, $p < .288$). The largest difference observed for enjoyment was between Format 1 and Format 4 and was small (Hedges' $g = 0.18$); the largest difference observed for design quality was also between Format 1 and Format 4 and was, again, small (Hedges' $g = 0.15$).

3.5. Exploratory and Post Hoc Analyses

We ran a number of exploratory analyses examining whether the type of device a participant used affected their average responses, response times, and subjective experiences completing the survey. There was no effect of device type on a participant's average response ($\chi^2(2) = 0.12$, $p = .942$). The largest difference observed was between desktop/laptop and tablet. The effect was negligible (Hedges' $g = 0.02$). Device type also did not moderate the effect of item format on a participant's average response (χ^2 (6) = 1.62, $p = .951$).

The type of device a participant used did have an effect on their response times ($\chi^2(2) = 9.29$, $p = .010$): participants using desktops/laptops responded to the survey significantly faster than those using tablets, with participants using cellphones falling somewhere in the middle. The largest difference, which was between desktop/laptop and mobile, was negligible (Hedges' $g = 0.02$). The type of device a par-

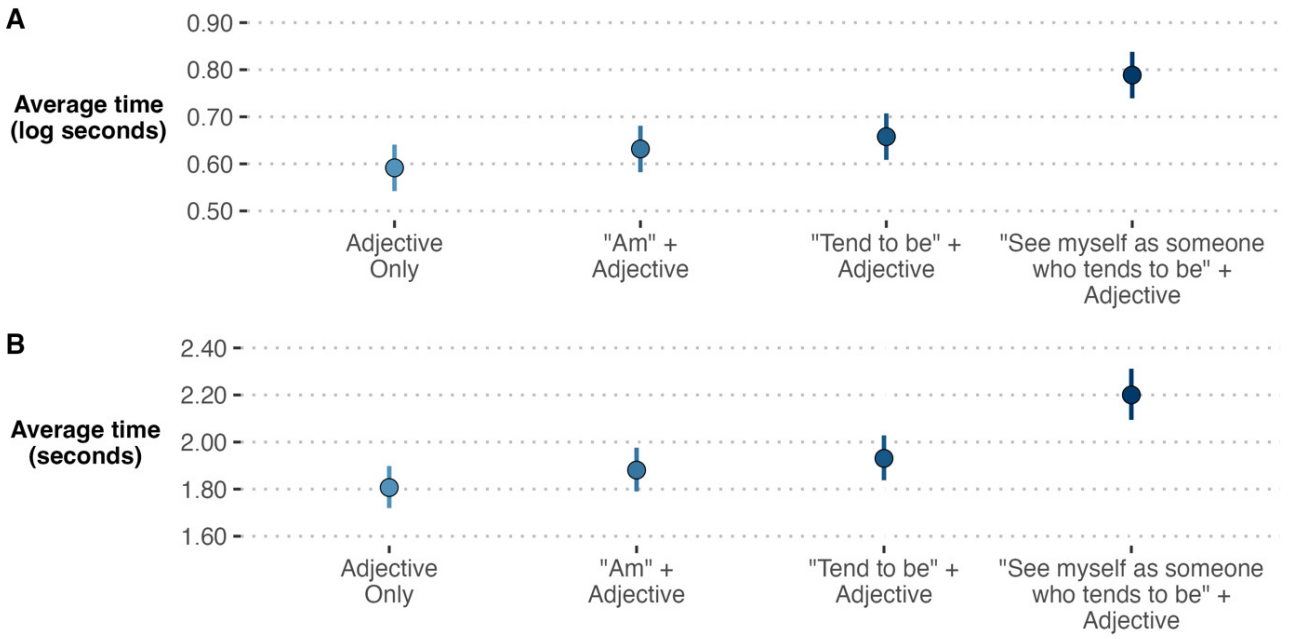

Figure 4. Effects of item format on (A) response times in log-seconds and (B) response times in seconds.

ticipant used also interacted with item format to predict response times ($\chi^2(6) = 13.52$, $p = .035$). That said, few significant differences appeared among the types of devices when considering each of the four formats individually.

The type of device a participant used did not influence their enjoyment of the survey ($F(2, 972) = 1.41$, $p = .245$) nor did it moderate the effect of item format on their enjoyment of the survey ($F(6, 963) = 0.88$, $p = .509$). The largest difference among the devices, which was between desktop/laptop and tablet, was small (Hedges' $g = 0.14$).

The type of device a participant used *did* significantly influence the perceived quality of the study ($F(2, 972) = 3.11$, $p = .045$). However, follow-up pairwise comparisons revealed no significant differences. The largest difference among the devices, which was between tablet and mobile, was small (Hedges' $g = 0.31$). The type of device a participant used did not significantly moderate the effect of item format on the perceived quality of the study ($F(6, 963) = 0.41$, $p = .871$).

4. Discussion

Researchers have commonly operated under the assumption that the general wording of items doesn't matter. If this assumption is correct, it suggests that researchers can subtly reword their items without distorting their results. If this assumption is incorrect, it suggests researchers are overlooking a source of variation in their personality assessments. The purpose of the present study was to test this assumption.

We considered four item formats: adjectives presented alone (e.g., "Talkative"; Format 1), with the linking verb "am" (e.g., "Am talkative"; Format 2), with the additional verb "tend" (e.g., "Tend to be talkative"; Format 3), and with the additional indefinite pronoun "someone" (e.g., "See myself as someone who tends to be talkative"; Format 4). We also considered the effect of adding the personal pronoun "I" to the latter three formats. The formats were examined for their effects on item means, extreme responding, acquiescent responding, internal consistency, test-retest reliability, and item response durations, as well as on the participants' subjective experiences completing the survey. An overview of the findings is provided in Table 4.

There was little evidence that item wording meaningfully influenced how participants responded to the survey. Item wording had a statistically significant effect on the average responses to the items, but few pairwise differences emerged, even when considered at the item level, and the effects were all exceptionally small (Hedges' *g*s = 0.00 to 0.05). Similar results were observed for the effect of item wording on extreme responding: while the overall effect was significant, few pairwise differences emerged, and the effects were all negligible (Hedges' *g*s = 0.00 to 0.04). The overall effect of item wording on acquiescent responding departed from this pattern by being both not statistically significant and negligible (Hedges' *g*s = 0.00 to 0.02). Based on these findings, we conclude that a researcher's choice of format, at least in the forms tested here, is unlikely to be of much consequence for their observed response distributions.

Despite prior evidence suggesting that longer item formats are more reliable than shorter item formats (Chmielewski & Watson, 2009; DeYoung, 2006; Watson, 2004), we found little support for this in our results. Turning first to the internal consistencies of the scales, only one significant difference was found. Namely, presenting an adjective with "Am" (Format 2) resulted in a significantly lower Cronbach's alpha for openness than presenting an adjective with "Tend to be" (Format 3). Given that this occurred for only one format-scale pairing, we suspect that

Table 4. Summary of findings and overall conclusions.

Research Question	Answer
Primary	
Does item wording affect item means?	Yes, but the differences are unlikely to be consequential. • There was an overall effect of item wording on item means but few pairwise differences. The effects were negligible (Hedges' *g*s = 0.00 to 0.05). • The inclusion of "I" did not affect the item means. The effect was negligible (Hedges' *g* = 0.00).
Does item wording affect extreme responding?	Yes, but the differences are unlikely to be consequential. • There was an overall effect of item wording on extreme responding but few pairwise differences. The effects were negligible (Hedges' *g*s = 0.00 to 0.04).
Does item wording affect acquiescent responding?	No. • Item wording did not affect acquiescent responding. The effects were negligible (Hedges' *g*s = 0.00 to 0.02).
Does item wording affect internal consistency?	Yes, but the differences are unlikely to be consequential. • The internal consistency of Openness was greater when adjectives were presented with "Am" (Format 2) than with "Tend to be" (Format 3). The difference was small (Δα = .09)
Does item wording affect test-retest reliability?	No. • Item wording did not affect test-retest reliability assessed across five minutes nor across two weeks. In both cases, the effects were negligible ($\Delta\beta_{T1\text{-}T2}$s = .001 to .017; $\Delta\beta_{T1\text{-}T3}$s = .006 to .019).
Does item wording affect response durations?	Yes, and the differences could be consequential. • Response times were faster when the adjectives were presented alone (Format 1), with "Am" (Format 2), or with "Tend to be" (Format 3) than when presented with "See myself as someone who tends to be" (Format 4). The effects were negligible to small (Hedges' *g*s = 0.04 to 0.31). • The inclusion of "I" resulted in faster response durations. The effect was negligible (Hedges' *g* = 0.02).
Does item wording affect how enjoyable a survey is?	No. • Item wording did not affect how enjoyable people found the survey. The effects were negligible to small (Hedges' *g*s = 0.04 to 0.18).
Does item wording affect the perceived design quality of a survey?	No. • Item wording did not affect the perceived design quality of the survey. The effects were negligible to small (Hedges' *g*s = 0.02 to 0.15).
Exploratory	
Does device type affect item means?	No. • Device type did not affect item means. The effects were negligible (*g*s = 0.00 to 0.02).
Does device type affect response durations?	Yes, but the differences are unlikely to be consequential. • Response durations were faster for computer respondents than survey respondents, with cellphone respondents falling in the middle. The effects were negligible (Hedges' *g*s = 0.00 to 0.02).
Does device type affect how enjoyable a survey is?	No. • Device type did not affect how enjoyable participants found the survey. The effects were negligible to small (Hedges' *g*s = 0.02 to 0.14).
Does device type affect the perceived design quality of a survey?	Yes, and the differences could be consequential. • Device type influenced the perceived design quality of the survey but follow-up pairwise comparisons revealed no significant differences. The effects were small (Hedges' *g*s = 0.14 to 0.31) but comparable to those seen for the effect of item wording on response durations.

it simply reflects random variation, though it is possible that aspects of openness co-occur more strongly when presented as tendencies than as parts of one's identity. Item formatting also had little effect on item-level test-retest reliability, whether it was assessed after a five-minute or two-week delay. Interestingly, the memory abilities of the participants *did* impact test-retest reliability. Participants who were able to recall more to-be-remembered words on a recall task exhibited a greater consistency in responses to personality items administered both five minutes and two weeks apart. Given that people have a seemingly innate desire to be consistent (e.g., Festinger, 1957), it is understandable that those with slightly better recall would show greater consistency in responding. Still, the apparent link between memory and response consistency is an issue for assessing test-retest reliability as it conflates recall ability with the stability of the construct being assessed. Though the possibility of a "memory effect" has been proposed as a potential issue with test-retest reliability (Kuder & Richardson, 1937; Lowman et al., 2018), the present study is the first, to our knowledge, to find evidence directly linking greater recall abilities to greater observed test-retest reliabilities.

One of the few outcome variables that showed clear differences across the four formats was item response duration. The shorter an item was (e.g., an adjective presented alone versus an adjective presented with "am"), the faster a participant could respond to it. We did not expect this. We had assumed that most of the time a participant spends responding to an item would be directed toward deciding how to respond to the item rather than reading the item. This finding does, however, align with the contention from a number of researchers that adjectives are more efficient to administer than phrased items (e.g., Briggs, 1992; Hamby & Ickes, 2015; but not Hendriks et al., 1999). This is an important finding, at least to the extent that researchers are interested in minimizing data collection costs and increasing the speed of data collection. Of course, it is only one of several contextual factors to consider, but the difference is sizable—about 22% more time is needed to complete a survey using the "See myself as someone who tends to be" format (Format 4) than a comparable survey made up of adjectives (Format 1).

Counterintuitively, we also found that including the pronoun "I" resulted in *shorter* response durations. Again, this finding was unexpected, but it may be due to "I" increasing the readability of items. For example, the item "I am talkative" is presumably more familiar to participants and, therefore, easier to read than "Am talkative". Whatever the reason for the difference, this finding provides evidence in favor of including "I" in one's items and against the common practice of dropping the pronoun (e.g., Goldberg et al., 2006). However, the time savings were quite modest, with only about 2% more time being required to administer an item without "I".

Finally, there was no evidence that item wording had an effect on the participants' enjoyment of the survey or their beliefs that the survey was well-designed. This suggests that the formatting of items (at least in the forms considered here) does not influence a participant's subjective experience completing a survey.

Taken together, the present findings suggest that item formatting has little effect on a scale's psychometric qualities. This supports the idea that researchers will not adversely affect their results by introducing slight variations in the wording of their items (at least along the lines of the variations tested here). That being said, if we were going to recommend one format, it would be Format 1 (i.e., presenting adjectives alone), solely because responding to adjectives alone takes less time than responding to adjectives in any of the other three formats. For a similar reason, if a researcher were going to use one of the longer formats, we would recommend including the personal pronoun "I". The time savings are modest, but its inclusion comes with no apparent drawbacks.

4.1. Implications

This study is part of programmatic research that seeks to inform and expand the taxonomy of personality traits (Condon et al., 2021; Mõttus et al., 2020). As such, there are several related topics to pursue in subsequent research. The most prominent is to situate comprehensive sets of personality descriptors (e.g., Condon et al., 2022; Goldberg, 1982) among existing pools of phrased items (e.g., the IPIP), preferably with a better understanding of the effects of wording formats. The implications of this work are relevant to survey-based evaluations of personality structure, as well as natural-language-processing-based methodologies using transformer architectures (Cutler & Condon, 2022; Hommel & Arslan, 2025). The goals of this taxonomic work extend beyond empirical documentation of these associations to include better integration of existing frameworks, a more informed understanding of personality structure, and the development of new personality assessment tools.

4.2. Limitations

Multiple limitations of this study should be noted. First, the study design could only address the influence of format on adjective-based items (e.g., "Talkative" versus "I tend to be talkative."). It could not address how adjective-based items perform relative to non-adjective-based items (e.g., "Talkative" versus "I enjoy chatting with strangers at parties"). This is a by-product of our intention to study the effects of formatting specifically. Without further investigation, it cannot be assumed that the present results would generalize across items containing dissimilar content or even across items containing additional changes in wording beyond those evaluated here. As a case in point, Rammstedt and colleagues (2022) found that removing content from items to create simplified items can hurt factorial validity (despite having little effect on average responses, internal consistency, and criterion validity).

A second limitation of the study design is particularly relevant to evaluating the effect of item formatting on response times. Specifically, it is yet unclear whether some (or all) participants increasingly ignore content that is repeated across items. For example, it is possible that some participants in the present study ignored the "See myself as someone who tends to be" portion of the items after the first few presentations and focused only on the personality descriptors at the end. The differences in response times among the formats in the present study may, therefore, actually be an underestimate of the true differences. Addressing this limitation may require a more complex study design or the use of non-survey-based assessment techniques (e.g., eye-tracking, cognitive interviewing).

Another limitation is that our efforts to control for memory effects using word recall scores only partially addressed the issue of participants remembering their previous responses to items. Prior evidence suggests that respondents are particularly likely to recall and provide consistent responses for traits that they deem salient, surprising, or otherwise important (e.g., Brunot & Sanitioso, 2004; Hastie & Kumar, 1979; Leyens et al., 1997; E. R. Smith & Henry, 1996). Recall effects are inevitable in all studies involving repeated measurement of the same stimuli, but the relevance of this concern to item wording effects in particular warrants further research with a variety of study designs.

Fourth, the study examined the effect of item format on the consistency (i.e., reliability) and efficiency (i.e., time duration) of responses, but it did not examine the accuracy

(i.e., validity) of responses. For example, we did not test whether a measure of neuroticism presented in Format 1 was more associated with anxiety than a measure of neuroticism presented in Format 2, nor whether a measure of extraversion presented in Format 3 was more associated with talkativeness than a measure of extraversion presented in Format 4. These are certainly important questions to explore, but they were beyond the scope of the preliminary investigation outlined here.

Fifth, our investigation was mostly centered on a single measure of the Big Five (i.e., the MIDI). As noted in the Introduction and Method sections, the MIDI has been used widely and demonstrates substantial overlap with other adjective-based Big Five measures. Nevertheless, it is possible that some of the present findings could be specific to the MIDI. In the present study, we tested this possibility, in part, by examining whether the results for each trait-descriptive adjective differed from the overall pool of adjectives. We found few such differences, indicating that item formatting may influence non-MIDI items in a similar manner to MIDI items.

A final limitation is in relation to our sample. Although we used quotas to ensure that we had a representative range of ages and roughly equal proportions of people reporting their biological sex assigned at birth as female and male, our sample was still comprised entirely of participants from the US. As such, our results should not be assumed to be generalizable to samples recruited from other countries or cultures. Additionally, our results should not be assumed to be generalizable to samples that are not drawn from online data collection platforms. Participants recruited from online data collection platforms, such as Prolific, may be better versed in taking surveys than participants who do not frequent these platforms, which could, in turn, influence how they react to differences in item wording.

5. Conclusion

The present study examined the effects of four item wording formats on the psychometric properties of a scale. Item wording appeared to have very little effect on the scale's psychometric properties, other than shorter wording formats translating to shorter response durations. This set of findings is an important first step in the path toward creating more generalized, unified, and comprehensive models of personality.

Author Contributions

C.S.K.: Conceptualization, Investigation, Methodology, Project administration, Visualization, Writing - original draft, and Writing - review & editing.

S.J.W.: Data curation, Formal analysis, Methodology, Validation, Visualization, Writing - original draft, and Writing - review & editing.

D.M.C.: Conceptualization, Data curation, Investigation, Methodology, Project administration, Software, Supervision, Validation, Visualization, Writing - original draft, and Writing - review & editing.

Competing Interests

We have no known conflicts of interest to disclose.

Acknowledgements

We gratefully acknowledge the assistance of Kelsey L. Condon with software coding for the pilot data collection.

Data Accessibility Statement

The Stage 1 Registered Report was registered in its approved state on OSF (https://osf.io/mfupw/overview?view_only=fae419e322dc44f8b218eae9da61d266). The deidentified data and analytic code for this study are also provided on OSF (https://osf.io/5nry8/overview?view_only=df1fd2cad0e84a749ac4f0515cda9925). An accompanying website with analytic code and results for this project can be found on GitHub (https://pie-lab.github.io/wording-effects/).

Editors: Brent Donnellan (Senior Editor)

Submitted: May 20, 2025 PST. Accepted: November 10, 2025 PST. Published: January 05, 2026 PST.

References

Allport, G., & Odbert, H. (1936). Trait names: A psycho-lexical study. *Psychological Monographs*, *47*(1, Whole No. 211). https://doi.org/10.1037/h0093360

Arel-Bundock, V., Greifer, N., & Heiss, A. (2024). How to interpret statistical models using {marginaleffects} for {R} and {Python}. *Journal of Statistical Software*, *111*(9), 1–32. https://doi.org/10.18637/jss.v111.i09

Ashton, M. C., Lee, K., & Goldberg, L. R. (2004). A hierarchical analysis of 1,710 English personality-descriptive adjectives. *Journal of Personality and Social Psychology*, *87*(5), 707. https://doi.org/10.1037/0022-3514.87.5.707

Ashton, M. C., Lee, K., & Goldberg, L. R. (2007). The IPIP–HEXACO scales: An alternative, public-domain measure of the personality constructs in the HEXACO model. *Personality and Individual Differences*, *42*(8), 1515–1526. https://doi.org/10.1016/j.paid.2006.10.027

Ashton, M. C., Lee, K., Perugini, M., Szarota, P., de Vries, R. E., Di Blas, L., Boies, K., & De Raad, B. (2004). A six-factor structure of personality-descriptive adjectives: Solutions from psycholexical studies in seven languages. *Journal of Personality and Social Psychology*, *86*(2), 356–366. https://doi.org/10.1037/0022-3514.86.2.356

Bowling, N. A., Gibson, A. M., Houpt, J. W., & Brower, C. K. (2021). Will the questions ever end? Person-level increases in careless responding during questionnaire completion. *Organizational Research Methods*, *24*(4), 718–738. https://doi.org/10.1177/1094428120947794

Briggs, S. R. (1992). Assessing the Five-Factor Model of personality description. *Journal of Personality*, *60*(2), 253–293. https://doi.org/10.1111/j.1467-6494.1992.tb00974.x

Brooks, M. E., Kristensen, K., van Bentham, K. J., Magnusson, A., Berg, C. W., Nielsenm, A., Skaug, H. J., Maechler, M., & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*(2), 378–400. https://doi.org/10.3929/ethz-b-000240890

Brunot, S., & Sanitioso, R. B. (2004). Motivational influence on the quality of memories: Recall of general autobiographical memories related to desired attributes. *European Journal of Social Psychology*, *34*(5), 627–635. https://doi.org/10.1002/ejsp.220

Carlson, E. B., Field, N. P., Ruzek, J. I., Bryant, R. A., Dalenberg, C. J., Keane, T. M., & Spain, D. A. (2016). Advantages and psychometric validation of proximal intensive assessments of patient-reported outcomes collected in daily life. *Quality of Life Research*, *25*, 507–516. https://doi.org/10.1007/s11136-015-1170-9

Cattell, R. B. (1943). The description of personality: 2. Basic traits resolved into clusters. *Journal of Abnormal and Social Psychology*, *38*, 476–507. https://doi.org/10.1037/h0054116

Charalambides, N. (2021). *We recently went viral on TikTok - here's what we learned*. Prolific. https://blog.prolific.co/we-recently-went-viral-on-tiktok-heres-what-we-learned/

Chmielewski, M., & Watson, D. (2009). What is being assessed and why it matters: The impact of transient error on trait research. *Journal of Personality and Social Psychology*, *97*(1), 186–202. https://doi.org/10.1037/a0015618

Condon, D. M., Chapman, R., Shaunfield, S., Kallen, M. A., Beaumont, J. L., Eek, D., … Cella, D. (2020). Does recall period matter? Comparing PROMIS® physical function with no recall, 24-hr recall, and 7-day recall. *Quality of Life Research*, *29*, 745–753. https://doi.org/10.1007/s11136-019-02344-0

Condon, D. M., Coughlin, J., & Weston, S. J. (2022). Personality trait descriptors: 2,818 trait descriptive adjectives characterized by familiarity, frequency of use, and prior use in psycholexical research. *Journal of Open Psychology Data*, *10*(1), 1–9. https://doi.org/10.5334/jopd.57

Condon, D. M., Wood, D., Mõttus, R., Booth, T., Costantini, G., Greiff, S., … Zimmermann, J. (2021). Bottom up construction of a personality taxonomy. *European Journal of Psychological Assessment*, *36*(6), 923–934. https://doi.org/10.1027/1015-5759/a000626

Conley, M. N., & Saucier, G. (2019). An initial broad-level mapping of personality-situation contingencies in self-report data. *Personality and Individual Differences*, *136*, 166–172. https://doi.org/10.1016/j.paid.2017.07.013

Craig, R. J. (2005). Assessing personality and mood with adjective check list methodology: A review. *International Journal of Testing*, *5*(3), 177–196. https://doi.org/10.1207/s15327574ijt0503_1

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Crowe, M., Carter, N. T., Campbell, W. K., & Miller, J. D. (2016). Validation of the Narcissistic Grandiosity Scale and creation of reduced item variants. *Psychological Assessment*, *28*(12), 1550–1560. https://doi.org/10.1037/pas0000281

Crowe, M. L., Edershile, E. A., Wright, A. G. C., Campbell, W. K., Lynam, D. R., & Miller, J. D. (2018). Development and validation of the Narcissistic Vulnerability Scale: An adjective rating scale. *Psychological Assessment*, *30*(7), 978–983. https://doi.org/10.1037/pas0000578

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4–19. https://doi.org/10.1016/j.jesp.2015.07.006

Cutler, A., & Condon, D. M. (2022). Deep Lexical Hypothesis: Identifying personality structure in natural language. *PsyArXiv*. https://doi.org/10.31234/osf.io/gdm5v

De Raad, B., Barelds, D. P., Levert, E., Ostendorf, F., Mlačić, B., Blas, L. D., … Katigbak, M. S. (2010). Only three factors of personality description are fully replicable across languages: A comparison of 14 trait taxonomies. *Journal of Personality and Social Psychology*, *98*(1), 160. https://doi.org/10.1037/a0017184

De Raad, B., Barelds, D. P., Timmerman, M. E., De Roover, K., Mlačić, B., & Church, A. T. (2014). Towards a pan–cultural personality structure: Input from 11 psycholexical studies. *European Journal of Personality*, *28*(5), 497–510. https://doi.org/10.1002/per.1953

Denissen, J. J., Geenen, R., Van Aken, M. A., Gosling, S. D., & Potter, J. (2008). Development and validation of a Dutch translation of the Big Five Inventory (BFI). *Journal of Personality Assessment*, *90*(2), 152–157. https://doi.org/10.1080/00223890701845229

DeYoung, C. G. (2006). *Higher-order factors of the Big Five in a multi-informant sample*. *91*(6), 1138–1151. https://doi.org/10.1037/0022-3514.91.6.1138

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the Big Five. *Journal of Personality and Social Psychology*, *93*(5), 880–896. https://doi.org/10.1037/0022-3514.93.5.880

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The mini-IPIP scales: tiny-yet-effective measures of the Big Five factors of personality. *Psychological Assessment*, *18*(2), 192. https://doi.org/10.1037/1040-3590.18.2.192

Edershile, E. A., Woods, W. C., Sharpe, B. M., Crowe, M. L., Miller, J. D., & Wright, A. G. C. (2019). A day in the life of narcissus: Measuring narcissistic grandiosity and vulnerability in daily life. *Psychological Assessment*, *31*(7), 913–924. https://doi.org/10.1037/pas0000717

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Stanford University Press. https://doi.org/10.1515/9781503620766

Fossati, A., Borroni, S., Marchione, D., & Maffei, C. (2011). The Big Five Inventory (BFI): Reliability and validity of its Italian translation in three independent nonclinical samples. *European Journal of Psychological Assessment*, *27*(1), 50–58. https://doi.org/10.1027/1015-5759/a000043

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). Sage.

Gatz, M., Reynolds, C. A., Finkel, D., Hahn, C. J., Zhou, Y., & Zavala, C. (2015). Data harmonization in aging research: Not so fast. *Experimental Aging Research*, *41*(5), 475–495. https://doi.org/10.1080/0361073X.2015.1085748

Gibby, R. E., & Zickar, M. J. (2008). A history of the early days of personality testing in American industry: An obsession with adjustment. *History of Psychology*, *11*(3), 164. https://doi.org/10.1037/a0013041

Goldberg, L. R. (1971). A historical survey of personality scales and inventories. In P. McReynolds (Ed.), *Advances in Psychological Assessment* (Vol. 2). Science and Behavior Books.

Goldberg, L. R. (1982). From Ace to Zombie: Some explorations in the language of personality. In C. D. Spielberger & J. N. Butcher (Eds.), *Advances in Personality Assessment* (Vol. 1, pp. 203–234). Erlbaum.

Goldberg, L. R. (1990). An alternative "description of personality": The Big Five factor structure. *Journal of Personality and Social Psychology*, *59*, 1216–1229. https://doi.org/10.1037/0022-3514.59.6.1216

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, *4*, 26–42. https://doi.org/10.1037/1040-3590.4.1.26

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several Five-Factor Models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality and Individual Differences* (pp. 7–28). Tilburg University Press.

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

Goldberg, L. R., & Kilkowski, J. M. (1985). The prediction of semantic consistency in self-descriptions: Characteristics of persons and of terms that affect the consistency of responses to synonym and antonym pairs. *Journal of Personality and Social Psychology*, *48*(1), 82–98. https://doi.org/10.1037/0022-3514.48.1.82

Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*(6), 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1

Gough, H. G. (1960). The adjective check list as a personality assessment research technique. *Psychological Reports*, *6*(1), 107–122. https://doi.org/10.2466/pr0.1960.6.1.107

Graham, E. K., Rutsohn, J. P., Turiano, N. A., Bendayan, R., Batterham, P. J., Gerstorf, D., … Mroczek, D. K. (2017). Personality predicts mortality risk: An integrative data analysis of 15 international longitudinal studies. *Journal of Research in Personality*, *70*, 174–186. https://doi.org/10.1016/j.jrp.2017.07.005

Hamby, T., & Ickes, W. (2015). Do the readability and average item length of personality scales affect their reliability? Some meta-analytic answers. *Journal of Individual Differences*, *36*(1), 54–63. https://doi.org/10.1027/1614-0001/a000154

Hastie, R., & Kumar, P. A. (1979). Person memory: Personality traits as organizing principles in memory for behaviors. *Journal of Personality and Social Psychology*, *37*(1), 25–38. https://doi.org/10.1037//0022-3514.37.1.25

Hendriks, A. A. J. (1997). *The construction of the five-factor personality inventory (FFPI)* [Doctoral dissertation]. University of Groningen.

Hendriks, A. A. J., Hofstee, W. K., & De Raad, B. (1999). The five-factor personality inventory (FFPI). *Personality and Individual Differences*, *27*(2), 307–325. https://doi.org/10.1016/S0191-8869(98)00245-1

Henry, S., Wood, D., Condon, D. M., Lowman, G. H., & Mõttus, R. (2024). Using multi-rater and test-retest data to detect overlap within and between psychological scales. *Journal of Research in Personality*, *113*, 104530. https://doi.org/10.1016/j.jrp.2024.104530

Hill, P. L., Turiano, N. A., Hurd, M. D., Mroczek, D. K., & Roberts, B. W. (2011). Conscientiousness and longevity: an examination of possible mediators. *Health Psychology*, *30*(5), 536. https://doi.org/10.1037/a0023859

Hogan, R. T., & Hogan, J. (1998). Theoretical Frameworks for Assessment. In R. Jeanneret & R. Silzer (Eds.), *Individual Psychological Assessment: Predicting behavior in organizational settings* (pp. 27–53). Jossey-Bass.

Horsch, A. C., & Davis, R. A. (1935). Topical summaries of current literature: Mental hygiene and personality tests. *American Journal of Sociology*, *40*(5), 646–658. https://doi.org/10.1086/216902

Iveniuk, J., Laumann, E. O., Waite, L. J., McClintock, M. K., & Tiedt, A. (2014). Personality measures in the national social life, health, and aging project. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *69*(Suppl_2), S117–S124. https://doi.org/10.1093/geronb/gbu073

John, O. P., & Srivastava, S. (1999). The Big-Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (Vol. 2, pp. 102–138). Guilford Press.

Juster, F. T., & Suzman, R. (1995). An overview of the Health and Retirement Study. *Journal of Human Resources*, S7–S56. https://doi.org/10.2307/146277

Kasper, J. D., & Freedman, V. A. (2021). *National Health and Aging Trends Study user guide: Rounds 1-10 final release*. Johns Hopkins University School of Public Health.

Kay, C. S. (2024). Validating the IDRIS and IDRIA: Two infrequency/frequency scales for detecting careless and insufficient effort survey responders. *Behavior Research Methods*, 1–24. https://doi.org/10.3758/s13428-024-02452-x

Kay, C. S., & Saucier, G. (2023). Measuring personality traits in context: Four approaches to situations in self-report measures of personality. In P. K. Jonason (Ed.), *Shining Light on the Dark Side of Personality: Measurement Properties and Theoretical Advances* (pp. 261–273). Hogrefe.

Kern, M. L., Hampson, S. E., Goldberg, L. R., & Friedman, H. S. (2014). Integrating prospective longitudinal data: Modeling personality and health in the Terman Life Cycle and Hawaii Longitudinal Studies. *Developmental Psychology*, *50*(5), 1390. https://doi.org/10.1037/a0030874

Keuleers, E. (2013). *vwr: Useful functions for visual word recognition research*. R package version 0.3.0. https://CRAN.R-project.org/package=vwr

Kindley, E. (2016). *Questionnaire*. Bloomsbury Academic. https://doi.org/10.5040/9781501314803

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*(3), 151–160. https://doi.org/10.1007/BF02288391

Lachman, M. E., & Weaver, S. L. (1997). *The Midlife Development Inventory (MIDI) Personality Scales: Scale construction and scoring*. https://www.brandeis.edu/psychology/lachman/pdfs/midi-personality-scales.pdf

Lang, F. R., Lüdtke, O., & Asendorpf, J. B. (2001). Test quality and psychometric equivalence of the German Big Five Inventory (BFI) for young, middle-aged, and old adults. *Diagnostica*, *47*, 111–121. https://doi.org/10.1026/0012-1924.47.3.111

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO personality inventory. *Multivariate Behavioral Research*, *39*(2), 329–358. https://doi.org/10.1207/s15327906mbr3902_8

Leyens, J. P., Yzerbyt, V. Y., & Rogier, A. (1997). Personality traits that distinguish you and me are better memorized. *European Journal of Social Psychology*, *27*(5), 511–522. https://doi.org/10.1002/(sici)1099-0992(199709/10)27:5%3C511::aid-ejsp827%3E3.0.co;2-7

Lowman, G. H., Wood, D., Armstrong, B. F., III, Harms, P. D., & Watson, D. (2018). Estimating the reliability of emotion measures over very short intervals: The utility of within-session retest correlations. *Emotion*, *18*(6), 896. https://doi.org/10.1037/emo0000370

Lüdecke, D. (2025). *sjPlot: Data visualization for statistics in social science* (2.9.0).

Marcus, B. (2009). "Faking" from the applicant's perspective: A theory of self-presentation in personnel selection settings. *International Journal of Selection and Assessment*, *17*(4), 417–430. https://doi.org/10.1111/j.1468-2389.2009.00483.x

McCrae, R. R., Costa, P. T., Jr, & Martin, T. A. (2005). The NEO–PI–3: A more readable revised NEO personality inventory. *Journal of Personality Assessment*, *84*(3), 261–270. https://doi.org/10.1207/s15327752jpa8403_05

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/a0028085

Mõttus, R., Bates, T., Condon, D. M., Mroczek, D., & Revelle, W. (2017). Leveraging a more nuanced view of personality: Narrow characteristics predict and explain variance in life outcomes. *PsyArXiv*. https://doi.org/10.31234/osf.io/4q9gv

Mõttus, R., Wood, D., Condon, D. M., Back, M., Baumert, A., Costantini, G., Epskamp, S., Greiff, S., Johnson, W., Lukaszewski, A., Murray, A., Revelle, W., Wright, A. G. C., Yarkoni, T., Ziegler, M., & Zimmermann, J. (2020). Descriptive, predictive and explanatory personality research: Different goals, different approaches, but a shared need to move beyond the Big Few traits. *European Journal of Personality*, *34*(6), 1175–1201. https://doi.org/10.1002/per.2311

National Institute on Aging. (2020). *Expert Meeting on the Harmonization and Coordinated Analysis of Behavioral and Psychological Phenotypes*. https://www.nia.nih.gov/research/dbsr/expert-meeting-harmonization-and-coordinated-analysis-behavioral-and-psychological

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *Journal of Abnormal and Social Psychology*, *66*(6), 574–583. https://doi.org/10.1037/h0040291

Program on Global Aging, Health, and Policy. (2021, July 27). *Gateway to Global Aging Data*. https://g2aging.org/

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

Rammstedt, B., Roemer, L., Danner, D., & Lechner, C. M. (2022). Don't keep it too simple: Simplified items do not improve measurement quality. *European Journal of Psychological Assessment*, 1–11.

Raskin, R. N., & Hall, C. S. (1979). A narcissistic personality inventory. *Psychological Reports*, *45*, 590. https://doi.org/10.2466/pr0.1979.45.2.590

Revelle, W., & Condon, D. M. (2019). Reliability from α to ω: A tutorial. *Psychological Assessment*, *31*(12), 1395. https://doi.org/10.1037/pas0000754

Rosenthal, S. A., Hooley, J. M., Montoya, R. M., van der Linden, S. L., & Steshenko, Y. (2020). The Narcissistic Grandiosity Scale: A measure to distinguish narcissistic grandiosity from high self-esteem. *Assessment*, *27*(3), 487–507. https://doi.org/10.1177/1073191119858410

Rosenthal, S. A., Hooley, J. M., & Steshenko, Y. (2007). *Distinguishing grandiosity from self-esteem: Development of the Narcissistic Grandiosity Scale* [Unpublished Manuscript].

Runge, S. K., Craig, B. M., & Jim, H. S. (2015). Word recall: Cognitive performance within internet surveys. *JMIR Mental Health*, *2*(2), e3969. https://doi.org/10.2196/mental.3969

Ryff, C. D., Almeida, D., Ayanian, J., Carr, D. S., Cleary, P. D., Coe, C., Davidson, R., Krueger, R. F., Lachman, M. E., Marks, N. F., & Mroczek, D. K. (2019). *Midlife in the United States (MIDUS 2), 2004-2006 (ICPSR 4652)*.

Ryff, C. D., Kitayama, S., Karasawa, M., Markus, H., Kawakami, N., & Coe, C. (2018). *Survey of Midlife in Japan (MIDJA 2), May-October 2012 (ICPSR 36427)*.

Saucier, G. (1994). Mini-Markers: A brief version of Goldberg's unipolar Big-Five Markers. *Journal of Personality Assessment*, *63*(3), 506–516. https://doi.org/10.1207/s15327752jpa6303_8

Saucier, G. (2020). Language, Subjectivity, Culture, Comprehensiveness, and Structure: Considerations for a Classification of Situations. In J. F. Rauthmann, R. A. Sherman, & D. C. Funder (Eds.), *The Oxford Handbook of Psychological Situations*. https://doi.org/10.1093/oxfordhb/9780190263348.013.22

Saucier, G., & Goldberg, L. R. (2001). Lexical studies of indigenous personality factors: Premises, products, and prospects. *Journal of Personality*, *69*(6), 847–879. https://doi.org/10.1111/1467-6494.696167

Saucier, G., Thalmayer, A. G., Payne, D. L., Carlson, R., Sanogo, L., Ole-Kotikash, L., … Zhou, X. (2014). A basic bivariate structure of personality attributes evident across nine languages. *Journal of Personality*, *82*(1), 1–14. https://doi.org/10.1111/jopy.12028

Schalet, B. D., Lim, S., Cella, D., & Choi, S. W. (2021). Linking scores with patient-reported health outcome instruments: A validation study and comparison of three linking methods. *Psychometrika*, 1–30. https://doi.org/10.1007/s11336-021-09776-z

Smith, E. R., & Henry, S. (1996). An in-group becomes part of the self: Response time evidence. *Personality and Social Psychology Bulletin*, *22*(6), 635–642. https://doi.org/10.1177/0146167296226008

Smith, J., Ryan, L., Fisher, G. G., Sonnega, A., & Weir, D. (2017). *Psychosocial and Lifestyle Questionnaire (2006-2016): Documentation Report, Core Section LB*. Health and Retirement Study, Survey Research Center, Institute for Social Research. University of Michigan, Ann Arbor. https://hrs.isr.umich.edu/sites/default/files/biblio/HRS%202006-2016%20SAQ%20Documentation_07.06.17_0.pdf

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, *113*, 117–143. https://doi.org/10.1037/pspp0000096

Steptoe, A., Breeze, E., Banks, J., & Nazroo, J. (2013). Cohort profile: the English longitudinal study of ageing. *International Journal of Epidemiology*, *42*(6), 1640–1648. https://doi.org/10.1093/ije/dys168

Thalmayer, A. G., Job, S., Shino, E. N., Robinson, S. L., & Saucier, G. (2020). ǂŪsigu: A mixed-method lexical study of character description in Khoekhoegowab. *Journal of Personality and Social Psychology*. Advance online publication. https://doi.org/10.1037/pspp0000372

Thalmayer, A. G., Mather, K. A., Saucier, G., Naudé, L., Florence, M., Adonis, T.-A., Shino, E. N., Asatsa, S., Witzlack-Makarevich, A., Bächlin, L. Z. M., & Condon, D. M. (2024). The cross-cultural big two: A culturally decentered theoretical and measurement model for personality traits. *Journal of Personality and Social Psychology*. Advance online publication. https://doi.org/10.1037/pspp0000528

Thalmayer, A. G., Saucier, G., & Eigenhuis, A. (2011). Comparative validity of brief to medium-length Big Five and Big Six personality questionnaires. *Psychological Assessment*, *23*(4), 995–1009. https://doi.org/10.1037/a0024165

Thalmayer, A. G., Saucier, G., Ole-Kotikash, L., & Payne, D. (2020). Personality structure in east and west Africa: Lexical studies of personality in Maa and Supyire-Senufo. *Journal of Personality and Social Psychology*, *119*(5), 1132. https://doi.org/10.1037/pspp0000264

Thurstone, L. L. (1934). The Vectors of the Mind. *Psychological Review*, *41*, 1–32. https://doi.org/10.1037/h0075959

Tupes, E. C., & Christal, R. E. (1961). *Recurrent personality factors based on trait ratings* [Technical report]. Personnel Laboratory, United States Air Force. https://doi.org/10.21236/AD0267778

Van den Berg, S. M., De Moor, M. H., McGue, M., Pettersson, E., Terracciano, A., Verweij, K. J., … Boomsma, D. I. (2014). Harmonization of Neuroticism and Extraversion phenotypes across inventories and cohorts in the Genetics of Personality Consortium: an application of Item Response Theory. *Behavior Genetics*, *44*(4), 295–313. https://doi.org/10.1007/s10519-014-9654-x

Walton, K. E., Radunzel, J., Moore, R., Burrus, J., Anguiano-Carrasco, C., & Murano, D. (2021). Adjectives vs. statements in forced choice and Likert item types: Which is more resistant to impression management in personality assessment? *Journal of Personality Assessment*, 1–35. https://doi.org/10.1080/00223891.2021.1878523

Watson, D. (2004). Stability versus change, dependability versus error: Issues in the assessment of personality over time. *Journal of Research in Personality*, *38*(4), 319–350. https://doi.org/10.1016/j.jrp.2004.03.001

Weston, S. J., Graham, E. K., Turiano, N. A., Aschwanden, D., Booth, T., Harrison, F., … Mroczek, D. K. (2020). Is healthy neuroticism associated with chronic conditions? A coordinated integrative data analysis. *Collabra: Psychology*, *6*(1). https://doi.org/10.1525/collabra.268

Witt, E. A., Brent Donnellan, M., & Blonigen, D. M. (2009). Using existing self-report inventories to measure the psychopathic personality traits of Fearless Dominance and Impulsive Antisociality. *Journal of Research in Personality*, *43*(6), 1006–1016. https://doi.org/10.1016/j.jrp.2009.06.010

Wood, D., Harms, P. D., Lowman, G. H., & DeSimone, J. A. (2017). Response speed and response consistency as mutually validating indicators of data quality in online samples. *Social Psychological and Personality Science*, *8*(4), 454–464. https://doi.org/10.1177/1948550617703168

Wood, D., Qiu, L., Lu, J., Lin, H., & Tov, W. (2018). Adjusting Bilingual ratings by retest reliability improves estimation of translation quality. *Journal of Cross-Cultural Psychology*, *49*(9), 1325–1339. https://doi.org/10.1177/0022022118789773

Wood, J. K., Gurven, M., & Goldberg, L. R. (2020). Ubiquitous personality-trait concepts in 13 diverse and isolated languages: A cluster–classification approach. *European Journal of Personality*, *34*(2), 164–179. https://doi.org/10.1002/per.2246

Ziegler, M., & Bensch, D. (2013). Lost in translation: Thoughts regarding the translation of existing psychological measures into other languages. *European Journal of Psychological Assessment*, *29*(2), 81–83. https://doi.org/10.1027/1015-5759/a000167

Zola, A., Condon, D. M., & Revelle, W. (2021). The convergence of self and informant reports in a large online sample. *Collabra: Psychology*.

Supplementary Materials

Supplemental Material

Download: https://collabra.scholasticahq.com/article/150375-to-be-or-i-am-someone-who-tends-to-be-does-the-wording-of-personality-items-matter/attachment/314531.pdf?auth_token=ZkOZshZHGkRARmUNDJ1S

Peer Review Communication

Download: https://collabra.scholasticahq.com/article/150375-to-be-or-i-am-someone-who-tends-to-be-does-the-wording-of-personality-items-matter/attachment/314532.docx?auth_token=ZkOZshZHGkRARmUNDJ1S