Soto, C. J., & John, O. P. (in press). Optimizing the length, width, and balance of a personality scale: How do internal characteristics affect external validity? *Psychological Assessment.*

Optimizing the Length, Width, and Balance of a Personality Scale: How Do Internal Characteristics Affect External Validity?

Christopher J. Soto

Colby College

Oliver P. John

University of California, Berkeley

Author Note

Christopher J. Soto, Department of Psychology, Colby College; Oliver P. John, Department of Psychology, University of California, Berkeley.

Supporting materials for the present research, including the complete correlation matrices, are available at https://osf.io/nxakd. This research was supported by faculty research grants from Colby College to Christopher J. Soto, and from the University of California, Berkeley to Oliver P. John. Correspondence concerning this article should be addressed to Christopher J. Soto, Department of Psychology, Colby College, 5550 Mayflower Hill, Waterville, ME 04901. Email: christopher.soto@colby.edu.

Abstract

How well can scores on a personality scale predict criteria such as behaviors and life outcomes? This question concerns external validity, which is a core aspect of personality assessment. The present research was conducted to examine how external validity is influenced by a trait scale's internal characteristics, such as its length (number of items), width (breadth of content), and balance (between positively and negatively keyed items). Participants completed the Big Five Inventory–2 (BFI-2), and were also assessed on a set of self-reported and peer-reported validity criteria. We used the BFI-2 item pool to construct multiple versions, or iterations, of each Big Five trait scale that varied in terms of length, width, and balance. We then identified systematic effects of these internal scale characteristics on external validity associations. Regarding length, we find that longer trait scales tend to have greater validity, with a scale length "sweet spot" of approximately six to nine items. Regarding width, we find that broad trait scales tend to have slightly stronger, and much more consistent, associations with external validity criteria than do narrow scales; broad scales thus represent relatively safe bets for personality assessment, whereas narrow scales carry greater risks but offer potentially greater rewards. Regarding balance, we find that associations between imbalanced trait and criterion scales can be substantially inflated or suppressed by acquiescent responding; trait scales that include an equal number of positively and negatively keyed items can minimize such acquiescence bias. We conclude by translating these findings into practical advice regarding psychological assessment.

Keywords: five factor personality model; personality assessment; personality traits; psychometrics; test validity

Public Significance Statement

This study finds that the internal features of a personality questionnaire scale (such as its number of items, breadth of content, and balance between positively and negatively keyed items) can substantially affect the capacity of personality trait scores to predict external, real-world criteria, such as behaviors and life outcomes.

Optimizing the Length, Width, and Balance of a Personality Scale: How Do Internal Characteristics Affect External Validity?

Personality traits are most often assessed using scales that aggregate responses to individual questionnaire items into a simple linear composite. The development of a personality trait scale should be grounded in a conceptual understanding of the construct to be measured: How is the trait defined, and what are the key components of this definition (John & Soto, 2007; Loevinger, 1957; Messick, 1995)? However, a scale developer must also consider more pragmatic questions: Should the scale include many items, or only a few (cf. Gosling, Rentfrow, & Swann, 2003; McCrae & Costa, 2010)? Should these items assess a broad or narrow range of trait-relevant content (cf. Jonason & Webster, 2010; Jones & Paulhus, 2014)? Should the items be positively-keyed (with agreement indicating high standing on the trait), negatively-keyed, or a mix of both (cf. Janis & Field, 1959; Rosenberg, 1965)?

The present research was conducted to examine how these internal characteristics of a personality trait scale affect external validity: the capacity of scale scores to predict other measures and nontest criteria. We first derive hypotheses about how scale length, width, and balance may affect trait-criterion validity associations. We then test these hypotheses empirically, using the item pool of the Big Five Inventory–2 (BFI-2; Soto & John, 2017a) to operationalize scale length, width, and balance. Finally, we translate our findings into practical advice regarding personality scale development (i.e., constructing a new scale) and selection (i.e., choosing from among existing scales).

Scale Length: How Many Items Is Enough?

Both classical test theory (CTT) and item response theory (IRT) imply that, when it comes to psychological measurement, longer is better: all else being equal, scale scores

computed by aggregating a larger number of items should show greater validity. From the perspective of CTT, each additional item enhances reliability by increasing the extent to which test takers' observed scores reflect their true scores (Lord & Novick, 1968). From an IRT perspective, each additional item provides additional information about the trait being assessed (Lord, 1980).[1] From either perspective, the greater precision and reduced error afforded by each additional item implies that longer trait scales should generally show stronger validity associations with external criteria (John & Soto, 2007; Lord, 1980; Lord & Novick, 1968).

Both CTT and IRT also suggest that the positive relation between scale length and validity should be monotonic but not strictly linear. As the length of a scale increases, each additional item will provide a proportionally smaller boost to measurement precision, and thus to validity; adding one more item to a five-item scale will generally yield larger measurement gains than would adding one more item to a fifty-item scale. Moreover, each additional item entails costs in terms of increased administration time and respondent fatigue (Burisch, 1984).

Thus, our first research hypothesis is that scores from longer trait scales will generally demonstrate stronger external validity associations, but that as scale length increases there will be diminishing returns for each additional item. This hypothesis also raises our first practical question: Is there a "sweet spot" for scale length—a point that optimally balances efficiency with validity? In other words, how many items is enough for a typical personality scale?

Scale Width and the Variability of Trait-Criterion Associations

When measuring a particular personality trait, a researcher can choose to include items that collectively represent either a broad or narrow range of trait-relevant content. Scales that

[1] IRT approaches allow for the use of adaptive tests, in which the items presented over the course of a testing session depend on the test-taker's responses to previous items. Shorter adaptive tests can provide greater measurement precision than longer fixed tests; however, longer adaptive tests are generally more precise than shorter adaptive tests (Embretson, 1996).

sample broadly from a trait domain are said to have high bandwidth, whereas those that sample narrowly have high fidelity. This classic bandwidth-fidelity tradeoff has received considerable attention within personality and industrial-organization psychology (Cronbach & Gleser, 1965; John, Hampson, & Goldberg, 1991). In recent years, it has taken on even greater importance as interest in personality assessment has spread to economic and educational policymakers and stakeholders (OECD, 2015). In many policy-relevant contexts, only a few minutes can be devoted to measuring personality. Would this limited time be better spent administering a small number of broad or narrow trait scales?

Figure 1 illustrates some important implications of the bandwidth-fidelity tradeoff. Each figure panel includes a schematic representation of a personality trait (T), the universe of content within this trait's conceptual boundaries, and six questionnaire items (I1–I6) that assess specific components of the overall trait (often referred to as "facets" or "nuances" of the trait; Costa & McCrae, 1995; McCrae, 2015). In this figure, physical distance represents conceptual and empirical distance. For example, the content of item 1 is most similar to that of item 2, and most different from item 6. Similarly, the content of items 3 and 4 is quite central to the meaning of the overall trait, whereas items 1 and 6 are more peripheral. Figure 1a shows two relatively broad *trait scale iterations* (i.e., possible versions of a trait scale) that could be constructed from this item set (B1, B2). Each scale iteration includes three items, and neither includes any directly adjacent item pairs. Thus, each scale iteration includes a broad sample of item content that spans most of the trait's conceptual space. In contrast, Figure 1b shows two relatively narrow trait scale iterations (N1, N2). Each of these scales includes three adjacent items; therefore, each focuses on a rather narrow range of content within the overall trait domain.

How should scores from broad vs. narrow trait scales differ in their associations with external validity criteria? Previous research has produced mixed results. Some studies have concluded that narrow trait scales' specificity generally provides greater predictive power (e.g., Paunonen & Ashton, 2001; Paunonen, Rothstein, & Jackson, 1999). However, others have concluded that broad trait scales often out-predict narrow scales after accounting for model complexity (Grucza & Goldberg, 2007; Morey et al., 2007; Ones & Viswesvaran, 1996).

To help consider this issue, Figures 1a and 1b also include one example of a trait-relevant criterion (C). The degree of alignment between a trait scale and this criterion represents their expected degree of association (i.e., trait-criterion validity correlation). Thus, some trait scale iterations (N2, B2) include content that aligns more closely with the criterion, and would therefore show relatively strong trait-criterion associations. Other scale iterations (N1, B1) include less relevant content, and would therefore show weaker associations.

Comparing Figures 1a and 1b suggests that, on average, broad and narrow trait scales should show approximately equal associations with any particular external criterion: on average, the criterion should align equally well with broad scale iterations (B1, B2) and narrow scale iterations (N1, N2). However, Figure 1 also illustrates one important difference between the external validity of broad vs. narrow trait scales. Specifically, there is more *variability* to the degree of trait-criterion alignment between narrow scale iterations than between broad scale iterations. Because the content of each broad scale is distributed similarly across the overall trait space, different broad iterations should show similar associations with any particular external criterion; criterion C will correlate only a bit more strongly with trait scale B2 than with B1. In contrast, narrow scale iterations will sometimes focus on quite different facets of the overall trait

space, and therefore show rather distinct criterion associations; criterion C will correlate much more strongly with trait scale N2 than with N1.

Thus, our second hypothesis is that trait-criterion validity associations will tend to be similar in their average magnitude for broad vs. narrow trait scales, but will tend to be more consistent across broad trait scale iterations and more variable across narrow scale iterations. This hypothesis also raises our second practical question: Given any systematic differences in their validity associations, when should researchers administer broad vs. narrow trait scales?

Scale Balance and the Effects of Acquiescent Responding

Acquiescent responding is an individual's tendency to consistently agree (yea-saying) or disagree (nay-saying) with questionnaire items, regardless of their content (Cronbach, 1946; Jackson & Messick, 1958). When left uncontrolled, individual differences in acquiescence tend to positively bias inter-item correlations. That is, positive correlations (e.g., between two positively-keyed scale items) tend to be inflated, whereas negative correlations (e.g., between a positively-keyed item and a negatively-keyed item) tend to be suppressed (Rammstedt & Farmer, 2013; Soto, John, Gosling, & Potter, 2008; Winkler, Kanouse, & Ware, 1982).

Although acquiescence effects occur at the item level, they can compound as items are aggregated into scales. Scores on a fully balanced scale (i.e., one with an equal number of positively and negatively keyed items) should not be biased by acquiescence, because each respondent's tendency to acquiesce with the scale's positively-keyed items will be canceled out by their tendency to acquiesce with the negatively-keyed items. Thus, scores on a key-balanced scale will reflect the trait being measured while controlling for individual differences in acquiescence. However, scales with an imbalance of positively and negatively keyed items are susceptible to acquiescence bias, because observed scale scores will confound individual

differences in the trait being measured with individual differences in acquiescence. Specifically, yea-saying will tend to inflate scores on a scale with more positively-keyed items, and suppress scores on a scale with more negatively-keyed items; nay-saying will have the opposite effect (Soto, John, Gosling, & Potter, 2011; Ware, 1978). As the degree of scale imbalance increases, so should the influence of acquiescence bias on observed scores.

Moreover, because individual differences in acquiescent responding generalize across measures (Danner, Aichholzer, & Rammstedt, 2015; John, Naumann, & Soto, 2008), acquiescence variance may bias validity associations when the trait and criterion measures (a) both include an imbalance of positively and negatively keyed items, and (b) are both obtained using the same method (e.g., self-report). An observed trait-criterion association should be positively biased when the trait and criterion scales are imbalanced in the same direction (e.g., because both scales include more positively-keyed than negatively-keyed items); the association should be negatively biased when the two scales are imbalanced in opposite directions (because one scale includes more positively-keyed items while the other includes more negatively-keyed items). These biases may either inflate or suppress the trait-criterion association, depending on the expected sign of the association and whether the trait and criterion scales are imbalanced in the same direction. For example, acquiescence variance should inflate an expected positive association if the trait and criterion scales are imbalanced in the same direction, or supress this positive association if the two scales are imbalanced in opposite directions. The reverse should hold for an expected negative trait-criterion association.

Thus, our third hypothesis is that imbalanced personality trait scales will show biased validity associations with imbalanced criterion measures. This hypothesis raises our final practical question: Are the effects of acquiescence bias on trait-criterion associations large

enough that researchers should consider key balance when developing or selecting trait scales, or small enough that they may be safely ignored?

Overview of the Present Research

In sum, the present research was conducted to (a) test hypotheses about how three key internal characteristics—length, width, and balance—affect the external validity of personality trait scales, and (b) translate these findings into practical advice regarding personality assessment. To address these goals, we analyzed data from a sample of participants who self-reported their personality traits using the BFI-2, and were also assessed using self-reported and peer-reported criterion measures. Specifically, for each Big Five trait and each internal scale characteristic (length, width, or balance), we first constructed a set of trait scale iterations from the BFI-2 item pool that varied the target scale characteristic while controlling the other two characteristics. Next, we computed the correlation of each individual trait scale iteration with each external criterion. We then aggregated these trait-criterion associations to identify systematic effects of scale length, width, and balance on external validity. Finally, we examined these aggregated effects to test our key hypotheses and translate our findings into practical advice regarding personality assessment.

Method

Participants and Procedure

Participants were 536 students enrolled in undergraduate psychology courses at the University of California, Berkeley, who completed a number of psychological measures online in exchange for partial fulfillment of a course requirement.[2] Approximately 67% were women, 32%

[2] Responses were screened for data quality on the basis of (a) completion (i.e., responding to at least 95% of items), (b) response variability (i.e., a within-person standard deviation of at least 0.50 across the 60 BFI-2 items), and (c) uniqueness (i.e., repeat participation was prohibited). Cases that failed any of these three quality checks were deleted.

were men, and 1% did not report gender; the median age was 21 years old (*M* = 21.80, *SD* = 3.84). Regarding ethnicity, 51% described themselves as Asian/Asian-American, 25% as white/Caucasian, 11% as Hispanic/Latino, 2% as black/African-American, and 8% as another ethnicity, with 3% not reporting ethnicity. Most participants (447) were also included in a previous study conducted to validate the BFI-2 (Soto & John, 2017a, Study 3); however, none were used to select items for the BFI-2. Therefore, the present data should provide unbiased estimates of trait-criterion validity associations. This research was approved by the Committee for the Protection of Human Subjects at the University of California, Berkeley.

Measures

All of the present participants completed the BFI-2, and many were also assessed on a set of 20 self-reported and peer-reported validity criteria (see also Soto & John, 2017a).

The Big Five Inventory-2. The BFI-2 (Soto & John, 2017a) is a hierarchical measure of the Big Five personality traits and 15 more-specific facets: Extraversion (with facets of Sociability, Assertiveness, and Energy Level), Agreeableness (Compassion, Respectfulness, and Trust), Conscientiousness (Organization, Productiveness, and Responsibility), Negative Emotionality (Anxiety, Depression, and Emotional Volatility), and Open-Mindedness (Intellectual Curiosity, Aesthetic Sensitivity, and Creative Imagination). Its 60 items are short, descriptive phrases with the common item stem “I am someone who...” (e.g., “Is outgoing, sociable,” “Tends to be quiet”). Respondents rate each item on a 5-point scale ranging from *disagree strongly* to *agree strongly*. In the present sample, alpha reliabilities averaged .86 (range = .82–.89) for the Big Five trait scales and .75 (range = .63–.85) for the facet scales.

Self-reported behavioral criteria. Approximately two weeks after completing the BFI-2, 524 participants described their behavior during the previous six months using a set of 80

items (Bardi & Schwartz, 2003). Each item was rated on a 5-point frequency scale ranging from *never* to *all the time*. Following this measure's standard scoring procedure, each respondent's complete set of ratings was within-person centered to remove individual differences in acquiescent responding. The centered items were then aggregated into scales corresponding with the 10 values of the Schwartz values circumplex: conformity, tradition, benevolence, power, universalism, hedonism, security, stimulation, achievement, and self-direction. As in previous research using this measure, the alpha reliabilities of these scales varied considerably, and some were rather low (M = .48, range = .28–.71). We therefore note that some of these scales' trait-criterion validity associations may be substantially attenuated by unreliability.

Self-reported well-being criteria. Approximately two weeks before completing the BFI-2, 265 participants completed the 14-item Psychological Well-Being Scales (Ryff, Lee, Essex, & Schmutte, 1994), which assess six aspects of psychological functioning: positive relations with others, purpose in life, environmental mastery, self-acceptance, autonomy, and personal growth. These scales include a total of 84 items that respondents rate on a 5-point agreement scale. In the present sample, the scales' alpha reliabilities averaged .87 (range = .84–.91).

Peer-reported criteria. Approximately two months after completing the BFI-2, 232 participants were rated by a knowledgeable peer. Most peers were friends (62%) or romantic partners (31%). Using a 5-point agreement scale, each peer rated the target participant on a set of items assessing four criteria: social connectedness (4 items), likability (2 items), stress resistance (5 items), and positive affect (2 items). In the present sample, alpha reliabilities were .75 for social connectedness, .76 for likability, .79 for stress resistance, and .67 for positive affect.

Analytic Strategy

To examine the effects of length, width, and balance on external validity, we first defined a set of constraints that varied one internal scale characteristic while controlling the other two. For each Big Five trait, we then identified all subsets of BFI-2 items that satisfied these constraints, and scored each item subset as an iteration of the trait scale. Next, we computed the correlation of each trait scale iteration with the complete set of self-reported and peer-reported criteria. (The full correlation matrices are available at https://osf.io/nxakd.) Finally, we aggregated these trait-criterion correlations across the Big Five traits and across the criteria to identity systematic effects of scale length, width, and balance on validity.

Evaluating the statistical significance of these effects was complicated by the complex pattern of dependencies between the individual trait-criterion correlations (due to varying degrees of item overlap between the trait scale iterations). However, aggregating these correlations within each combination of a Big Five trait and a criterion measure (e.g., the mean correlation of Extraversion scale iterations with power-seeking behavior) would eliminate these complex dependencies. Therefore, within each trait-criterion combination and at each level of scale length, width, or balance, we aggregated the individual trait-criterion correlations across scale iterations. We then conducted sign tests comparing these aggregated correlations between levels of scale length, width, or balance (e.g., comparing the absolute mean trait-criterion correlations for one-item vs. two-item scale iterations) to test our key hypotheses.[3]

Results and Discussion

Scale Length

[3] We conducted sign tests to avoid imposing assumptions about the distribution of aggregated trait-criterion correlations. However, additional analyses revealed that sign tests and paired-samples *t*-tests yielded identical conclusions.

We expected that scores from longer trait scales would tend to show greater external validity, but that there would be diminishing returns for each additional increase in scale length. To test this hypothesis, for each Big Five trait we constructed scale iterations that varied in length from one to twelve items, with these items distributed across the trait's three facet subscales as evenly as possible. For example, each six-item Extraversion scale iteration included two Sociability items, two Assertiveness items, and two Energy Level items. The balance between positively and negatively keyed items was allowed to vary across individual scale iterations. However, because the BFI-2 item pool includes an equal number of positively and negatively keyed items for each Big Five trait and facet, scale balance was controlled in the aggregate (i.e., when aggregated across the complete set of trait scale iterations).

Within these constraints, we identified 1 twelve-item, 12 eleven-item, 48 ten-item, 64 nine-item, 288 eight-item, 432 seven-item, 216 six-item, 432 five-item, 288 four-item, 64 three-item, 48 two-item, and 12 one-item scale iterations per Big Five trait, with a total of 9,525 scale iterations across the five traits. We then scored each trait scale iteration and correlated these scores with each of the 20 self-reported and peer-reported criteria, yielding 190,500 trait-criterion validity correlations. Next, for each level of scale length, we computed the mean absolute validity correlation (using Fisher's transformation) across all 20 criteria, as well as within each Big Five trait and each criterion category (self-reported behavior, self-reported well-being, and peer-reported criteria). These mean correlations are presented in Table 1.[4] Finally, we rescaled the mean validity correlations in terms of relative predictive power (i.e., the mean

[4] Note that aggregating correlations across traits and criteria meant that our overall estimates of the average validity correlations were based on the mean of 100 trait-criterion correlations (5 traits × 20 criteria) per scale iteration. Thus, even though there was only one twelve-item scale iteration per Big Five trait, our overall estimate of the average validity correlation for a twelve-item trait scale was still based on the mean of 100 correlations.

proportion of criterion variance explained by a trait scale), with the twelve-item trait scales providing a baseline of 100% power. These predictive power ratios are shown in Figure 2.

Table 1 and Figure 2 indicate that, as hypothesized, longer scales showed greater external validity, with diminishing returns for each additional increase in scale length. This pattern was quite consistent across the Big Five traits (Figure 2a), and across the behavioral, well-being, and peer criteria (Figure 2b). Validity declined rather gradually as scale length decreased from twelve items to six items (89% relative predictive power), then more rapidly to three-item scales (73% power), and very rapidly to two-item (60% power) and single-item (39% power) scales. Sign tests indicated that each of these differences in mean predictive power was statistically significant. Specifically, for each set of 100 independent comparisons (5 traits × 20 criteria) between adjacent levels of scale length (e.g., one-item vs. two-item scales), the mean validity correlation was stronger for the greater scale length in at least 98 cases (all ps < .001).

These results support our first key hypothesis, and also address our first practical question: How long is long enough for a typical personality trait scale? The present findings suggest a scale length sweet spot of approximately six to nine items: validity decreased rapidly below this range, and increased only gradually above it.

Scale Width

We expected that scores from broad and narrow personality scales would show similar average external validity associations, but that these associations would be more consistent across broad scale iterations and more variable across narrow scale iterations. To operationalize scale width, for each Big Five trait we identified and scored all three-item scale iterations that either (a) included one item from each of the trait's three facet subscales (and therefore assessed a broad range of trait-relevant content), or (b) included three items from the same facet subscale

(and therefore assessed a narrow range of trait-relevant content). For example, we contrasted three-item Extraversion scales that included one Sociability item, one Assertiveness item, and one Energy Level item with scales that included three items from one of these three subscales. The balance between positively and negatively keyed items was allowed to vary across individual scale iterations, but was again controlled in the aggregate due to the balanced keying of the full BFI-2 item pool. These constraints yielded 64 broad and 12 narrow scale iterations per Big Five trait, with a total of 380 iterations and 7,600 validity associations across the five traits.

To examine the effect of trait scale width on average validity, we computed the mean absolute trait-criterion correlation separately for broad and narrow trait scales, both across all 7,600 validity correlations and within each trait and each criterion category. We then rescaled these mean correlations in terms of relative predictive power, with narrow scales providing the baseline of 100% power. These results are presented in Table 2 and Figure 3. On average, broad trait scales had approximately 15% greater predictive power than did narrow scales, and this difference was quite consistent across traits (Figure 3a) and criterion categories (Figure 3b). A sign test indicated that the overall difference in predictive power was statistically significant: out of 100 comparisons (5 traits × 20 criteria) between broad vs. narrow trait scale iterations, the mean validity correlation was stronger for the broad iterations in 91 cases ($p < .001$).

To examine the effect of scale width on the consistency vs. variability of validity associations, for each trait-criterion correlation we first computed the variance of this correlation separately across the broad trait scale iterations and across the narrow iterations. For example, for the association between Extraversion and power-seeking behavior, we computed the 64 possible correlations of a broad Extraversion scale iteration with the power-seeking behavior criterion, and computed the variance of these 64 correlations; we then computed the 12 possible

correlations of a narrow Extraversion scale iteration with this same criterion, and computed the variance of these 12 correlations. Next, we averaged these variances separately for broad and narrow trait scales, both across all 100 trait-criterion combinations and within each trait and criterion category. Finally, we converted the mean variances to standard deviations (presented in Table 2) and variance ratios (presented in Figure 4), with broad scales providing the baseline of 100% variability. Scores from broad trait scales showed substantially more-consistent validity associations than did scores from narrow scales. Specifically, across scale iterations, narrow scales showed about three-and-a-half times as much variability in their trait-criterion correlations as did broad scales, although the magnitude of this effect varied across the Big Five traits (Figure 4a) and three criterion categories (Figure 4b).[5] A sign test indicated that the overall difference in variability was statistically significant: out of 100 independent comparisons between the variance of a trait-criterion correlation across broad vs. narrow trait scale iterations, the variance was greater across the narrow scale iterations in 84 cases ($p < .001$).

Taken together, these results partially support our second key hypothesis: broad trait scales tended to show much more consistent, but also somewhat stronger, external validity associations than did narrow scales. These results also help answer our second practical question: When should researchers administer broad vs. narrow trait scales? The present findings indicate that researchers should assess personality using broad scales when they wish to make a relatively safe bet, and should use narrow scales when they are willing to take on a greater degree of risk in exchange for potentially greater rewards: for a given external criterion, most narrow trait scales

[5] Inspection of the individual trait-criterion correlations revealed that the especially dramatic effect of scale width on variability for Negative Emotionality reflects the fact that most of the well-being criteria correlated much more strongly with the Depression facet of Negative Emotionality than with the Anxiety and Emotional Volatility facets.

will yield weaker validity associations than would the typical broad scale, but a few narrow scales will out-predict their broader counterparts.

Scale Balance

We expected that, due to individual differences in acquiescent responding, trait-criterion validity associations would tend to be biased when the trait and criterion measures both include an imbalance of positively and negatively keyed items, and are both obtained using the same method (e.g., self-report). Specifically, we expected that trait-criterion associations would be positively biased when the trait and criterion scales are both imbalanced in the same direction, and negatively biased when the two scales are imbalanced in opposite directions.

To operationalize scale balance, for each Big Five trait we identified and scored all six-item scale iterations that included two items from each of the trait's three facet subscales, and categorized each scale iteration as including either six, five, four, three, two, one, or zero positively-keyed items. This procedure varied scale balance while holding length and width constant. Each Big Five trait yielded 1 scale iteration with six positively-keyed items, 12 with five positively-keyed items, 51 with four positively-keyed items, 88 with three positively-keyed items, 51 with two positively-keyed items, 12 with one positively-keyed item, and 1 with no positively-keyed items, with a total of 1,080 scale iterations across the five traits.[6] Next, to create a set of imbalanced criterion measures, we separated each of the six psychological well-being scales into a positively-keyed subscale and a negatively-keyed subscale. We then correlated each trait scale iteration with the 12 imbalanced well-being subscales, yielding a total of 12,960 validity correlations. Next, we computed the mean absolute trait-criterion validity correlation for

[6] For these analyses we reversed the orientation of the BFI-2 Negative Emotionality scale, such that items indicating greater emotional stability were considered positively-keyed. This reversal yielded a uniformly positive trait-criterion correlation matrix, which was consistent with the other four Big Five traits and therefore simplified presentation of the results.

seven degrees of "matched" keying ranging from fully matched (e.g., both scales have only positively-keyed items) to fully mismatched (i.e., one scale has only positively-keyed items while the other has only negatively-keyed items). These mean correlations are presented in Table 3. Finally, we rescaled these mean correlations in terms of relative trait-criterion overlap (i.e., mean proportion of variance shared between the trait and criterion scales), with the balanced trait scales providing a baseline of 100% overlap; these ratios are presented in Figure 5.

As hypothesized, Table 3 and Figure 5 show that, relative to the baseline of key-balanced trait scales, validity correlations between trait and criterion scales imbalanced in the same direction were positively biased, whereas correlations between scales imbalanced in opposite directions were negatively biased. Moreover, the degree of bias increased with the degree of scale imbalance. Averaged across the Big Five traits (Figure 5a) and six well-being criteria (Figure 5b), fully matched keying inflated the degree of trait-criterion overlap by 11%, whereas fully mismatched keying suppressed overlap by 25%. Another way to view these results is that, compared with the baseline of fully mismatched scales, the average degree of trait-criterion overlap was 47% greater for fully matched scales.[7] Sign tests indicated that almost all of the increases in trait-criterion overlap from zero to six key-matched items were all statistically significant, although the size of these increases became smaller with each additional matched

[7] Table 3 and Figure 5 appear to suggest that mismatched keying introduces greater bias than does matched keying. However, it is important to note that the direction of this asymmetry is largely arbitrary. If the orientation of either the trait scales or the criterion scales were reversed, such that one would expect negative rather than positive trait-criterion correlations, then the asymmetrical effects of key balance on trait-criterion overlap would also be reversed. In this case, the average degree of trait-criterion overlap would be 25% smaller for fully *matched* scales than for balanced trait scales (because matched keying tends to suppress negative correlations), 11% greater for fully *mismatched* scales than for balanced trait scales (because mismatched keying tends to inflate negative correlations), and 47% greater for fully *mismatched* scales than for fully matched scales. Therefore, we propose that the present results are best interpreted as indicating that scale imbalance generally biases trait-criterion validity associations, rather than indicating that either matched or mismatched keying introduces greater bias.

item. Out of 60 comparisons (5 traits × 6 criteria × 2 keying directions) between adjacent levels of key-matching, the mean trait-criterion correlation was stronger in 52 of 60 cases for the increase from zero to one matched items ($p < .001$), 50 of 60 cases for the increase to two matched items ($p < .001$), 48 of 60 cases for the increase to three matched items ($p < .001$), 45 of 60 cases for the increase to four matched items ($p < .001$), 38 of 60 cases for the increase to five matched items ($p = .026$), and 35 of 60 cases for the increase to six matched items ($p = .123$).

These results support our third key hypothesis, and also address our final practical question: Are the biases introduced by imbalanced keying and acquiescence generally small enough that they may be safely ignored? We conclude that these biases are substantial enough that key balance should be an important consideration when developing or selecting personality and criterion measures.

General Discussion

The present findings support several conclusions about how the length, width, and balance of personality scales affect external validity. These findings have practical implications for personality assessment, and suggest promising directions for future research.

Scale Length: A Sweet Spot for Measurement Efficiency

Our first conclusion is that longer trait scales show stronger validity associations with external criteria, but that there are diminishing returns for each additional increase in scale length. Conceptually, this pattern reflects the phenomenon that aggregating a larger number of items increases measurement precision, but that each additional item provides a proportionally smaller increment in precision (cf. John & Soto, 2007; Lord, 1980; Lord & Novick, 1968). Practically, our findings also suggest a scale length “sweet spot” of approximately six to nine items. Big Five trait scales below this range appeared too short, in that validity declined rapidly

below six items. By contrast, validity increased only modestly from nine to twelve items; presumably, per-item validity gains would be even smaller beyond this point.

We propose that, in many personality assessment contexts, this sweet spot represents a reasonable balance between measurement efficiency and validity, and can therefore serve as a useful heuristic. This implies that the domain scales of the BFI-2 and its 30-item short form (Soto & John, 2017b), as well as other Big Five measures of similar length (e.g., Saucier, 1994), can provide considerable validity while remaining quite efficient. However, shorter measures entail substantial validity costs, which implies that the 15-item BFI-2 extra-short form (Soto & John, 2017b) and other ultra-brief Big Five measures (e.g., Donnellan, Oswald, Baird, & Lucas, 2006; Gosling et al., 2003) are shorter than optimal. The present findings also suggest that, compared with the four-item BFI-2 facet scales, longer measures could provide greater facet-level validity (e.g., Goldberg, 1999; McCrae & Costa, 2010).

We also caution that the scale length heuristic of six to nine items should not be applied in *all* contexts. In some circumstances, such as when making personnel or diagnostic decisions, or when participants can devote considerable time to personality assessment, even modest increments in validity can be valuable; in these contexts, the greater validity offered by longer scales can outweigh the cost of increased assessment time (Cronbach & Gleser, 1965). Other contexts impose severe constraints on personality assessment. For example, large-scale panel studies that interview nationally representative samples may be able to devote only one or two minutes to personality assessment (Donnellan & Lucas, 2008). Other studies may ask each participant to complete the same personality measure multiple times within the same session (e.g., Srivastava, Guglielmo, & Beer, 2010; Wood & Roberts, 2006). In such cases, ultra-brief scales may be the only feasible option. However, in the absence of such severe constraints, we

discourage researchers from undermining validity by administering ultra-brief personality measures (see also Credé, Harms, Niehorster, & Gaye-Valentine, 2012; McCrae, 2015).

Scale Width: Making Relatively Safe vs. Risky Validity Bets

Our second conclusion is that scores from broad trait scales tend to have somewhat stronger, and much more consistent, external validity associations than do scores from narrow scales. The finding that broader scales show more-consistent validity associations agrees with our conceptual analysis of bandwidth and fidelity (Figure 1; cf. Cronbach & Gleser, 1965; John et al., 1991). Within the universe of scales that could be constructed to measure a particular trait, scale iterations that include a broad range of item content will more-consistently represent the trait's overall meaning, and therefore show relatively similar patterns of validity associations. In contrast, scale iterations that each include a narrow range of content will sometimes assess quite different parts of the overall trait, and therefore show more-distinct criterion associations.

Our additional finding that broader trait scales tend to show somewhat stronger average validity associations likely reflects the phenomenon that items on a broad scale tend to be less homogenous. They therefore capture a greater total amount of unique personality information, which should enhance the scale scores' average validity across external criteria (John & Soto, 2007). This interpretation is consistent with the observation that excessive item redundancy can inflate reliability while weakening validity—except for associations with very closely matched criteria (Boyle, 1991; Loevinger, 1954).

Taken together, our scale width findings can help inform decisions about when to assess personality using broad vs. narrow trait scales. Specifically, they suggest that assessment using broad trait scales is a relatively safe bet. Broad scales will generally yield stronger external validity associations, and these associations will usually be robust across alternative measures of

the same trait. Thus, broad scales appear best suited for research exploring the general relations of personality traits with behaviors and life outcomes (Ozer & Benet-Martínez, 2006), and for prediction contexts where the goal is to predict a diverse set of criteria using a small number of trait scales (Cronbach & Gleser, 1965; Grucza & Goldberg, 2007; Morey et al., 2007).

In contrast, personality assessment using narrow trait scales is a riskier proposition. On average, this strategy will produce somewhat weaker validity associations. However, the greater variability of these associations implies that, for any particular criterion, scores from a narrow trait scale with especially relevant items can show greater validity than would scores from a broad scale (Paunonen & Ashton, 2001; Paunonen et al., 1999). Thus, narrow trait scales appear particularly well suited for two contexts. The first is when researchers have well-founded hypotheses regarding which specific facets of an overall trait will relate most closely with the target criterion (e.g., Samuel & Widiger, 2008); if these hypotheses are correct then they will be rewarded with especially strong validity associations, but if they are incorrect then validity will suffer. The second context is when researchers can administer many narrow trait scales to a very large sample of respondents, so that they can identify and cross-validate the strongest trait-criterion associations without excessive capitalization on chance (Browne, 2000).

Our scale width findings also have implications for scale development. When selecting items, some commonly used criteria—such as maximizing internal consistency, simple factor structure, or total IRT information—will tend to produce narrow scales (John & Soto, 2007; Smith, McCarthy, & Anderson, 2000). Other criteria—such as preferring moderate inter-item correlations and using domain sampling methods to avoid redundant content—will tend to produce broader scales (Gosling et al., 2003; Haynes, Richard, & Kubany, 1995; Soto & John, 2017b). We therefore recommend that scale developers (a) consider the meaning of the target

construct, as well as the validity tradeoffs between broad and narrow scales, (b) decide upon their preferred degree of scale breadth, and then (c) employ item-selection criteria that will promote this degree of breadth.

Note that our general conclusions about scale width should we qualified by some caveats. First, although personality assessment using broad trait scales is a relatively safe bet, it is not a sure thing. Each of the Big Five traits can be defined and operationalized in different ways, which can lead to quite distinct trait-criterion associations (e.g., Heller, Watson, & Ilies, 2004; Steel, Schmidt, & Shultz, 2008). Second, the bandwidth-matching hypothesis proposes that predictive accuracy can be maximized by assessing personality traits and external criteria at similar levels of content breadth (Smith, 1976). Some evidence is consistent with this hypothesis (Hogan & Roberts, 1996; Samuel & Widiger, 2008), but to our knowledge it has not yet been directly tested by systematically manipulating content breadth (O'Neill & Paunonen, 2013). Additional research is therefore needed to further investigate whether the predictive power of broad and narrow trait scales depends on the breadth of the target criterion.

We also note that broad and narrow trait scales are not always mutually exclusive. One way to resolve the bandwidth-fidelity tradeoff is by administering hierarchical personality measures, like the BFI-2, that nest narrow, facet-level scales within broad, domain-level scales (cf. DeYoung, Quilty, & Peterson, 2007; Goldberg, 1999; McCrae & Costa, 2010). These measures' broad domain scales can be used to explore general patterns of trait-criterion associations, while their narrow facet scales can enhance the prediction of specific criteria. We therefore encourage hierarchical personality assessment whenever feasible.

Scale Balance: The Importance of Minimizing Acquiescence Bias

Our third conclusion is that individual differences in acquiescent responding tend to bias trait-criterion validity associations when the trait and criterion measures both include an imbalance of positively and negatively keyed items, and are both obtained using the same method. Such validity associations tend to be positively biased when the trait and criterion scales are imbalanced in the same direction, and negatively biased when the scales are imbalanced in opposite directions. This bias can either inflate or suppress trait-criterion associations, depending on the expected direction of the association and the direction of the key imbalance. Moreover, greater imbalance leads to stronger bias. When both the trait and criterion scales were fully imbalanced (i.e., each included only positively-keyed or only negatively-keyed items), we found that the relative degree of trait-criterion overlap could be affected by as much as 47%.

These results highlight the importance of administering key-balanced trait and criterion scales. Ideally, a scale will have an equal number of items keyed in each direction, so that observed scale scores will reflect the trait or criterion being measured while controlling for individual differences in acquiescence. In practice, this ideal of perfect key balance is not always attainable. Some traits and criteria are better conceptualized as unipolar than bipolar, and thus easier to measure using items keyed in one direction. However, the present results indicate that even approximate key balance can substantially reduce the effects of acquiescence bias on validity associations, compared with the alternative of fully imbalanced scales.

It is important to note that key-balanced scales are not without complexities of their own. For example, adequately modeling item-level responses to a scale with both positively and negatively keyed items often necessitates the inclusion of a method factor alongside the substantive factors (e.g., Billiet & McClendon, 2000; Soto & John, 2017a). However, we propose that such complexities are a feature rather than a bug of balanced scales. A scale with

only positively or only negatively keyed items can often be adequately modeled (in terms of fit statistics) as a single factor. However, this factor will perfectly confound individual differences in the trait being measured with individual differences in acquiescence, leaving no way to separate the two. In contrast, scales with a mix of positively and negatively keyed items allow researchers to measure, model, and thus control the influence of acquiescence. For example, a researcher can construct an observed acquiescence index by identifying pairs of items with conceptually opposite content, and then computing each respondent's acquiescence score as their mean response (without reversing the negatively-keyed items) across all of the opposite-item pairs (Rammstedt & Farmer, 2013; Soto et al., 2008). This index can then be used to control acquiescence in subsequent item-level (by within-person centering participants' item responses around their acquiescence scores) or scale-level (by using the acquiescence index as a control variable) analyses. Similarly, a researcher can specify a latent variable model that includes an acquiescence method factor: all positively and negatively keyed items are constrained to load equally on this acquiescence factor, which is constrained to be uncorrelated with the substantive trait factors (Aichholzer, 2014; Billiet & McClendon, 2000; Soto & John, 2017a). This bifactor acquiescence model can also be extended to examine the associations of the trait and acquiescence factors with external variables (Danner, Aichholzer, & Rammstedt, 2015; John et al., 2008). Thus, both observed and latent variable approaches can be used to separate acquiescence variance from substantive personality information, estimate unbiased trait-criterion associations, and test hypotheses about the phenomenon of acquiescent responding itself.

Our balance findings also have implications for evaluating the validity of scores from some popular psychological measures. For example, the Grit Scale (Duckworth, Peterson, Matthews, & Kelly, 2007) includes two facet subscales: one (perseverance of effort) that

includes only positively-keyed items, and another (consistency of interests) that includes only negatively-keyed items. Because these two subscales are both fully imbalanced—in opposite directions—acquiescence variance will tend to suppress the correlation between the two facets, and can also bias—again in opposite directions—the facets' external validity associations. Acquiescence bias will thus lead the two grit facets to appear more empirically distinct than is conceptually warranted (cf. Credé, Tynan, & Harms, in press). Similar concerns also apply to some widely used criterion measures, which include only positively-keyed or only negatively-keyed items (e.g., Diener, Emmons, Larsen, & Griffin, 1985; Watson, Clark, & Tellegen, 1988). This further highlights the importance of developing and administering balanced scales.

Limitations and Future Directions

The present research has important strengths, including its systematic manipulation of scale length, width, and balance, and its examination of both self-reported and peer-reported validity criteria. However, the present research also has limitations that highlight important future directions. One such limitation concerns generalizability. We analyzed data from a particular sample, personality measure, and set of external criteria. Additional research is therefore needed to test whether the present findings generalize to other populations (e.g., children and adolescents, community adults), other personality measures (assessing the Big Five and other traits), and other criteria (e.g., directly observed behavior, long-term life outcomes).

For example, compared with the undergraduate students assessed in the present research, children, early adolescents, and adults with lower levels of educational attainment tend to have greater difficulty rating negatively-keyed items, which has led some researchers to suggest avoiding such items (Marsh, 1986; Ware, 1978). However, these same populations also tend to show greater individual differences in acquiescent responding, which suggests that administering

negatively-keyed items may be especially important so that acquiescence variance can be measured and separated from substantive personality information (Rammstedt & Farmer, 2013; Soto et al., 2008). Future research is therefore needed to investigate the effects of scale balance on validity in younger and less well educated populations.

Another generalizability issue concerns the BFI-2 item pool. The highly symmetrical structure of this pool made it well suited for manipulating scale length, width, and balance. However, each BFI-2 item has already passed a series of reliability and validity checks (Soto & John, 2017a). Previous research has shown that item characteristics, such as face validity, social desirability, extremity, length, and complexity, can affect external validity (Angleitner, John, & Loehr, 1986; Holden & Fekken, 1990; Holden, Fekken, & Jackson, 1985). Therefore, additional research is needed to test whether the present findings generalize to larger and more heterogeneous item sets. Such research could also manipulate scale length, width, and balance more dramatically than was possible here. For example, our analyses of scale width were limited to a maximum of three facets per Big Five domain, while holding scale length constant at only three items; future research could expand these analyses to include more facets and items.

Research using a larger item set could also test additional hypotheses. For example, CTT implies that the effects of scale length and width may interact. Because item homogeneity (i.e., inter-item correlations) tends to be lower for broad scales than for narrow scales, reliable assessment of a broad trait (such as one of the Big Five) may require more items than would assessment of a narrow trait (such as a Big Five facet). The present research could not adequately test this hypothesis due to the brevity of the BFI-2 facet scales. Thus, future research using a larger item set is needed to investigate the possibility of scale length × width interactions, and

thereby examine whether the scale length sweet spot of six to nine items observed here should be adjusted when researchers wish to measure broader or narrower traits.

More generally, we note that all of the present findings and recommendations should be viewed not as hard and fast rules, but as heuristics: potentially useful starting points that can help inform decisions. Researchers should interpret and modify these heuristics in light of their particular assessment context. For example, what trait is being measured, and how should this affect scale width? Is the trait unipolar or bipolar, and how should this affect scale balance? How much time can be devoted to personality assessment, and how should this affect scale length? The optimal scale characteristics for any particular context should be identified by thoughtfully considering the relevant conceptual and practical factors, not just by general rules of thumb.

Looking beyond scale length, width, and balance, future research can also examine whether additional characteristics affect external validity. For example, scales differ in their use of dichotomous (e.g., true-false) or polytomous (e.g., Likert scale) response options. How does response format affect trait-criterion associations (see Simms, this issue)? Because dichotomous items may provide less information than polytomous items, does optimal scale length depend on response format (Embretson & Reise, 2000)? As a second example, personality tests also differ in the simplicity vs. complexity of their factor structure. Do deviations from simple structure generally decrease or enhance validity (Hopwood & Donnellan, 2010; Loevinger, 1957)?

Future research can also look beyond external validity, which is only one key aspect of measurement quality (John & Soto, 2007; Loevinger, 1957; Messick, 1995). We focused on external validity because of its prominence in personality research, which often examines the capacity of personality measures to predict important behaviors and life outcomes (Kenrick & Funder, 1988; Ozer & Benet-Martínez, 2006). However, additional research is needed to

investigate how internal scale characteristics relate with other important aspects of measurement quality, such as temporal stability and inter-rater agreement.

A final limitation is that the present research operationalized external validity in terms of observed trait-criterion correlations between simple composite scales. We adopted this approach due to its feasibility (given our need to estimate more than 200,000 trait-criterion correlations), and because much previous research examining the associations of personality traits with behaviors and life outcomes has employed similar observed-variable methods (Ozer & Benet-Martínez, 2006; Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007). However, latent-variable approaches such as IRT and structural equation modeling continue to gain in popularity (Embretson & Reise, 2000; Kline, 2016). Thus, additional research is needed to test whether the present findings extend to latent-variable methods.

Conclusion

When developing a new personality trait scale, or selecting from among existing measures, researchers face a number of seemingly straightforward but deceptively nuanced considerations: How many items should the scale include? Should these items assess a broad or narrow range of trait-relevant content? How many items should be positively-keyed, and how many negatively-keyed? The present research indicates that these decisions about internal scale characteristics have important consequences for external validity. Practically, our findings suggest that personality scales should generally include at least six items, with a balanced number of positively and negatively keyed items, and that the appropriate degree of content breadth depends on the particular assessment context—as well as the researcher's preference for making a relatively safe or risky bet.

References

Aichholzer, J. (2014). Random intercept EFA of personality scales. *Journal of Research in Personality, 53,* 1-4.

Angleitner, A., John, Q P., & Lohr, E J. (1986). It's what you ask and how you ask it: An itemmetric analysis of personality questionnaires. In A. Angleitner & J. S. Wiggins (Eds.), *Personality assessment via questionnaires* (pp. 61-107). Berlin: Springer-Verlag.

Bardi, A., & Schwartz, S. H. (2003). Values and behavior: Strength and structure of relations. *Personality and Social Psychology Bulletin, 29,* 1207-1220.

Billiet, J. B., & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling, 7,* 608-628.

Boyle, G. J. (1991). Does item homogeneity indicate internal consistency or item redundancy in psychometric scales? *Personality and Individual Differences, 12,* 291-294.

Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology, 44,* 108-132.

Burisch, M. (1984). You don't always get what you pay for: Measuring depression with short and simple versus long and sophisticated scales. *Journal of Research in Personality, 18,* 81-98.

Costa, P. T., & McCrae, R. R. (1995). Domains and facets: Hierarchical personality assessment using the revised NEO personality inventory. *Journal of Personality Assessment, 64,* 21-50.

Credé, M., Harms, P., Niehorster, S., & Gaye-Valentine, A. (2012). An evaluation of the consequences of using short measures of the Big Five personality traits. *Journal of Personality and Social Psychology, 102,* 874-888.

Credé, M., Tynan, M. C., & Harms, P. D. (in press). Much ado about grit: A meta-analytic synthesis of the grit literature. *Journal of Personality and Social Psychology.*

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement, 6,* 475-494.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana: University of Illinois Press.

Danner, D., Aichholzer, J., & Rammstedt, B. (2015). Acquiescence in personality questionnaires: Relevance, domain specificity, and stability. *Journal of Research in Personality, 57,* 119-130.

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the Big Five. *Journal of Personality and Social Psychology, 93,* 880-896.

Diener, E. D., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The Satisfaction With Life Scale. *Journal of Personality Assessment, 49,* 71-75.

Donnellan, M. B., & Lucas, R. E. (2008). Age differences in the Big Five across the life span: Evidence from two national samples. *Psychology and Aging, 23,* 558-566.

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The mini-IPIP scales: Tiny-yet-effective measures of the Big Five factors of personality. *Psychological Assessment, 18,* 192-203.

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology, 92,* 1087-1101.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8,* 341-349.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Goldberg, L. R. (1999). A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7, pp. 7-28). Tilburg, The Netherlands: Tilburg University Press.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big- Five personality domains. *Journal of Research in Personality, 37,* 504-528.

Grucza, R. A., & Goldberg, L. R. (2007). The comparative validity of 11 modern personality inventories: Predictions of behavioral acts, informant reports, and clinical indicators. *Journal of Personality Assessment, 89,* 167-187.

Haynes, S. N., Richard, D., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7,* 238-247.

Heller, D., Watson, D., & Ilies, R. (2004). The role of person versus situation in life satisfaction: A critical examination. *Psychological Bulletin, 130,* 574-600.

Hogan, J., & Roberts, B. W. (1996). Issues and non-issues in the fidelity-bandwidth trade-off. *Journal of Organizational Behavior, 17,* 627-637.

Holden, R. R., & Fekken, G. C. (1990). Structured psychopathological test item characteristics and validity. *Psychological Assessment, 2,* 35-40.

Holden, R. R., Fekken, G. C., & Jackson, D. N. (1985). Structured personality test item characteristics and validity. *Journal of Research in Personality, 19,* 386-394.

Hopwood, C. J., & Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated? *Personality and Social Psychology Review, 14,* 332-346.

Jackson, D. N., & Messick, S. (1958). Content and style in personality assessment. *Psychological Bulletin, 55,* 243-252.

Janis, I. L., & Field, P. B. (1959). Sex differences and personality factors related to persuasibility. In C. I. Hovland & I. L. Janis (Eds.), *Personality and persuasibility* (pp. 55-68). New Haven, CT: Yale University Press.

John, O. P., Hampson, S. E., & Goldberg, L. R. (1991). The basic level in personality-trait hierarchies: Studies of trait use and accessibility in different contexts. *Journal of Personality and Social Psychology, 60,* 348-361.

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (3rd ed., pp. 114-158). New York, NY: Guilford.

John, O. P., & Soto, C. J. (2007). The importance of being valid. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 461-494). New York, NY: Guilford.

Jonason, P. K., & Webster, G. D. (2010). The Dirty Dozen: A concise measure of the Dark Triad. *Psychological Assessment, 22,* 420-432.

Jones, D. N., & Paulhus, D. L. (2014). Introducing the Short Dark Triad (SD3): A brief measure of dark personality traits. *Assessment, 21,* 28-41.

Kenrick, D. T., & Funder, D. C. (1988). Profiting from controversy: Lessons from the person-situation debate. *American Psychologist, 43,* 23-34.

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York, NY: Guilford.

Loevinger, J. (1954). The attenuation paradox in test theory. *Psychological Bulletin, 51,* 493-504.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3,* 635-694.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental tests.* New York, NY: Addison-Wesley.

Marsh, H. W. (1986). Negative item bias in ratings scales for preadolescent children: A cognitive-developmental phenomenon. *Developmental Psychology, 22,* 37-49.

McCrae, R. R. (2015). A more nuanced view of reliability: Specificity in the trait hierarchy. *Personality and Social Psychology Review, 19,* 97-112.

McCrae, R. R., & Costa, P. T. (2010). *NEO Inventories professional manual.* Lutz, FL: Psychological Assessment Resources.

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality, 60,* 175-215.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741-749.

Morey, L. C., Hopwood, C. J., Gunderson, J. G., Skodol, A. E., Shea, M. T., Yen, S., Stout, R. L., Zanarini, M. C., Grilo, C. M., Sanislow, C.A., & McGlashan, T. H. (2007).

Comparison of alternative models for personality disorders. *Psychological Medicine, 37,* 983-994.

O'Neill, T. A., & Paunonen, S. V. (2013). Breadth in personality assessment: Implications for the understanding and prediction of work behavior. In N. D. Christiansen & R. P. Tett (Eds.), *Handbook of personality at work.* New York, NY: Routledge.

OECD (2015). *Skills for social progress: The power of social and emotional skills.* Paris, France: OECD Publishing.

Ones, D. S., & Viswesvaran, C. (1996). Bandwidth-fidelity dilemma in personality measurement for personnel selection. *Journal of Organizational Behavior, 17,* 609-626.

Ozer, D. J., & Benet-Martínez, V. (2006). Personality and the prediction of consequential outcomes. *Annual Review of Psychology, 57,* 401-421.

Paunonen, S. V., & Ashton, M. C. (2001). Big Five factors and facets and the prediction of behavior. *Journal of Personality and Social Psychology, 81,* 524-539.

Paunonen, S. V., Rothstein, M. G., & Jackson, D. N. (1999). Narrow reasoning about the use of broad personality measures for personnel selection. *Journal of Organizational Behavior, 20,* 389-405.

Rammstedt, B., & Farmer, R. F. (2013). The impact of acquiescence on the evaluation of personality structure. *Psychological Assessment, 25,* 1137-1145.

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science, 2,* 313-345.

Rosenberg, M. (1965). *Society and the adolescent self-image.* Princeton, NJ: Princeton University Press.

Ryff, C. D., Lee, Y. H., Essex, M. J., & Schmutte, P. S. (1994). My children and me: Midlife evaluations of grown children and of self. *Psychology and Aging, 9,* 195-205.

Samuel, D. B., & Widiger, T. A. (2008). A meta-analytic review of the relationships between the five-factor model and DSM-IV-TR personality disorders: A facet level analysis. *Clinical Psychology Review, 28,* 1326-1342.

Saucier, G. (1994). Mini-markers: A brief version of Goldberg's unipolar Big-Five markers. *Journal of Personality Assessment,* 63, 506-516.

Simms, L. J. (in press). Significance of the number of response options in personality Likert-scale data. *Psychological Assessment.*

Smith, P. C. (1976). Behaviors, results, and organizational effectiveness: The problem of criteria. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 745-775). Chicago, IL: Rand McNally.

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment, 12,* 102-111.

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,* 117-143.

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.

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2008). The developmental psychometrics of Big Five self-reports: Acquiescence, factor structure, coherence, and differentiation from ages ten to twenty. *Journal of Personality and Social Psychology, 94,* 718-737.

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2011). Age differences in personality traits from 10 to 65: Big Five domains and facets in a large cross-sectional sample. *Journal of Personality and Social Psychology, 100,* 330-348.

Srivastava, S., Guglielmo, S., & Beer, J. S. (2010). Perceiving others' personalities: Examining the dimensionality, assumed similarity to the self, and stability of perceiver effects. *Journal of Personality and Social Psychology, 98,* 520-534.

Steel, P., Schmidt, J., & Shultz, J. (2008). Refining the relationship between personality and subjective well-being. *Psychological Bulletin, 134,* 138-161.

Ware, J. E., Jr. (1978). Effects of acquiescent response set on patient satisfaction ratings. *Medical Care, 16,* 327-336.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54,* 1063-1070.

Winkler, J. D., Kanouse, D. E., & Ware, J. E., (1982). Controlling for acquiescence response set in scale development. *Journal of Applied Psychology, 67,* 555-561.

Wood, D., & Roberts, B. W. (2006). Cross-sectional and longitudinal tests of the Personality and Role Identity Structural Model (PRISM). *Journal of Personality, 74,* 779-809.

Table 1

Effects of Scale Length on Mean Absolute Trait-Criterion Validity Associations

		Mean correlation by trait					Mean correlation by criterion		
Scale length	All criteria and traits	E	A	C	N	O	Self-reported behavior	Self-reported well-being	Peer-reported
1	.15	.17	.15	.16	.15	.11	.10	.24	.14
2	.18	.21	.18	.20	.18	.14	.12	.30	.17
3	.20	.23	.21	.22	.20	.15	.13	.33	.18
4	.21	.24	.21	.22	.21	.16	.14	.34	.19
5	.22	.25	.22	.23	.21	.16	.14	.35	.20
6	.22	.25	.23	.24	.22	.17	.15	.36	.20
7	.23	.26	.24	.24	.22	.17	.15	.36	.20
8	.23	.26	.24	.25	.22	.17	.15	.37	.21
9	.23	.26	.25	.25	.23	.17	.15	.37	.21
10	.23	.26	.25	.25	.23	.18	.15	.37	.21
11	.24	.27	.25	.25	.23	.18	.15	.38	.21
12	.24	.27	.25	.25	.23	.18	.16	.38	.21

Note. Scale length = Number of items on each trait scale iteration. Values are mean absolute trait-criterion validity correlations, computed using Fisher's *r*-to-*z* transformation. For each Big Five trait, there were 12 one-item, 48 two-item, 64 three-item, 288 four-item, 432 five-item, 216 six-item, 432 seven-item, 288 eight-item, 64 nine-item, 48 ten-item, 12 eleven-item, and 1 twelve-item scale iterations. For self-reported behavioral criteria, $N = 524$. For self-reported well-being criteria, $N = 265$. For peer-reported criteria, $N = 232$. E = Extraversion. A = Agreeableness. C = Conscientiousness. N = Negative Emotionality. O = Open-Mindedness.

Table 2

Effects of Scale Width on the Strength and Variability of Trait-Criterion Validity Associations

			Mean or standard deviation by trait				Mean or standard deviation by criterion		
Scale width	All criteria and traits	Extra-version	Agree-ableness	Conscien-tiousness	Negative Emotionality	Open-Mindedness	Self-reported behavior	Self-reported well-being	Peer-reported
Mean absolute trait-criterion validity correlation									
Narrow	.19	.22	.19	.20	.19	.14	.31	.17	.12
Broad	.20	.23	.21	.22	.20	.15	.33	.18	.13
Average standard deviation of trait-criterion validity correlations									
Narrow	.070	.072	.072	.059	.084	.058	.066	.084	.055
Broad	.037	.037	.040	.037	.032	.037	.031	.044	.037

Note. Narrow scale iterations include three items from the same BFI-2 facet subscale. Broad scale iterations include one item from each of the three BFI-2 facet subscales within a Big Five trait. Mean correlations were computed using Fisher's *r*-to-*z* transformation. Average standard deviations were computed by (a) computing the variance of each trait-criterion correlation across scale iterations, (b) computing the mean of these variances, and then (c) taking the square root of this mean variance. For each Big Five trait, there were 64 broad and 12 narrow scale iterations. For self-reported behavioral criteria, $N = 524$. For self-reported well-being criteria, $N = 265$. For peer-reported criteria, $N = 232$.

Table 3

Effects of Scale Balance on Mean Absolute Trait-Criterion Validity Associations

		Mean correlation by trait					Mean correlation by criterion					
Key-matched items	All traits and criteria	E	A	C	N	O	Self-acceptance	Environ. mastery	Purpose in life	Positive relations	Personal growth	Autonomy
0 (Fully mismatched)	.28	.32	.25	.33	.33	.18	.32	.32	.30	.27	.26	.23
1	.30	.34	.27	.34	.35	.20	.34	.34	.31	.28	.28	.24
2	.31	.35	.28	.36	.36	.21	.36	.36	.32	.30	.29	.25
3 (Balanced)	.33	.36	.30	.37	.37	.23	.37	.37	.34	.32	.31	.26
4	.34	.37	.31	.38	.37	.24	.38	.38	.34	.33	.32	.27
5	.34	.38	.32	.38	.37	.24	.38	.38	.35	.34	.32	.27
6 (Fully matched)	.34	.39	.32	.38	.38	.25	.39	.38	.35	.34	.33	.27

Note. Key-matched items = Number of items on the trait scale keyed in the same direction as the items on the fully imbalanced criterion scale. Positive relations = Positive relations with others. Values are mean absolute trait-criterion validity correlations, computed using Fisher's *r*-to-*z* transformation. For each Big Five trait, there was 1 scale iteration with six positively-keyed items, 12 with five positively-keyed items, 51 with four positively-keyed items, 88 with three positively-keyed items, 51 with two positively-keyed items, 12 with one positively-keyed item, and 1 with no positively-keyed items. $N = 265$. E = Extraversion. A = Agreeableness. C = Conscientiousness. N = Negative Emotionality. O = Open-Mindedness. Environ. = Environmental.

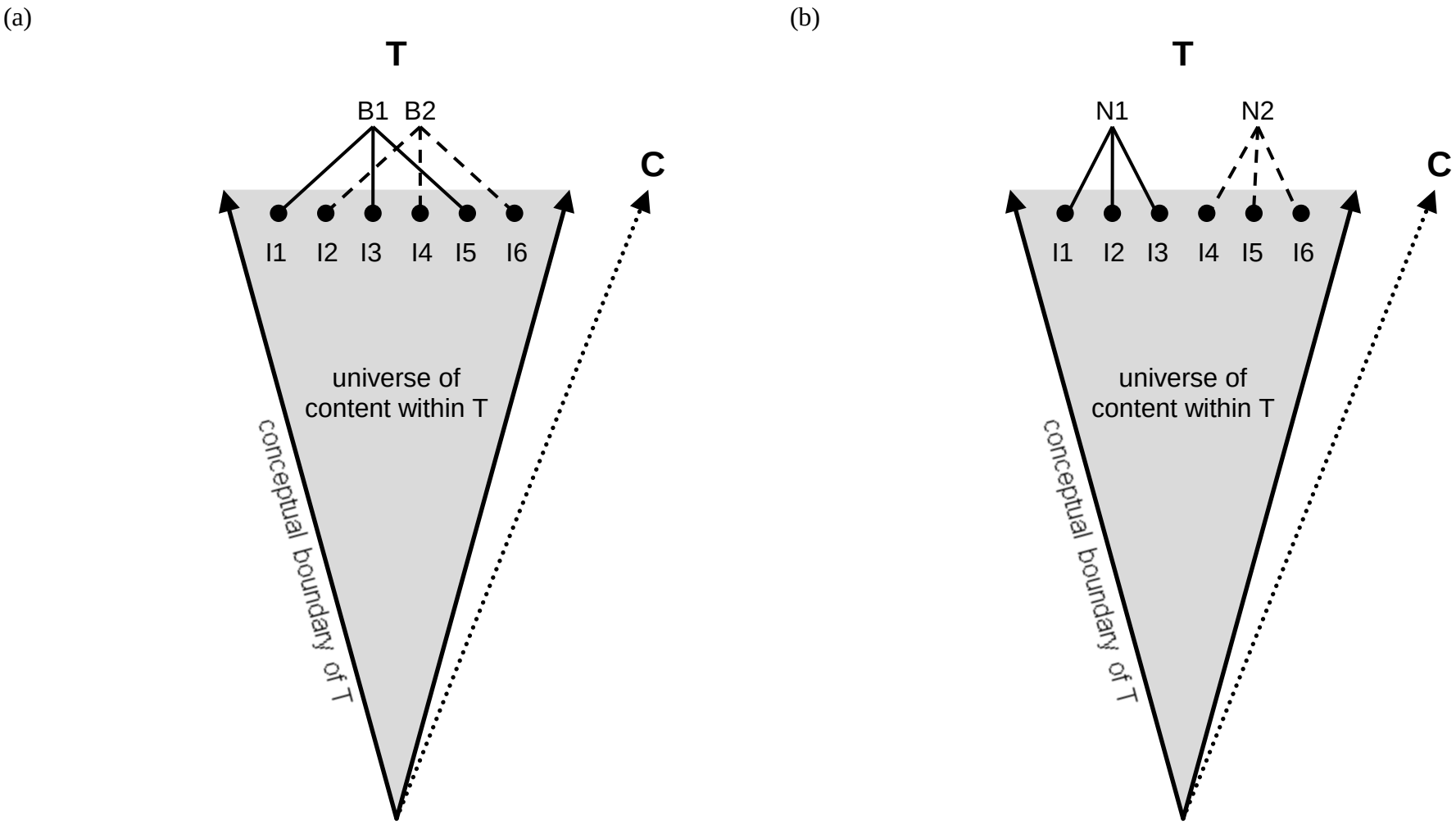
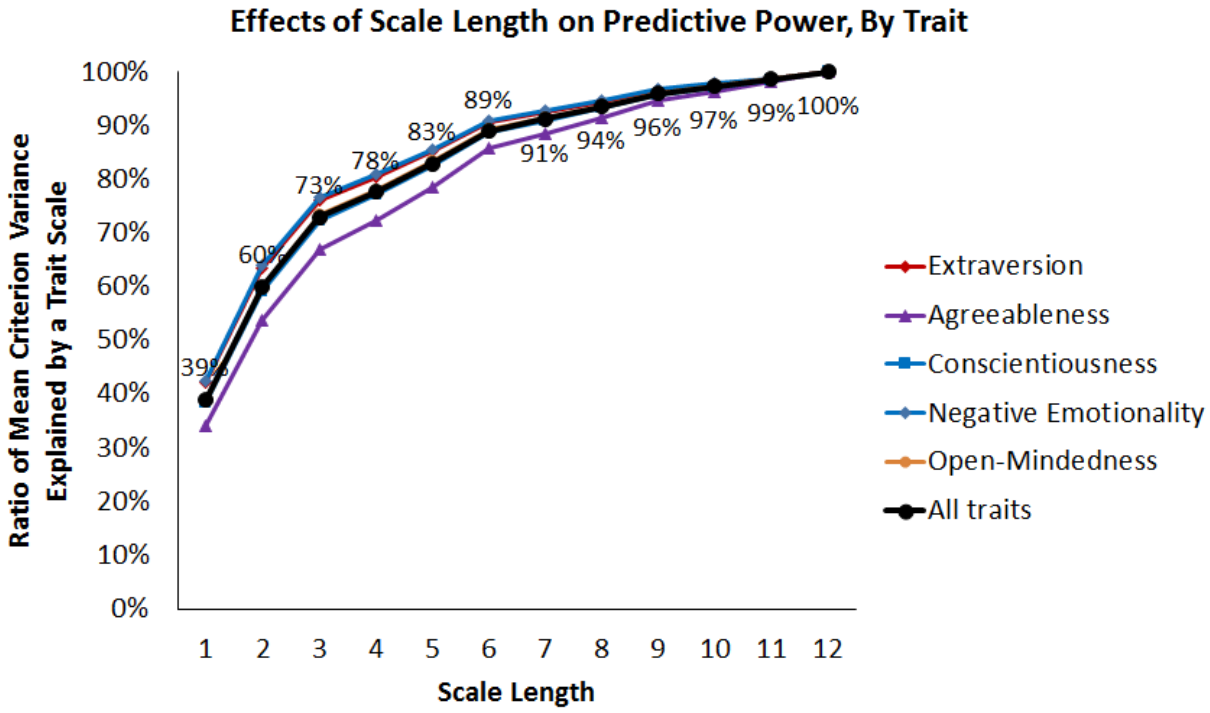


Figure 1. A schematic illustration of broad and narrow trait scales, which shows how scale width can affect trait-criterion validity associations. T = a personality trait. I1–I6 = six items assessing different facets of personality trait T. B1, B2 = two broad scale iterations assessing personality trait T. N1, N2 = two narrow scale iterations assessing personality trait T. C = a trait-relevant external criterion.

(a)



(b)

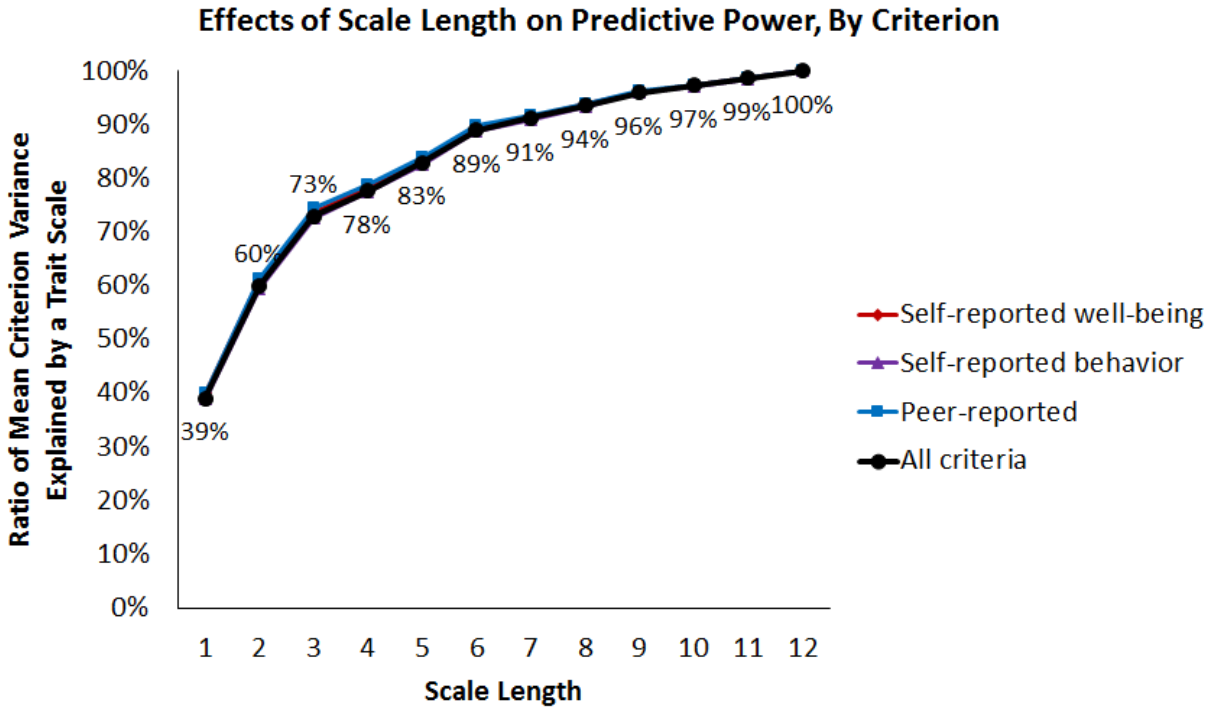
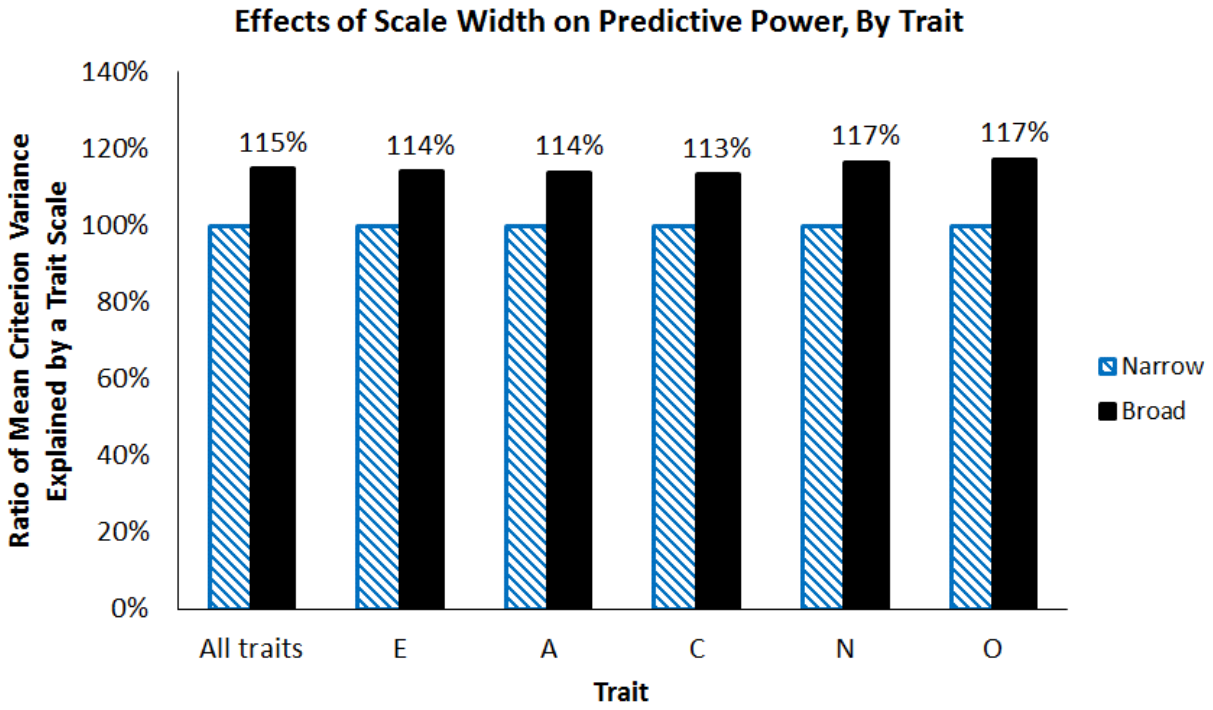


Figure 2. Effects of scale length on mean predictive power (i.e., proportion of criterion variance explained by a trait scale), relative to a baseline provided by the twelve-item trait scales. Data labels represent relative predictive power, averaged across all traits and criteria.

(a)



(b)

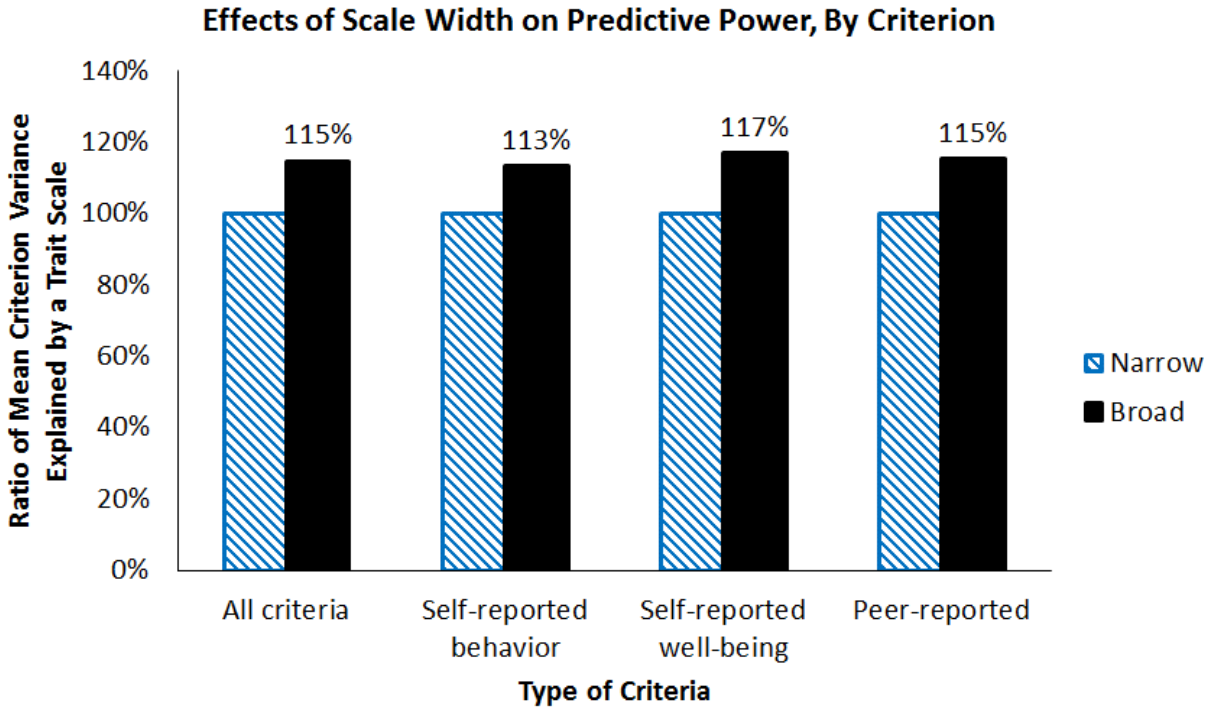
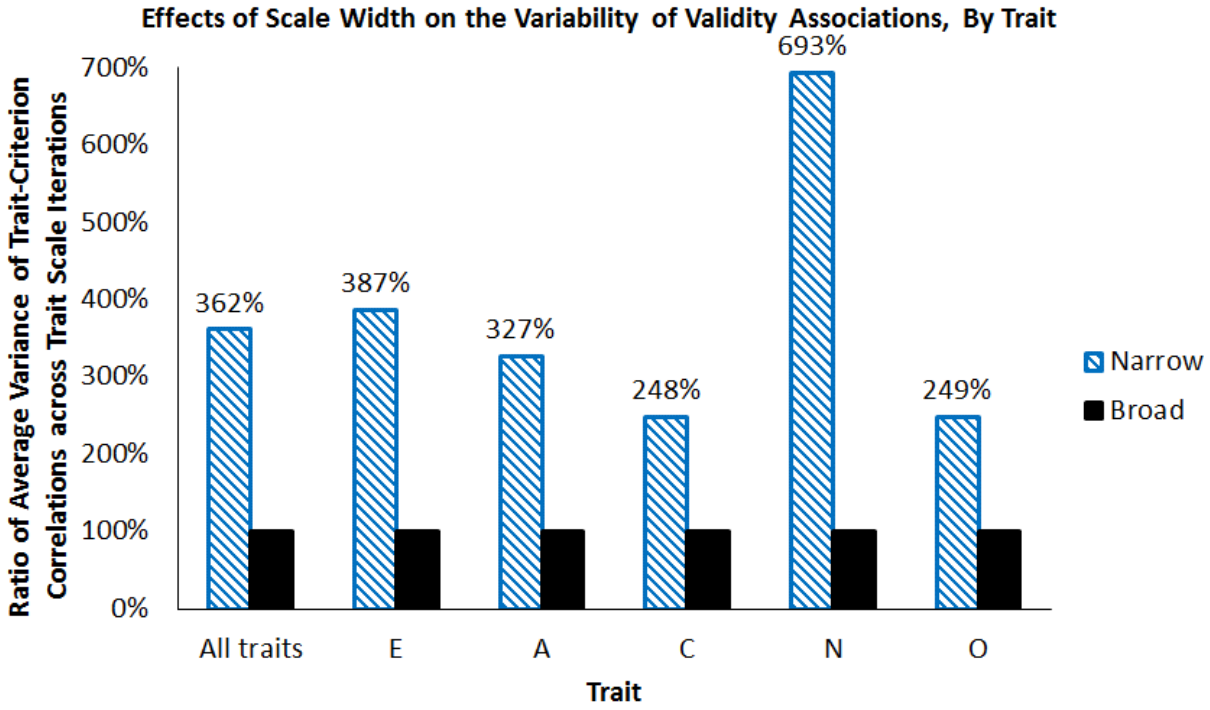


Figure 3. Effects of scale width on mean predictive power (i.e., proportion of criterion variance explained by a trait scale), relative to a baseline provided by the narrow trait scales. E = Extraversion. A = Agreeableness. C = Conscientiousness. N = Negative Emotionality. O = Open-Mindedness.

(a)



(b)

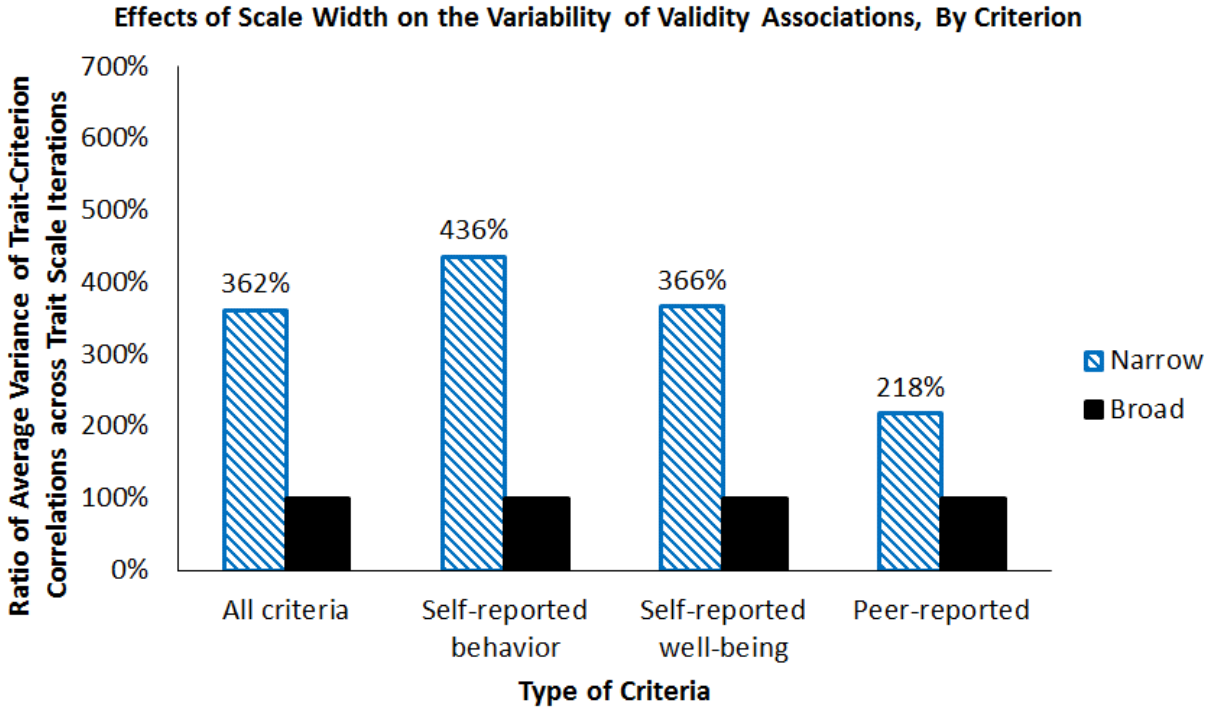
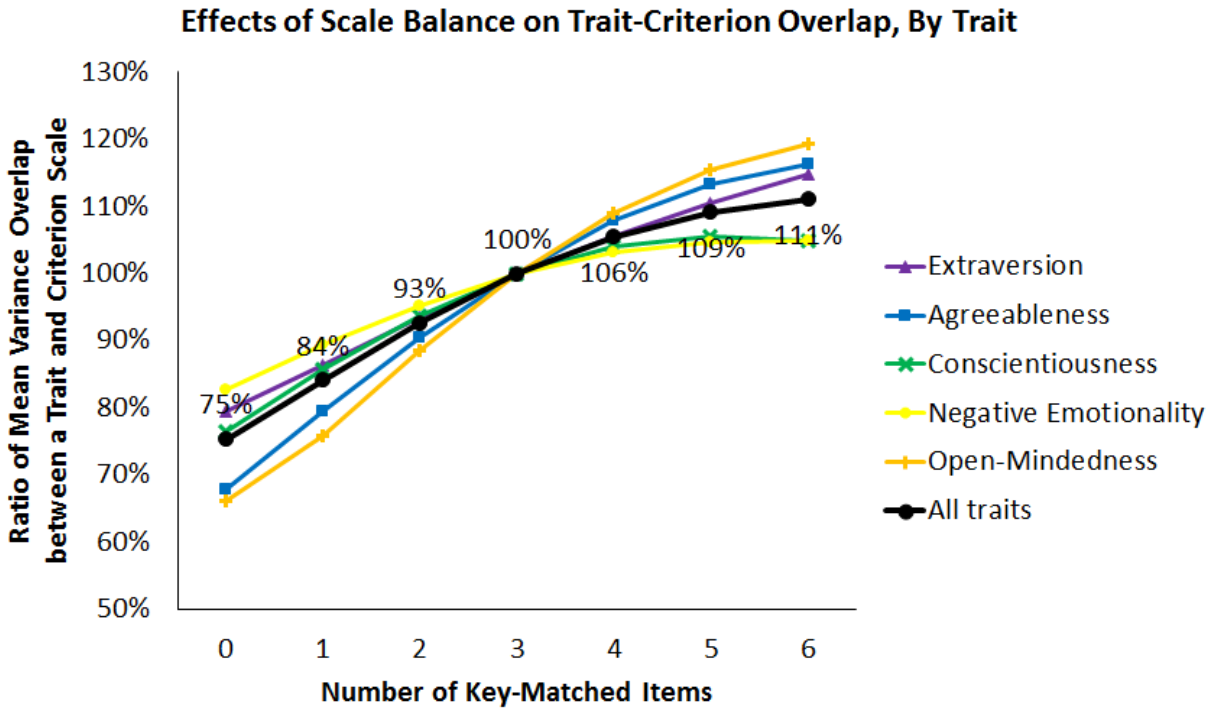


Figure 4. Effects of scale width on the variability of trait-criterion validity correlations across scale iterations, relative to a baseline provided by the broad trait scales. E = Extraversion. A = Agreeableness. C = Conscientiousness. N = Negative Emotionality. O = Open-Mindedness.

(a)



(b)

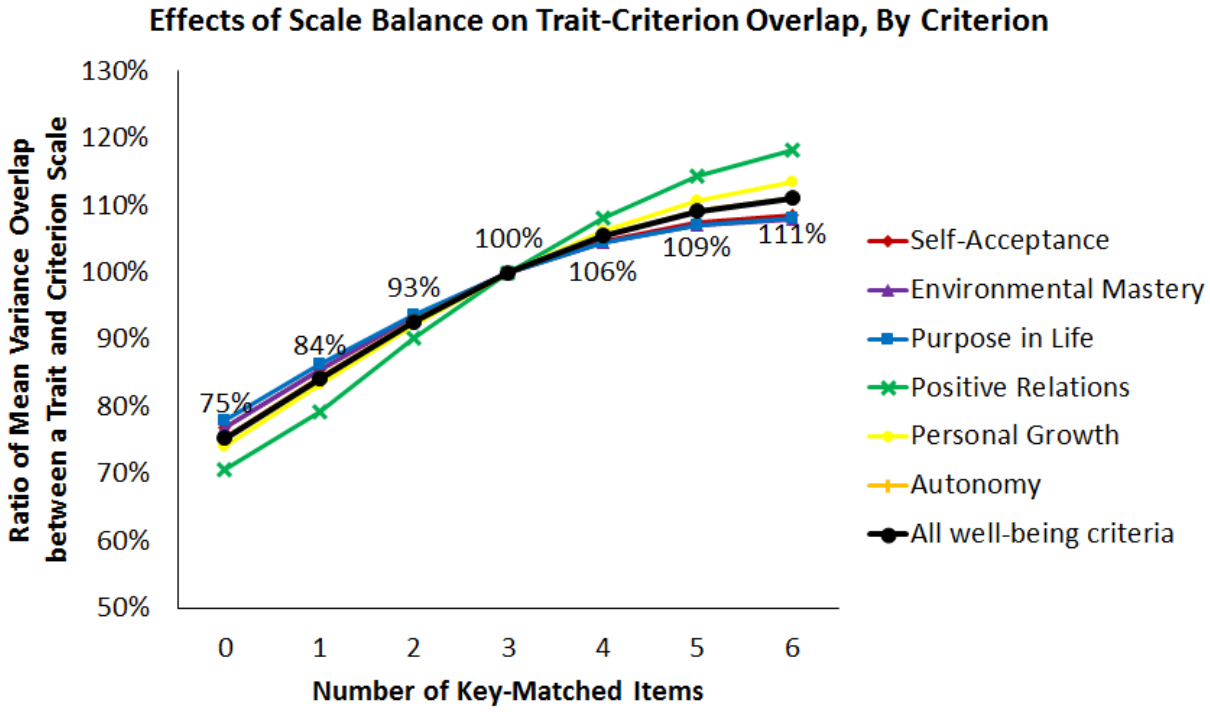


Figure 5. Effects of scale balance on mean trait-criterion overlap (i.e., proportion of overlapping variance between a trait and criterion scale), relative to a baseline provided by the key-balanced trait scales. Number of key-matched items = Number of items on the six-item trait scale keyed in the same direction as the items on the fully imbalanced criterion scale. Data labels represent relative trait-criterion overlap, averaged across all traits and well-being criteria.