Improving the Measurement of the Big Five via Alternative Formats for the BFI-2

Xijuan Zhang[1], Muhua Huang[2], Jessie Sun[3], and Victoria Savalei[4]

[1]York University

[2]University of Chicago

[3]Washington University in St. Louis

[4]University of British Columbia

Author Note

Correspondence concerning this article should be addressed to Xijuan Zhang, Department of Psychology, York University, 4700 Keele St, North York, ON, Canada, M3J 1P3. Contact: cathyxijuan@gmail.com

Abstract

The Big Five Inventory-2 (BFI-2; Soto & John, 2017) was developed to improve upon the limitations of the original BFI by balancing the number of positively- and negatively-worded items and establishing a hierarchical structure for the Big Five traits. However, as the BFI-2 employs a Likert format with agree-disagree options, it suffers from common problems of the Likert format, including acquiescence bias and method effects due to the negatively-worded items. In this research, we converted the BFI-2 into three alternative formats: Expanded, Item-Specific-Full, and Item-Specific-Light. These formats have tailored response options for each item and avoid the use of negatively-worded items, thereby addressing the issues associated with the Likert format. Across two studies (Ns = 1,335/1,451), we randomly assigned Canadian undergraduate students to complete the BFI-2 in the original Likert format or one of three alternative formats. Results showed that the Likert and alternative formats exhibit similar predictive validity. However, the alternative formats—particularly the Expanded format—showed better psychometric properties, including enhanced factor structure, increased reliability, and possibly reduced careless responding. We recommend that researchers consider adopting the BFI-2 in these alternative formats and adapting other Likert scales to these alternative formats.

Keywords: Big Five Inventory, BFI-2, Likert format, Expanded format, Item-Specific format, alternative formats, factor structure, validity, reliability

Improving the Measurement of the Big Five via Alternative Formats for the BFI-2

The Big Five Personality model has been the predominant model of personality over the past quarter century. According to this model, individual differences in people's patterns of behaviour, feeling, and cognition can be summarized by five major personality domains: Extraversion, Agreeableness, Conscientiousness, Negative Emotionality (a.k.a., Neuroticism), and Open-Mindedness (a.k.a., Openness to Experiences, Intellect, or Imagination). The Big Five Inventory (BFI) by John et al. (1991) is one of the most widely-used measures of the Big Five. Recently, Soto and John (2017a) developed a revised version of the BFI–the BFI-2. The BFI-2 addresses two limitations of the original BFI by 1) balancing the number of positively- and negatively-worded items for each personality domain and 2) specifying a hierarchical structure of the Big Five, such that each personality domain contains three facets.

Problems with the Likert Format

Despite BFI-2's improvement, it still suffers from several shortcomings due to its use of a Likert response format, in which participants rate their level of agreement or disagreement with various statements. There are generally two types of issues associated with the Likert format.

Ambiguity of Response Options

First, the agree-disagree options in typical Likert scales can be ambiguous and open to many interpretations (e.g., Dykema et al., 2022; Saris et al., 2010; Zhang et al., 2019). To respond to a Likert item with agree-disagree options, respondents need to identify the concept that represents the underlying response dimension along which they can disagree or agree. They also need to judge how various degrees of agreement map onto the underlying dimension that they have in mind. However, participants can differ both in which concept they select and in their interpretation of how various degrees of agreement map onto the underlying dimension. For example, for the BFI-2 item "I tend to find fault with others", the underlying dimension can either be "*how often* the respondent finds faults with others" or "*how many* faults the respondent finds with

others". Although agreement with the statement indicates low level of agreeableness in either case, the possible differential selection of the dimension along which to agree or disagree may increase measurement error, reducing the scale's reliability and validity. For another BFI-2 item, "I am someone who can be somewhat careless", the underlying dimension is obviously the extent to which the respondent can be careless. However, one respondent could answer "disagree" because they believe they are "very careless" (not just "somewhat careless"), whereas another respondent could answer "disagree" because they think they are "not careless at all". In this case, the interpretation of the option "disagree" can vary across respondents, increasing measurement error.

Issues Related to Negatively-Worded Items

Second, negatively-worded items are commonly used in the Likert format. Unlike positively-worded items, for which agreement indicates greater endorsement of the construct being measured, negatively-worded items indicate lower endorsement of the construct. There are two types of negatively-worded items: 1) negation items, created by adding negative particles such as "not" or "no" or by adding affixal negation such as "un-" or "-less"; and 2) polar opposite items, created by using words with an opposite meaning. The main rationale for including an equal number of positively- and negatively-items (known as a balanced scale), as is done in the BFI-2, is to control for acquiescence bias – the tendency of participants to agree with statements regardless of their content. Acquiescence bias can arise from several causes, summarized by Pasek and Krosnick (2010), including: 1) a general inclination to please others (associated with high agreeableness), which is particularly pronounced in certain cultures (e.g., East Asian); 2) a tendency to rely on cognitive shortcuts and select responses that seem "good enough," known as satisficing; and 3) a tendency to comply with researchers who are perceived as authority figures.

However, a balanced number of positively- and negatively-worded items only effectively control for acquiescence bias with regard to the average or sum scale scores but not with respect to the underlying factor structure. The reason for this is that even with a balanced scale, acquiescence bias remains a latent factor affecting the

correlations or covariances between items, whether the items are positively- or negatively-worded. As a result, statistical analyses based on the covariances, such as factor analysis, will continue to be affected by acquiescence bias even in a balanced scale (Savalei & Falk, 2014).

Furthermore, in a balanced scale such as BFI-2, the constant switching between positively- and negatively-worded items can cause confusion and elicit erroneous responses from some participants (Swain et al., 2008; Zhang et al., 2019). Particularly, with negation items, participants may accidentally skip these negative participles when reading the items.

Together, the issue related to (1) the acquiescence bias and (2) the issue of careless responding and confusion due to the negatively-worded items, may explain why one or two additional unintended factors emerge when data from a Likert scale are subjected to factor analyses (e.g., Swain et al., 2008; Woods, 2006). Such unintended extra factors are commonly referred to as "method factors," since they primarily reflect measurement-related artifacts rather than the construct researchers intend to assess. Alternatively, the unintended factor may also be interpreted as an "acquiescence factor," which is similarly uncorrelated with the intended construct. Indeed, when developing the BFI-2, Soto and John (2017a) found six factors when the raw data on the BFI-2 were subjected to exploratory factor analyses. The sixth factor can be thought of either as an acquiescence factor or a method factor.

Alternatives to the Likert Format

Given the problems of the Likert format, many researchers have advocated using alternative formats (e.g., Saris et al., 2010; Schuman & Presser, 1981; Wong et al., 2003; Zhang & Savalei, 2016, 2024; Zhang et al., 2019). Two noticeable alternative formats are the Expanded and Item-Specific formats. The Expanded format, introduced by Zhang and Savalei (2016), involves expanding and replacing the agree-disagree options in the Likert format with options written in full sentences. For example, when converting the item "I tend to find fault with others" to the Expanded format, the participant will be asked to choose the response option that best describes themselves

out of response options such as: 1) “I almost always find fault with others”, 2) “I often find fault with others”, 3) “I sometimes find fault with others”, 4) “I rarely find fault with others ”, 5) “I almost never find fault with other” (see Table 1 for more examples). In other words, when creating an Expanded item, the researcher must select an underlying dimension that varies across the response options, such as the *frequency* of finding faults with others. This underlying dimension can be tailored to each item.

The Item-Specific format, proposed by Saris et al. (2010), also has response options that are tailored to each item. Unlike the Expanded format, however, the response options are short phrases, rather than complete sentences, to a question stem. For example, the item “I tend to find fault with others” can be rewritten as “How often do you find faults in others?” with the response options: 1) almost always, 2) often, 3) sometimes, 4) rarely, 5) almost never (see Table 1 for more examples). We refer to this as the Item-Specific-Full format. The Item-Specific format can also be simplified such that only the first and last options are written out, with intermediate options presented as checkboxes (see Table 1 for examples). We refer to this as the Item-Specific-Light format.

These three alternative formats (i.e., Expanded, Item-Specific-Full, and Item-Specific-Light) alleviate the problems associated with the Likert format in four different ways. First, by replacing the ambiguous agree-disagree options with more specific response options tailored to each item, the three alternative formats reduce ambiguity in the interpretation of the item and response options. Second, in cases where participants engage in acquiescence due to a tendency to be agreeable or compliant toward perceived authority, the alternative formats can theoretically minimize the acquiescence bias because the items do not involve participants agreeing with any statement. Third, when acquiescence stems from cognitive shortcuts such as satisficing, the alternative formats should help reduce the acquiescence bias because varying the response options across items makes it more difficult for participants to repeatedly rely on the same cognitive shortcut.

Fourth, these alternative formats integrate both positively- and negatively-worded anchors directly into each item’s response options. For example, the

item "How quickly/slowly do you get things done?" is anchored by "Very slowly" and "Very quickly." As a result, there is no switching between positively- and negatively-worded items across the scale, reducing confusion and careless responding. Indeed, previous studies on these alternative formats have demonstrated that relative to the Likert format with the presence of negatively-worded items, the alternative format yielded a factor structure that was more consistent with the theoretically intended structure (Kam, 2020; Zhang & Savalei, 2016; Zhang et al., 2019, 2023). However, it is important to note that if a scale in the Likert format comprises only positively-worded items, then the corresponding scale in the Expanded format may show little to no improvement in fit to the theoretically intended structure (e.g., Brauer & Proyer, 2021; Zhang et al., 2019). The reason is that, in a Likert scale with all positively-worded items, the domain factor and the acquiescence factor are conflated with each other; that is, without negatively-worded items, we can no longer tell if a participant strongly agrees with an item due to their underlying high score on the domain or due to acquiescence bias.

However, it is important to note that, although the alternative formats eliminate the method effect associated with the negative-worded items, the alternative formats may still be affected by the order effect, which is the tendency of participants to choose either the first or last response option for each item. The Likert format may also be affected by order effects because the agree-disagree options in the Likert format can also be ordered differently. Fortunately, previous research showed that the order effect is not very pronounced in either the Likert or the alternative format. For the Likert format, studies have shown that altering the order of response options across all items does not affect the scale mean difference, the distribution of response option endorsement, or the factor structure of the scale (e.g., Robie et al., 2022; Weng & Cheng, 2000). For the Expanded format, Zhang et al. (2019) examined the order effect by manipulating the response option order in three conditions: 1) all items' response options ordered from those indicating low endorsement of the construct to those with high endorsement; 2) all items' response options ordered from high to low endorsement; 3) half of the items ordered from low to high and the other half ordered from high to low. Zhang et al.

(2019) demonstrated no significant difference among these three conditions regarding the distribution of response option endorsement and the factor structure of the original BFI by John et al. (1991).

Regarding scale reliability and validity of the alternative formats, previous studies had somewhat mixed results (e.g., Kam et al., 2024; Kuru & Pasek, 2016; Zhang & Savalei, 2016; Zhang et al., 2019). On one hand, Kuru and Pasek (2016) compared social media usage scales such as Facebook Intensity scale and Facebook Intrusion Scale in the Item-Specific-Full and the Likert formats and found that the scales in the Item-Specific-Full had a higher average correlation with criterion measures including the number of friends and time spent on Facebook. Similarly, Dykema et al. (2022) reviewed 20 experimental studies comparing these two formats and concluded that the Item-Specific-Full format generally yields better reliability and validity. On the other hand, studies by Zhang and Savalei (2016) and Zhang et al. (2019) reported that the alternative formats produced scale reliability and convergent validity comparable to that of the traditional Likert format.

The Present Research

Across two studies, we aimed to convert the BFI-2 to the alternative formats and compare the properties of the scale across multiple formats. Study 1 focuses on evaluating the factor structure, reliability, and predictive validity of BFI-2 in the alternative formats compared to the Likert format. Based on previous research (Kam, 2020; Kam et al., 2024; Kuru & Pasek, 2016; Zhang & Savalei, 2016; Zhang et al., 2019), we hypothesize that compared to the original BFI-2 in the Likert format, the BFI-2 in the alternative formats would yield a factor structure more consistent with the intended structure and at least comparable reliability and validity. In Study 2, we aim to investigate the effect of scale format on participants' careless responding while replicating results from Study 1. We hypothesize that, due to the reduced confusion among participants when responding to the alternative formats, careless responding will be lower in these formats relative to the Likert format.

Ethics and Open Science Practices Statement

Data collection procedures for Study 1 were approved by the Ethics Review Board at the University of British Columbia (UBC; ID: H19-03371, Study Title: "Alternative scale formats of Big Five Inventory with Peer Ratings" and ID: H19-02297, Study Title: "Alternative scale formats of Big Five Inventory"). Data collection procedures for Study 2 were approved by the Ethics Review Board at York University (ID: e2023-288, Study Title: "Enhancing Self-Report Measures in Psychology"). We preregistered stopping rules, exclusions, and analysis plans at https://osf.io/x4dqa/?view_only=c9c947b874824a808177e1b101f7c17c (Study 1) and https://osf.io/fduzw/?view_only=9cb3ff9c34224af09dcef1aa2a133df6 (Study 2). The Supplementary Materials, the data and analyses code are available at https://osf.io/pabfj/?view_only=565db6ba819a48dc8732ea0e2c93f0e3 (Study 1) and https://osf.io/yuv5b/?view_only=332b3798ee8d49579ead72e6b2b2cf9b (Study 2).

Study 1

Method

Participants and Procedure

We recruited 1,335 participants from UBC through the Human Subject Pool (HSP) who were compensated with course credit. Participants completed the study online through Qualtrics. There were two versions of Study 1 posted on HSP at the same time. Version 1 ($n = 720$) included informant reports, and Version 2 ($n = 615$) did not. Participants were only able to sign up to participate in one version of the study. In both versions, participants were first randomly assigned to complete the BFI-2 in one of the four scale formats: the original Likert version or one of three alternative formats (Expanded, Item-Specific-Full, or Item-Specific-Light; described below). Next, all participants completed a series of criterion measures related to the life outcomes (see Table 2), which were adapted from Soto (2019)'s study on how the BFI-2 predicts life outcomes. All participants then reported their age, gender, and ethnicity. Finally, in Version 1 only, each participant was further asked to send emails to two of their friends ("informants"), inviting them to participate in a peer-rating survey in

exchange for a five-dollar Amazon gift card. [1] The peer-rating survey asked the informants to rate the original participant's personality. For example, for the item "I am dominant, act like a leader", the corresponding peer-rating item is "Your friend is dominant, acts like a leader". Peers completed the same format of the BFI-2 that the original participant was randomly assigned to.

A majority of the participants (76.25%) identified themselves as female; 22.54% identified themselves as male and 0.37% as gender variant. The mean age of the participants was 21.05 with a standard deviation of 3.40. In terms of ethnicity, 23.04% were European, 49.25% were East Asian, and 27.71% identified with other ethnic groups.

Of the 720 participants who completed Version 1 and emailed two friends for the follow-up study, 345 had at least one informant report; among these 345 participants, 131 received two informant reports. The demographics of the informants were similar to those of the participants but with greater age variability, with a mean age of 22.38 and a standard deviation of 7.8. In terms of gender, 32.76% of the informants identified themselves as male and 66.95% as female. Regarding ethnicity, 20.40% were European, 48.85% were East Asian, and 30.75% identified with other ethnic groups.

Across the two versions of the study, an approximately equal number of participants completed the BFI-2 in each of the four formats (Likert = 341, Expanded = 329, Item-Specific-Full = 331, and Item-Specific-Light = 334). Given that only participants in Version 1 recruited informants, and that only approximately half of these informants responded, the numbers of participants who had at least one peer complete the ratings were much lower (Likert = 75, Expanded = 94, Item-Specific-Full = 81, and Item-Specific-Light = 95).

Measures

Original and Alternative Formats of the BFI-2. The original BFI-2 in the Likert format comprises 60 items, with 12 items for each personality domain subscale

[1] We offered two versions of the study because, based on past experiences, we knew that many students would be reluctant to sign up for a study that required them to email their friends about participating in a follow-up study. Therefore, we allowed participants to choose either a version that required them to email their friends or a version that did not.

(Soto & John, 2017a). Each domain has three facets, each measured by two positively-worded and two negatively-worded items. In other words, each domain subscale is a balanced scale with six positively-worded and six negatively-worded items. Each item is measured on a five-point response scale, ranging from "disagree strongly" to "agree strongly".

We converted the BFI-2 to the Expanded, Item-Specific-Full, and Item-Specific-Light formats. Two sample items are shown in Table 1. The complete scale in each format is available in the Supplementary Materials Section 1.

When converting the BFI-2 to alternative formats, we implemented the following principles. First, we aimed to preserve the wording of the original items and keep consistent wording across formats, as much as possible. For the Expanded format, each item's response options are written in a complete sentence, using wording similar to the original item (see Table 1). For the Item-Specific-Full and -Light formats, each item presents a question stem written with wording similar to the original items; the response options are answers tailored to each question stem, using wording similar to the options in the Expanded format (see Table 1). For each item in each alternative format, the response options were arranged from the option indicating the lowest endorsement of the personality domain to the option indicating the highest endorsement. We chose not to vary the order of response options across items because previous studies showed that the order of the response options does not significantly affect the psychometric properties of the scale (e.g., Robie et al., 2022; Weng & Cheng, 2000; Zhang et al., 2019).

Second, as mentioned previously, when converting an item in the Likert format to one of the alternative formats, we need to choose a dimension along which the response options vary. For a given Likert item, there were often several dimensions we could choose from (e.g., frequency or degree of a behaviour). For each item, we chose a dimension that facilitated clear and concrete response options. For example, it made more sense to rephrase "I tend to be quiet" as "How often are you quiet?" (frequency), as opposed to "How quiet are you?" (intensity). This means that we tailored the type of dimension to each item, which is a key advantage of the alternative formats.

Third, some items on the original BFI-2 were constructed with an adjective followed by a phrase that elaborates on its content. Consider, for example, the extraversion item, "I am dominant, act like a leader" (i.e., Item 21 in Table 1). When converting such an item to the Expanded format, we retained this wording structure for each response option (see Table 1). However, when converting the item to the Item-Specific-Full and -Light formats, we had to choose either the adjective or the phrase for the item's question stem. For the above example, the question stem in the Item-Specific-Full and -Light formats says "How often do you act as a leader or a follower?"; that is, we chose the phrase for the question stem. Consequently, compared to the Item-Specific-Full and -Light formats, the Expanded format contains more descriptive response options, each more similar to the wording of the original items.

Finally, the items in the original Likert BFI-2 in Soto and John (2017a) were presented in a grid format, where the response options are presented once at the top of all items. However, response options varied for each item in our alternative scale formats. To control for item presentation format (item-by-item vs. grid), we, therefore, chose to present each item separately across all formats, including the Likert format (see Table 1).

Criterion Measures. Criterion measures were selected from those used in Soto (2019)'s study, which examined how the BFI-2 predicts life outcomes measured by previously developed questionnaires such as the Behaviour Report Form by Paunonen (2003). Table 2 shows the details of the criterion measures used in our study, including their original sources and the anticipated correlations with the Big Five personality traits.

Data Analyses

In Study 1, our analyses focused on comparing the factor structure, reliability and validity of the BFI-2 across the different formats.

Factor Analyses. To compare the factor structure across the different formats, we conducted CFAs using the `lavaan` package (version .6-17) (Rosseel, 2012) in the R programming system. For each personality domain subscale in each format, we fit five CFA models to the BFI-2 data across the two versions of the study. These same five

models were also fit for each domain in Soto and John (2017a)'s original study on the BFI-2 in the Likert format. Our goal is to compare the results of the Likert format to those of the alternative formats. Figure 1 illustrates the diagrams for the five models.

Model 1 (a.k.a, single-domain model) is a one-factor model with all items loading on their respective personality domain. Model 2 is a correlated-two-factors model, where all items that were negatively-worded in the original Likert format load on one factor and all items that were positively-worded in the original Likert format load on another factor (see Figure 1). Model 2 (a.k.a, positive-and-negative-factors model) is a popular factor model for Likert scales because it accounts for the negatively- vs. positively-worded structural distinction in Likert scales (e.g., Carmines & Zeller, 1979; Hensley & Roberts, 1976; Horan et al., 2003; Zhang et al., 2023).

In the pre-registration of Study 1, we only included Models 1 and 2 because these are the most commonly employed factor models in past research examining factor structures across different scale formats (e.g., Kam, 2020; Zhang & Savalei, 2016; Zhang et al., 2023). However, during the data analysis phase, we decided to add three additional models (Models 3-5), which, as we will explain, are more suitable for the BFI-2 and provide deeper insights into its factor structure. To confirm the results for these models, we pre-registered them for Study 2.

Model 3 (a.k.a, single-domain-plus-acquiescence model) is another popular model for Likert scales because it explicitly models the acquiescence bias factor (e.g., Billiet & McClendon, 2000; Maydeu-Olivares & Coffman, 2006; Savalei & Falk, 2014). Model 3 features a personality domain factor (which all items load onto) and an orthogonal acquiescence bias factor (which has all positively-worded items loading 1 and all reverse-scored negatively-worded items loading -1 onto it) (see Figure 1). Models 4 and 5 incorporate the three facets within the personality domain (see Figure 1). Model 4 (a.k.a, three-facets model) is a correlated-three-factors model modelling each facet in the personality domain as a separate factor. This model is the most consistent with the theoretical factor structure that Soto and John (2017a) had in mind when developing BFI-2 since they intentionally wanted to include three facets for each factor. Model 5 (a.k.a, three-facets-plus-acquiescence model) combines features of Models 3 and

4, adding an orthogonal factor representing the acquiescence bias (similar to Model 3) atop three correlated facet factors (as in Model 4).

In summary, out of the five models in Figure 1, Models 4-5 account for the structure of the facets within each personality domain, whereas Models 1-3 do not. Models 1 and 4 do not account for the acquiescence bias nor the negatively- vs. positively-worded structure in the Likert scale, whereas Models 2, 3 and 5 do. It is important to note that for the alternative formats, the negative and positive factors in Model 2, as well as the acquiescence bias factor in Models 3 and 5 are not theoretically meaningful, but we fit the same five models across all four formats for comparison.

As mentioned previously, Soto and John (2017a)'s initial paper on BFI-2 also included these five models. They found that Models 1-3 yielded very poor fit due to their inconsistency with the facet structure inherent in the BFI-2. Conversely, Models 4 and 5 demonstrated a reasonable fit. Notably, Model 5, which incorporated an acquiescence bias factor, achieved a much better fit than Model 4, which did not include the acquiescence factor, a result that aligns with our expectation that the original Likert BFI-2 is influenced by acquiescence bias. Although Soto and John (2017a) most likely did not intend to include an acquiescence bias factor when designing BFI-2, they recommended researchers fit Model 5 for the original BFI-2 in the Likert format since it achieved the best fit out of the five models.

Based on Soto and John (2017a)'s results, we hypothesize that Models 4-5 (including the facet factors) will yield considerably better fit than Models 1-3 (without the facet factors). We also expect that Models 1 and 4 will have poorer fit for the Likert format relative to the alternative formats, as they do not account for acquiescence bias or the negatively- vs. positively-worded structure. In contrast, we expect that Models 2, 3, and 5 should show similar fit across formats, as they incorporate factors accounting for acquiescence bias or the negatively- vs. positively-worded structure.

To estimate each of the five models, we used the full information maximum likelihood (FIML) estimation as approximately 10% participants had some missing data. To evaluate the fit of the two CFA models, we used the chi-square test of fit and three approximate fit indices with common cutoff points: 1) the comparative fit index

(CFI), with a value above .90 indicating a reasonable fit and above .95 indicating a very good fit; 2) the root mean square error of approximation (RMSEA), with a value of less than .08 indicating reasonable fit and less than .05 indicating very good fit; and 3) the standardized root mean square residual (SRMR), with a value of less than .08 indicating reasonable fit and less than .05 indicating very good fit. [2]

In addition to evaluating the CFA models' fit, we also compared the fit between nested models. Model 1 is nested within Models 2, 3, and 4, whereas Models 3 and 4 are nested within Model 5. Therefore, we compared the fit between five pairs of models: a) Models 1 vs. 2, b) Models 1 vs. 3, c) Models 1 vs. 4, d) Models 3 vs. 5, and e) Models 4 vs. 5. Following Savalei et al. (2024), we used a fit index called RMSEA_D. RMSEA_D is an RMSEA value associated with the chi-square difference test, designed for nested models where the less restricted model only achieves approximate fit rather than exact fit. RMSEA_D measures the amount of increase in misfit due to the constraints introduced by the more restricted model, with a value greater than .08 indicating a substantial increase in misfit.

We expect that for the Likert format, when the acquiescence factor or the negatively- vs. positively-worded structure was taken into account, the model fit would be substantially different. Specifically, we expect large RMSEA_D values between Models 1 vs. 2, Models 1 vs. 3, and Models 4 vs. 5 for the Likert format. On the other hand, we hypothesize that the alternative formats, which are not affected by acquiescence bias nor the negatively- vs. positively-worded structure, would show relatively stable model fit (i.e., small RMSEA_D values) between these three sets of model comparisons. Furthermore, since the BFI-2 in all formats was constructed to include three facets for each personality domain, we expect that accounting for these three facets would lead to substantial differences in model fit for all formats. In other words, we expect large RMSEA_D values between Models 1 vs. 4 and Models 3 vs. 5 across all formats.

Reliability Analyses. For each domain subscale and scale format, we computed omega reliability coefficients based on Model 4 in Figure 1, which, as

[2] For RMSEA and CFI, we used the `rmsea.robust` and `cfi.robust` versions in lavaan, which are corrected for missing data, proposed by Zhang and Savalei (2023).

mentioned previously, is the most consistent with the intended construction of the subscale using the `compRealSEM()` function in the `semTools` package (version 0.5-6.941). Since each domain subscale has three facets, we computed the facet-specific reliability and the total reliability. The facet-specific reliability coefficient represents the proportion of variance in facet items explained by their respective facet factor, whereas the total reliability coefficient represents the proportion of variance in all domain items explained by all three facet factors.

Validity Analyses. For our validity analyses, we assessed two main types of correlations: the correlation between self- and peer-ratings (i.e., self-other agreement) and the correlation between traits and criterion measures. For participants who had two peer ratings, we averaged the two peer-rating scores before computing self-other agreement correlations. We tested the significance of all validity correlations while controlling for gender, age, and ethnicity (see Table 2 for the expected correlations between the traits and criterion measures). Because we focused on comparisons across four scale formats for each criterion measure, we applied Bonferroni corrections by multiplying each original p-value by four to control for family-wise Type-I error. Additionally, pairwise differences in both self-other agreement and trait-criterion correlations across the four scale formats were statistically compared using the `cocor.indep.groups` function from the `cocor` package in R (Diedenhofen & Musch, 2015). For these pairwise comparisons, Bonferroni corrections were applied by multiplying each original p-value by six.

Results

Descriptive Statistics

Descriptive statistics including items' means, variances, correlation matrices, and distribution of response option endorsements for each format and subscale are provided in Section 2.1 of the Supplementary Materials.

Means and Correlations. Across different formats, item means and correlation matrices for each subscale were similar. On the other hand, on average, item variances were higher in the Likert format compared to the alternative formats for all

subscales. Among the three alternative formats, the Item-Specific-Light format consistently exhibited higher item variances than the Expanded and Item-Specific-Full formats, which makes sense given that the Item-Specific-Light format featured less detailed response options compared to the Expanded and Item-Specific-Full formats. For example, in the negative emotionality subscale, the Likert format had the highest average item variance (1.50), followed by the Item-Specific-Light format (1.17), and finally, by the Expanded and Item-Specific-Full formats, which each had similar average item variances (0.91 and 0.90, respectively). This reduction in item variance for the alternative formats may be attributed to a reduction in confusion and acquiescence bias, as demonstrated in the subsequent factor analyses.

Distribution of Response Option Endorsements. In response to a reviewer's recommendation, we also explored possible differences in the distribution of response option endorsements across the four formats (see Supplementary Materials, Section 2.1.4). In general, the distributions for the alternative formats align better with the expected bell-shaped distribution of the personality trait compared to those for the Likert format. Figure 2 illustrates this distribution for the Extraversion subscale across various formats. As shown in Figure 2, a higher percentage of participants selected the mid-point option in the alternative formats than in the Likert format. The response distributions were unimodal in each of the alternative formats. In contrast, for the Likert format, the responses often exhibited a bimodal distribution with peaks on each side of the mid-point (see Figure 2). Non-preregistered pairwise chi-square tests of homogeneity showed that for all BFI-2 subscales, the endorsement distributions for the Likert format were significantly different than those for each alternative format; however, the endorsement distributions for each pair of the alternative formats were not significantly different from one another (see Supplementary Materials Table 36 for full results).

Factor Analyses

Our first substantive question was whether the alternative formats better capture the theorized factor structure of the BFI-2, compared to the original Likert

format. To do so, we fit five models (see Figure 1) for each of the Big Five domains and for each of the four formats. Fit measures of the CFA models are provided in Table 3. Table 4 presents the $RMSEA_D$ values for comparing nested models. Selected results for the factor loadings and correlations are provided in Table 5 (see Section 2.2 of the Supplementary Materials for complete results).

Consistent with our hypotheses and past findings (Soto & John, 2017a), Models 4 and 5 (i.e., three-facets and three-facets-plus-acquiescence models) consistently showed better fit across all formats and subscales compared to Models 1-3, which did not account for the facets' structure (see Table 3). Models 1-3 did not yield fit indices that met the criteria for a good fit, whereas Models 4-5 produced fit indices indicating a reasonable fit for most subscales and formats. The substantial $RMSEA_D$ values when comparing Models 1 and 4, and Models 3 and 5, further indicated the large differences in model fit per degree of freedom (see Table 4). Given that the BFI-2 was designed to include facets within each personality domain, these patterns of results provided evidence for construct validity across all formats, including the newly developed alternative formats.

Furthermore, consistent with our hypothesis, Models 1 and 4 (single-domain and three-facets models), which did not account for the acquiescence bias nor the negatively- vs positively-worded distinction, tended to have a worse fit for the Likert format than for the three alternative formats. Specifically, across the five subscales, on average, Model 1 or 4 had lower CFI and higher RMSEA and SRMR (i.e., worse fit) for the Likert format than for the alternative formats (see the "Average Across Subscales" panel in Table 4). Notably, although Model 4 (three-facets model) is the most consistent with the theoretical model of BFI-2, its CFI and RMSEA values for the Agreeableness and Conscientiousness subscales in the Likert format did not pass the cutoffs for a reasonable fit. In contrast, CFI and RMSEA for the same subscales in the alternative formats passed the cutoffs for a reasonable fit, with some surpassing the cutoffs for a very good fit.

For Models 2, 3, and 5, which account for acquiescence bias (Models 3 and 5) or the negatively- vs positively-worded distinction (Model 2), the Likert format's model fit

was similar to the alternative formats. In other words, for the Likert format, changing from a model that did not account for acquiescence bias or the negatively- vs. positively-worded distinction to a model that did account for these factors improved the fit substantially, as also indicated by the large $RMSEA_D$ values comparing nested Models 1 vs. 2, Models 1 vs. 3, and Models 4 vs. 5 for the Likert format (see Table 4). On the other hand, for the alternative formats, the fit measures were very similar between nested Models 1 vs. 2, Models 1 vs. 3, and Models 4 vs. 5, as evidenced by the small $RMSEA_D$ values for the alternative formats in Table 4. This is consistent with our expectation that the alternative formats are not affected by the same response bias as the Likert format. It is noteworthy that for the alternative formats, Models 2, 3 and 5 often yielded improper parameter estimate solutions (a.k.a., Heywood cases), as indicated by asterisks in Table 3). This is a further indication of the incompatibility of these models with the data under the alternative formats.

Finally, examining the factor loadings and correlations across subscales and formats revealed two problematic items with very small loading sizes across all models for the Likert format. These two items were Item 5 ("I have few artistic interests"; for the Open-Mindedness subscale) and Item 17 ("I have little sympathy for others"; for the Agreeableness subscale), both of which are negatively-worded. Specifically, as shown in Table 5, for Model 4, the loading sizes for Items 5 and 17 were very low at .199 and .167, respectively, in Study 1. These two factor loadings were similarly low in the other models (see Section 2.2 of the Supplementary Materials for factor loading tables). In CFA, the square of the standardized factor loading represents a single item's reliability; thus, these two items' low loadings under the Likert format indicate poor item reliability. Such phenomena, where negatively-worded items demonstrate poor reliability, are often observed in scales in the Likert format (Sonderen et al., 2013). In contrast, for the alternative formats, the loading sizes for these two items in Model 4 were all very reasonable, ranging from .644 to .774.

Reliability Analyses

Next, we compared the reliability of the scores for the domain and facet scales across the four formats. Table 6 presents the composite reliability measures for the total scores of each domain subscale (i.e., "Total" row for each domain in Table 6), as well as for each facet within each domain. Overall, each domain subscale and each facet scores had reasonable reliability, with most values above .70 (see the shaded cells in Table 6). When comparing the domains, the Agreeableness domain had slightly lower reliability for its three facets across all formats, compared to the other Big Five domains.

Compared to the alternative formats, the Likert format tended to have lower reliabilities with an average of .84 for total score reliabilities across domains. Compared to the alternative formats, the Likert format had particularly lower facet reliabilities for the Agreeableness and Open-Mindedness domains. These low reliabilities for the Agreeableness and Open-Mindedness domains were partly caused by the low loadings for Items 5 and 17, as explained previously.

Among the three alternative formats, the Expanded format consistently showed high reliabilities across domains and facets, with all but one value above .70 and an average of .88 for total score reliabilities across domains. The Item-Specific-Full and the Item-Specific-Light had similar reliabilities, both with an average of .85 for total score reliabilities across domains.

Validity Analyses

Finally, we compared the validity of the four formats by examining self-other agreement and correlations with relevant criterion variables (see Table 7).

Despite the relatively low response rate for the peer-ratings, the correlations between self- and peer-ratings (i.e., self-other agreement) were consistently high and significant across almost all formats and subscale. The only exception was the Agreeableness subscale in the original Likert format, where self-other agreement was not significant. In fact, compared to the alternative formats, the self-other agreement for the Agreeableness subscale in the Likert format was much lower (see Table 7).

Similarly, for criterion measures, most correlations were in the anticipated

direction across scale formats, indicating high self-other agreement and predictive validity (see Table 7). However, there were a few exceptions. For example, across all formats, the correlations between conscientiousness and substance abuse, between volunteerism and agreeableness, and between negative emotionality and affectual solidarity with parents were not significant. These non-significant correlations might be attributable to our sample, which consisted exclusively of UBC students, potentially limiting variability on some criterion variables (e.g., 97% of our sample reported no substance abuse).

In short, the correlations for self-other agreement and criterion variables were very similar across scale formats (see Table 7). Among the 76 pairs of correlations between formats, 73 were not significantly different. Consistent with previous research (e.g., Zhang et al., 2019, 2023), this result indicates that the alternative formats provided comparable self-other agreement and predictive validity to the Likert format.

Summary of Results

In summary, the results of Study 1 were generally consistent with our hypotheses. The alternative formats of the BFI-2 tended to have better model fit than the Likert format — particularly for Model 4, which is most consistent with the intended structure of BFI-2. All scale formats' scores demonstrated high and comparable levels of reliability and validity. The Expanded format had the best overall performance, consistently showing better model fit measures, reliability coefficients, and distribution of response options endorsement as well as demonstrating comparable validity.

Study 2

Method

The primary goal of Study 2 is to investigate the effect of scale format on careless responding, and the secondary goal is to replicate the factor and reliability analyses from Study 1. As the focus of Study 2 is on the measures of careless responding (described below), we did not include peer ratings or criterion measures.

Participants and Procedure

We recruited 1,451 participants from York University in Canada who completed an online Qualtrics survey in exchange for course credit. As in Study 1, participants were randomly assigned to complete the BFI-2 in one of the four scale formats (using the same measures described in Study 1): Likert (n = 365), Expanded (n = 356), Item-Specific-Full (n = 366), and Item-Specific-Light (n = 364). Participants then provided basic demographic information, including their age (M = 19.48, SD = 3.62), gender (22.04% male, 76.79% female, 1.10% gender variant), and ethnic identity (21.76% European, 21.42% South Asian, 12.67% East Asian, 9.57% African, and 35.27% from other ethnic backgrounds).

Careless Responding Measures

To evaluate careless responding, we focused on three measures of careless responding that had been used in past research studies (Johnson, 2005; Meade & Craig, 2012): time taken to complete the BFI-2, odd-even consistency, average long-string, maximum long-string, and the Mahalanobis distance.

Completion Time. We recorded the time it took participants to complete the BFI-2 in different formats by inserting a timer tool in Qualtrics at the start and end of the BFI-2 scale, ensuring that only the time spent completing the BFI-2 was recorded. The recorded time in seconds was then converted to minutes by dividing by 60. Longer completion times indicate less careless responding because they suggest that participants took more time to consider and process the scale items.

Odd-Even Consistency. To compute the even-odd consistency, we divided items in each domain into even or odd items and then computed two sum scores—one for even item and one for odd items—for each domain. In other words, in total, across domains, there were five sum scores for even items and five sum scores for odd items. For each participant, we then calculated a within-person correlation between five sum scores for the even items and five sum scores for the odd items. A high score on the even-odd consistency has been interpreted as indicating low levels of careless responding because it shows that a participant's responses are consistent across even and odd items

(Johnson, 2005; Meade & Craig, 2012).

Average and Maximum Long-String. Long-string refers to the number of identical consecutive responses. Thus, the average long-string and maximum long-string represent, respectively, the average and maximum number of identical consecutive responses calculated for each participant. [3] According to Meade and Craig (2012), higher average or maximum long-string scores represent a response heuristic called straightlining (e.g., consistently choosing the first option), which can be an indicator of inattentive or careless responding.

Mahalanobis Distance. The Mahalanobis Distance is a multivariate distance between a participant's response vector and the vector of the sample means. While often used as an outlier indicator, Mahalanobis distance has also been used to detect careless responding (Ehlers et al., 2009; Meade & Craig, 2012). Higher values (and therefore, more unusual responses) indicate more careless responding because unusual responses could be attributable to inattention, response biases, and differential interpretation of the items or response options (Ehlers et al., 2009; Meade & Craig, 2012).

Data Analyses

We used the same analytic strategy as in Study 1 to replicate the factor and reliability analyses. For each measure of careless responding, we tested the differences across the four scale formats using omnibus F-tests as well as pairwise t-tests with Bonferroni corrections. We expect that the alternative formats will yield less careless responding than the Likert format.

Results

Descriptive Statistics

The descriptive statistics including item means, item variances, correlation matrices, and distribution of response option endorsements are provided in Section 3.1 of the Supplementary Materials. Overall, the patterns of results for the descriptive statistics in Study 2 were similar to those in Study 1. For example, consistent with

[3] For the Likert format, the average and maximum long-string indices were computed using participants' original scores—that is, before reverse-keying the negatively-worded items. In contrast, all other careless responding measures were calculated after reverse-keying.

Study 1, Study 2 also showed that the item variances were larger for the Likert format relative to the alternative formats. Moreover, the endorsement distributions for the Likert format in Study 2 were also significantly different from those in the alternative formats, with notably fewer mid-option endorsements in the Likert format compared to the alternative formats. A main difference between the descriptive statistics of Studies 1 and 2 is that Item 5 ("I have few artistic interests"; from the Open-Mindedness domain) in the Likert format showed low and slightly negative correlations with other items in the same domain. This issue affected the factor loadings of other items in the same domain, as we will explain in the subsequent section.

Factor Analyses

Many patterns of factor analytic results in Study 2 were similar to those in Study 1. Consistent with our hypothesis and Study 1's results, Models 4 and 5 (which include the structure of the facets in each subscale) showed better model fit than Models 1-3 (which do not include facets). Most fit measures for Models 1-3 demonstrated poor fit and did not pass the cutoffs for a reasonable fit across formats, whereas most fit measures for Models 4-5 indicated at least a reasonable fit. Table 8 shows selected fit measures results for Models 4 and 5; for the full results of all models, please refer to the Supplementary Materials.

Consistent with Study 1, changing from a model without the acquiescence bias (Model 4) to a model with (Model 5) resulted in a substantially improved model fit for the Likert format (as indicated by large RMSEA_D values) but barely changed the model fit for the alternative formats (see Table 8). However, one noticeable difference between Studies 1 and 2's fit measure results was that the fit measures for the Open-Mindedness domain in the alternative formats were considerably worse in Study 2 compared to Study 1. Specifically, for Study 1, Models 4-5's RMSEA, CFI, and SRMR values for the Open-Mindedness domain in the alternative formats passed the cutoffs for a reasonable fit, whereas in Study 2, only the SRMR value did. In addition, in Study 2, contrary to our expectations, Models 4-5 had *better* model fit for the Likert format than for the alternative formats (an issue we will turn to in the General Discussion).

Consistent with Study 1, most items yielded reasonable loadings across models and subscales. However, as in Study 1, Items 5 ("I have few artistic interests") and 17 ("I feel little sympathy for others") had low loadings in the Likert format (see Table 5). Notably, the loadings for Items 5 and 17 were lower in Study 2 than in Study 1, with values of .091 and .031, respectively (see Table 5). Furthermore, an unexpected result was that the slight negative correlations between Item 5 and other Open-Mindedness items led to negative loadings for other items (i.e., Items 20, 35, and 50) measuring the same facet as Item 5. It also resulted in negative factor correlations between the "Aesthetic Sensitivity" and other facets of the Open-Mindedness subscale, essentially transforming what was supposed to represent "Aesthetic Sensitivity" into what could be described as "Aesthetic Insensitivity" (see red-shaded cells in Table 5). This pattern emphasizes a broader issue with negatively-worded items: even when they do not impact the overall model fit, they can affect the loading sizes and directions of both their own and related items, ultimately changing the meaning of the latent construct.

Reliability Analyses

The results for the reliability analyses are shown in Table 6. The general patterns of the reliability coefficients in Study 2 are similar to those in Study 1. For example, as in Study 1, the Agreeableness and the Open-Mindedness subscales in Study 2 showed lower reliabilities than other subscales. Furthermore, similar to Study 1, among the three alternative formats, the Expanded format tended to have higher reliabilities than the Item-Specific-Full and Item-Specific-Light formats.

However, Study 2's reliability coefficients were generally lower than those in Study 1. This difference was particularly evident in the Likert format, where all corresponding reliabilities were lower in Study 2 than in Study 1. The reliability coefficient for the Compassion facet of the Agreeableness subscale (which includes the very low-loading Item 17) was exceptionally low, at .091.

Careless Responding Analyses

Next, we examined indicators of careless responding using previously explained measures: completion time, even-odd consistency, average long-string, maximum

long-string, and Mahalanobis distances for each domain. Table 9 shows the correlations among the careless responding measures and Figure 3 shows each measure's mean differences across the four formats.

Correlations Among the Careless Responding Measures. Before comparing careless responding across formats, we calculated correlations among all careless responding measures to assess their validity (see Table 9). As shown in Table 9, while some correlations are weak, their signs are consistent with our interpretation of these measures as indirect indicators of careless responding. Specifically, the Mahalanobis distances across domains were moderately positively correlated, with values ranging from $r = .364$ to $r = .522$. The correlation between completion time and even-odd consistency was slightly positive and significant ($r = .114$, $p < .001$), indicating that respondents who spent more time completing the questionnaire provided more consistent responses. Additionally, both completion time and even-odd consistency showed weak but generally significant negative correlations with Mahalanobis distances (ranging from $r = -.178$ to $r = -.054$), suggesting that higher consistency generally corresponds to lower Mahalanobis distances. Furthermore, the average and maximum long-string scores strongly correlated with each other ($r = .832$) and showed small but significant negative correlations with even-odd consistency ($r = -.188$ and $r = -.238$, respectively). This finding indicates that increased consecutive identical responses corresponded to lower consistency. However, these long-string measures generally showed no significant correlations with either completion time or Mahalanobis distances.

Completion Times. Comparing the careless responding measures across the four formats, we found that completion times significantly differed across formats ($F = 42.47$, $p < .001$, see Figure 3a). Consistent with our hypothesis, the Expanded format resulted in the highest completion time, with an average completion time of 8.89 minutes with 95% CI=[8.35, 9.43], followed by the Item-Specific-Full ($M = 6.87$, 95% CI=[6.47, 7.27]), Item-Specific-Light ($M = 6.06$, 95% CI=[5.67, 6.35]), and lastly the Likert ($M = 5.72$, 95% CI=[5.36, 6.08]). Pairwise, the mean completion time of the Expanded was significantly different from that of the Item-Specific-Full, Item-Specific-Light, and the Likert formats (see Figure 3a). However, there were no

significant differences in mean completion times among the Item-Specific-Full, Item-Specific-Light, and the Likert formats (see Figure 3a). In other words, out of the three alternative formats, only the Expanded format resulted in a significantly increased completion time compared to the Likert format.

Even-Odd Consistency. For even-odd consistency measure, as shown in Figure 3b, the Expanded format showed a somewhat higher mean even-odd consistency (M = .73, 95%CI=[.70, .76]) than the Likert format (M = .68, 95%CI=[.65 .71]), Item-Specific-Full (M = .68, 95%CI=[.65 .70]), and Item-Specific-Light (M = .67, 95%CI=[.64 .71]). An F-test showed significant differences across formats (F = 3.18, p = .02). Contrary to our expectations, however, pairwise comparisons did not reveal any significant differences in even-odd consistency between the Likert format and any of the alternative formats (see Figure 3b). In other words, there was no evidence that participants' responses were more consistent in the alternative formats compared to the Likert format. The only significant pairwise comparison in even-odd consistency was between the Expanded and the Item-Specific-Full formats.

Average and Maximum Long-String. The average and maximum long-string scores did not show significant differences across formats with F = 1.5, p = 0.21 for average long-string, and F = 1.78, p = 0.21 for maximum long-string (see Figure 3c and d). In other words, contrary to our hypothesis, the alternative formats did not show less careless responding as measured by the average and maximum long-string. Interestingly, the Item-Specific-Light format had much more variability in the average and maximum long-string scores compared to the other formats (see Figure 3c and d). Further examination revealed that, in the Item-Specific-Light format, three participants provided 48 or more identical consecutive responses. We will explore possible reasons for these findings in the General Discussion.

Mahalanobis Distances. Mahalanobis distances' for each domain significantly differed across formats (see the footnote of Figure 3 for F-test results). Across all subscales, the Likert format showed the largest Mahalanobis distance followed by the Item-Specific-Light. The smallest distances were either in the Expanded or the Item-Specific-Full formats, depending on the subscale (see Figure 3e-i). In addition, for

all subscales, there were significant differences in the Mahalanobis distance between the Likert format and any of the alternative formats (see Figure 3e-i). This pattern parallels the results for the item variances, which demonstrated that the Likert format had higher average item variances for all subscales. These results demonstrated that relative to the Likert format, the alternative formats yielded less unusual and more consistent responses across participants.

Summary of Study 2

In summary, Study 2 replicated most of the factor analyses and reliability findings of Study 1. However, Study 2 showed lower reliability of the BFI-2, especially for the Likert format, and highlighted even more pronounced issues with Items 5 and 17. The careless responding measures showed somewhat mixed results. Specifically, participants spent more time completing the BFI-2 using the alternative formats—particularly the Expanded format—and exhibited less unusual responses. However, participants did not show higher even-odd consistency or less straightlining in the alternative formats compared to the Likert format.

Discussion

The Big Five are commonly measured by asking participants about the extent to which they agree or disagree with various statements (e.g., "I am outgoing, sociable"; Soto & John, 2017a). However, items in the Likert format are susceptible to ambiguity in the interpretation of the item and response scale, acquiescence bias, and careless responding due to confusion when switching between positively- and negatively-worded items. Here, we investigated whether the BFI-2 in three alternative formats can address these issues and thereby yield superior psychometric properties compared to the original Likert format. Across two studies, we found that, compared to the Likert format, the alternative BFI-2 formats generally yielded factor structure more consistent with the theoretical expectation, higher reliability coefficients, response endorsement distributions more bell-shaped, longer completion times, fewer extreme responses, and similar validity.

Specifically, the BFI-2 in the original Likert format was more consistent with the

three-facets model with an acquiescence bias factor (Model 5), whereas the alternative formats were more consistent with the intended model of the BFI-2, which is the three-facets model but no acquiescence bias (Model 4). Indeed, for Model 4, the alternative formats had a better fit than the Likert format for four out of the Big Five domains. However, for the Open-Mindedness domain, the Likert format outperformed the alternative formats, especially for Study 2. One possible explanation for this unexpected result is the "reliability paradox" (Hancock & Mueller, 2011; McNeish & Hancock, 2018), which occurs when a less reliable measurement model with lower loading sizes yields a better fit than a more reliable model with higher loading sizes. The reason for the reliability paradox is that a less reliable measurement model is less likely to detect model misspecifications. Using McNeish and Hancock (2018)'s simile, a less reliable measurement model is like "a dirtier window" through which it is harder to discern misspecifications, whereas a more reliable model with high loadings is more sensitive to any misspecification, yielding a poorer fit. Indeed, the negatively-worded Item 5 ("I have few artistic interests") in the Open-Mindedness subscale yielded very small loadings (.199 in Study 1 and .031 in Study 2) in the Likert format, but ranged from .64 to .71 for the alternative formats across the two studies. Furthermore, in Study 2, not only did Item 5 yield a low loading, it also caused other items belonging to the same latent facet to have negative loadings and the latent facet to have negative correlations with other facets, essentially altering the meaning of the latent facet. Perhaps, due to the poor reliability of Item 5 in the Likert format, misspecifications are less likely to be detected when using the BFI-2 in the Likert format, which yields an apparent better fit than the BFI-2 in the alternative formats.

Other items in the Likert format also had lower item reliabilities compared to those in the alternative formats. In particular, the negatively-worded Item 17 ("I feel little sympathy for others") in the Agreeableness scale had a very low loading size in both studies (.167 in Study 1 and .091 in Study 2). Furthermore, overall, the Likert format tended to have lower reliability coefficients at the facet and the domain levels compared to the alternative formats—especially in Study 2. The low reliabilities—especially for some negatively-worded items—for the Likert formats are

consistent with past research that scales that use the Likert format often have a few negatively-worded items with low correlations with other items in the same scale (e.g., Sonderen et al., 2013). Given that the reliabilities for the Likert format remained low even after accounting for the acquiescence factor in Model 5, this phenomenon is likely attributable to the confusion and careless responding caused by the negatively-worded items or the ambiguous agree-disagree response options.

Among the alternative formats, across Studies 1 and 2, we found that the Expanded format tended to have the highest reliabilities, followed by the Item-Specific-Full and then the Item-Specific-Light formats. This result aligns with the expectation that more detailed response options can enhance participants' understanding and processing of the items, thereby leading to higher reliability. The Expanded format, which offers the most detailed response options, appears to facilitate better comprehension and item processing, resulting in better reliability compared to the other formats.

Furthermore, in both Studies 1 and 2, we found that the distributions of response option endorsement were significantly different between the Likert and the alternative formats. Specifically, for the Likert format, the responses often exhibited a bimodal distribution with peaks on each side of the mid-point; that is, fewer participants selected the mid-point relative to adjacent response options. In contrast, responses in the alternative formats exhibited a unimodal distribution, with most participants selecting the mid-point option. These patterns of results may suggest that participants perceive the mid-point option in the Likert format as ambiguous or as an "undecided" choice. This might explain why participants select the mid-point option less frequently, instead favoring more extreme responses for each item. In contrast, the mid-point option in the alternative formats has a clear meaning; as a result, the distribution of endorsement follows the expected bell-shaped distribution.

In terms of validity, we found that subscales in the alternative format had similar self-other agreements and correlations with the criterion variables as those in the Likert format. Although these results did not favour the alternative format over the Likert format, these results are consistent with most of the past research studies, which

also found comparable but not higher validities for the alternative formats (e.g., Zhang & Savalei, 2016; Zhang et al., 2019). This implies that the validity of the BFI-2 items is relatively robust to minor changes in wording and response format. Moreover, these results show that the alternative formats are at least as valid as the Likert format (while performing better on other psychometric properties).

Regarding the careless responding analyses, we found significant differences between the Likert and the alternative formats for the completion times and the Mahalanobis distances, but not for the even-odd consistency and average and maximum long-string. One notable result is that the Expanded format required a significantly longer time than the other formats to complete. Specifically, on average, the BFI-2 in the Expanded format required around 1.5 minutes longer to complete than the Item-Specific-Full and 2.5 minutes longer than the Item-Specific-Light and the Likert format. Furthermore, we found that all alternative formats yielded significantly lower Mahalanobis distances than the Likert format, with the Expanded and Item-Specific-Full formats consistently having the lowest two scores. These results demonstrate that the Expanded and Item-Specific-Full formats yielded fewer outliers and unusual responses, possibly because their detailed response options reduced participant confusion or careless responding. The higher Mahalanobis distances in the Likert format might result from confusion caused by frequent switching between positively- and negatively-worded items as well as the varied interpretations of agree-disagree response options, as discussed in the Introduction; these factors could potentially increase measurement error and response variability. However, it is important to emphasize that completion time and Mahalanobis distances are indirect indicators of careless responding. Thus, these results cannot confirm that participants indeed experienced less confusion or were more attentive with the Expanded and Item-Specific-Full formats.

We did not find any significant differences between response formats for the average and maximum long-string, which measures the number of identical consecutive responses. The directionality of the items may be a confounding variable for the effect of scale format on the long-string scores. Specifically, in the Likert format, the adjacent

items not only vary in the personality domain they assess but may also vary in the item directionality (positively- or negatively-worded items). In contrast, items in the alternative formats do not incorporate directionality, and the personality domains themselves are often positively correlated. Consequently, an attentive participant, who, for example, is high on extraversion, agreeableness, and conscientiousness, might consecutively choose the fourth response option for three items measuring these traits in an alternative format. This can lead to slightly higher long-string scores for the alternative formats compared to the Likert format. Indeed, Figure 3c and d depict a slight, although non-significant, trend toward higher average and maximum long-string scores in the alternative formats compared to the Likert format. Second, as mentioned in the Result section of Study 2, the Item-Specific-Light had much more variability in the average and maximum long string scores. In addition, although non-significant, the Item-Specific-Light had the highest average and maximum long-string among the four formats, with three participants having more than 48 identical consecutive responses. A plausible explanation for these results is that all response options for a given item in the Item-Specific-Light format appeared in a single row. This may have made it easier for inattentive participants to straightline their responses without scrolling down the page or moving their mouse too much (see Table 1). In contrast, in the other three formats, each response item for each item appeared in a separate row. This required participants to actively scroll down the page and move their mouse between items, thus discouraging straightlining.

Limitations and Future Directions

There are several limitations to our study. First, we will consider developing short-form versions of the BFI-2 in alternative formats, similar to the shorter versions of the Likert BFI-2 (i.e., BFI-2-S and BFI-2-XS) created Soto and John (2017b). A shorter scale with reduced completion time would likely encourage more applied researchers to adopt alternative response formats of the BFI-2.

Second, the measures of careless responding used in Study 2 were indirect. For example, a longer completion time does not necessarily indicate less careless responding,

as it could also reflect the time required to process the lengthier or more varied response options in our alternative formats. Furthermore, the even-odd consistency involves computing a correlation using only ten scores for each participant (five even-item sum scores and five odd-item sum scores), which raises concerns regarding the reliability of this measure. Moreover, the long-string measures were confounded with the directionality of the items across formats. Future research should employ more direct measures of careless responding, such as explicitly asking participants about their engagement during the survey (Meade & Craig, 2012).

Furthermore, our assumption that improved psychometric properties from alternative formats stem from reduced ambiguity, acquiescence bias, and confusion needs more direct evidence. Relying only on indirect measures of careless responding and quantitative analyses does not allow us to confirm that participants indeed experienced less confusion or responded more carefully under the alternative formats, nor does it clarify why participants respond differently across formats or whether the detailed response options in alternative formats aid participants' understanding of the items. To gain such insight, a mixed-method approach that includes both qualitative and quantitative analyses is necessary. This mixed-method approach is also suitable for incorporating explicit measures of careless responding as suggested previously. Several previous studies have employed this approach to explore participants' responses to scale items in greater depth. For example, Cabooter et al. (2016) utilized the think-aloud method to investigate how participants interpret varying wordings of response options. Building on these existing methodologies, we plan to conduct a follow-up study that examines participants' interpretations of items across different scale formats.

Conclusion and Recommendations

We converted the BFI-2 into three alternative formats and assessed whether they had better psychometric properties compared to the original Likert format. We found that the alternative formats demonstrated better psychometric properties, such as factor structure that is more consistent with the intended structure, increased reliability, unimodal (rather than bimodal) response endorsement distributions, longer

completion time, and reduced extreme or unusual responses (i.e., lower Mahalanobis distances). However, self-other agreement and predictive validity were similar across all formats. Among the three alternative formats, the Expanded format yielded the highest reliability, longest completion time, and least unusual responses, followed by the Item-Specific-Full format. On average, the BFI-2 in the Expanded format took 1.5 minutes longer to complete compared to the BFI-2 in the Item-Specific-Full format (and 2.5 minutes longer compared to the original Likert format).

Based on these findings, if researchers can accommodate a longer survey duration, we recommend using the BFI-2 in the Expanded format—particularly when research analyses involve modelling the factor structure of the BFI-2. If time is a constraint, the Item-Specific-Full format might be a preferable alternative, as it took only one minute longer than the Likert format while still offering improved psychometric properties, providing a good balance between measurement quality and survey time. On the other hand, if research primarily involves examining scale-level relationships between variables and requires brevity, the original BFI-2 in the Likert format remains an appropriate choice, given its high predictive validity and self-other agreement, comparable to the Expanded and Item-Specific-Full formats.

Moreover, we suggest that researchers consider converting existing scales into these formats or using the Expanded or Item-Specific-Full formats when developing new psychological scales so that we can build upon the current and previous studies on the alternative formats (e.g., Dykema et al., 2022; Kam, 2020; Zhang & Savalei, 2016; Zhang et al., 2023) to learn more about their advantages and limitations.

Acknowledgement

We extend our appreciation to all the research assistants who contributed to this study. We acknowledge Sam Phuong Can for her assistance with designing and conducting Study 1, as well as Nikki Lombardo, Victoria Celio, and Zeyi Xu for their assistance with the design, data collection, and data analyses of Study 2.

References

Belsky, J., Jaffee, S. R., Caspi, A., Moffitt, T., & Silva, P. A. (2003). Intergenerational relationships in young adulthood and their life course, mental health, and personality correlates. *Journal of Family Psychology*, *17*(4), 460. https://doi.org/10.1037/0893-3200.17.4.460

Billiet, J. B., & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural equation modeling*, *7*(4), 608–628. https://doi.org/10.1207/S15328007SEM0704_5

Brauer, K., & Proyer, R. T. (2021). Analyzing a german-language expanded form of the phophikat-45: Psychometric properties, factorial structure, measurement invariance with the likert-version, and self-peer convergence. *Journal of Personality Assessment*, *103*(2), 267–277. https://doi.org/10.1080/00223891.2020.1720699

Cabooter, E., Weijters, B., Geuens, M., & Vermeir, I. (2016). Scale format effects on response option interpretation and use. *Journal of Business Research*, *69*(7), 2574–2584. https://doi.org/10.1016/j.jbusres.2015.10.1380148-2963/

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Sage publications.

Diedenhofen, B., & Musch, J. (2015). Cocor: A comprehensive solution for the statistical comparison of correlations. *PloS one*, *10*(4), e0121945. https://doi.org/10.1371/journal.pone.0121945

Dykema, J., Schaeffer, N. C., Garbarski, D., Assad, N., & Blixt, S. (2022). Towards a reconsideration of the use of agree-disagree questions in measuring subjective evaluations. *Research in Social and Administrative Pharmacy*, *18*(2), 2335–2344. https://doi.org/10.1016/j.sapharm.2021.06.014

Ehlers, C., Greene-Shortridge, T., Weekley, J., & Zajack, M. (2009). The exploration of statistical methods in detecting random responding. *Annual Meeting of the Society for Industrial/Organizational Psychology, Atlanta, GA*.

Godin, G., & Shephard, R. J. (1997). Godin Leisure-Time Exercise Questionnaire. *Medicine & Science in Sports & Exercise*, *26*, S36–S38.

Greene, K., Krcmar, M., Walters, L. H., Rubin, D. L., Hale, L., et al. (2000). Targeting adolescent risk-taking behaviors: The contributions of egocentrism and sensation-seeking. *Journal of adolescence*, *23*(4), 439–461. https://doi.org/10.1006/jado.2000.0330

Hancock, G. R., & Mueller, R. O. (2011). The reliability paradox in assessing structural relations within covariance structure models. *Educational and Psychological Measurement*, *71*(2), 306–324. https://doi.org/10.1177/0013164410384856

Hensley, W. E., & Roberts, M. K. (1976). Dimensions of Rosenburg's Self-Esteem Scale. *Psychological Reports*, *38*(2), 583–584. https://doi.org/10.2466/pr0.1976.38.2.583

Horan, P. M., DiStefano, C., & Motl, R. W. (2003). Wording effects in self-esteem scales: Methodological artifact or response style? *Structural Equation Modeling*, *10*(3), 435–455. https://doi.org/10.1207/S15328007SEM1003_6

John, O. P., Donabue, E. M., & Kentle, R. L. (1991). *Big Five Inventory (BFI)*. APA PsycTests. https://doi.org/10.1037/t07550-000

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality*, *39*(1), 103–129. https://doi.org/10.1016/j.jrp.2004.09.009

Kam, C. C. S. (2020). Expanded format shows better response consistency than Likert-scale format in the measurement of optimism. *Personality and Individual Differences*, *152*, 109606. https://doi.org/https://doi.org/10.1016/j.paid.2019.109606

Kam, C. C. S., Cheng, E. H., & Cui, T. (2024). Measuring self-esteem with Expanded format in a fraction of time: ESE-S and ESE-US. *Journal of Personality Assessment*, *106*(2), 196–207. https://doi.org//10.1080/00223891.2023.2259990

Kuru, O., & Pasek, J. (2016). Improving social media measurement in surveys: Avoiding acquiescence bias in Facebook research. *Computers in Human Behavior*, *57*, 82–92. https://doi.org/https://doi.org/10.1016/j.chb.2015.12.008

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological methods*, *11*(4), 344. https://doi.org/https://doi.org/10.1037/1082-989X.11.4.344

McNeish, D., & Hancock, G. R. (2018). The effect of measurement quality on targeted structural model fit indices: A comment on Lance, Beck, Fan, and Carter (2016). *Psychological Methods*, *23*(1), 184–190. https://doi.org/10.1037/met0000157

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

Pasek, J., & Krosnick, J. A. (2010). Optimizing survey questionnaire design in political science: Insights from psychology. In J. E. Leighley (Ed.), *The Oxford handbook of American elections and political behaviour* (pp. 27–50). Oxford University Press.

Paunonen, S. V. (2003). Big Five factors of personality and replicated predictions of behavior. *Journal of Personality and Social Psychology*, *84*(2), 411. https://doi.org/10.1037/0022-3514.84.2.411

Robie, C., Meade, A. W., Risavy, S. D., & Rasheed, S. (2022). Effects of response option order on likert-type psychometric properties and reactions. *Educational and Psychological Measurement*, *82*(6), 1107–1129. https://doi.org/10.1177/00131644211069406

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://www.jstatsoft.org/v48/i02/

Saris, W., Revilla, M. A., Krosnick, J. A., & Shaeffer, E. M. (2010). Comparing questions with Agree/Disagree response options to questions with Item-Specific response options. *Survey Research Methods*, *4*(1). https://doi.org/10.18148/srm/2010

Savalei, V., Brace, J. C., & Fouladi, R. T. (2024). We need to change how we compute RMSEA for nested model comparisons in structural equation modeling. *Psychological Methods*, *29*, 480–493. https://doi.org/https://doi.org/10.1037/met0000537

Savalei, V., & Falk, C. F. (2014). Recovering Substantive Factor Loadings in the Presence of Acquiescence Bias: A Comparison of Three Approaches. *Multivariate Behavioral Research*, *49*(5), 407–424. https://doi.org/10.1080/00273171.2014.931800

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Academic Press.

Schwaba, T., Luhmann, M., Denissen, J. J., Chung, J. M., & Bleidorn, W. (2018). Openness to experience and culture-openness transactions across the lifespan. *Journal of Personality and Social Psychology*, *115*(1), 118. https://doi.org/10.1037/pspp0000150

Sonderen, E. v., Sanderman, R., & Coyne, J. C. (2013). Ineffectiveness of reverse wording of questionnaire items: Let's learn from cows in the rain. *PloS one*, *8*(7), e68967. https://doi.org/10.1371/journal.pone.0068967

Soto, C. J. (2019). How Replicable Are Links Between Personality Traits and Consequential Life Outcomes? The Life Outcomes of Personality Replication Project. *Psychological Science*, *30*(5), 711–727. https://doi.org/10.1177/0956797619831612

Soto, C. J., & John, O. P. (2017a). 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*(1), 117–143. https://doi.org/10.1037/pspp0000096

Soto, C. J., & John, O. P. (2017b). Short and extra-short forms of the Big Five Inventory–2: The BFI-2-S and BFI-2-XS. *Journal of Research in Personality*, *68*, 69–81. https://doi.org/10.1016/j.jrp.2017.02.004

Swain, S. D., Weathers, D., & Niedrich, R. W. (2008). Assessing three sources of misresponse to reversed Likert items. *Journal of Marketing Research*, *45*(1), 116–131. https://doi.org/10.1509/jmkr.45.1.116

Trull, T. J., & Sher, K. J. (1994). Relationship between the five-factor model of personality and Axis I disorders in a nonclinical sample. *Journal of abnormal psychology*, *103*(2), 350. https://doi.org/10.1037/0021-843X.103.2.350

Weng, L.-J., & Cheng, C.-P. (2000). Effects of response order on Likert-type scales. *Educational and Psychological Measurement*, *60*(6), 908–924. https://doi.org/10.1177/00131640021970989

Wong, N., Rindfleisch, A., & Burroughs, J. E. (2003). Do reverse-worded items confound measures in cross-cultural consumer research? The case of the Material Values Scale. *Journal of Consumer Research*, *30*(1), 72–91. https://doi.org/10.1086/374697

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, *28*(3), 186. https://doi.org/10.1007/s10862-005-9004-7

Zhang, X., & Savalei, V. (2016). Improving the Factor Structure of Psychological Scales: The Expanded Format as an Alternative to the Likert Scale Format. *Educational and Psychological Measurement*, *76*(3), 357–386. https://doi.org/10.1177/0013164415596421

Zhang, X., & Savalei, V. (2023). New computations for RMSEA and CFI following FIML and TS estimation with missing data. *Psychological Methods*, *28*(2), 263. https://doi.org/10.1037/met0000445

Zhang, X., & Savalei, V. (2024). An overview of alternative formats to the Likert format: A comment on Wilson et al.(2022). *Psychological Methods*, *29*(3), 606–612. https://doi.org/10.1037/met0000631

Zhang, X., Tse, W. W.-Y., & Savalei, V. (2019). Improved properties of the Big Five Inventory and the Rosenberg Self-Esteem Scale in the Expanded format relative to the Likert format. *Frontiers in Psychology*, *10*, 1286. https://doi.org/10.3389/fpsyg.2019.01286

Zhang, X., Zhou, L., & Savalei, V. (2023). Comparing the Psychometric Properties of a Scale Across Three Likert and Three Alternative Formats: An Application to the Rosenberg Self-Esteem Scale. *Educational and Psychological Measurement*, *83*(4), 649–683. https://doi.org/10.1177/00131644221111402

Table 1

Example Items for Each Format

Item Number	**Scale Formats**
	Likert
Item 21	I am dominant, act as a leader. ○ Disagree strongly ○ Disagree a little ○ Neutral; no opinion ○ Agree a little ○ Agree strongly
Item 23	I have difficulty getting started on tasks. ○ Disagree strongly ○ Disagree a little ○ Neutral; no opinion ○ Agree a little ○ Agree strongly
	Expanded
Item 21	Choose one that best describes you. ○ I am very submissive, almost always act as a follower. ○ I am fairly submissive, often act as a follower. ○ I am neither particularly submissive nor dominant, sometimes a follower and sometimes a leader. ○ I am fairly dominant, often act as a leader. ○ I am very dominant, almost always act as a leader.
Item 23	Choose one that best describes you. ○ I have a lot of difficulty getting started on tasks. ○ I have a fair amount of difficulty getting started on tasks. ○ I have some difficulty getting started on tasks. ○ I have little difficulty getting started on tasks. ○ I have no difficulty getting started on tasks.
	Item-Specific-Full
Item 21	How often do you act as a leader or a follower? ○ Almost always act as a follower ○ Often act as a follower ○ Sometimes a follower, sometimes a leader ○ Often act as a leader ○ Almost always act as a leader
Item 23	How much difficulty do you have getting started on tasks? ○ A lot ○ A fair amount ○ Some ○ Little ○ None
	Item-Specific-Light
Item 21	How often do you act as a leader or a follower? ○ ○ ○ ○ ○ Almost always act as a follower — Almost always act as a leader
Item 23	How much difficulty do you have getting started on tasks? ○ ○ ○ ○ ○ A lot — None

Table 2
Criterion Measures in Study 1

Criterion Measures	**Scale Details**	**Predicted Direction With Big Five Traits**
Peer Acceptance	Single item regarding popularity compared to peers in the Behaviour Report Form (BRF; Paunonen, 2003)	Positively Related to Extraversion
Peer Status	Single item regarding social status in the BRF (Paunonen, 2003)	Positively Related to Extraversion
Dating Variety	Single item regarding the number of dating partners in the BRF (Paunonen, 2003)	Positively Related to Extraversion
Party Attendance	Single item regarding the number of party attended in the BRF (Paunonen, 2003)	Positively Related to Extraversion
Volunteerism	Single item regarding the average number of days doing volunteering work in the BRF (Paunonen, 2003)	Positively Related to Agreeableness
Academic Performance	Single item regarding University GPA in the BRF (Paunonen, 2003)	Positively Related to Conscientiousness
Risky Physical Activity	Godin Leisure-Time Exercise Questionnaire (Godin & Shephard, 1997)	Negatively Related to Conscientiousness
Criminal Behaviour	Delinquent Behaviour Scale (Greene et al., 2000)	Negatively Related to Conscientiousness
Substance Abuse	Three items in the Psychopathology Scale (Trull & Sher, 1994)	Negatively Related to Conscientiousness
Pathological Anxiety	Three items in the Psychopathology Scale (Trull & Sher, 1994)	Positively Related to Negative Emotionality
Pathological Depression	Three items in the Psychopathology Scale (Trull & Sher, 1994)	Positively Related to Negative Emotionality
Affectual Solidarity with Parents	Intergenerational Relationship Scale (Belsky et al., 2003)	Positively Related to Negative Emotionality
Cultural Activity	Single item regarding the amount of cultural activities attended (Schwaba et al., 2018)	Positively Related to Open-Mindedness
Playing Music Instruments	Single item regarding whether participants play any musical instrument in the BRF (Paunonen, 2003)	Positively Related to Open-Mindedness

Table 3

Fit Measures for CFA Models in Study 1

	Model 1 (df = 54)				Model 2 (df = 53)				Model 3 (df = 53)				Model 4 (df = 51)				Model 5 (df = 50)			
Fit Measure	L	E	ISF	ISL	L	E	ISF	ISL	L	E	ISF	ISL	L	E	ISF	ISL	L	E	ISF	ISL
Extraversion																				
Chi-Squared	446.155	347.406	353.235	403.780	432.128	344.334*	345.491*	397.632*	435.806	343.075*	335.572*	395.903*	160.538	95.742	96.994	133.272	125.357	95.254	92.985*	132.964*
CFI	.774	.789	.758	.764	.782	.791	.764	.768	.779	.792	.772	.769	.937	.968	.963	.944	.957	.968	.965	.944
RMSEA	.147	.129	.130	.139	.146	.130	.129	.140	.146	.130	.127	.139	.080	.052	.052	.070	.067	.052	.051	.071
SRMR	.082	.079	.082	.082	.082	.078	.079	.081	.082	.078	.079	.081	.058	.041	.043	.049	.056	.041	.042	.049
Agreeableness																				
Chi-Squared	245.370	354.454	254.938	239.125	237.552	353.770*	252.112*	229.860*	230.396	354.247*	253.799*	231.221*	162.396	133.421	114.843	94.629	133.474	127.468	114.327	93.628*
CFI	.777	.731	.782	.787	.785	.731	.784	.797	.794	.731	.782	.795	.870	.926	.931	.950	.903	.930	.930	.950
RMSEA	.102	.130	.106	.102	.101	.132	.107	.101	.099	.132	.107	.101	.080	.070	.062	.051	.070	.069	.062	.051
SRMR	.073	.088	.070	.072	.072	.088	.070	.071	.070	.088	.070	.071	.063	.052	.047	.047	.056	.051	.047	.046
Conscientiousness																				
Chi-Squared	388.251	291.401	252.637	349.085	372.562	290.095	245.494	349.011*	365.138	286.261	236.032	349.045*	194.913	153.465	149.100	118.907	122.907	129.507	116.850	113.839
CFI	.731	.820	.834	.790	.743	.820	.839	.789	.749	.823	.847	.789	.884	.922	.918	.952	.942	.939	.944	.955
RMSEA	.135	.116	.105	.128	.134	.117	.105	.130	.132	.116	.102	.130	.091	.078	.076	.063	.066	.070	.064	.062
SRMR	.081	.071	.072	.092	.078	.070	.071	.092	.078	.070	.070	.092	.057	.047	.053	.048	.050	.043	.046	.047
Negative Emotionality																				
Chi-Squared	317.015	301.768	266.677	319.851	279.697	301.191*	255.335	316.374	286.726	301.681*	257.781	315.743	149.749	120.982	143.400	158.255	87.640	118.210	130.368	151.766
CFI	.858	.857	.853	.825	.878	.857	.860	.826	.874	.856	.859	.827	.947	.960	.936	.929	.980	.961	.945	.933
RMSEA	.120	.118	.109	.122	.112	.119	.107	.122	.114	.120	.108	.122	.076	.065	.074	.079	.047	.064	.070	.078
SRMR	.058	.059	.061	.066	.057	.059	.061	.066	.056	.059	.060	.065	.043	.040	.051	.056	.035	.038	.047	.052
Open-Mindedness																				
Chi-Squared	326.522	283.656	307.066	312.310	322.161	282.405*	301.205*	312.289	320.775	283.162*	301.827*	311.771	129.602	135.356	152.085	162.237	106.440	132.375	152.021	151.254
CFI	.725	.812	.782	.792	.728	.812	.786	.791	.730	.812	.786	.792	.920	.931	.913	.911	.944	.932	.912	.918
RMSEA	.123	.114	.119	.120	.123	.115	.119	.121	.123	.115	.119	.121	.068	.071	.077	.080	.058	.071	.079	.078
SRMR	.079	.070	.073	.074	.078	.070	.073	.074	.078	.070	.073	.074	.056	.051	.053	.054	.054	.051	.053	.054
Average Across Subscales																				
Chi-Squared	344.663	315.737	286.911	324.830	328.820	314.359	279.927	321.033	327.768	313.685	277.002	320.736	159.440	127.793	131.284	133.460	115.164	120.563	121.310	128.690
CFI	.773	.802	.802	.792	.783	.802	.807	.794	.785	.803	.809	.794	.912	.941	.932	.937	.945	.946	.939	.940
RMSEA	.125	.121	.114	.122	.123	.123	.113	.123	.123	.122	.113	.123	.079	.067	.068	.069	.061	.065	.065	.068
SRMR	.075	.074	.072	.077	.073	.073	.071	.077	.073	.073	.070	.077	.055	.046	.049	.051	.050	.045	.047	.050

Note. L=Likert, E=Expanded, ISF=Item-Specific-F, ISL=Item-Specific-L. See Figure 1 for specifications of Models 1–5. The asterisk beside a chi-squared value indicates that the model fit yielded a warning message regarding obtaining an improper solution (a.k.a., Heywood case). The light gray shade indicates that the fit measure values passed the cutoffs for indicating an acceptable fit (i.e., CFI > .9, RMSEA < .08, SRMR < 0.08); the dark gray shade indicates that the fit measure values passed the cutoffs for indicating very good fit (i.e., CFI > .95, RMSEA < .05, SRMR < 0.05).

Table 4
$RMSEA_D$ Values for Comparing CFA Models in Study 1

Subscales	Likert	Expanded	Item-Specific-Full	Item-Specific-Light
			Models 1 vs. 2	
Extraversion	.196	.079	.143	.124
Agreeableness	.142	.000	.074	.158
Conscientiousness	.208	.031	.136	.000
Negative Emotionality	.327	.000	.177	.086
Openness	.099	.028	.121	.000
Average	.194	0.028	.130	.074
			Models 1 vs. 3	
Extraversion	.166	.101	.225	.144
Agreeableness	.203	0.000	.021	.144
Conscientiousness	.255	.112	.217	.000
Negative Emotionality	.294	.000	.155	.097
Openness	.118	.000	.113	.000
Average	.207	.043	.146	.077
			Models 1 vs. 4	
Extraversion	.526	.503	.506	.517
Agreeableness	.280	.471	.372	.376
Conscientiousness	.432	.370	.319	.477
Negative Emotionality	.401	.425	.349	.398
Openness	.436	.384	.392	.384
Average	.415	.431	.387	.431
			Models 3 vs. 5	
Extraversion	.549	.499	.492	.510
Agreeableness	.303	.477	.371	.367
Conscientiousness	.484	.395	.343	.482
Negative Emotionality	.438	.428	.354	.401
Openness	.455	.388	.385	.397
Average	.446	.437	.389	.432
			Models 4 vs. 5	
Extraversion	.317	.000	.095	0.000
Agreeableness	.287	.123	.000	.002
Conscientiousness	.457	.265	.308	.111
Negative Emotionality	.424	.073	.191	.128
Openness	.255	.078	.000	.173
Average	.348	.108	.119	.083

Note: See Figure 1 for specifications of Models 1–5. The smaller the $RMSEA_D$ value is, the more equivalent the models are. $RMSEA_D$ values less than .05 are shaded in dark gray to indicate high equivalence, those between .05 and .08 in light gray to indicate acceptable equivalence, and values greater than .08 are unshaded to indicate non-equivalence.

Table 5

Factor Loadings for Agreeableness and Open-Mindedness for Model 4 in Studies 1 and 2

	Study 1				Study 2			
Subscales and Facets	**L**	**E**	**ISF**	**ISL**	**L**	**E**	**ISF**	**ISL**
Agreeableness								
Facet 1: Compassion								
Item 2 (PW)	.655	.709	.682	.747	.703	.738	.804	.774
Item 17 (NW)	.167	.718	.734	.745	.091	.665	.696	.744
Item 32 (PW)	.537	.575	.446	.400	.454	.609	.410	.368
Item 47 (NW)	.732	.764	.780	.678	.521	.780	.703	.827
Facet 2: Respectfulness								
Item 7 (PW)	.722	.716	.654	.623	.589	.748	.593	.564
Item 22 (NW)	.347	.504	.365	.410	.497	.418	.542	.409
Item 37 (NW)	.454	.698	.554	.684	.545	.672	.650	.736
Item 52 (PW)	.721	.777	.644	.682	.568	.719	.544	.569
Facet 3: Trust								
Item 12 (NW)	.611	.483	.381	.351	.389	.346	.355	.321
Item 27 (PW)	.605	.542	.491	.600	.647	.553	.521	.583
Item 42 (NW)	.441	.630	.648	.483	.209	.508	.521	.440
Item 57 (PW)	.646	.694	.789	.768	.535	.748	.767	.704
Average Loading Size	.553	.651	.597	.594	.479	.625	.592	.587
Correlation Between Facets 1 and 2	.804	.680	.717	.707	.655	.755	.638	.565
Correlation Between Facets 1 and 3	.644	.438	.538	.518	.683	.544	.640	.651
Correlation Between Facets 2 and 3	.594	.457	.592	.483	.745	.461	.486	.518
Open-Mindedness								
Facet 1: Aesthetic Sensitivity								
Item 5 (NW)	.199	.721	.644	.709	.031	.754	.643	.703
Item 20 (PW)	.815	.809	.813	.867	-.816	.757	.836	.763
Item 35 (PW)	.761	.687	.704	.704	-.835	.732	.696	.616
Item 50 (NW)	.588	.538	.601	.624	-.552	.579	.620	.614
Facet 2: Intellectual Curiosity								
Item 10 (PW)	.531	.635	.590	.569	.531	.551	.549	.498
Item 25 (NW)	.601	.583	.682	.605	.590	.507	.621	.538
Item 40 (PW)	.585	.632	.597	.515	.535	.559	.452	.406
Item 55 (NW)	.671	.716	.748	.758	.547	.641	.654	.672
Facet 3: Creative Imagination								
Item 15 (PW)	.632	.606	.552	.532	.545	.581	.594	.362
Item 30 (NW)	.696	.758	.705	.741	.459	.717	.685	.844
Item 45 (NW)	.533	.423	.385	.412	.430	.364	.304	.305
Item 60 (PW)	.731	.731	.688	.664	.660	.648	.698	.632
Average Loading Size	.613	.656	.654	.649	.205	.626	.625	.589
Correlation Between Facets 1 and 2	.618	.683	.681	.666	-.732	.526	.577	.625
Correlation Between Facets 1 and 3	.434	.634	.611	.640	-.546	.544	.548	.704
Correlation Between Facets 2 and 3	.652	.755	.650	.811	.628	.651	.660	.673

Note: L=Likert, E=Expanded, ISF=Item-Specific-Full, ISL=Item-Specific-Light, PW=positively-worded in the original Likert version, NW=Negatively-worded in the original Likert version. Red cells indicate loadings and factor correlations that were very low or unusual.

Table 6

Composite Reliability Coefficients in Studies 1 and 2

	Study 1				Study 2			
Subscales and Facets	**L**	**E**	**ISF**	**ISL**	**L**	**E**	**ISF**	**ISL**
Extraversion								
Sociability Facet Reliability	.851	.848	.842	.802	.798	.862	.860	.855
Assertiveness Facet Reliability	.822	.784	.682	.800	.738	.711	.734	.729
Energy-Level Facet Reliability	.728	.716	.731	.719	.607	.687	.636	.655
Average Facet Reliability	.801	.783	.752	.774	.714	.753	.743	.746
Total Subscale Reliability	.886	.882	.845	.869	.825	.874	.857	.857
Agreeableness								
Compassion Facet Reliability	.524	.790	.761	.744	.382	.789	.743	.771
Respectfulness Facet Reliability	.539	.748	.584	.656	.587	.700	.662	.645
Trust Facet Reliability	.663	.683	.684	.633	.492	.638	.635	.607
Average Facet Reliability	.575	.740	.676	.678	.487	.709	.680	.674
Total Subscale Reliability	.751	.855	.820	.813	.694	.838	.825	0.815
Conscientiousness								
Organization Facet Reliability	.815	.847	.866	.881	.782	.808	0.826	.842
Productiveness Facet Reliability	.752	.720	.657	.691	.711	.705	.700	.720
Responsibility Facet Reliability	.673	.712	.595	.704	.621	.569	.654	.681
Average Facet Reliability	.747	.760	.706	.759	.705	.694	.727	.748
Total Subscale Reliability	.853	.887	.859	.876	.847	.847	.849	.864
Negative Emotionality								
Anxiety Facet Reliability	.800	.829	.806	.794	.768	.825	.820	.761
Depression Facet Reliability	.840	.756	.746	.764	.803	.762	.716	.695
Emotional-Volatility Facet Reliability	.806	.810	.708	.746	.797	.804	.702	.681
Average Facet Reliability	.816	.798	.753	.768	.789	.797	.746	.712
Total Subscale Reliability	.910	.895	.875	.856	.871	.891	.859	.830
Open-Mindedness								
Aesthetic-Sensitivity Facet Reliability	.674	.781	.790	.821	.616	.787	.797	.770
Intellectual-Curiosity Facet Reliability	.690	.733	.753	.708	.635	.646	.658	.611
Creative-Imagination Facet Reliability	.744	.727	.684	.675	.608	.669	.666	.652
Average Facet Reliability	.703	.747	.742	.735	.619	.701	.707	.678
Total Subscale Reliability	.799	.869	.859	.869	.739	.839	.831	.831
Average Reliability Across Subscales								
Average Facet Reliability	.728	.766	.726	.743	.663	.731	.721	.712
Average Total Reliability	.840	.878	.852	.857	.795	.858	.844	.839

Note: L=Likert, E=Expanded, ISF=Item-Specific-Full, ISL=Item-Specific-Light. These reliability coefficients were computed based on Model 4 in Figure 1. Coefficients greater than .80 are shaded in dark gray to indicate very good reliability, those between .70 and .79 in light gray for acceptable reliability, and those less than .70 are unshaded, indicating low reliability.

Table 7
Validity Correlations in Study 1

	Likert	Expanded	Item-Specific-Full	Item-Specific-Light
Extraversion				
Self-Other Agreement	.680***	.560***	.411**	.508***
Criterion Measures:				
Peer Acceptance	.561***	.448***	.426***	.451***
Peer Status	.509***	.457***	.436***	.463***
Dating Variety	.226***	.132	.192**	.221***
Party Attendance	***.338******	*.124*	.268***	**.081**
Average Correlation with Criterion Variables	.409	.290	.331	.304
Agreeableness				
Self-Other Agreement	.269	.504***	.507***	.471***
Criterion Measure:				
Volunteerism	-.075	.129	.025	.064
Conscientiousness				
Self-Other Agreement	.570***	.499***	.498***	.385***
Criterion Measures:				
Academic Performance	.206**	.253***	.213***	.197**
Risky Physical Activity	.157*	.149*	.118	.153*
Criminal Behaviour	-.188**	-.166*	-.091	-.159*
Substance Abuse	-.075	-.069	.042	-.098
Average Correlation with Criterion Variables	.157	.159	.116	.152
Negative Emotionality				
Self-Other Agreement	.474***	.467***	.611***	.469***
Criterion Measures:				
Pathological Anxiety	.240***	.338***	.301***	.351***
Pathological Depression	.324***	.262***	.225***	.249***
Affectual Solidarity with Parents	-.033	-.074	-.028	.126
Average Correlation with Criterion Variables	.199	.225	.185	.242
Open-Mindness				
Self-Other Agreement	.479***	.381***	.512***	.513***
Criterion Measures:				
Cultural Activity	.110	.179*	**.287*****	**.008**
Playing Music Instrument	.218***	.139	.233***	.130
Average Correlation with Criterion Variables	.164	.159	.260	.069
Average Across Subscales				
Self-Other Agreement	.494	.482	.508	.484
Correlation with Criterion Variables	.233	.208	.206	.197

Note: *= $p < .05$; ** = $p < .01$; *** = $p < .001$. Except for self-other agreement correlations, all other correlations were partial-correlations controlling for the gender, age, and ethnicity of the participants. For each validity measure, a pair of bolded values indicates that the difference between the correlations is significant at $p < .01$; a pair of italicized values indicates that the difference is significant at $p < .05$. The average correlations were computed with absolute correlation values.

Table 8

Fit Measures for CFA Models 4 and 5 for Study 2

	Model 4 (df = 51)				Model 5 (df = 50)				$RMSEA_D$ (Models 4 vs.5)			
Fit Measure	**L**	**E**	**ISF**	**ISL**	**L**	**E**	**ISF**	**ISL**	**L**	**E**	**ISF**	**ISL**
Extraversion												
Chi-Squared	149.537	134.773	108.881	126.412	132.069	134.710	108.163*	122.076*				
CFI	.915	.942	.959	.944	.929	.942	.959	.946				
RMSEA	.073	.068	.056	.064	.067	.069	.056	.064	.213	.000	.000	.096
SRMR	.058	.052	.041	.048	.057	.052	.041	.047				
Agreeableness												
Chi-Squared	142.656	147.212	172.867	104.760	108.482	146.400*	167.881	99.065				
CFI	.848	.918	.881	.946	.903	.918	.885	.950				
RMSEA	.070	.073	.080	.054	.057	.074	.080	.052	.302	.000	.105	.114
SRMR	.057	.055	.061	.045	.051	.055	.062	.044				
Conscientiousness												
Chi-Squared	162.791	144.177	172.347	128.856	105.855	141.840	140.122	117.695				
CFI	.892	.911	.903	.939	.947	.912	.928	.946				
RMSEA	.078	.072	.080	.066	.055	.072	.070	.062	.392	.061	.292	.167
SRMR	.057	.050	.052	.046	.045	.050	.045	.044				
Negative Emotionality												
Chi-Squared	205.104	140.035	109.101	135.138	154.749	119.487	107.668	124.424				
CFI	.913	.950	.958	.937	.941	.961	.959	.944				
RMSEA	.091	.070	.056	.068	.076	.063	.056	.064	.368	.235	.034	.164
SRMR	.060	.044	.046	.053	.051	.040	.045	.049				
Open-Mindedness												
Chi-Squared	126.807	182.896	193.004	212.957	101.486	182.506	192.523	192.973				
CFI	.915	.878	.875	.849	.942	.878	.875	.867				
RMSEA	.064	.086	.087	.094	.053	.087	.088	.089	.258	.000	.000	.229
SRMR	.056	.060	.063	.061	.052	.060	.063	.059				
Average Across Subscales												
Chi-Squared	165.956	161.465	151.053	174.048	128.118	150.997	150.095	158.699				
CFI	.914	.914	.917	.893	.941	.919	.917	.905				
RMSEA	.077	.078	.072	.080	.065	.075	.072	.077	.307	.059	.086	.154
SRMR	.058	.052	.055	.057	.052	.050	.054	.054				

Note. L=Likert, E=Expanded, ISF=Item-Specific-Full, ISL=Item-Specific-Light. For CFI, RMSEA, and SRMR, The light gray shade indicates that the fit measure values passed the cutoffs for indicating an acceptable fit (i.e., CFI > .9, RMSEA < .08, SRMR < .08); the dark gray shade indicates that the fit measure values passed the cutoffs for indicating very good fit (i.e., CFI > .95, RMSEA < .05, SRMR < .05). For $RMSEA_D$ values, the light gray shade indicates acceptable equivalence (i.e., .05<$RMSEA_D$ <.08); the dark gray shade indicates high equivalence (i.e., $RMSEA_D$ <.05).

Table 9

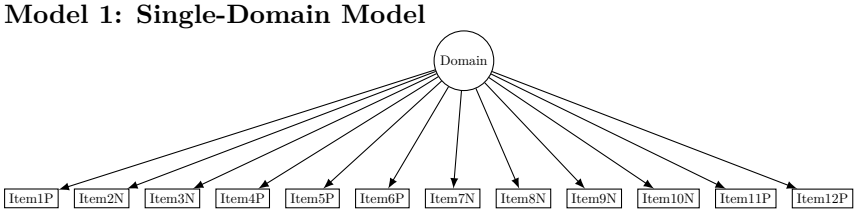
Correlations Between the Careless Responding Measures in Study 2

	Time	Consistency	Average Long-String	Max Long-String	Distance (E)	Distance (A)	Distance (C)	Distance (N)	Distance (O)
Time	1.000								
Consistency	.114***	1.000							
Average Long-String	-.086	-.188***	1.000						
Max Long-String	-.099*	-.238***	.832***	1.000					
Distance (E)	-.114***	-.068	-.079*	-.072	1.000				
Distance (A)	-.152***	-.178***	-.020	.010	.465***	1.000			
Distance (C)	-.077	-.085	-.071	-.067	.522***	.470***	1.000		
Distance (N)	-.095*	-.080	-.086*	-.077	.463***	.364***	.428***	1.000	
Distance (O)	-.054	-.108**	-.079	-.068	.419***	.387***	.416***	.404***	1.000

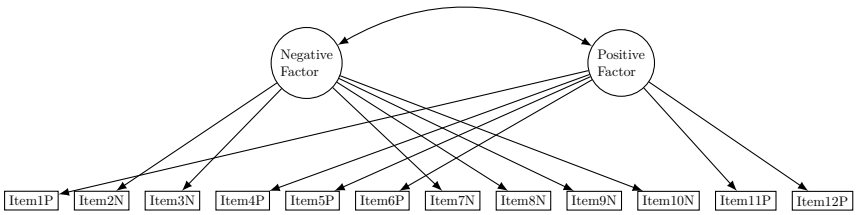
Note. *= $p < .05$; ** = $p < .01$; *** = $p < .001$. Time=Completion Time; Consistency=Even-Odd Consistency; Distance=Mahalanobis Distance; E=Extraversion; A=Agreeableness; C=Conscientiousness; N=Negative Emotionality; O=Open-Mindedness. To control for family-wise Type-I error in the 36 pairwise correlations, the p-values were adjusted using the Bonferroni correction by multiplying each original p-value by 36.

Figure 1
CFA Models

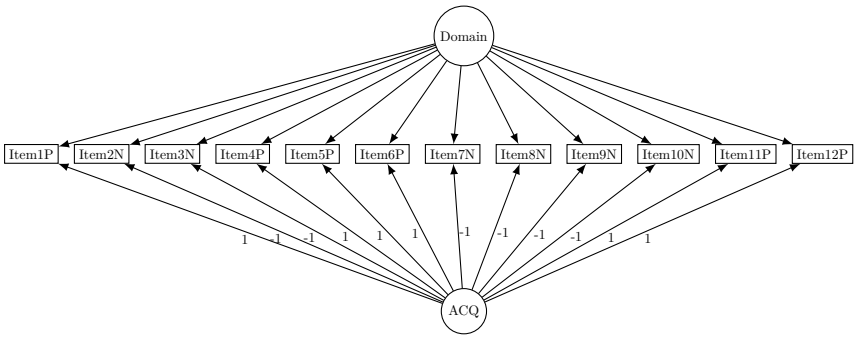
Model 1: Single-Domain Model



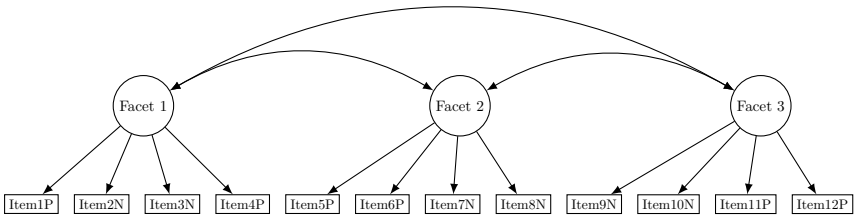
Model 2: Positive-and-Negative-Factors Model



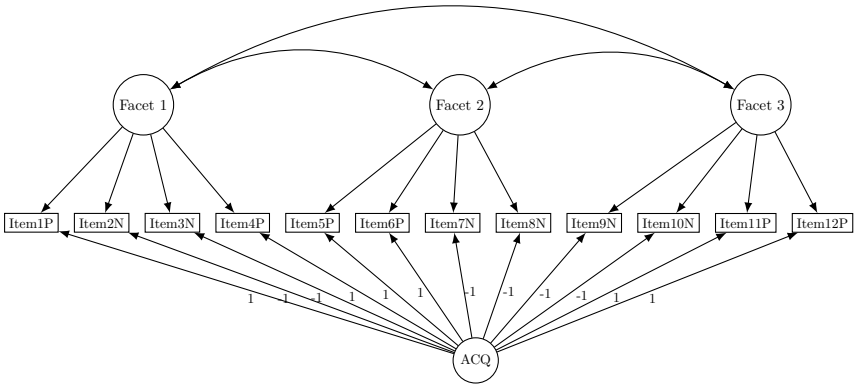
Model 3: Single-Domain-Plus-Acquiescence Model



Model 4: Three-Facets Model



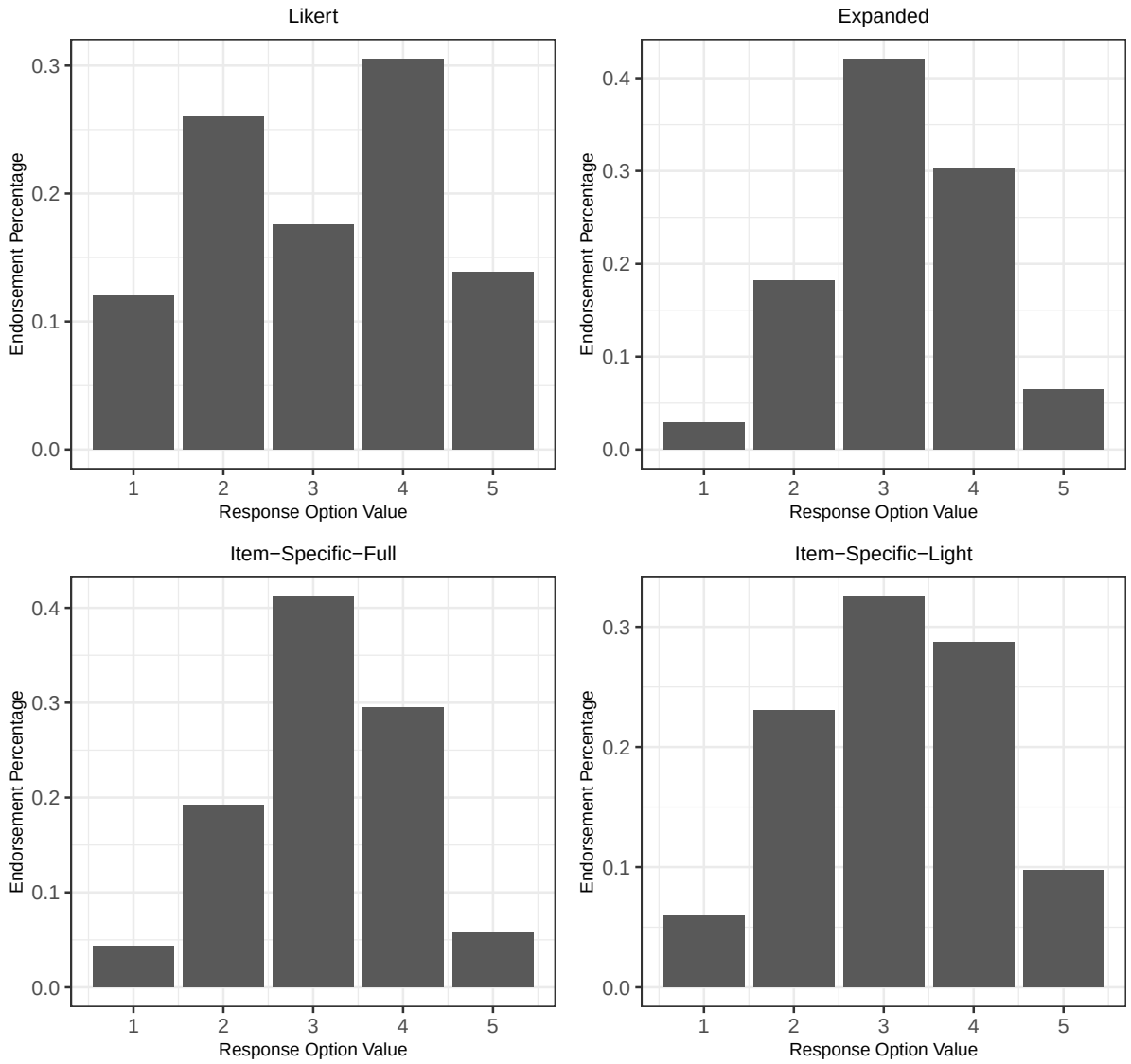
Model 5: Three-Facets-Plus-Acquiescence Model



Note: P=Positively-Worded (PW) Item for the Likert Version; N=Negatively-Worded (NW) Item for the Likert format; ACQ=Acquiescence Bias. For the Likert format, the negatively-worded items were reverse-scored before the factor analyses.

Figure 2

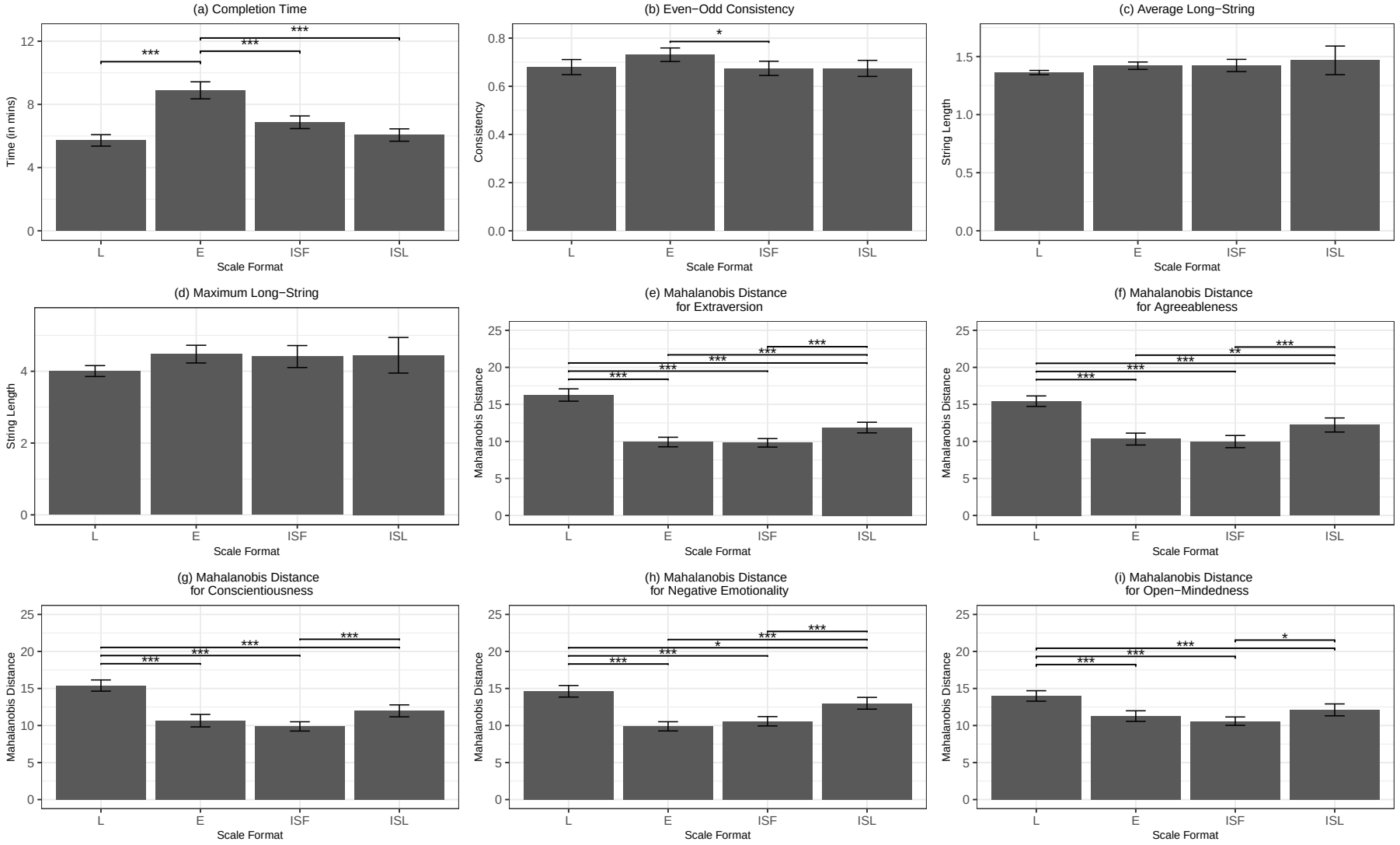
Distribution of Response Option Endorsements For the Extraversion Subscale Across Formats in Study 1



Note. For each format, the endorsement percentages were averaged across all extraversion items in the BFI-2. For the Likert format, the response option values are those after the negative-worded items were reverse-scored; therefore, across all four formats, a higher response option value indicates higher extraversion.

Figure 3

Careless Measures Across Scale Formats in Study 2



Note. *= $p < .05$; ** = $p < .01$; *** = $p < .001$. The vertical bars indicate 95% confidence intervals of the means. The significance of the p-values was corrected using Bonferroni corrections. For plots (a)-(i), the F-values were 42.47 ($p < .001$), 3.18 ($p = .02$), 1.50 ($p = 0.21$) , 1.78 ($p = 0.15$) 75.98 ($p < .001$), 38.49 ($p < .001$), 42.36 ($p < .001$), 38.38 ($p < .001$), and 18.34 ($p < .001$), respectively.