**Running Head: INDEXING ITEM SIMILARITY**

# Using Retest-Adjusted Correlations as Indicators of the Semantic Similarity of Items

Dustin Wood
University of Alabama

Graham H. Lowman
Kennesaw State University

Benjamin F. Armstrong III
McGill University

P.D. Harms
University of Alabama

## Author Note

Dustin Wood, Department of Management, University of Alabama. Graham H. Lowman, Michael A. Leven School of Management, Entrepreneurship, and Hospitality, Kennesaw State University. Benjamin F. Armstrong III, Department of Psychology, McGill University. P.D. Harms, Department of Management, University of Alabama.

This research was supported by the National Science Foundation under award #2121275. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Science Foundation.

Data and materials from this project are available at https://osf.io/vp6kr/. We thank Justin DeSimone and Qiuyu Su for comments on an earlier draft of the manuscript. Some of the data used here has previously been used to explore univariate properties of within-study retest correlations (Lowman et al., 2018) or to evaluate response speed and consistency indices as indicators of data quality (Wood et al., 2017). However, the perceived semantic similarity ratings are new data collected and used for the first time in this study, and the analyses exploring how different correlational indices of inter-item relationships relate to the perceived semantic similarity of the item-pairs represent new analyses.

**ABSTRACT**

Determining whether different items *provide the same information* or *mean the same thing* is a central concern when determining whether different scales or constructs are overlapping or redundant. In the present study, we suggest that *retest-adjusted correlations,* $\hat{\rho}_{\mathbb{XY}|d|}$, provide a valuable means of adjusting for item-level unreliability. More exactly: we suggest dividing the estimated correlation between items $X$ and $Y$ measured over measurement interval $|d|$ by the average retest correlations of the items over the same interval. For instance: if we correlate scores from items $X$ and $Y$ measured one week apart, their retest-adjusted correlation is estimated by using their one-week retest correlations. Using data from four inventories, we provide evidence that retest-adjusted correlations are significantly better predictors of whether two items are consensually regarded as "meaning the same thing" by judges than raw-score correlations. The results may provide the first empirical evidence that Spearman's (1904, 1910) suggested reliability adjustment do – in certain (perhaps very constrained!) circumstances – improve upon raw-score correlations as indicators of the informational or semantic equivalence of different tests.

**Keywords:** *semantic similarity; test equivalence; reliability; nuances; item-level analysis*

**Using Retest-Adjusted Correlations as Indicators of the Semantic Similarity of Items**

It is often appreciated that large correlations between self-report measures can reflect *tautological* relationships or *redundancies*, where a non-trivial proportion of the items in the two measures quite literally 'mean the same thing'. For instance, in the pithily titled "Some dangers of using personality questionnaires to study personality," Nicholls and colleagues (1982) argued that the large positive associations observed between widely-used tests of *masculinity* and *self-esteem* was due in large part to the fact that commonly used tests of each construct include nearly identical items about the person's level of *assertiveness* and *confidence*. As a more recent example, Credé and colleagues (2017) noted that the very high correlation between Conscientiousness scales and the widely-used Grit scale (Duckworth et al., 2007) is due in part to the scales regularly containing very similar items – for instance, the Grit scale item "I am a hard worker" and the IPIP Conscientiousness item "I work hard." More generally, the fact that scales of nominally distinct constructs regularly contain similar items represents a problem for the behavioral sciences, as it makes it more difficult to interpret whether the correlations between multi-item scales should be interpreted as indicating a *functional* relationship (e.g., X causes Y), a *tautological* relationship (X and Y measure the same thing), or some mixture of both (Mõttus, 2016; Wilt & Revelle, 2015; Wood et al., 2015). This issue is considered under various labels, such as *construct overlap, construct redundancy, construct identity fallacies, construct proliferation, jangle fallacies* and *tautological relationships across psychological scales* (or *TRAPS*)(Bainbridge et al., 2022; Kelley, 1927; Larsen & Bong, 2016; Rosenbusch et al., 2020; Shaffer et al., 2016; Singh, 1991; Wood & Harms, 2016).

As these examples illustrate, determining whether two *scales* overlap often involves determining whether particular *items* are similar or even redundant in meaning. But questions regarding item similarity can also be important when there is no aspiration to create multi-item scales at all. For instance, Block (1961) noted that items were replaced through iterations of his Q-sort inventories when they were determined to be too similar to others in the inventory. This strategy continues to play a role in the development of inventories aiming to maximize breadth or comprehensiveness (e.g., Funder, 2016; Furr et al., 2010; Shedler & Westen, 2007; Wood et al., 2010). More generally, there is considerable current

interest in creating inventories which measure a greater number of important *nuances* of personality, which involves, in part, identifying items which contribute distinct information beyond others in the set (Condon et al., 2020; McCrae, 2015; Mõttus et al., 2020; Saucier et al., 2020).

However, there remain considerable questions regarding how to establish whether two items are either informationally or semantically equivalent. Addressing these issues is in turn a fundamental step to better establishing when a correlation between measures of conceptually distinct constructs should be attributed to redundant content. In the current article, we present evidence that *retest-adjusted correlations* serve as particularly valuable indices of the level of information similarity of items. We also describe a straightforward method for operationalizing the *semantic similarity* of items through rater judgments. We then conduct empirical tests to explore whether retest-adjusted correlations outperform raw-score correlations as indicators of the rated semantic similarity of item-pairs.

## How are Levels of Item Similarity Estimated?

The level of similarity between two items can be estimated through *information similarity* and *semantic similarity* approaches (Larsen & Bong, 2016; Le et al., 2010).[1] *Information similarity* approaches involve estimating the degree to which items X and Y provide the same information within a population. In contrast, *semantic similarity* approaches concern the degree to which items are understood as 'meaning the same thing' within a population. We describe approaches to quantifying both below.

### Estimating Information Similarity

To evaluate whether two items X and Y provide distinct information, we would ideally evaluate whether their scores correlate at a level near unity – i.e., evaluate the hypothesis $r_{XY} = 1$. Correlations less than 1 would indicate that the items provide unique information. However, the fact that respondents have some tendency to provide different scores *even when rating the same item twice* usually precludes this from being a meaningful test.

[1] These are sometimes referred to as *quantitative/empirical* versus *qualitative/theoretical* approaches, respectively (e.g., Le, Schmidt, Harter, & Lauver, 2010; Shaffer, DeGeest, & Li, 2016). We do not use these labels here, as the semantic similarity of item content can be estimated in a quantitative manner. *Information similarity* versus *semantic similarity* may also be termed *score (or response) similarity* versus *content similarity,* respectively.

***The Logic of Reliability-Adjusted Correlations***

Over a century ago, Spearman (1904, 1910) suggested that the ability to evaluate the information similarity of two tests (or in his terms, their *correspondence*) via correlational indices can be restored by adjusting for score unreliability in the manner below:

**Equation 1.** $\rho_{\mathbb{XY}} = \frac{\rho_{XY}}{\sqrt{\rho_{XX}\rho_{YY}}}$

We use the 'double-struck' $\mathbb{X}$ and $\mathbb{Y}$ in the present notation as a more compact means of indicating the *expected values of X and Y*, which are sometimes notated as $\mathbb{E}[X]$ and $\mathbb{E}[Y]$). The parameter resulting from this calculation, $\rho_{\mathbb{XY}}$, can be understood as an estimate of how the expected values of X and Y are correlated (Borsboom, 2005; Lazarsfeld, 1959; Lord & Novick, 1968).

It is useful to note that Spearman's equation operationalizes $\rho_{\mathbb{XY}}$ as a *ratio* of potentially estimable quantities. Specifically, it indexes the degree to which the raw-score correlation between two tests is smaller than the geometric mean of the raw-score correlations of the tests with themselves (Revelle & Condon, 2019). For instance: $\rho_{\mathbb{XY}}$ = .80 would indicate that the correlation between X and Y is 4/5$^{th}$ the size of the correlation that X and Y have with themselves. This fact is important as while an empirically-estimated raw-score correlation $r_{XY}$ *cannot* fall outside the range of $[-1,1]$, reliability-adjusted correlations, $\hat{\rho}_{\mathbb{XY}}$, *can* and occasionally *do* to fall outside this range, due to sample fluctuations in $r_{XY}$, $r_{XX}$ and $r_{YY}$ estimates (Charles, 2005). The range $[-1,1]$ nonetheless serves as the *expected* limits of $\rho_{\mathbb{XY}}$ values: the population-level correlation between two measures – if properly estimated – cannot exceed the population geometric-mean reliability of the measures (i.e., $\rho_{XY} \leq \sqrt{\rho_{XX}\rho_{YY}}$). Consequently, an advantage of reliability-adjusted correlations is that they restore the expectation that we should observe values near the limits of the $[-1,1]$ correlational range when tests are extremely similar.

However, there continues to be considerable debate regarding how to correctly operationalize the reliabilities in the denominator of Spearman's equation (e.g., Le et al., 2010; McCrae, 2015; Revelle & Condon, 2019). For instance, it is widely understood that the statistic most commonly provided by researchers as a reliability estimate, coefficient alpha (Cronbach, 1951), systematically underestimates a

measure's reliability, i.e., $\mathbb{E}(\alpha_X) < \rho_{XX}$ (e.g., (John & Soto, 2007; McDonald, 1999; Osburn, 2000; Sijtsma, 2009). Furthermore, the specific *degree* to which $\rho_{XX}$ is underestimated by any particular internal consistency statistic is difficult to establish, in turn making it difficult to compensate for this problem. Consequently, using coefficient alpha estimates as the reliability estimates within Spearman's equation should produce $\hat{\rho}_{\mathbb{XY}}$ estimates which are *expected* to exceed the correct value in magnitude; i.e., $\mathbb{E}(\hat{\rho}_{\mathbb{XY}}) > \rho_{\mathbb{XY}}$. This has considerable implications for meta-analysis, structural equation modeling, and other research areas were reliability adjustments are routinely conducted (LeBreton et al., 2014; Sackett et al., 2021). Among other things, this will result in overly liberal estimates that two or more scales 'measure the same thing' when estimating their reliability-adjusted correlation.

The problem becomes even more acute when evaluating the reliability of tests consisting of a single item: the prevailing strategy over the last several decades in behavioral research has been to estimate a measure's reliability through internal consistency statistics, which utilize information about associations between items within the test (e.g., Nunnally & Bernstein, 1991). Single-item tests by definition have no inter-item correlations to utilize for this purpose.

***Using Retest Correlations to Adjust for Measurement Unreliability***

There has been increasing interest in using retest correlations to adjust for measurement unreliability (e.g., Dragostinov & Mõttus, 2021; Mueller et al., 2015; Wood et al., 2018; Wood & Wortman, 2012). This suggestion has been based in part on recent evidence that retest correlations over intervals such as a couple weeks or months outperform more commonly employed reliability estimates toward predicting validity-related criteria that should be attenuated by unreliability, such as correlational estimates of self-other agreement or heritability (Henry et al., 2022; Lowman et al., 2018; McCrae et al., 2011; Mõttus et al., 2017). Further, it is possible to estimate the retest correlation for scales of any length, including single items.

**Problem: What retest interval to use?** Although investigators have suggested some applications of retest correlations for reliability adjustments (e.g., Green, 2003; Le et al., 2010; McCrae, 2015; McCrae et al., 2011), there are not well-accepted guidelines for how best to do so. A question

researchers have raised regarding the use of retest correlations for reliability adjustments is: *which retest correlation to use*? Retest correlations are typically expected to decrease as the retest interval increases – for instance, from one day, to one week, to one month, and so on (Fraley et al., 2011; Fraley & Roberts, 2005; Gnambs, 2014; Lucas & Donnellan, 2007; Revelle & Condon, 2019). Consequently, there is no single 'retest reliability' value of a measure; rather there are at least as many retest reliability values as there are retest intervals (Cronbach, 1947).[2]

To address this problem, we propose that *the most conceptually appropriate reliability estimates for adjusting the raw-score correlation between X and Y for measurement unreliability are the retest correlations of tests X and Y over the same measurement interval*. To understand how this idea can be operationalized, it is useful to appreciate that for each participant, the measurements of two items X and Y are made at particular moments in time, which we can denote as $t_{X_p}$ and $t_{Y_p}$, respectively. We can take the absolute difference of these estimates for each participant, $\left|t_{X_p} - t_{Y_p}\right| = \left|d_{XY_p}\right|$. The average of these estimates across participants, $|d_{XY}|$, provides the average time interval separating the measurements of $X$ and $Y$ forming the observed $r_{XY}$ correlation.[3] Similarly, $|d_{XX}|$ and $|d_{YY}|$ provide the average time intervals separating X and Y used to form the $r_{XX}$ and $r_{YY}$ retest correlations. Given that correlations have been clearly demonstrated to be a function of their associated measurement interval, we suggest it can be useful to note the measurement interval explicitly within the notation of the correlation, as $r_{XY|d_{XY}|}$.

When the measurement intervals associated with the correlations in Spearman's formula can be constrained to a single common value $|d|$ – or more formally: when we cannot reject the hypothesis $|d_{XY}| = |d_{XX}| = |d_{YY}| = |d|$ – the equation can be re-represented in this form:

---

[2] The "at least" qualifier indicates that rater expected values (or 'true scores'; Lord & Novick, 1968) – will also differ meaningfully across rating conditions (e.g., fatigue, cognitive load, frequency of prior testing). Reliability estimates will also differ across rater populations (Borsboom & Mellenbergh, 2002; Lazarsfeld, 1959).

[3] We can also operationalize $t_{X_p}$ and $t_{Y_p}$ as providing the *order* items were presented to participants. For instance, $t_{X_p}$ = 3 and $t_{Y_p}$ = 40 indicates that X and Y were measured $d_{XY_p}$ = 37 items apart for participant *p*. This can be useful when X and Y are measured multiple times within a single survey (as with some of the present samples), as contemporary survey software (e.g., Qualtrics, SurveyMonkey) often does not readily provide records of the time in which every survey item was rated by the participant but can readily supply item order information.

**Equation 2.** $\rho_{\mathbb{XY}|d|} = \frac{\rho_{XY|d|}}{\sqrt{\rho_{XX|d|}\rho_{YY|d|}}}$

We will refer to the resulting parameter, $\rho_{\mathbb{XY}|d|}$, as a *matched-retest-adjusted correlation,* or more simply as a *retest-adjusted correlation*. The magnitude of $\rho_{\mathbb{XY}|d|}$ can be interpreted as a proportion indexing the extent to which the observed correlation between $X$ and $Y$ over interval $|d|$ is lower than the geometric mean correlation we would have observed by simply retesting $X$ and retesting $Y$ over that same interval. For instance, $\rho_{\mathbb{XY}|1day|}$ = .80 would indicate that the correlation between X and Y when measured one day apart is 80% the size of these tests' retest correlations over that interval.

**Strategy: Repeated measures studies.** It is much easier to form proper estimates of the $\rho_{\mathbb{XY}|d|}$ parameter shown in Equation 2 than it might initially seem. To ensure the measurements forming the numerators and denominators of the equation are formed over matching measurement intervals, we simply need to have $X$ and $Y$ be items embedded within a broader inventory (or instrument, test battery, survey) which is administered at least twice. We can notate each administration of the broader inventory as $m$, and the score from the $m$'th administration of some item $X$ as $X_m$. For instance, $X_1$ and $Y_1$ represent the scores from the first administrations of a pair of items, and $X_2$ and $Y_2$ represent the scores from their second administration.

In Table 1, we have provided hypothetical values for the estimated *score correlations* and *measurement intervals* of items X and Y which are contained within some broader inventory that has been administered to participants twice, an average of one week apart. When the inventory has been administered twice, this will result in the data structure represented in Table 1A, which will have four estimates of the correlation between $X$ and $Y$. First, there are two correlations which concern how their scores correlate when measured within the *same administration* of the inventory, $r_{X_1Y_1}$ and $r_{X_2Y_2}$; these can be averaged to estimate the items' *same-administration correlation*, or how highly the scores correlate when obtained in the same wave of a multi-wave study; $r_{XY|d_0|} = M\left(r_{X_1Y_1}, r_{X_2Y_2}\right)$. Second, there are two correlations which concern how their scores correlate when X and Y are measured in *different administrations* of the inventory, $r_{X_1Y_2}$ and $r_{X_2Y_1}$; these can be averaged to estimate their *different-*

*administration correlation*; $r_{XY|d_1|} = M\left(r_{X_1Y_2}, r_{X_2Y_1}\right)$. The $|d_0|$ and $|d_1|$ notation indicates the *same-administration* and *different-administration* measurement intervals are equivalent to a *lag-0* and *lag-1* interval in a time-series analysis, respectively (Stadnitski, 2020). The values given in Table 1A indicate the general expectation that in a two-wave study, the different-administration (lag-1) correlations should be smaller in magnitude than the same-administration (lag-0) correlations: $\left|r_{XY|d_1|}\right| < \left|r_{XY|d_0|}\right|$.

Further, the values in Table 1B indicate the general expectation that in a two-wave study, measurement intervals associated with all 'lag-1' correlations should be nearly equivalent. For instance, in Table 1B, $\left|d_{X_1Y_2}\right| = \left|d_{X_2Y_1}\right| = \left|d_{X_1X_2}\right| = \left|d_{Y_1Y_2}\right| = 1$ week. Crucially: this is *not* true for the same-administration correlations: whereas the $r_{XY|d_0|}$ values indicate how scores on X and Y will correlate when they are administered a couple items or minutes apart, we almost invariably have no idea how highly scores on X and Y would correlate with themselves if they had somehow been measured twice over the same interval. For instance, in Table 1B, the measurement interval separating 'same-administration' measurements of X and Y is 5 minutes; however the diagonals of this matrix reveal there are no corresponding retest correlations for X and Y over the corresponding 5-minute interval.

This indicates we *cannot* form appropriate estimates of Equation 2's retest-adjusted correlation from estimates of the same-administration (lag-0) correlations of X and Y; i.e., parameter $\hat{\rho}_{\mathbb{XY}|d_0|}$ cannot be correctly formed.[4] However, when we have two-wave data, we *can* correctly specify the parameter given in Equation 2 using solely the different-administration (lag-1) correlations. This is indicated below:

**Equation 3.** $\hat{\rho}_{\mathbb{XY}|d_1|} = \frac{M\left(r_{X_1Y_2|d_{X1Y2}|}, r_{X_2Y_1|d_{X2Y1}|}\right)}{\sqrt{r_{X_1X_2|d_{X1X2}|} r_{Y_1Y_2|d_{Y1Y2}|}}} = \frac{M\left(r_{X_1Y_2|d_1|}, r_{X_2Y_1|d_1|}\right)}{\sqrt{r_{X_1X_2|d_1|} r_{Y_1Y_2|d_1|}}} = \frac{r_{XY|d_1|}}{\sqrt{r_{XX|d_1|} r_{YY|d_1|}}}$

[4] It is actually possible to obtain appropriate estimates of the same-session diagonals for $r_{XX|d_0|}$ by administering X and Y twice within a broader inventory, with all items – *including both repetitions of X and Y* – presented in a randomized order to each participant. In such a design, some participants will occasionally rate item X immediately after rating the same item X. This design is expected to result in $|d_0| = \left|d_{X_1X_1}\right| = \left|d_{Y_1Y_1}\right| = \left|d_{X_1Y_1}\right|$, which would allow $\hat{\rho}_{\mathbb{XY}|d_0|}$ to be correctly formed. However we expect most researchers will view the design as too obnoxious for participants to complete, or otherwise an overly inefficient use of researcher and participant resources (Cronbach, 1951). It will also certainly increase reactivity effects to an almost maximal degree – where responses are impacted by the fact that many participants remember or actually *see* their previous rating to an item they are rating.

This formula is actually very close to a form of the reliability-adjusted formula suggested in Spearman's original 1904 article, in which all four of the XY correlations available when X and Y are measured twice are averaged to estimate $r_{XY}$ in the numerator.[5] However, to ensure that the measurement intervals are equivalent in the numerator and denominator as required to operationalize Equation 2, *only* the correlations linking scores of X and Y across *different administrations of the inventory* – i.e., the lag-1 or $|d_1|$ interval – should be used within the computation. That is: the estimates of how X and Y correlate when measured in the same session (i.e., $r_{X_mY_m}$; or estimates $r_{X_1Y_1}$ and $r_{X_2Y_2}$ in a two-wave study) should be excluded.

Table 1 shows the simplest case where two items X and Y are administered twice within some inventory. For instance, applying Equation 3 to the values in Table 1A, we would estimate the $\hat{\rho}_{\mathbb{XY}|d_1|}$ correlation between items X and Y to be $M(.45, .45)/\sqrt{.50 \times .60}$ = .818. However, the calculations given in Equation 2 or 3 can be easily done for *all* item-pairs within an inventory. When this is done, it will create a *retest-adjusted correlation matrix,* which can be regarded as estimating the proportion that each inter-item correlation is smaller than the associated items' retest correlations over the measurement internval $|d|$ for all pairs of items within the inventory.

**Estimating Semantic Similarity**

Judgments about the level of semantic similarity between verbal statements almost necessarily involve judgments from human judges (Nicholls et al., 1982).[6] For instance, item-pairs of the sort given in the opening paragraph are often identified by researchers 'eyeballing' the items and listing item-pairs

[5] Specifically, Spearman provides this equation in his original 1904 article (see page 90). Note that we have altered his variables *p* and *q* to match the present *X* and *Y*, and renotated what he referred to as the 'true correlation' between the variables to the present $\hat{\rho}_{\mathbb{XY}}$:

$$\hat{\rho}_{\mathbb{XY}} = \frac{\left(\boldsymbol{r_{X_1Y_1}} + r_{X_1Y_2} + r_{X_2Y_1} + \boldsymbol{r_{X_2Y_2}}\right)}{4\sqrt{r_{X_1X_2}r_{Y_1Y_2}}}$$

Spearman's equation only differs from the present Equation 3 by including the bolded "same administration" estimates of the X-Y correlation within the values averaged to form the numerator.

[6] At the very least: such judgments serve as a particularly valuable criterion for training and evaluating the performance of *natural language processing (NLP)* algorithms, which are quickly improving in their ability to approximate human judgments of the semantic similarity of verbal text (Arnulf & Larsen, 2021; Christensen & Kenett, 2021; Cutler & Condon, 2022; Rosenbusch et al., 2020).

they have determined to be semantically similar (e.g., Banks et al., 2016; Christensen et al., 2020; Mõttus, 2016; Newman et al., 2016; Nicholls et al., 1982). However, there are problems with relying on a single person's semantic similarity judgments. As Le and colleagues (2010) note, "the implicit assumption is often that if researchers can make a conceptual, theoretical, or logical distinction between constructs then this distinction will also exist in the minds of employees or survey respondents… This assumption may not hold" (p. 113). To which we add the reverse problem can also occur: researchers may judge different items or constructs to be redundant when others are able to make reliable distinctions between them.

However, the tendencies for some people to overly "lump" or "split" different verbal statements does not in turn make the use of semantic similarity judgments an intractably subjective affair. The semantic similarity of two items can be operationalized by asking multiple raters the extent to which the items 'mean the same thing,' and then reporting the average of these judgments across raters. As in person perception research, the aggregation of multiple ratings helps to reduce the role of idiosyncratic biases (Hofstee, 1994; Paunonen, 1984). Semantic similarity ratings are rarely collected in personality and social psychological research (with scattered exceptions: e.g., Block et al., 1979; Shweder & D'Andrade, 1979; Weidman et al., 2018), however they are collected with greater regularity in cognitive and linguistics research (e.g., Chaffin & Herrmann, 1984; Miller & Charles, 1991; Resnik, 1999; Rubenstein & Goodenough, 1965; Whitten et al., 1979).

## Do Retest-Adjusted Correlations Outperform Raw-Score Correlations as Indicators of Semantic Similarity?

As we have detailed, it is possible to estimate the *information similarity* of items using raw-score correlations or retest-adjusted correlations, and to estimate the *semantic similarity* of items using aggregated rater judgments. Although the information similarity and semantic similarity of items can thus be operationalized independently, we should expect these estimates to track with one another. That is, it would be extremely surprising to find that estimates of the correlations between items had *no* association with judgments of their semantic similarity. However, we should also expect *better* indices of the

information similarity of items to *better* track their judged semantic similarity. This understanding underlies the hypotheses we will explore here.

As noted above, score reliability should decrease as the time interval $|d|$ increases; this should result in raw-score correlations between items becoming worse predictors of their judged semantic similarity as the time interval $|d|$ separating the measurements of the items increases. This leads to the first hypothesis of the study:

> **H1**: As the time interval $|d|$ separating measurements of $X$ and $Y$ increases, raw-score $r_{XY|d|}$ correlation estimates will become worse predictors of their perceived semantic similarity.
>
> Restated in the terms of a two-wave study: same-administration (lag-0) inter-item correlations will be better predictors of the items' perceived semantic similarity than corresponding different-administration (lag-1) correlations.
>
> Formally, H1: $q\left(r_{XY|d_0|}, SemSim_{XY}\right) > q\left(r_{XY|d_1|}, SemSim_{XY}\right)$

Note also that throughout this manuscript, we will use $q$ to denote correlations estimated at the '*between-item-pair*' level of analysis, and will restrict $r$ to indicate *between-person* correlations (Cattell, 1952; Stephenson, 1953). The central study hypotheses can thus be understood as evaluating the ability of different between-person correlational indices of information similarity to predict variation in semantic similarity judgments *across item-pairs*.

The remaining hypotheses concern whether adjusting for measurement unreliability actually results in better indicators of whether two items 'measure the same thing':

> **H2**: If sample sizes are sufficiently large, then retest-adjusted correlations ($\hat{\rho}_{\mathbb{XY}|d|}$; Equations 2 and 3) will outperform raw-score correlations over the same measurement interval $|d|$ as predictors of semantic similarity ratings.
>
> Formally, H2: $q\left(\hat{\rho}_{\mathbb{XY}|d_1|}, SemSim_{XY}\right) > q\left(r_{XY|d_1|}, SemSim_{XY}\right)$

The qualifier that this is only expected '*if sample sizes are sufficiently large'* is important, as it recognizes that empirically-estimated correlations can fluctuate dramatically in small samples (Carter et al., 2019;

Cohen, 1992), and consequently the three components involved in forming $\hat{\rho}_{\mathbb{XY}|d|}$ estimates via (Equation 2) may be so unstable in small samples that they infuse more unreliable variance than they remove.

Note that this hypothesis most directly evaluates the conventional understanding that reliability adjustments should outperform raw-score correlations as indices of whether two tests 'measure the same thing' (Le et al., 2010; Spearman, 1910). Specifically: it evaluates whether dividing $r_{XY|d_1|}$ estimates by the associated tests' retest reliabilities over that interval results in better indicators of semantic similarity than the $r_{XY|d_1|}$ estimates alone. Finally, we will explore a closely related variant of the last hypothesis:

> **H3**: If sample sizes are sufficiently large, then retest-adjusted correlations ($\hat{\rho}_{\mathbb{XY}|d_1|}$) will outperform 'same-administration' raw-score correlations ($r_{XY|d_0|}$) as predictors of semantic similarity ratings.
>
> Formally, H3: $q\left(\hat{\rho}_{\mathbb{XY}|d_1|}, SemSim_{XY}\right) > q\left(r_{XY|d_0|}, SemSim_{XY}\right)$

Whereas H2 serves as the most direct test of the relative value of retest-adjusted correlations over corresponding unadjusted correlations, H3 is likely the hypothesis with greater *practical* significance. Many researchers already use unadjusted same-administration inter-item correlations, $r_{XY|d_0|}$, to support determinations of whether two or more items 'mean the same thing' (Block, 1961; Cattell & Tsujioka, 1964; Weidman et al., 2018; Wiggins, 2003; Wood et al., 2010). However, as we have detailed, it is not possible to properly estimate $\hat{\rho}_{\mathbb{XY}|d_0|}$ from ordinary same-administration data. H3 concerns the more practical question of whether there is value in making the two administrations of the inventory necessary to form retest-adjusted correlations (Equation 3), given the greater expenditure of participant and researcher resources generally required to administer an inventory twice.

It is also useful to consider what it would mean for H1, H2, and H3 to *all* be supported – and perhaps H1 and H3 in particular. Support for H1 would mean that inter-item correlations formed by measuring X and Y over the longer interval $|d_1|$ are inferior indicators of item similarity relative to their intercorrelation over the shorter interval $|d_0|$. We imagine most researchers would find this unsurprising. However, finding support for H3 would indicate a sort of leap-frogging: dividing the *inferior* different-

administration indices of item similarity by their retest correlations over that interval creates indices which are *superior* to the same-administration correlations more typically employed index item similarity or redundancy.

## Study Overview

Adjustments for unreliability were originally proposed as a means of improving correlations as indices of whether two variables '*measure* the same thing' (Spearman, 1904, 1910), and continue to be used for this function (Le et al., 2010; Schmidt, 2010; Shaffer et al., 2016). Given the numerous concerns about how to appropriately estimate reliability, it is noteworthy that there appears to have never been an *empirical* demonstration – using real participant responses to real measures – that reliability adjustments actually improve the extent to which correlations track judgments of whether the two measures '*mean* the same thing'.

To clarify this point: although there are certainly *simulations* and *algebraic models* indicating that a variety of reliability adjustments should result in better indices of test similarity or redundancy (e.g., Charles, 2005; Le et al., 2009; Schmidt et al., 2013), these demonstrations *assume* or *specify* conditions whereby the reliability parameter is the correct one. Many of these conditions or assumptions – such as unidimensionality, uncorrelated errors, and independence of measurements – are typically unrealistic in real data (e.g., Epskamp et al., 2017; Fried et al., 2016; McDonald, 1999). But in the 115+ years since Spearman first proposed reliability adjustments, *we are aware of no empirical investigations that have shown that adjusting correlations by reliability estimates improves upon raw-score correlations toward indexing whether tests 'measure the same thing.'* A major aim of this investigation is to show that this can and should be treated as an empirical question – particularly given the understanding that adjustments for unreliability frequently introduce as much systematic error into estimates of the associations between variables as they remove (LeBreton et al., 2014; Sackett, 2014; Sackett et al., 2021).

To explore this question, we thus examine our hypotheses regarding the relative value of raw-score correlations compare to retest-adjusted correlations as predictors of the rated semantic similarity of item-pairs of items in four different inventories: the Positive and Negative Affect Schedule – Extended

Form (PANAS-X; Watson & Clark, 1999), the Big-Five Inventory-2 (Soto & John, 2017), the Inventory of Individual Differences in the Lexicon (Wood et al., 2010), and a combined set of two short measures of the Dark Triad (Jonason & Webster, 2010; Jones & Paulhus, 2014).

For each inventory, we conducted a common set of procedures and analyses to explore the critical hypotheses, which is summarized in Figure 1. First, we identified one or more datasets in which respondents had rated a particular inventory twice. In each case, we estimated the extent to which all items in the inventory correlated with one another when measured within the same testing session (i.e., a 'lag-0' interval; $r_{XY|d_0|}$), across two testing sessions (i.e., a 'lag-1' interval; $r_{XY|d_1|}$), and as indexed by retest-adjusted correlations ($\hat{\rho}_{\mathbb{XY}|d_1|}$; Equation 3). Second, we then identified a set of item-pairs consisting of an equal number of item-pairs estimated to have the highest same-administration raw-score correlations ($r_{XY|d_0|}$) and the highest retest-adjusted correlation ($\hat{\rho}_{\mathbb{XY}|d_1|}$). Third, for each of the resulting item sets we then obtained ratings from about ten or more raters of the semantic similarity of each item-pair. Fourth, to test the central study hypotheses (H1-3), we examined how each index of the association between variables fare in predicting the judged semantic similarity of the item-pairs, as rated by an independent sample of raters.

## Method

We will describe how self-ratings and semantic similarity ratings of the items within the inventories were collected separately for each sample. This study was not preregistered, and was conducted under University of Alabama IRB study #8370. Note that the estimated correlations between all possible item-pairs within the inventory, the rated semantic similarity of selected item-pairs within the inventory, and code to replicate the present analyses, are available at https://osf.io/vp6kr/.

### Self-Rating Participants and Procedure

Much of the data used for these analyses came from study designs in which participants completed the inventory twice within a single testing session, where the repeated administrations were

separated by the administration of additional measures (Lowman et al., 2018). Following screening procedures recommended by Wood, Harms, Lowman, and DeSimone (2017), participants in some of the samples were excluded by failing a *speed screen,* where they completed either administration of the inventory at a rate faster than 1 second-per-item, or a *consistency screen,* in which the profile correlation of their responses across the two administrations was estimated at a level below $q < .25$. The speed screen was only used in samples where response time data was available. The consistency screen was only used when respondents completed the inventory twice within a single testing session, as less consistent responding is expected to become a less valid indicator of careless responding over longer retest intervals (Henry et al., 2022).

***Positive and Negative Affect Schedule – Extended Form (PANAS-X; Watson & Clark, 1999)***

The PANAS-X is designed to measure 11 distinct types of emotion with subscales ranging between three and eight items each. The PANAS-X can be used flexibly to measure self-reports of momentary mood states or more general emotional tendencies.

**MTurk-1 Sample.** We utilized a dataset originally described by Wood et al. (2017), in which participants were recruited from Mechanical Turk and were randomly assigned to complete either the PANAS-X, the BFI-2, or the IIDL twice, with administrations of the repeated inventory separated by administering 138 additional items related to personality and workplace attitudes and behaviors. These additional materials were completed in approximately 12 minutes for the average participant. Participants received $1.25 for completing this survey. We will refer to this as the MTurk-1 study.

An initial 142 participants were assigned to complete the PANAS-X twice within the MTurk-1 study. At both administrations, participants were instructed to "Indicate to what extent you are feeling like this in general" (1 = *very slightly or not at all* to 5=*extremely*). Speed and consistency screens resulted in a final sample size of 115 participants; $M(SD)_{age}$ = 35.6(10.2); 57 (49%) female.

**MTurk-2 Sample.** 160 participants recruited from MTurk completed the PANAS-X twice over an interval of about 10 minutes for the average participant. The item order was not randomized. Participants rated the items under the instruction "Indicate to what extent you have felt this way during

the past few weeks." Speed and consistency screens resulted in a final sample size of 154 participants; $M(SD)_{age}$ = 37.4(11.6); 81 (53%) female.

**Student Sample.** 88 undergraduate students provided rated the PANAS-X items twice over approximately a 20-25 minute retest interval, in which participants completed 206 other items. The item order was not randomized. Participants rated the items under the instruction "Indicate *to what extent you have felt this way TODAY*." Speed and inconsistency screens resulted in a final sample size of 83 participants; $M(SD)_{age}$ = 22.7(4.5), 53 (64%) female.

***Big Five Inventory-2 (BFI-2; Soto & John, 2016)***

The BFI-2 is an inventory designed to measure three major facets of each of the Big Five personality dimensions with four items apiece, resulting in a 60-item inventory. In both samples, participants rated the BFI-2 items under the instruction "Please rate the extent to which you feel each characteristic describes how you see yourself" on a scale from "Very Uncharacteristic" (1) to "Very Characteristic" (5), with items presented in a randomized order for each participant and at each survey administration.

**MTurk-1 Sample.** An initial 128 participants were randomly assigned to rate the BFI-2 twice with the larger MTurk-1 study. Speed and consistency exclusions resulted in a final sample consisting of 110 participants, $M(SD)_{age}$ = 36.1(10.4), 46% female.

**BFI-2 Student Sample**. 470 undergraduate students completed the BFI-2 twice for a class research participation credit over a retest interval of approximately eight weeks (Soto & John, 2017).

***Inventory of Individual Differences in the Lexicon (IIDL; Wood, Nye, & Saucier, 2010)***

Each item within the IIDL consists of a pair of fairly synonymous person-descriptor adjective (e.g., "dependable, reliable") reflecting larger clusters of highly correlated terms found within the English language. Participants in both samples utilized here completed a 84-item set of the IIDL consisting of the standard 61-item set plus an additional 23 items given in Appendix A of the original article (Wood et al., 2010). However, analyses were limited to the 79 items that were administered in the same manner across both samples.

**MTurk-1 sample**. An initial 140 participants were randomly assigned to rate the IIDL twice within the larger MTurk-1 study. Speed and consistency exclusions resulted in a final sample consisting of 118 participants, $M(SD)_{age}$ = 35.2(10.7); 53% female.

**Student Sample**: A sample of undergraduate students at a university in Singapore completed the IIDL twice for course credit. Between administrations, participants rated 110 other items related to emotion and well-being (Wood et al., 2018). Following the response inconsistency screen, this resulted in a final sample contained 78 participants; $M(SD)_{age}$ = 20.6(1.6); 77% female. The order of the IIDL items was randomized for each participant and for each administration of the inventory.

***The Short Dark Triad (SDT; Jones & Paulhus, 2014) and Dirty Dozen (DD; Jonason & Webster, 2010)***

Both the SDT and DD inventories are designed to assess the 'Dark Triad' dimensions of narcissism, Machiavellianism, and psychopathy (Paulhus & Williams, 2002) through short scales. An initial 314 participants were recruited from MTurk and completed both the 27-item SDT and the 12-item DD measures. The 39 total items were rated together in a single set in which items were presented in a randomized order for each administration of the survey, and administrations were separated by a retest interval approximately 20 minutes apart (Wood et al., 2017). All items were rated on a scale from "Strongly Disagree" (1) to "Strongly Agree" (5) scale. Participants exclusions by speed and consistency screens resulted in a final sample of 242 participants, $M(SD)_{age}$ = 33.9(10.0); 51% female.

**Analyses**

***Indices of Information Similarity of Item-Pairs***

In all samples, we estimated several different indices of the inter-item associations within the focal inventory. Note that when these coefficients are calculated for each unique pair of items in an inventory, the number of unique item-pairs equals $\left(N_j^2 - N_j\right)/2$, where $N_j$ equals the number of items in the inventory. Consequently: the 60-item PANAS-X and BFI-2 inventories both have a total of 1770 distinct item-pairs, whereas the 79 items of the IIDL set had 3081 total item-pairs, and the 39 items of the

combined SDT and DD inventories had 741 total item-pairs. Note that for all these parameters, the item that is assigned as "item X" versus "item Y" is arbitrary.

**Same-administration raw-score correlation; $r_{XY|d_0|}$.** First, we estimated the *same-administration* (or *lag-0*) correlation between items X and Y. As all samples consisted of participants who rated an inventory twice, the two estimates of the same-administration inter-item correlation were averaged; i.e., $r_{XY|d_0|} = (r_{X_1Y_1} + r_{X_2Y_2})/2$.

**Different-administration raw-score correlation; $r_{XY|d_1|}$.** Second, we estimated the *different-administration* (or *lag-1*) correlation between item X and Y. The two estimates of the inter-item correlation over different administrations were averaged; i.e., $r_{XY|d_1|} = (r_{X_1Y_2} + r_{X_2Y_1})/2$.

**Average retest correlation.** The *average retest correlation* of the item pairs was estimated as the *geometric mean of the retest correlations of X and Y over the 'lag-1' interval*; i.e., $\sqrt{r_{XX|d_1|}r_{YY|d_1|}}$ or $\sqrt{r_{X_1X_2}r_{Y_1Y_2}}$.

**Retest-adjusted correlation.** Finally, we estimated the *retest-adjusted correlation* over the lag-1 interval, $\hat{\rho}_{\mathbb{XY}|d_1|}$. Following Equations 2 and 3, this was done by dividing estimates of the different-administration correlation between items X and Y by their average retest correlation over this interval; i.e., $r_{XY|d_1|}/\sqrt{r_{XX|d_1|}r_{YY|d_1|}}$.

**Cross-sample weighted averages.** As there were multiple samples used to estimate information-similarity indices for three of the four inventories, we computed a cross-sample weighted average, where estimates of some index $V_{XY}$ were weighted by their associated sample size, using an equation used in meta-analysis (Schmidt & Hunter, 2014, Equation 3.1):

$$\textbf{Equation 4. } V_{XY} = \frac{\Sigma(N_k \times V_{XY_k})}{\Sigma N_k}$$

Where $N_k$ equals the number of participants within sample *k,* and $V_{XY_k}$ indicates the correlational index of similarity estimated for that sample. This equation was used separately for all of the parameters discussed above, except for cross-sample estimates of the retest-adjusted correlations, $\hat{\rho}_{\mathbb{XY}|d_1|}$, which we computed

by dividing the cross-sample weighted averages of the different-administration correlation and average retest correlations.

***Indexing the Semantic Similarity of Item-Pairs***

For each inventory, only a small subset of all possible item-pairs contained within the inventory were rated for their semantic similarity. We were particularly interested in the relative ability of very high raw-score and retest-adjusted correlations to indicate the semantic redundancy of items. Consequently, for each inventory exampled we identified an equal number of item-pairs estimated to have the highest (1) same-administration raw-score correlations ($r_{XY|d_0|}$) and (2) retest-adjusted correlations ($\hat{\rho}_{\mathbb{XY}|d_1|}$). To do this, we ranked each item-pair by the size of its same-administration raw-score correlation, and separately by the size of its retest-adjusted correlation; minimum ranks of 1 were given to the item-pairs with the maximum $r_{XY|d_0|}$ and $\hat{\rho}_{\mathbb{XY}|d_1|}$ values. An equal number of the minimum ranking item-pairs were taken from each list to form a set of 100 or 101 item-pairs to be rated for semantic similarity.[7]

The item-pairs identified from each inventory were then rated for their perceived semantic similarity. For the PANAS-X, raters were first presented with these instructions:

> In the following section you will be presented two words or phrases. You will be asked to use the scale below to indicate how much you see the words or phrases as having the *same* vs. *different* meanings.
>
> Instructions: To what degree are these two words or phrases **similar in what they mean** when used to describe **how someone feels**?
>
> 0 – Have **completely different** meanings
> 1 – Have **slightly similar** meanings
> 2 – Have **fairly similar** meanings
> 3 – Have **very similar** meanings
> 4 – Have **essentially the same** meaning

[7] If the number of item-pairs found in both lists was odd, 101 item-pairs were rated to ensure an equal number of pairs were drawn from each list. For instance, 100 item-pairs would be selected if there were 50 common to both lists and 25 unique to each, whereas 101 item-pairs would be selected if there were 51 common to both lists and 25 unique to each.

For the BFI-2, IIDL, SDT/DT inventories, participants read the same passage except the "Instructions" sentence shown above was modified to read "Instructions: To what degree are the words or phrase on the left similar in what they mean to the words or phrase on the right when used to describe someone?"

The item-pairs were presented to each participant in a randomized order. Additionally, the order in which participants saw the two items within the pair was also randomized (e.g., whether participants saw the item pair as "*cheerful : happy*" or "*happy : cheerful*").

Raters were selected through a convenience sample of research assistants and acquaintances. We excluded raters whose scores loaded below .30 on the first principal axis factor (equivalent to showing less than $q$ = .09 average agreement with other raters). This exclusion was done as it indicated careless responding which would otherwise inhibit the ability for mean semantic similarity ratings to approach the scale maximum (indicating that items within a pair were consensually judged to "mean the same thing").

For the PANAS-X set, 22 raters were surveyed, and none were excluded. The average inter-rater agreement was $q_{pp'}$ = .52. Using coefficient alpha, the estimated internal consistency of average scores across 22 raters was $\alpha$ = .96, which can be interpreted as the expected correlation of the resulting average scores with averages formed from sampling a new 22 raters.

For the BFI-2 set: 10 raters were initially surveyed, and one was excluded. The average inter-rater agreement was $q_{pp'}$ = .45; the estimated internal consistency of the 9-rater average scores was α = .88. For the IIDL set: 11 raters were initially surveyed, and two were excluded. The average inter-rater agreement was $q_{pp'}$ = .45; the estimated internal consistency of the 9-rater average scores was α = .87. For the SDT/DD set: 13 raters were initially surveyed, and one was excluded. The average inter-rater agreement was $q_{pp'}$ = .42; the estimated internal consistency of the 12-rater average scores was α = .89.

## Results

Table 2 includes a listing of which five item-pairs were estimated to have (1) the highest same-administration correlations ($r_{XY|d_0|}$), (2) the highest retest-adjusted correlations ($\hat{\rho}_{\mathbb{XY}|d_1|}$), and (3) the highest judged semantic similarity, separately for each inventory. We also report the different-

administration ($r_{XY|d_1|}$) and average retest ($\sqrt{r_{XX}r_{YY|d_1|}}$) for these item-pairs. Tables 3.1-4 provides means, standard deviations, and inter-correlations across the 100 or 101 item-pair subsets for these five indices separately for each inventory.

We have additionally provided scatterplots detailing how the same-administration, different administration, and retest-adjusted correlations are associated with semantic similarity judgments separately for each of the four inventories (Figure 2). Note that in these figures, we have added a vertical line at $X = .90$, as this is a threshold sometimes used to support arguments that tests X and Y are informationally redundant (e.g., John & Benet-Martínez, 2000; Judge & Bono, 2001; Le et al., 2009). And we have added a horizontal line at $Y = 3.5$ – a threshold that could only be obtained by at least half of raters judging the two items to "have essentially the same meaning" via the scale used here. In other words, values of $X \geq .90$ and $Y \geq 3.5$ can be regarded as indicating that the items within the pair passed thresholds for being regarded as *informationally redundant* or *semantically redundant,* respectively.

We continue by briefly describing tests of the central hypotheses (H1-3) separately for each of the four inventories. We then conducted a mega-analysis (Beck & Jackson, 2022; Burke et al., 2017; Curran & Hussong, 2009) in which data for the separate inventories were combined into a single dataset to examine the hypotheses more generally. Finally, we discuss more specific additional themes suggested from the analysis.

**Inventory-Specific Relations between Correlational Indices with Perceived Semantic Similarity**

For each inventory, the associations between the (1) same-administration, (2) different-administration, and (3) retest-adjusted correlations of item-pairs with their judged semantic similarity were tested for statistical significance using Steiger's (1980) test of dependent correlations.

***PANAS-X, BFI-2, and IIDL Results***

Tables 3.1 through 3.3 indicate that the PANAS-X, BFI-2, and IIDL results showed identical patterns of support for the H1, H2, and H3 hypotheses. For each of these inventories, we found:

**Significant support for H1.** The different-administration inter-item correlations ($r_{XY|d_1|}$) were significantly worse predictors of the rated semantic similarity of the item-pair ratings than same-administration inter-item correlations, $r_{XY|d_0|}$.

**Significant support for H2.** We found that retest-adjusted correlations ($\hat{\rho}_{\mathbb{XY}|d_1|}$) were significantly more highly associated with the judged semantic similarity of item pairs than were raw-score different-administration correlations ($r_{XY|d_1|}$).

**Insufficient support for H3.** Although retest-adjusted correlations ($\hat{\rho}_{\mathbb{XY}|d_1|}$) were more highly associated with the judged semantic similarity of item-pairs than were raw-score same-administration correlations ($r_{XY|d_0|}$) in all three of these inventories, none of these differences reached statistical significance.

***SDT/DD Results***

As shown in Table 3.4, we observed a slightly different pattern for the set formed from short Dark Triad measures than the pattern found for the PANAS-X, BFI-2, and IIDL instruments. Surprisingly, the different-administration correlations, $r_{XY|d_1|}$, were slightly more predictive of semantic similarity judgments than were same-administration correlations, $r_{XY|d_0|}$ ($q$ =.627 vs. .583; Z = 1.34, $p$ = .18), but were slightly less predictive than retest-adjusted correlations, $\hat{\rho}_{\mathbb{XY}|d_1|}$ ($q$ = .627 vs. .659; $Z$ = 1.09, $p$ = .28), although neither of these differences reached statistical significance. However, retest-adjusted correlations were significantly better predictors of judged semantic similarity than same-administration raw-score correlations ($q$ = .659 vs. .583; $Z$ = 2.06, $p$ < .05).

**Mega-Analytic Relations between Correlational Indices and Perceived Semantic Similarity**

We combined the data from the four inventories into a single dataset to increase the statistical power to explore the central hypotheses. This resulted in a dataset with 402 item-pairs with both information and semantic similarity indices. For the focal analyses, we examined how these indices were associated with one another while controlling for the inventory through dummy-code variables, resulting in $df$ = 397 item-pairs associated with the reported correlations, which are given in Table 3.5.

As shown in the final row of Table 3.5, all study hypotheses were supported. Supporting H1: an item-pair's same-administration correlation ($r_{XY|d_0|}$) was a significantly better predictor than its different-administration correlation ($r_{XY|d_1|}$) of the items' judged semantic similarity ($q$ = .52 vs. .46; $Z$ = 3.70, $p$ < .05). Supporting H2: an item-pair's retest-adjusted correlation ($\hat{\rho}_{\mathbb{XY}|d_1|}$) was a significantly better predictor than its raw-score correlation over the different-administration interval ($r_{XY|d_1|}$) of the items' judged semantic similarity ($q$ = .58 vs. 46; $Z$ = 5.75, $p$ < .05). And supporting H3: an item-pair's retest-adjusted correlation ($\hat{\rho}_{\mathbb{XY}|d_1|}$) was a significantly better predictor than its raw-score same-administration correlation ($r_{XY|d_0|}$) of the items' judged semantic similarity ($q$ = .58 vs. 52; $Z$ = 2.83, $p$ < .05).

**Additional Themes**

***No Retest-Adjusted Correlations Exceeded 1.0***

Somewhat incredibly, across the combined 7362 distinct item-pairs examined in the present analyses across four inventories, estimates of the retest-adjusted correlation $\hat{\rho}_{\mathbb{XY}|d_1|}$ for the item-pair *never* fell outside the range of $[-1,1]$ – as sometimes occurs when adjusting correlations by reliability estimates (LeBreton et al., 2014; Sackett, 2014; Sackett et al., 2021). Nonetheless, numerous item-pairs *approached* this value. As shown in Table 2, the highest value was estimated for the BFI-2 item-pair "*Often feels sad : Tends to feel depressed, blue*" $\hat{\rho}_{\mathbb{XY}|d_1|}$ = .995.

The absence of $\hat{\rho}_{\mathbb{XY}|d_1|}$ estimates exceeding 1.0 was certainly due somewhat to luck – for instance, this BFI-2 item-pair was indeed indexed to have an $\hat{\rho}_{\mathbb{XY}|d_1|}$ estimate above 1.0 in one of the two subsamples prior to computing cross-sample averages. Nonetheless, this suggests retest-adjusted correlations may be more resistant to forming 'out-of-boundary' correlations than other common reliability estimators.

***Indications of 'Necessity' Relationships between Information and Semantic Similarity***

As detailed by Dul (2016), the signature that some threshold level of variable X is *necessary* for high levels of another variable Y is indicated by a scatterplot of the X-Y relationship in which some area of the *upper-left* quadrant is devoid of observations. As seen in Figures 2.1C-2.4C, there were indications

that some parts of the upper-left quadrants of the scatterplots linking retest-adjusted correlations to semantic similarity judgments were indeed empty in this manner. For instance, across the four inventories, the lowest $\hat{\rho}_{\mathbb{XY}|d_1|}$ value for any item-pair consensually judged to be "very similar in meaning" ($M_{SemSim} \geq 3$ on the present scale) was found for the SDT/DD item-pair "*It's not wise to tell your secrets : There are things you should hide from other people to preserve your reputation*" which showed a retest-adjusted correlation of $\hat{\rho}_{\mathbb{XY}|d_1|} = .69$ (Row #38 in Table 2). Although 230 of the 402 of the item-pairs included in this analysis were estimated to have $\hat{\rho}_{\mathbb{XY}[D_1]}$ values below .68, *none* of these pairs (0%) were consensually judged to be even "very similar in meaning" by the average rater whereas 18 of the remaining 172 items exceeding a $\hat{\rho}_{\mathbb{XY}|d_1|} > .68$ magnitude (10.5%) crossed this level of judged semantic similarity. Similarly, of the 387 items estimated to have $\hat{\rho}_{\mathbb{XY}[D_1]}$ values below .95, *none* of these pairs (0%) achieved a mean semantic similarity judgment reaching 3.5 – which we have suggested as sufficient evidence that the items within the pair were consensually judged to "mean the same thing" by raters on the present scale. In contrast 2 of the 15 items estimated to have values *exceeding* a $\hat{\rho}_{\mathbb{XY}|d_1|} > .95$ (13.3%) reached this threshold (specifically, the PANAS-X item-pairs *frightened:scared* and *afraid:scared*).

This indicates that certain levels of information similarity may be *necessary* to expect high levels of semantic similarity. We propose that researchers may test the hypotheses that $\rho_{\mathbb{XY}|d_1|}$ values exceeding .70 may be necessary for two terms to be consensually judged to be "very similar in meaning" (i.e., $M \geq 3.0$ by the present scale) whereas $\rho_{\mathbb{XY}|d_1|}$ values exceeding .90 may be necessary for two terms to be consensually judged as "meaning the same thing" (i.e., $M \geq 3.5$ by the present scale).

***Indications of 'Sufficiency' Relationships between Information and Semantic Similarity***

As detailed by Dul (2016), a signature that some threshold level of X is *sufficient* to expect certain levels of Y is a scatterplot of the X-Y relationship in which some area of the *lower-right* quadrant is devoid of observations. As seen in Figures 2.1C-2.4C, there were also indications that some parts of the lower-right quadrants of the scatterplots linking retest-adjusted correlations to semantic similarity

judgments were largely devoid of observations in this manner. For instance, whereas 315 of the 387 item-pairs estimated to have $\hat{\rho}_{\mathbb{XY}|d_1|}$ values below .90 (81%) failed to reach a mean semantic similarity rating of 2.0 – i.e., they were judged to not be even "fairly similar in meaning" – only 2 of the 15 items with values exceeding $\hat{\rho}_{\mathbb{XY}|d_1|} > .90$ (13%) failed to cross this threshold (specifically, the PANAS-X item-pairs *alone:lonely* and *enthusiastic:lively*).

This indicates that certain levels of information similarity may be *sufficient* to expect at least threshold levels of perceived semantic similarity. We propose that future researchers may attempt to formally test the hypotheses that $\rho_{\mathbb{XY}|d_1|}$ values exceeding .90 may be *sufficient* for the associated items to be consensually judged to be at least "fairly similar in meaning" (i.e., $M \geq 2.0$ using the present scale).

## General Discussion

Using data from four inventories, we demonstrate how the information similarity of items can be indexed by *retest-adjusted correlations,* $\hat{\rho}_{\mathbb{XY}|d|}$*,* in which the estimated correlation between two items $X$ and $Y$ over a particular measurement interval $|d|$ is divided by the average retest correlation of those items over that interval (Equations 2 and 3). Most crucially, we found that retest-adjusted correlations outperform raw-score correlations as predictors of the consensually judged semantic similarity of the associated items.

### The Information and Semantic Similarity of Item-Pairs is Conceptually and Empirically Distinct

The bivariate relationships we observed between information and semantic similarity estimates, shown in Figure 2, indicate there is considerable room for two items to be indexed as providing highly similar information – i.e., as having high $\hat{\rho}_{\mathbb{XY}|d|}$ estimates, or as having fairly negligible reliable specific variance – while being understood as semantically distinct. This has important implications for how reliability adjustments should be used to help adjudicate questions of construct proliferation or redundancy.

First, it is commonly suggested that reliability-adjusted correlations exceeding .90 – and sometimes even exceeding lower thresholds – can be interpreted as indicating that the two tests "measure

the same thing" (John & Benet-Martínez, 2000; Le et al., 2010; Shaffer et al., 2016). However, as illustrated in Figure 2, we only found 2 of the 15 item-pairs with retest-adjusted correlations exceeding .90 to be consensually judged to "mean the same thing" by judges. It may be that many of the remaining relationships are understood by people as having strong *functional* relationships – where X is a strong and regular cause of Y or vice versa – rather than as representing a tautology. As one example: the SDT/DD items "I like to use clever manipulation to get my way," and "I tend to manipulate others to get my way" were estimated to have a retest-adjusted correlation of $\hat{\rho}_{\mathbb{XY}|d_1|}$ = .91 but a semantic similarity level (3.3) falling under our threshold for regarding them as being consensually judged to "mean the same thing" (i.e., $M(SemSim_{XY}) \geq 3.5$). For this particular item-pair, a range of psychological models readily suggest that *liking* an activity can very strongly affect one's *tendency to do* that activity, while *liking* and the *tendency to do* that activity are nonetheless conceptually distinct (e.g., Ajzen, 1991; Bandura, 1999; Dweck, 2017; Feather, 1982; Mischel & Shoda, 1995; Wilt & Revelle, 2015; Wood et al., 2015).

Second, retest correlations are likely to produce systematically *higher* estimates of the reliability of test scores than the internal consistency estimates typically used for this purpose (Henry et al., 2022; Lowman et al., 2018; McCrae et al., 2011; Wood et al., 2018). This is important as Equation 1 shows that dividing by *higher* reliability estimates will tend to produce *lower* estimates of reliability-adjusted correlations. If even *very high* levels of a more *conservative* index of how tests correlate after adjusting for unreliability is insufficient for establishing that two tests "measure the same thing," this would indicate that the many investigations which have considered reliability-adjusted correlations exceeding .80 or .90 to provide sufficient evidence of scale or construct redundancy (Banks et al., 2016; Credé et al., 2017; Harrison et al., 2006; Judge & Bono, 2001; Newman et al., 2010; Shaffer et al., 2016) are premature and need to be reevaluated. More generally, the low frequency with which item-pairs were estimated to cross reasonable thresholds of being *either* informationally redundant ($\hat{\rho}_{\mathbb{XY}|d_1|} \geq .90$) or semantically redundant ($M_{SemSim} \geq 3.0$ or 3.5) is consistent with the understanding that there is a vast

space of fine distinctions or 'nuances' in how people vary from one another which can be better measured by researchers (Condon et al., 2020; Mõttus et al., 2017, 2019, 2020).

Finally, it is important to note that empirical estimates of the associations between raw-score correlations, retest-adjusted correlations, and perceived semantic similarity ratings (Table 3) of item-pairs are likely substantially underestimated by the present method. The selection of item-pairs indexed as having high raw-score or retest-adjusted correlations resulted in clear range restriction, which can be seen by the fact that no item-pairs indexed as having zero-order correlations below about $r_{XY|d_0|} = .20$ were selected to be rated for their perceived semantic similarity (Figure 2). This indicates that the approximately $q \approx .50$ association between the estimated informational and semantic similarity of item-pairs would likely be considerably higher if larger ranges of inter-item correlations were included.

**Limitations, Remaining Questions, and Future Directions**

***How Much Does the Measurement Interval Matter?***

As we have noted, when reliability is operationalized as a test's retest correlation, there are as many reliability coefficients as there are retest intervals (Cronbach, 1947). As the measurement interval $|d|$ increases, $r_{XY|d|}$ correlations will typically become less reflective of *transient* or *circumstantial* factors influencing scores – such as the mood at the time of testing (Chmielewski & Watson, 2009; McCrae, 2015; van Bork et al., 2022). In turn, $\hat{\rho}_{\mathbb{XY}|d|}$ values can be expected to vary over differing intervals of $|d|$, as more transient factors influencing X and Y may covary in a different manner than more stable factors.

However, we suggest that for the specific purpose of evaluating whether a pair of items *is redundant*, the interval $|d|$ used to form the $\hat{\rho}_{|d|}$ matrix should not generally matter. This is because if two items are perfectly redundant, this can be understood as meaning that the function of factors determining scores on item X and item Y is essentially equivalent. We can illustrate this idea with the item-pair indexed with the highest estimated retest-adjusted correlation in the present study: the BFI-2 items [I am someone who] "Often feels sad" versus "Tends to feel depressed, blue," $\hat{\rho}_{\mathbb{XY}|d|} = .995$ (Table 2, row #10).

Any factor which tends to cause people to provide high scores on one of these items – such as (a) a fairly objective record of low levels of positive affectivity over time (a trait factor), (b) having a recent bad or stressful interaction (a state factor), or (c) having already settled into mindlessly answering '5' to every question to get through the survey (use of a *response set*; Cronbach, 1946) – may tend to influence scores on the other, and to the same degree. If this is true, then the near-perfect $\hat{\rho}_{\mathbb{XY}|d|}$ correlation between these two items may be expected to be preserved over intervals resulting in differing ratios of 'state' versus 'trait' variance, or even of 'valid' versus 'invalid' variance. Indeed, effective equivalence of the *functions* people are using to respond to the two items may serve as a good conceptual definition of what it means for those items to be redundant within the surveyed population. For truly redundant items, the $r_{XY|d|}$, $r_{XX|d|}$, and $r_{YY|d|}$ values within Equation 2 may all be expected to change in concert to preserve the $\hat{\rho}_{\mathbb{XY}|d|}$ of 1 as the measurement interval $|d|$ is specified at any arbitrarily short or long interval.

It is worth noting that this idea represents an important change in how reliability is treated within the literature. Traditionally, researchers have argued that the reliability estimates that should be used for reliability adjustments are those in which the expected levels of test scores for respondents remain nearly stationary – i.e., where change in 'true-scores' is negligible (Cattell & Tsujioka, 1964; Chmielewski & Watson, 2009; Gnambs, 2014; Watson, 2004). However, we argue that the goal should be instead to *equate* the measurement intervals used in the numerator and denominator of Spearman's equation, so that the level of systematic change in the factors affecting scores is *matched*.

If this interpretation of how to conceptualize the role of unreliability on correlations is appropriate, and especially if differences in retest-adjusted correlations $\hat{\rho}_{\mathbb{XY}|d|}$ across different measurement intervals $|d|$ are relatively negligible, this would represent an enormous opportunity to explore questions of item or test equivalence across previously collected datasets. Essentially, it would indicate that *any time a researcher has administered the tests of interest twice within a larger inventory or survey, the reliability adjustment detailed in Equations 2 and 3 can be used to explore questions of item similarity*. This means that every longitudinal study in which participants have rated the same

inventory two or more times over the course of the study – whether over a span of days or years – becomes a source of data which may be usefully reanalyzed for such purposes.

**Possible problems at very short retest intervals.** The only interval $|d|$ we think might present problems for indexing the informational similarity via these indices are *exceptionally* short retest correlations – where items might be retested at spans as short as, say, 30 items apart or less (e.g., by using the design described in Footnote #4). There are reasons to suspect that if participants rate an item (such as "Often feels sad") and then are asked to rate the same item a few items later, many participants may feel compelled to provide the same rating they can still remember – or perhaps even can still *see* – as having given the item previously. However, seeing a very-semantically-similar-and-yet-different item (such as "Tends to feel depressed, blue") over the same measurement interval may resulting in many of the same participants reacting by providing a *more* distinct answer, perhaps due to generously assuming the researcher was 'trying to get at something different' with the second question. Such a pattern of reactivity effects would represent a problem as it would cause $r_{XY|d|}$ and $r_{XX|d|}$ estimates to move in opposite directions. However, we suspect this should only become a sizable problem for *exceptionally* short retest intervals. The present results indicate that even retest intervals as short as 15 minutes produce $\hat{\rho}_{\mathbb{XY}|d|}$ estimates serving as valuable indices of item redundancy.

**Possible problems at long retest intervals.** An important limitation of estimating $\hat{\rho}_{\mathbb{XY}|d|}$ over much longer measurement intervals than used in the present study – such as a year or more – is that all of the components going into the estimation of this parameter (i.e., $r_{XY|d|}, r_{XX|d|}$ and $r_{YY|d|}$; Equation 3) will be expected to decrease in magnitude toward 0 as the measurement interval $|d|$ increases (Fraley & Roberts, 2005; Kenny & Zautra, 1995; Lucas & Donnellan, 2007), making the resulting $\hat{\rho}_{\mathbb{XY}|d|}$ estimates noisier. But as long as these components do not drift to nearly zero (e.g., as might happen when correlating ratings of momentary moods collected years apart), this limitation may in principle be counteracted by estimating correlations through samples which are sufficiently large (e.g., $N > 1000$) to make the confidence intervals of the resulting $\hat{\rho}_{\mathbb{XY}|d|}$ estimates reasonably small.

***How Best to Rate the Semantic Similarity of Items?***

The findings presented here also indicate the value of having laypersons rate the semantic similarity of item-pairs. Although the current findings demonstrate the utility of such judgments, they may also indicate ways in which these judgments could be improved. For instance, if raters providing semantic similarity ratings interpreted the items in different ways than the participants providing the self-ratings used in reliability adjustments, it should decrease the extent to which semantic similarity and information similarity indices will track one another. As an example, the items *alone* and *lonely* were rated as quite semantically distinct (as only '*slightly* to *fairly* similar in meaning'; $M = 1.68$) despite being estimated as having a very high retest-adjusted correlation ($\hat{\rho}_{\mathbb{XY}|d_1|}$=.95; Row #3 in Table 2). Respondents rating the semantic similarity of these items may have been indicating that '*being* alone' is easily distinguishable from '*being* lonely.' (Indeed, a quick web search reveals a large number of articles, blog posts, songs, and other media roughly titled "*Alone But Not Lonely*.") In contrast, as detailed in the Method, participants who completed the self-ratings used to form retest-adjusted correlations were asked to describe the extent to which they '*felt* alone' and '*felt* lonely', which we suspect would also have tended to be judged as more semantically similar.

This indicates a way in which the observed relationships between inter-item correlations and semantic similarity judgments should probably be interpreted as lower-bound estimates. It is worth exploring how modifications to the instructions used to collect semantic similarity judgments could improve their ability to track information similarity estimates.

***Exploring Other Uses of Retest-Adjusted Correlation Matrices***

The identification of informationally or semantically redundant item-pairs has been discussed as being valuable for reducing scale length (Cattell & Tsujioka, 1964; Cortina et al., 2020; DeVellis, 2017), and for understanding whether correlations between conceptually distinct variables may be due scales operationalizing the variables containing redundant content (Mõttus, 2016; Nicholls et al., 1982; Wood & Harms, 2016). However, as we have noted, it is straightforward to adjust *all* correlations between items within an inventory by their retest reliability to form a complete *retest-adjusted correlation matrix*. We

expect that researchers will find working with these matrices useful for a wide range of purposes, even if they have little interest in identifying item-pairs with direct one-to-one redundancy.

For instance, some researchers have prescribed replacing items from a set if they have negligible unique variance beyond other items within the set (Block, 1961; Condon et al., 2020; Furr et al., 2010; Yarkoni, 2010) to increase the breadth or comprehensiveness of the total inventory. This can be done by regressing scores on a particular item on scores from other items within the inventory (e.g., Mõttus et al., 2017, 2019), and replacing items predicted at levels with $R^2$ values near 1.00. At a more theoretical level, many researchers have argued that the certain traits can be adequately regarded as *combinations* of other traits. For instance, Hough and Ones (2002) suggested a wide range of ways in which specific traits could be regarded as 'compounds' of Big Five dimensions, for instance: *Warmth* ≈ *Agreeableness* + *Extraversion*; and *Traditionalism* ≈ *Conscientiousness* – *Openness* (see also Credé et al., 2016; Hofstee et al., 1992; McCrae & Costa, 1989). And conversely, many researchers have suggested how more domain-general constructs may be usefully regarded as weighted sums of domain-contextualized constructs (e.g., *General Life Satisfaction* ≈ *Work Satisfaction* + *Relationship Satisfaction*; Rohrer & Schmukle, 2018; Wood & Roberts, 2006; *Academic Efficacy* ≈ *Verbal Efficacy* + *Math Efficacy*; Bong, 1997). For these and other purposes, the use of retest-adjusted correlation matrices would afford the ability to explore these questions with greater confidence that $R^2$ values very close to 1.00 are attainable (with appropriate measures taken to prevent overprediction through use of multiple predictors, Mõttus et al., 2020; Yarkoni & Westfall, 2017).

More generally, we suspect that the fact unreliability can be removed in a straightforward manner for all items in a set that has been rated twice will strike many researchers as surprising. This challenges the understanding that the creation of multi-item scales may almost be *necessary* to handle the unreliability of single item-measures, and consequently could help shift and sharpen discussions of how and when it is appropriate to adjust for 'internal consistency' versus temporal stability when estimating the reliable associations between variables (e.g. Le et al., 2010; McCrae, 2015).

## Conclusion

Researchers are increasingly concerned that large correlations between scales might be driven by item overlap, resulting in problems of *construct proliferation, overlap,* and *redundancy* across the behavioral sciences (Larsen & Bong, 2016; O'Boyle et al., 2015; Rosenbusch et al., 2020; Shaffer et al., 2016; Singh, 1991). As we have noted, these problems often ultimately involve the presence of redundant items, but procedures for indexing item similarity are not well-developed. We illustrate here how it is possible to make more systematic quantitative estimates of both the level of *information similarity* and *semantic similarity* of items within an inventory. Ultimately, we showed that these methods are mutually validating: across four different inventories, the relationship between retest-adjusted correlations and consensually judged semantic similarity consistently exceeded a $q$ = .50 magnitude. This indicates that both methods should be valuable toward addressing questions of construct or scale overlap.

But perhaps the largest contribution of the study concerns addressing more fundamental questions of how to appropriately adjust for measurement unreliability. Namely, an assumption underlying the common practice of adjusting correlations by reliability coefficients, as is regularly done within contemporary meta-analysis and structural equation modeling, is that this improves upon raw-score correlations as indices of the degree of overlap between measures (Banks et al., 2016; Credé & Harms, 2015; Le et al., 2009, 2009; McGrath et al., 2017; O'Boyle et al., 2015; Schmidt & Hunter, 2014). However, we understand that *the current study provides the first empirical evidence that reliability adjustments can improve the extent to which correlations indicate the semantic similarity of tests*. We believe the key to empirically evaluating this assumption comes from understanding that the *information similarity* and *semantic similarity* of test pairs can be operationalized independently and correlated.

Importantly, the current findings only indicate that retest-adjusted correlations, $\hat{\rho}_{\mathbb{XY}|d|}$, formed by dividing the correlation between test scores with their average retest correlation over the same measurement interval (Equations 2 and 3), improve upon raw-score correlations as indicators of the semantic similarity of tests. We have noted that this parameter can be understood as addressing the question: *how much lower is the observed correlation between X and Y over measurement interval* $|d|$ *from the average correlation we would have observed by just retesting X and Y over the same interval?*

As we have noted, this conceptual quantity can be estimated whenever X and Y are tests within an inventory or survey that has been administered twice.

The current findings do not speak to the much more common practice of adjusting correlations for unreliability using internal consistency estimates or other means (Le et al., 2009; Schmidt & Hunter, 2014). Unfortunately, due to the fact that conventional internal consistency statistics appear to track validity-related criteria considerably worse than do retest correlations (Henry et al., 2022; Lowman et al., 2018; McCrae et al., 2011) and are systematically affected by factors such as item breadth (John & Soto, 2007), we suspect that many commonly used procedures for adjusting for measurement unreliability likely infuse as much unwanted variance into estimates of the information similarity of tests as they remove (LeBreton et al., 2014; Sackett, 2014; Sackett et al., 2021). This in turn has implications for any place adjustments for measurement unreliability are used – such as structural equation modeling and meta-analysis. However, the present research makes clearer that such concerns about how best to adjust for measurement unreliability can and should be regarded as open questions that can be evaluated empirically.

**References**

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*, 179–211.

Arnulf, J. K., & Larsen, K. R. (2021). Semantic and ontological structures of psychological attributes. In D. Wood, S. J. Read, P. D. Harms, & A. Slaughter (Eds.), *Measuring and modeling persons and situations* (pp. 69–101). Academic Press.

Bainbridge, T. F., Ludeke, S. G., & Smillie, L. D. (2022). Evaluating the Big Five as an organizing framework for commonly used psychological trait scales. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/pspp0000395

Bandura, A. (1999). Social cognitive theory of personality. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research (2nd ed.).* (pp. 154–196). Guilford.

Banks, G. C., McCauley, K. D., Gardner, W. L., & Guler, C. E. (2016). A meta-analytic review of authentic and transformational leadership: A test for redundancy. *The Leadership Quarterly*, *27*, 634–652.

Beck, E. D., & Jackson, J. J. (2022). A mega-analysis of personality prediction: Robustness and boundary conditions. *Journal of Personality and Social Psychology*, *122*, 523–553. https://doi.org/10.1037/pspp0000386

Block, J. (1961). *The Q-sort method in personality assessment and psychiatric research*. Charles C Thomas.

Block, J., Weiss, D. S., & Thorne, A. (1979). How relevant is a semantic similarity interpretation of personality ratings? *Journal of Personality and Social Psychology*, *37*, 1055–1074. https://doi.org/10.1037/0022-3514.37.6.1055

Bong, M. (1997). Generality of academic self-efficacy judgments: Evidence of hierarchical relations. *Journal of Educational Psychology*, *89*, 696–709.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics.* Cambridge University Press.

Borsboom, D., & Mellenbergh, G. J. (2002). True scores, latent variables, and constructs: A comment on Schmidt and Hunter. *Intelligence*, *30*, 505–514.

Burke, D. L., Ensor, J., & Riley, R. D. (2017). Meta-analysis using individual participant data: One-stage and two-stage approaches, and why they may differ. *Statistics in Medicine*, *36*, 855–875.

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, *2*, 115–144.

Cattell, R. B. (1952). The three basic factor-analytic research designs—Their interrelations and derivatives. *Psychological Bulletin*, *49*, 499–520.

Cattell, R. B., & Tsujioka, B. (1964). The importance of factor-trueness and validity, versus homogeneity and orthogonality, in test scales1. *Educational and Psychological Measurement*, *24*, 3–30.

Chaffin, R., & Herrmann, D. J. (1984). The similarity and diversity of semantic relations. *Memory & Cognition*, *12*, 134–141.

Charles, E. P. (2005). The correction for attenuation due to measurement error: Clarifying concepts and creating confidence sets. *Psychological Methods*, *10*, 206–226.

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

Christensen, A. P., Garrido, L. E., & Golino, H. (2020). *Unique variable analysis: A novel approach for detecting redundant variables in multivariate data*.

Christensen, A. P., & Kenett, Y. N. (2021). Semantic network analysis (SemNA): A tutorial on preprocessing, estimating, and analyzing semantic networks. *Psychological Methods*. https://doi.org/10.1037/met0000463

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Condon, D. M., Wood, D., Möttus, R., Booth, T., Costantini, G., Greiff, S., Johnson, W., Lukaszweski, A., Murray, A., & Revelle, W. (2020). Bottom-up construction of a personality taxonomy. *European Journal of Psychological Assessment*, *36*, 923–934.

Cortina, J. M., Sheng, Z., Keener, S. K., Keeler, K. R., Grubb, L. K., Schmitt, N., Tonidandel, S., Summerville, K. M., Heggestad, E. D., & Banks, G. C. (2020). From alpha to omega and beyond! A look at the past, present, and (possible) future of psychometric soundness in the Journal of Applied Psychology. *Journal of Applied Psychology*, *105*, 1351–1381. https://doi.org/10.1037/apl0000815

Credé, M., & Harms, P. D. (2015). 25 years of higher-order confirmatory factor analysis in the organizational sciences: A critical review and development of reporting recommendations. *Journal of Organizational Behavior*, *36*, 845–872.

Credé, M., Harms, P. D., Blacksmith, N., & Wood, D. (2016). Assessing the utility of compound trait estimates of narrow personality traits. *Journal of Personality Assessment*, *98*, 503–513. https://doi.org/10.1080/00223891.2016.1170023

Credé, M., Tynan, M. C., & Harms, P. D. (2017). Much ado about grit: A meta-analytic synthesis of the grit literature. *Journal of Personality and Social Psychology*, *113*, 492–511. https://doi.org/10.1037/pspp0000102

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement*, *6*(4), 475–494.

Cronbach, L. J. (1947). Test “reliability”: Its meaning and determination. *Psychometrika*, *12*, 1–16.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334. https://doi.org/10.1007/BF02310555

Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: The Simultaneous analysis of multiple data sets. *Psychological Methods*, *14*, 81–100.

Cutler, A., & Condon, D. M. (2022). *Deep lexical hypothesis: Identifying personality structure in natural language* [Unpublished manuscript]. Department of Psychology, Boston University.

DeVellis, R. F. (2017). *Scale development: Theory and applications*. Sage.

Dragostinov, Y., & Mõttus, R. (2021). *Test-retest reliability and construct validity of the brief Dark Triad measurements* [Unpublished manuscript]. PsyArXiv; Department of Psychology, Boston University.

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.

Dul, J. (2016). Necessary Condition Analysis (NCA) logic and methodology of "necessary but not sufficient" causality. *Organizational Research Methods*, *19*, 10–52.

Dweck, C. S. (2017). From needs to goals and representations: Foundations for a unified theory of motivation, personality, and development. *Psychological Review*, *124*, 689–719.

Epskamp, S., Rhemtulla, M., & Borsboom, D. (2017). Generalized network psychometrics: Combining network and latent variable models. *Psychometrika, 82,* 904–927.

Feather, N. T. (1982). *Expectations and actions: Expectancy-value models in psychology*. Erlbaum.

Fraley, R. C., & Roberts, B. W. (2005). Patterns of continuity: A dynamic model for conceptualizing the stability of individual differences in psychological constructs across the life course. *Psychological Review*, *112*, 60.

Fraley, R. C., Vicary, A. M., Brumbaugh, C. C., & Roisman, G. I. (2011). Patterns of stability in adult attachment: An empirical test of two models of continuity and change. *Journal of Personality and Social Psychology*, *101*, 974–992.

Fried, E., van Borkulo, C., Epskamp, S., Schoevers, R., Tuerlinckx, F., & Borsboom, D. (2016). Measuring depression over time... Or not? Lack of unidimensionality and longitudinal measurement invariance in four common rating scales of depression. *Psychological Assessment*, *28*, 1354–1367.

Funder, D. C. (2016). Taking situations seriously: The situation construal model and the Riverside Situational Q-Sort. *Current Directions in Psychological Science*, *25*, 203–208.

Furr, R. M., Wagerman, S. A., & Funder, D. C. (2010). Personality as manifest in behavior: Direct behavioral observation using the Revised Riverside Behavioral Q-Sort (RBQ-3.0). In C. R.

Agnew, D. E. Carlston, W. G. Graziano, & J. R. Kelly (Eds.), *Then a miracle occurs: Focusing on behavior in social psychological theory and research.* (pp. 186–204). Oxford University Press.

Gnambs, T. (2014). A meta-analysis of dependability coefficients (test–retest reliabilities) for measures of the Big Five. *Journal of Research in Personality*, *52*, 20–28.

Green, S. B. (2003). A coefficient alpha for test-retest data. *Psychological Methods*, *8*, 88–101. https://doi.org/10.1037/1082-989X.8.1.88

Harrison, D. A., Newman, D. A., & Roth, P. L. (2006). How important are job attitudes? Meta-analytic comparisons of integrative behavioral outcomes and time sequences. *Academy of Management Journal*, *49*, 305–325.

Henry, S., Thielmann, I., Booth, T., & Mõttus, R. (2022). Test-retest reliability of the HEXACO-100—And the value of multiple measurements for assessing reliability. *PLOS ONE*, *17*, e0262465. https://doi.org/10.1371/journal.pone.0262465

Hofstee, W. K. (1994). Who should own the definition of personality? *European Journal of Personality*, *8*, 149–162.

Hofstee, W. K., de Raad, B., & Goldberg, L. R. (1992). Integration of the Big Five and circumplex approaches to trait structure. *Journal of Personality and Social Psychology*, *63*, 146–163.

Hough, L. M., & Ones, D. S. (2002). The structure, measurement, validity, and use of personality variables in industrial, work, and organizational psychology. In N. Anderson, D. S. Ones, H. K. Sinangil, C. Viswesvaran, N. Anderson (Ed), D. S. Ones (Ed), H. K. Sinangil (Ed), & C. Viswesvaran (Ed) (Eds.), *Handbook of industrial, work and organizational psychology, Volume 1: Personnel psychology.* (pp. 233–277). Sage Publications Ltd.

John, O. P., & Benet-Martínez, V. (2000). Measurement: Reliability, construct validation, and scale construction. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology.* (pp. 339–369). Cambridge University Press.

John, O. P., & Soto, C. J. (2007). The importance of being valid: Reliability and the process of construct validation. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology.* (pp. 461–494). Guilford.

Jonason, P. K., & Webster, G. D. (2010). The Dirty Dozen: A concise measure of the Dark Triad. *Psychological Assessment*, *22*, 420–432. https://doi.org/10.1037/a0019265

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

Judge, T. A., & Bono, J. E. (2001). Relationship of core self-evaluations traits—Self-esteem, generalized self-efficacy, locus of control, and emotional stability—With job satisfaction and job performance: A meta-analysis. *Journal of Applied Psychology*, *86*, 80–92. https://doi.org/10.1037/0021-9010.86.1.80

Kelley, T. L. (1927). *Interpretation of educational measurements.* World Book Company.

Kenny, D. A., & Zautra, A. (1995). The trait-state-error model for multiwave data. *Journal of Consulting and Clinical Psychology*, *63*, 52–59. https://doi.org/10.1037/0022-006X.63.1.52

Larsen, K. R., & Bong, C. H. (2016). A tool for addressing construct identity in literature reviews and meta-analyses. *MIS Quarterly*, *40*, 1–23.

Lazarsfeld, P. F. (1959). Latent structure analysis. In S. Koch (Ed.), *Psychology: A study of a science.* (Vol. 3, pp. 476–543). McGraw-Hill.

Le, H., Schmidt, F. L., Harter, J. K., & Lauver, K. J. (2010). The problem of empirical redundancy of constructs in organizational research: An empirical investigation. *Organizational Behavior and Human Decision Processes*, *112*, 112–125.

Le, H., Schmidt, F. L., & Putka, D. J. (2009). The multifaceted nature of measurement artifacts and its implications for estimating construct-level relationships. *Organizational Research Methods*, *12*, 165–200. https://doi.org/10.1177/1094428107302900

LeBreton, J. M., Scherer, K. T., & James, L. R. (2014). Corrections for criterion reliability in validity generalization: A false prophet in a land of suspended judgment. *Industrial and Organizational Psychology*, *7*, 478–500.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Lowman, G., Wood, D., Armstrong, B., Harms, P., & Watson, D. (2018). Estimating the reliability of emotion measures over very short intervals: The utility of within-session retest correlations. *Emotion*, *18*, 896–901.

Lucas, R. E., & Donnellan, M. B. (2007). How stable is happiness? Using the STARTS model to estimate the stability of life satisfaction. *Journal of Research in Personality*, *41*, 1091–1098. https://doi.org/10.1016/j.jrp.2006.11.005

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. Jr. (1989). The structure of interpersonal traits: Wiggins's circumplex and the five-factor model. *Journal of Personality and Social Psychology*, *56*, 586–595.

McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review*, *15*, 28–50.

McDonald, R. P. (1999). *Test theory: A unified treatment.* Erlbaum.

McGrath, R. E., Hall-Simmonds, A., & Goldberg, L. R. (2020). Are measures of character and personality distinct? Evidence from observed-score and true-score analyses. *Assessment, 27,* 117–135.

Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, *6*, 1–28.

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, *102*, 246–268.

Mõttus, R. (2016). Towards more rigorous personality trait-outcome research. *European Journal of Personality*, *30*, 292–303.

Mõttus, R., Kandler, C., Bleidorn, W., Riemann, R., & McCrae, R. R. (2017). Personality traits below facets: The consensual validity, longitudinal stability, heritability, and utility of personality nuances. *Journal of Personality and Social Psychology*, *112*, 474–490. https://doi.org/10.1037/pspp0000100

Mõttus, R., Sinick, J., Terracciano, A., Hřebíčková, M., Kandler, C., Ando, J., Mortensen, E. L., Colodro-Conde, L., & Jang, K. L. (2019). Personality characteristics below facets: A replication and meta-analysis of cross-rater agreement, rank-order stability, heritability, and utility of personality nuances. *Journal of Personality and Social Psychology*, *117*, e35–e50. https://doi.org/10.1037/pspp0000202

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

Mueller, S., Wang, D., Fox, M. D., Pan, R., Lu, J., Li, K., Sun, W., Buckner, R. L., & Liu, H. (2015). Reliability correction for functional connectivity: Theory and implementation. *Human Brain Mapping*, *36*, 4664–4680.

Newman, D. A., Harrison, D. A., Carpenter, N. C., & Rariden, S. M. (2016). Construct mixology: Forming new management constructs by combining old ones. *Academy of Management Annals*, *10*, 943–995.

Newman, D. A., Joseph, D. L., & Hulin, C. L. (2010). Job attitudes and employee engagement: Considering the attitude "A-factor." In S. L. Albrecht (Ed.), *The handbook of employee engagement: Perspectives, issues, research, and practice* (pp. 43–61). Edward Elgar Publishing.

Nicholls, J. G., Licht, B. G., & Pearl, R. A. (1982). Some dangers of using personality questionnaires to study personality. *Psychological Bulletin*, *92*, 572–580.

Nunnally, J. C., & Bernstein, I. H. (1991). *Psychometric theory.* McGraw.

O'Boyle, E. H., Forsyth, D. R., Banks, G. C., Story, P. A., & White, C. D. (2015). A meta-analytic test of redundancy and relative importance of the dark triad and five-factor model of personality. *Journal of Personality*, *83*, 644–664. https://doi.org/10.1111/jopy.12126

Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods*, *5*, 343–355. https://doi.org/10.1037/1082-989X.5.3.343

Paulhus, D. L., & Williams, K. M. (2002). The dark triad of personality: Narcissism, Machiavellianism, and psychopathy. *Journal of Research in Personality*, *36*, 556–563.

Paunonen, S. V. (1984). Optimizing the validity of personality assessments: The importance of aggregation and item content. *Journal of Research in Personality*, *18*, 411–431.

Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, *11*, 95–130.

Revelle, W., & Condon, D. M. (2019). Reliability from α to ω: A tutorial. *Psychological Assessment*, *31*, 1395–1411.

Rohrer, J. M., & Schmukle, S. C. (2018). Individual importance weighting of domain satisfaction ratings does not increase validity. *Collabra: Psychology*, *4*, 6. https://doi.org/10.1525/collabra.116

Rosenbusch, H., Wanders, F., & Pit, I. L. (2020). The Semantic Scale Network: An online tool to detect semantic overlap of psychological scales and prevent scale redundancies. *Psychological Methods*, *25*, 380–392. https://doi.org/10.1037/met0000244

Rubenstein, H., & Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, *8*, 627–633.

Sackett, P. R. (2014). When and why correcting validity coefficients for interrater reliability makes sense. *Industrial and Organizational Psychology*, *7*, 501–506.

Sackett, P. R., Zhang, C., Berry, C. M., & Lievens, F. (2021). Revisiting meta-analytic estimates of validity in personnel selection: Addressing systematic overcorrection for restriction of range. *Journal of Applied Psychology*.

Saucier, G., Iurino, K., & Thalmayer, A. G. (2020). Comparing predictive validity in a community sample: High-dimensionality and traditional domain-and-facet structures of personality variation. *European Journal of Personality*, *34*, 1120–1137.

Schmidt, F. L. (2010). Detecting and correcting the lies that data tell. *Perspectives on Psychological Science*, *5*, 233–242.

Schmidt, F. L., & Hunter, J. E. (2014). *Methods of meta-analysis: Correcting error and bias in research findings*. Sage.

Schmidt, F. L., Le, H., & Oh, I. (2013). Are true scores and construct scores the same? A critical examination of their substitutability and the implications for research results. *International Journal of Selection and Assessment*, *21*, 339–354.

Shaffer, J. A., DeGeest, D., & Li, A. (2016). Tackling the problem of construct proliferation: A guide to assessing the discriminant validity of conceptually related constructs. *Organizational Research Methods*, *19*, 80–110.

Shedler, J., & Westen, D. (2007). The Shedler–Westen assessment procedure (SWAP): Making personality diagnosis clinically meaningful. *Journal of Personality Assessment*, *89*, 41–55.

Shweder, R. A., & D'Andrade, R. G. (1979). Accurate reflection or systematic distortion? A reply to Block, Weiss, and Thorne. *Journal of Personality and Social Psychology*, *37*, 1075–1084. https://doi.org/10.1037/0022-3514.37.6.1075

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*, 107–120.

Singh, J. (1991). Redundancy in constructs: Problem, assessment, and an illustrative example. *Journal of Business Research*, *22*, 255–280.

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

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*, 72–101.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*, 271–295.

Stadnitski, T. (2020). Time series analyses with psychometric data. *PLOS ONE*, *15*, e0231785. https://doi.org/10.1371/journal.pone.0231785

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*, 245–251. https://doi.org/10.1037/0033-2909.87.2.245

Stephenson, W. (1953). *The study of behavior; Q-technique and its methodology.* University of Chicago Press.

van Bork, R., Rhemtulla, M., Sijtsma, K., & Borsboom, D. (2022). A causal theory of error scores. *Psychological Methods*. APA PsycInfo. https://doi.org/10.1037/met0000521

Watson, D. (2004). Stability versus change, dependability versus error: Issues in the assessment of personality over time. *Journal of Research in Personality*, *38*, 319–350.

Watson, D., & Clark, L. A. (1999). *The PANAS-X: Manual for the Positive and Negative Affect Schedule-Expanded Form.*

Weidman, A. C., Cheng, J. T., & Tracy, J. L. (2018). The psychological structure of humility. *Journal of Personality and Social Psychology*, *114*, 153–178.

Whitten, W. B., Suter, W. N., & Frank, M. L. (1979). Bidirectional synonym ratings of 464 noun pairs. *Journal of Verbal Learning and Verbal Behavior*, *18*, 109–127.

Wiggins, J. S. (2003). *Paradigms of personality assessment*. Guilford.

Wilt, J., & Revelle, W. (2015). Affect, behaviour, cognition and desire in the Big Five: An analysis of item content and structure. *European Journal of Personality*, *29*, 478–497.

Wood, D., Gardner, M. H., & Harms, P. D. (2015). How functionalist and process approaches to behavior can explain trait covariation. *Psychological Review*, *122*, 84–111.

Wood, D., & Harms, P. D. (2016). On the TRAPs that make it dangerous to study personality with personality questionnaires. *European Journal of Personality*, *30*, 327–328.

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

Wood, D., Nye, C. D., & Saucier, G. (2010). Identification and measurement of a more comprehensive set of person-descriptive trait markers from the English lexicon. *Journal of Research in Personality*, *44*, 258–272.

Wood, D., Qiu, L., Lu, J., Lin, H., & Tov, W. (2018). Adjusting bilingual ratings by retest reliability Improves estimation of translation quality. *Journal of Cross-Cultural Psychology*, *49*, 1325–1339.

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.

Wood, D., & Wortman, J. (2012). Trait means and desirabilities as artifactual and real sources of differential stability of personality traits. *Journal of Personality*, *80*, 665–701.

Yarkoni, T. (2010). The abbreviation of personality, or how to measure 200 personality scales with 200 items. *Journal of Research in Personality*, *44*, 180–198.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*, 1100–1122.

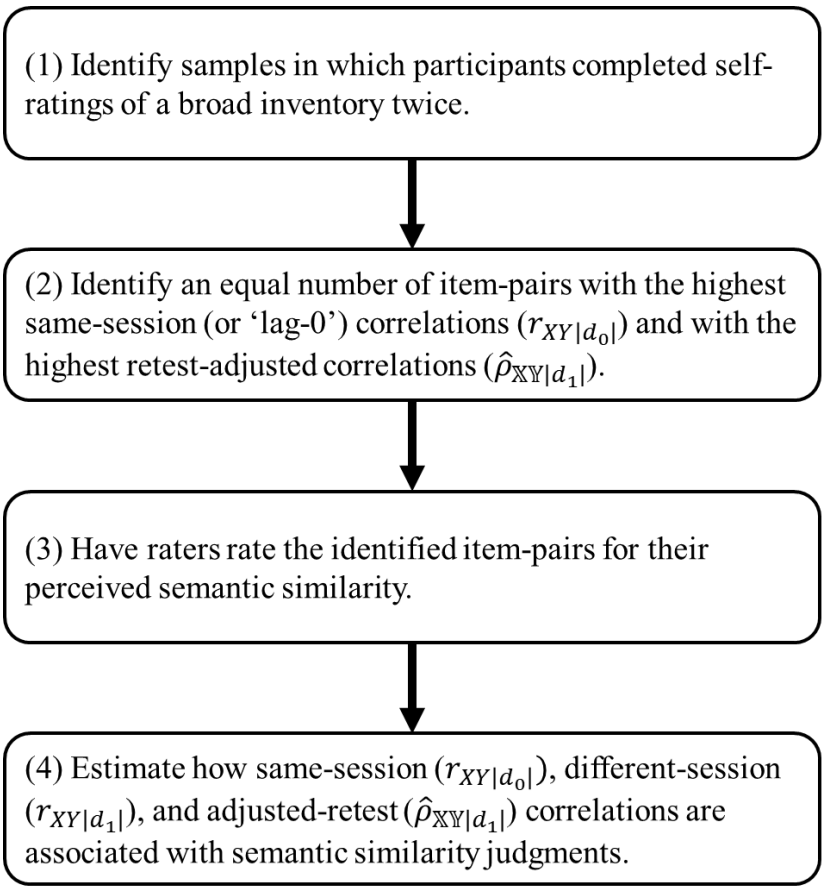


**Figure 1.** Analysis strategy for current study. Code for replicating the analyses relevant to Step #4 is available at https://osf.io/vp6kr/.

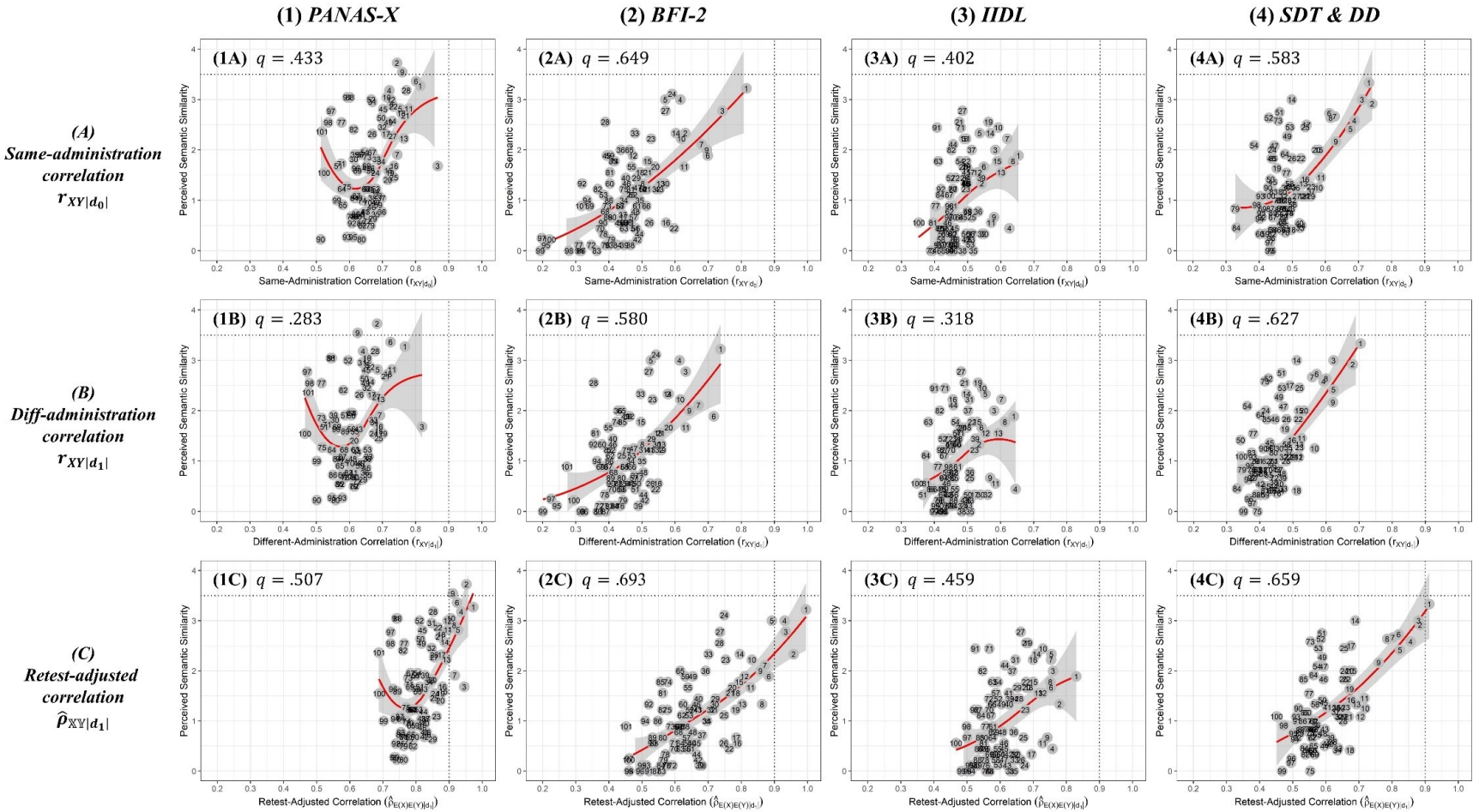


**Figure 2.** Scatter-plots illustrating estimated associations between *estimated informational similarity (X-axis)* and *semantic similarity judgments (Y-axis)*. For each figure, lines have been added at $X = .9$ to indicate a proposed lower-bound for considering items "informational redundant" and at $Y = 3.5$ to indicate a proposed lower-bound for considering item-pairs as being "semantically redundant." Numbers for item-pairs within the scatterplot correspond to the rank of their retest-adjusted correlation (e.g., #1 indicates the item-pair with this highest $\hat{\rho}_{\mathbb{XY}|d_1|}$ correlation.

**Table 1.** Structure of hypothetical *score correlations* and *measurement intervals* for associations between two items X and Y when administered inside an inventory administered twice over a one-week retest interval.

(1A) Matrix of *score correlations,* $r_{X_m Y_m}$, in a two-wave repeated measures study

| | $X_1$ | $Y_1$ | $X_2$ | $Y_2$ |
|---|---|---|---|---|
| $X_1$ | **1** | | | |
| $Y_1$ | **.60** | **1** | | |
| $X_2$ | **.60** | **.45** | **1** | |
| $Y_2$ | **.45** | **.50** | **.60** | **1** |

(1B) Matrix of *measurement intervals,* $|d_{X_m Y_m}|$, in a two-wave repeated measures study

| | $X_1$ | $Y_1$ | $X_2$ | $Y_2$ |
|---|---|---|---|---|
| $X_1$ | 0 seconds | | | |
| $Y_1$ | 5 minutes | 0 seconds | | |
| $X_2$ | **1 week** | **1 week** | 0 seconds | |
| $Y_2$ | **1 week** | **1 week** | 5 minutes | 0 seconds |

**Note.** In both matrices 1A and 1B, higher values are shown in darker cells, whereas values close to zero are shown in lighter values. The values forming the 'lag-1' or the 'different administration,' $|d_1|$, submatrix are additionally outlined by a black box.

**Table 2.** Subset of Item-Pairs with Top Five Highest Estimated *Retest-Adjusted Correlations* ($\hat{\rho}_{\mathbb{XY}[D_1]}$), *Same-administration Correlations* $r_{XY[D_0]}$, and *Judged Semantic Similarity*

| # | Item X | Item Y | $r_{XY\|d_0\|}$ | $r_{XY\|d_1\|}$ | $\hat{\rho}_{\mathbb{XY}\|d_1\|}$ | $\sqrt{r_{XX}r_{YY\|d_1\|}}$ | SemSimilarity |
|---|---|---|---|---|---|---|---|
| ***Positive and Negative Affect Schedule – Extended Form (PANAS-X)*** | | | | | | | |
| *1* | sleepy | tired | **.814 (2)** | **.768 (2)** | **.972 (1)** | .789 (25) | **3.273 (4)** |
| *2* | frightened | scared | .744 (12) | .684 (14) | **.952 (2)** | .718 (82) | **3.727 (1)** |
| *3* | alone | lonely | **.866 (1)** | **.819 (1)** | **.946 (3)** | **.866 (1)** | 1.682 (44) |
| *4* | afraid | frightened | .721 (21) | .641 (37) | **.937 (4)** | .685 (90) | **3.182 (5)** |
| *5* | energetic | lively | .753 (10) | .685 (13) | **.926 (5)** | .740 (69) | 2.818 (15) |
| *6* | blue | sad | **.801 (3)** | **.724 (4)** | .924 (6) | .784 (29) | **3.364 (3)** |
| *7* | afraid | scared | .760 (8) | .625 (45) | .911 (9) | .686 (89) | **3.545 (2)** |
| *8* | cheerful | happy | **.779 (4)** | **.729 (3)** | .896 (11) | **.813 (4)** | 2.818 (15) |
| *9* | drowsy | sleepy | **.772 (5)** | .678 (16) | .852 (28) | .796 (15) | **3.182 (5)** |
| ***Big Five Inventory – 2 (BFI-2)*** | | | | | | | |
| *10* | Often feels sad. | Tends to feel depressed, blue. | **.815 (1)** | **.737 (1)** | **.995 (1)** | .745 (22) | **3.222 (1)** |
| *11* | Is polite, courteous to others. | Is respectful, treats others with respect. | .631 (6) | .581 (10) | **.958 (2)** | .604 (89) | 2.333 (8) |
| *12* | Is fascinated by art, music, or literature. | Values art and beauty. | **.743 (2)** | **.632 (5)** | **.934 (3)** | .669 (69) | **2.778 (5)** |
| *13* | Is inventive, finds clever ways to do things. | Is original, comes up with new ideas. | .616 (9) | .613 (8) | **.931 (4)** | .660 (74) | **3.000 (3)** |
| *14* | Is dependable, steady. | Is reliable, can always be counted on. | .571 (15) | .527 (24) | **.892 (5)** | .583 (94) | **3.000 (3)** |
| *15* | Tends to be quiet. | Is sometimes shy, introverted. | **.698 (3)** | **.717 (2)** | .883 (6) | **.808 (1)** | 1.889 (17) |
| *16* | Is outgoing, sociable. | Is talkative. | **.680 (5)** | **.671 (3)** | .871 (7) | .775 (6) | 2.111 (13) |
| *17* | Is full of energy. | Shows a lot of enthusiasm. | **.696 (4)** | **.643 (4)** | .858 (9) | .753 (19) | 2.000 (14) |
| *18* | Is moody, has up and down mood swings. | Is temperamental, gets emotional easily. | .591 (12) | .542 (19) | .748 (24) | .700 (55) | **3.111 (2)** |
| *19* | Is systematic, likes to keep things in order. | Keeps things neat and tidy. | .571 (14) | .520 (27) | .736 (27) | .684 (62) | **2.778 (5)** |
| ***Inventory of Individual Differences in the Lexicon (IIDL)*** | | | | | | | |
| *20* | excited, enthusiastic | lively, playful | **.653 (1)** | **.640 (2)** | **.829 (1)** | .766 (48) | 1.889 (16) |
| *21* | kind-hearted, caring | polite, courteous | .543 (16) | .535 (15) | **.779 (2)** | .680 (94) | 1.333 (35) |
| *22* | efficient, thorough | hard-working, productive | .584 (8) | .581 (7) | **.760 (3)** | .755 (58) | 2.000 (13) |
| *23* | happy, joyful | likeable, well-liked | **.626 (3)** | **.645 (1)** | **.758 (4)** | .843 (8) | .444 (63) |
| *24* | dependable, reliable | efficient, thorough | .536 (18) | .554 (11) | **.755 (5)** | .706 (85) | 2.333 (7) |
| *25* | afraid, scared | tense, anxious | **.620 (4)** | **.603 (4)** | .754 (7) | .793 (28) | 2.222 (9) |
| *26* | excited, enthusiastic | happy, joyful | **.637 (2)** | **.613 (3)** | .751 (8) | .809 (20) | 1.778 (17) |
| *27* | determined, persistent | hard-working, productive | .592 (6) | .545 (12) | .729 (10) | .740 (70) | **2.444 (4)** |
| *28* | happy, joyful | lively, playful | **.601 (5)** | **.593 (5)** | .716 (13) | .821 (14) | 1.556 (26) |
| *29* | angry, hostile | hot-tempered, short-tempered | .563 (12) | .531 (16) | .684 (19) | .779 (41) | **2.556 (2)** |
| *30* | dependable, reliable | faithful, loyal | .475 (54) | .491 (34) | .679 (21) | .705 (88) | **2.556 (2)** |
| *31* | competent, capable | skilled, skillful | .484 (50) | .481 (38) | .662 (27) | .740 (71) | **2.778 (1)** |
| *32* | assertive, bold | direct, straight-forward | .479 (51) | .432 (70) | .567 (71) | .760 (52) | **2.444 (4)** |
| *32* | dependable, reliable | prompt, punctual | .407 (95) | .400 (92) | .525 (91) | .763 (51) | **2.444 (4)** |
| ***Short Dark Triad and Dirty Dozen (SDT/DD)*** | | | | | | | |
| *33* | I tend to manipulate others to get my way. | I like to use clever manipulation to get my way. | **.731 (2)** | **.704 (1)** | **.912 (1)** | .772 (30) | **3.333 (1)** |
| *34* | I tend to manipulate others to get my way. | I'll say anything to get what I want. | **.741 (1)** | **.680 (2)** | **.885 (2)** | .769 (31) | **2.917 (4)** |
| *35* | I tend to manipulate others to get my way. | I tend to exploit others toward my own end. | **.711 (3)** | **.623 (3)** | **.880 (3)** | .712 (81) | **3.000 (2)** |
| *36* | I tend to exploit others toward my own end. | I'll say anything to get what I want. | **.687 (4)** | .593 (7) | **.856 (4)** | .695 (90) | 2.583 (10) |
| *37* | I like to use clever manipulation to get my way. | I'll say anything to get what I want. | **.676 (5)** | **.622 (4)** | **.824 (5)** | .755 (45) | 2.417 (15) |
| *38* | It's not wise to tell your secrets | There are things you should hide from other people to preserve your reputation. | .498 (34) | .512 (20) | .689 (14) | .746 (55) | **3.000 (2)** |
| *39* | I have used deceit or lied to get my way. | I'll say anything to get what I want. | .461 (64) | .465 (39) | .588 (51) | .793 (13) | **2.750 (5)** |

**Note.** Cell values are shown in bold if estimated to be within the five highest values for the column variable for that inventory; the rank of the item-pair for the column property is then given in parentheses. The item labeled "Item X" versus "Item Y" in a given row is arbitrary. A complete list of all item-pairs within the inventory, including item-pairs not rated for semantic similarity, is available at https://osf.io/vp6kr/.

**Table 3.** Estimated Relationships (*q*-correlations) between Different Inter-Item Similarity Estimates.

| **(3.1) PANAS-X; $N_{pairs}$ = 101** | **M** | **(SD)** | $r_{XY\|d_0\|}$ | $r_{XY\|d_1\|}$ | $\widehat{\rho}_{\mathbb{XY}\|d_1\|}$ | $\sqrt{r_{XX}r_{YY\|d_1\|}}$ |
|---|---|---|---|---|---|---|
| Same-administration correlation; $r_{XY\|d_0\|}$ | .664 | (.066) | | | | |
| Different-administration correlation; $r_{XY\|d_1\|}$ | .615 | (.064) | .932 | | | |
| Retest-adjusted correlation; $\widehat{\rho}_{\mathbb{XY}\|d_1\|}$ | .815 | (.062) | .869 | .837 | | |
| Average retest correlation; $\sqrt{r_{XX}r_{YY\|d_1\|}}$ | .754 | (.044) | .557 | .718 | .223 | |
| *Perceived Semantic Similarity (**SemSim**)* | *1.685* | *(.897)* | *$.433_{1}$* | *$.283_{12}$* | *$.507_{2}$* | *$-.152_{ns}$* |

| **(3.2) BFI-2; $N_{pairs}$ = 101** | **M** | **(SD)** | $r_{XY\|d_0\|}$ | $r_{XY\|d_1\|}$ | $\widehat{\rho}_{\mathbb{XY}\|d_1\|}$ | $\sqrt{r_{XX}r_{YY\|d_1\|}}$ |
|---|---|---|---|---|---|---|
| Same-administration correlation; $r_{XY\|d_0\|}$ | .459 | (.110) | | | | |
| Different-administration correlation; $r_{XY\|d_1\|}$ | .461 | (.096) | .938 | | | |
| Retest-adjusted correlation; $\widehat{\rho}_{\mathbb{XY}\|d_1\|}$ | .664 | (.120) | .878 | .875 | | |
| Average retest correlation; $\sqrt{r_{XX}r_{YY\|d_1\|}}$ | .689 | (.072) | .400 | .540 | $.084_{ns}$ | |
| *Perceived Semantic Similarity (**SemSim**)* | *1.115* | *(.811)* | *$.649_{1}$* | *$.580_{12}$* | *$.693_{2}$* | *$-.040_{ns}$* |

| **(3.3) IIDL; $N_{pairs}$ = 100** | **M** | **(SD)** | $r_{XY\|d_0\|}$ | $r_{XY\|d_1\|}$ | $\widehat{\rho}_{\mathbb{XY}\|d_1\|}$ | $\sqrt{r_{XX}r_{YY\|d_1\|}}$ |
|---|---|---|---|---|---|---|
| Same-administration correlation; $r_{XY\|d_0\|}$ | .483 | (.059) | | | | |
| Different-administration correlation; $r_{XY\|d_1\|}$ | .471 | (.060) | .926 | | | |
| Retest-adjusted correlation; $\widehat{\rho}_{\mathbb{XY}\|d_1\|}$ | .615 | (.076) | .837 | .865 | | |
| Average retest correlation; $\sqrt{r_{XX}r_{YY\|d_1\|}}$ | .763 | (.053) | $.173_{ns}$ | .273 | -.229 | |
| *Perceived Semantic Similarity (**SemSim**)* | *.997* | *(.784)* | *$.402_{1}$* | *$.318_{12}$* | *$.459_{2}$* | *-.263* |

| **(3.4) SDT & DD; $N_{pairs}$ = 100** | **M** | **(SD)** | $r_{XY\|d_0\|}$ | $r_{XY\|d_1\|}$ | $\widehat{\rho}_{\mathbb{XY}\|d_1\|}$ | $\sqrt{r_{XX}r_{YY\|d_1\|}}$ |
|---|---|---|---|---|---|---|
| Same-administration correlation; $r_{XY\|d_0\|}$ | .491 | (.076) | | | | |
| Different-administration correlation; $r_{XY\|d_1\|}$ | .454 | (.073) | .915 | | | |
| Retest-adjusted correlation; $\widehat{\rho}_{\mathbb{XY}\|d_1\|}$ | .610 | (.093) | .886 | .926 | | |
| Average retest correlation; $\sqrt{r_{XX}r_{YY\|d_1\|}}$ | .746 | (.045) | .212 | .334 | $-.043_{ns}$ | |
| *Perceived Semantic Similarity (**SemSim**)* | *1.268* | *(.787)* | *$.583_{3}$* | *.627* | *$.659_{3}$* | *$.035_{ns}$* |

| **(3.5) Across all four inventories; $N_{pairs}$ = 402** | **M** | **(SD)** | $r_{XY\|d_0\|}$ | $r_{XY\|d_1\|}$ | $\widehat{\rho}_{\mathbb{XY}\|d_1\|}$ | $\sqrt{r_{XX}r_{YY\|d_1\|}}$ |
|---|---|---|---|---|---|---|
| Same-administration correlation; $r_{XY\|d_0\|}$ | -- | -- | | | | |
| Different-administration correlation; $r_{XY\|d_1\|}$ | -- | -- | .928 | | | |
| Retest-adjusted correlation; $\widehat{\rho}_{\mathbb{XY}\|d_1\|}$ | -- | -- | .868 | .876 | | |
| Average retest correlation; $\sqrt{r_{XX}r_{YY\|d_1\|}}$ | -- | -- | .346 | .474 | .011 | |
| *Perceived Semantic Similarity (**SemSim**)* | -- | -- | *$.520_{13}$* | *$.459_{12}$* | *$.577_{23,ns}$* | *-.100* |

***Note.*** Subscript *ns* indicates that correlations are not statistically significantly different from zero ($p <$ .05). Shared subscripts *1, 2,* and *3* on a row indicate significant support for *H1, H2,* and *H3,* respectively – that the associated column variables showed evidence of significantly different ($p < .05$) associations with semantic similarity judgments by Steiger's (1980) test of dependent correlations. For cross-sample analyses (3.5), partial correlations are reported after controlling for the inventory containing the item-pair via dummy-codes.