Journal of Research in Personality 113 (2024) 104530

Contents lists available at ScienceDirect

Journal of Research in Personality

journal homepage: www.elsevier.com/locate/jrp

Using multi-rater and test-retest data to detect overlap within and between psychological scales

Sam Henry [a,*], Dustin Wood [b], David M. Condon [c], Graham H. Lowman [d], René Mõttus [a,e]

[a] Department of Psychology, University of Edinburgh, 7 George Square, Edinburgh EH8 9JZ, UK
[b] Culverhouse College of Business, University of Alabama, Alston Hall, Tuscaloosa, AL 35401, USA
[c] Department of Psychology, University of Oregon, Straub Hall, Eugene, OR, USA
[d] Coles College of Business, Kennesaw State University, 560 Parliament Garden Way NW, Kennesaw, GA 30144, USA
[e] Institute of Psychology, University of Tartu, Estonia, Näituse 2, 50409 Tartu, Estonia

ARTICLE INFO

Keywords:
Jingle-jangle
Semantic similarity
Reliability
Cross-rater agreement
Item-level analysis

ABSTRACT

Correlations estimated in single-source data provide uninterpretable estimates of empirical overlap between scales. We describe a model to adjust correlations for errors and biases using test–retest and multi-rater data and compare adjusted correlations among individual items with their human-rated semantic similarity (*SS*). We expected adjusted correlations to predict *SS* better than unadjusted correlations and exceed *SS* in absolute magnitude. While unadjusted and adjusted correlations predicted *SS* rankings equally well across all items, adjusted correlations were superior where items were judged most semantically redundant in meaning. Retest- and agreement-adjusted correlations were usually higher than *SS*, whereas unadjusted correlations often underestimated *SS*. We discuss uses of test–retest and multi-rater data for identifying construct redundancy and argue *SS* often underestimates variables' empirical overlap.

1. Introduction

Construct overlap is among the more pernicious issues in modern psychological assessment, with increasing attention given to so-called *jingle* and *jangle* fallacies and their consequences. The term *jingle fallacy*, first attributed to Aikins (1902), refers to attributing the same label to two or more empirically distinct ideas, while *jangle fallacy* (Kelley, 1927; Anastasi, 1984) means giving different labels to indistinguishable concepts. There are few well-established procedures to identify and resolve these fallacies; the resulting proliferation of scales and constructs that differ primarily in labels rather than content or meaning has created artificial divisions within and between research areas and has contributed to the replicability crisis (Flake & Fried, 2020; Leising et al., 2022; Lilienfeld & Strother, 2020).

To address this, Lawson and Robins (2021) proposed a framework to identify "sibling constructs" that are "conceptually and/or empirically related, but distinct" (p. 345), and distinguish them from truly identical "twin" constructs. As one criterion, they recommended examining correlations between constructs' scales, interpreting two scales that correlate $r = 0.80–1.00$ as twin or identical constructs, $r = 0.00-0.20$ as unrelated, and $r = 0.20–0.60$ as sibling constructs. However, they acknowledged that these were not "hard cutoffs" but rather guidelines that should depend "on the research area, the degree of measurement error (which can attenuate correlations between measures of sibling constructs), and the possibility that correlations may be inflated or attenuated by nonshared method variance" (p. 350).[1]

Indeed, in typical single-method (e.g., self-report) studies, random measurement error and systematic method biases constitute a large proportion of assessed constructs' variance and can both inflate and deflate their correlations. This makes raw, unadjusted correlations near 1 unattainable even for scales of highly similar constructs (McCrae, 2015, 2018) and correlations near 0 also less common than otherwise. To address these biases, Wood et al. (2023) adjusted correlations using retest correlations to account for random and assessment occasion-specific measurement errors. They presented evidence that these retest-adjusted correlations significantly outperformed raw-score

* Corresponding author.
E-mail addresses: samuel.henry@ed.ac.uk (S. Henry), dustin.wood@cba.ua.edu (D. Wood), dcondon@oregon.edu (D.M. Condon), glowman@kennesaw.edu (G.H. Lowman), rene.mottus@ed.ac.uk (R. Mõttus).

[1] Correlation strength is just one of ten criteria that Lawson & Robins (2021) describe, and they recommend that researchers consider all 10 criteria rather than simply focusing only any given criterion. Here, we focus on just one criterion in the sibling construct framework to help clarify it further.

https://doi.org/10.1016/j.jrp.2024.104530
Received 29 May 2023; Received in revised form 6 August 2024; Accepted 28 August 2024
Available online 1 September 2024

unadjusted correlations as indicators of an independent measure of construct overlap – human-rated semantic similarity.

Here, we replicate and extend this work by going beyond adjustments for random and occasion-specific (i.e., transient) measurement errors to also remove systematic method-specific biases from correlations. We argue that personality variables' empirical overlap, free of both transient and systematic measurement issues, can be most accurately estimated by using multi-source (i.e., self- and informant-report) data. We also argue that the adjustments should not only align the *rankings* of variables' correlations with their semantic similarity rankings but should also make the former match or even exceed the latter in *absolute terms*. Specifically, for item pairs which are nearly semantically identical (e.g., I am sleepy, I am drowsy), adjusted correlations should match the similarity, whereas for pairs of items that are semantically similar but not identical (e.g., I am assertive, I am confident), correlations should exceed semantic similarity. This is because items' correlations also reflect functional relations among their underlying personality traits, either due to reflecting a common underlying cause or direct causal relations between them (Mõttus & Allerhand, 2018; Wood et al., 2023), and this should add to their overlap over and above semantic similarity. Our work provides researchers with a clearer understanding of what a given correlation between two constructs means and how to assess it most accurately, thereby allowing for more interpretable estimates of constructs' true overlap and, subsequently, providing more confidence when identifying and resolving jingle-jangle fallacies.

1.1. Evaluating overlap between scales

Lawson and Robins (2021) and others (e.g., John & Benet-Martínez, 2000; Judge & Bono, 2001; Le et al., 2009) have suggested that a correlation > 0.80 or 0.90 between two scales is sufficient to consider them twin constructs. But psychological measurements are noisy, making their correlations difficult to interpret at face value and leading to loose standards for interpreting them (Funder & Ozer, 2019). For example, even broad constructs measured by many items, such as the Big Few personality domains, typically only reach short-term (e.g., over 1–2 weeks) test–retest reliabilities around 0.90 (e.g., Henry et al., 2022; McCrae et al., 2011), so their correlations with other variables can never exceed this (over a given timespan, a scale's reliability is a ceiling for its correlation with any other scale over that same timespan); shorter scales will be even less reliable and hence their correlations even more limited even for perfectly identical constructs. As such, the recommended threshold for redundancy applies to correlations adjusted for measurement error.

Adjusting correlations to better approximate the extent to which they "measure the same thing" is not a new idea (Spearman, 1904, 1910). However, most prior work evaluating this approach has adjusted correlations by scales' internal consistency (e.g., Cronbach's alpha; Banks et al., 2016; Le et al., 2010; McGrath et al., 2020; Schmidt & Hunter, 2014), which has several issues. First, adjusting at the scale level inherently ignores single items, thus masking any potential jingle or jangle effects related to scales' item content. Second, internal consistency statistics systematically underestimate reliability, leading to overcorrections (John & Soto, 2007; McCrae et al., 2011; McDonald, 1999; Sijtsma, 2009). Third, about 40 % of scale scores' variability in single-method data reflects systematic method-specific and hence unverifiable information [test–retest correlations minus cross-rater correlations; (McCrae et al., 2019; McCrae & Mõttus, 2019)]. Most of this method-specific variance likely stems from raters' stable response styles such as extreme, acquiescent, and socially desirable responding (Credé, 2010), implicit understanding of personality structure (McCrae et al., 2019), and unique item/trait interpretations (McCrae et al., 1998), besides some degree of asymmetry in raters' knowledge about the rating targets. Such method-specific effects can both inflate (e.g., response styles) or attenuate (idiosyncratic interpretations) scales' correlations, but since they are systematic, adjusting correlations for unreliability does not address them. Other approaches for removing measurement error, such as *meta*-analysis and structural equation modelling (Lebreton et al., 2014; Sackett, 2014) are also susceptible to these issues. Hence, adjustments for unreliability often misestimate scales' empirical similarity and mask their true overlap.

1.2. Evaluating overlap between items

Alternatively, one can compare the items used to assess each construct: items are, after all, how a construct is functionally defined and empirically instantiated. Grit (Duckworth et al., 2007) and conscientiousness (Soto & John, 2017) provide an illustrative example: though conceptually distinct, their assessments are often difficult to distinguish empirically (Ponnock et al., 2020), because the items used to measure grit are often very similar to those in the self-discipline and achievement striving facets of conscientiousness (Credé et al., 2017). For example, items within the most commonly used grit scale, such as "I finish whatever I begin" and "I am a hard worker," are effectively indistinguishable in content from items used to operationalize the conscientiousness facets of achievement striving and persistence facets within the International Personality Item Pool (IPIP; Goldberg, 1999), such as "I carry out my plans" and "I work hard."

One approach to evaluating item redundancy across scales is to use human-rated estimates of the semantic similarity of their items. Such ratings have long been used in research areas such as linguistics and cognitive science (Miller & Charles, 1991; Rubenstein & Goodenough, 1965; Whitten et al., 1979), and more recently, form a central aspect in training natural language processing algorithms (Christensen & Kenett, 2021; Cutler & Condon, 2022; Rosenbusch et al., 2020). Likewise, psychologists regularly perform informal assessments of semantic similarity when deciding upon which items to select for a scale (e.g., Banks et al., 2016; Christensen et al., 2023; Mõttus, 2016; Newman et al., 2016). Humans' proficiency in understanding language and its nuances thus makes their subjective judgment of semantic similarity a reasonable criterion to use to evaluate the extent to which two items mean the same thing – this is especially true when ratings are averaged across many individuals, which accounts for idiosyncratic interpretations to provide a consensus estimate of semantic similarity.

However, while judging the semantic similarity of a few items is a relatively straightforward task, this becomes time- and resource-intensive as the number of items increases. For example, if we were interested in evaluating the overlap of a 10-item grit and 10-item perseverance scales, we should collect 10*10 = 100 semantic similarity judgments per rater to index all possible redundancies.[2] The number of judgments to be made can become much larger if we are interested in all items within an item-pool, where the number of distinct judgments equals $N_k(N_k\text{-}1)/2$, with N_k equaling the number of items. For instance, a 100-item pool will have 4,950 distinct judgments to make, and very large item pools such as those used within the IPIP or SAPA (both N_ks ≥ 2,500; Condon et al., 2017) have an astronomical number of item-pairs.

Moreover, shared semantic information is only one reason that items can correlate with one another and be empirically redundant. Empirical correlations reflect not just semantic overlap, but also veridical processes underlying personality structure, such as the influences of common latent causes or direct causal influence between the traits denoted by the items. That items correlate due to some underlying cause is a fundamental assumption in nearly all psychological measurement (e.g., a correlation between "Fear for the worst" and "Get stressed out easily" would usually be understood to result not from their being semantically redundant, but rather from their shared indexing of a latent anxiety

[2] Alternatively, some investigators rate the similarity of the definitions of the constructs the scale is designed to measure (e.g., Larsen & Bong, 2016). But placing faith in this system involves trusting the very questionable assumption that the items closely operationalize the intended construct.

facet). As such, empirical correlations among items that are semantically not perfectly redundant should generally exceed the items' semantic similarity – especially when these correlations are disattenuated for error and method biases: even semantically non-redundant scale content can be empirically redundant, and semantic similarity may be a sufficient but not necessary condition for empirical redundancy.

As one example, Wood et al. (2023) argued that adjusting inter-item correlations for unreliability using the items' test–retest reliabilities in self-report data provides more accurate population estimates of the empirical association[3] among items. Specifically, they adjusted "lagged" correlations between items by their retest correlations over that interval (e.g., dividing the correlation between X and Y when measured 1 week apart by their respective retest correlations over the same interval). This approach is highly scalable provided that test–retest data are available; it also provides adjusted estimates that incorporate both semantic similarity among items as well as their overlaps due to substantive reasons. The authors found the rankings of adjusted correlations to better align with the rankings of human-rated semantic similarity than the rankings of unadjusted correlations, in support of their hypothesis that the former would more effectively detect item redundancy than the standard raw-score or unadjusted correlations often used for this purpose.

Despite examining only the item pairs with highest empirical similarity, Wood et al. (2023) found that semantic similarity ratings were comparably quite low: of 402 inter-item pairs, only 18 were consensually rated as being at least "very similar in meaning," of which just two were judged to "mean the same thing." This led them to propose various thresholds above which retest-adjusted correlations may be necessary and/or sufficient to index semantic redundancy (e.g., > 0.90 as *necessary* for two items to be rated as "meaning the same thing" and *sufficient* for them to be rated "fairly similar in meaning"). Conversely, they suggested that items with high adjusted correlations but comparatively low semantic similarity ratings may indicate "functional relationships" between them (i.e., reasons beyond simple semantic redundancy discussed earlier). Both proposals imply that inter-item correlations indicate redundancy not captured solely by semantic overlap, especially when accounting for measurement error.

If so, then researchers need to ensure that correlations index shared information as accurately as possible – to avoid both the jingle fallacy at scale level and "bloated specifics" or narrowly defined item content within scales (Cattell & Tsujioka, 1964; Cortina et al., 2020). Researchers usually want to assess the full breadth of constructs with maximal efficiency, making empirically redundant content is costly and wasteful. Here, we propose a more pointed examination of whether and when adjusted correlations can match and even exceed semantic similarity. Specifically, we suggest that the approach used by Wood et al. does not address systematic method biases that can distort correlations, and that the most scalable and accurate approach to adjusting correlations is by combining information from multiple raters.

1.3. Using multiple raters to account for method biases

Adjusting correlations for test–retest unreliability alone does not account for the effects of method-specific (e.g., within-rater) biases that are 1) stable over time, 2) make up much of items' variance, and 3) can both attenuate and/or inflate items' correlations (McCrae, 2018; McCrae, 2015). To also adjust correlations for single method-specific effects, we need multi-method data, which allows us to estimate correlations between items free of not only random and occasions-specific errors but also method-specific biases (Mõttus et al., 2024).

Besides valid trait variance ('true' score in Classical Test Theory), single sources of information for a target (e.g., a self- or informant-report) reflect stable response biases (e.g., response styles like socially desirable, acquiescent or extreme responding), raters' unique views about the rating target that generalizes across items (Credé, 2010; Wood et al., 2017), and an implicit hierarchical model of how items assessing similar trait content hang together (Implicit Personality Theory [IPT]; Borkenau, 1992; McCrae et al., 2019), all of which affect their responses to multiple items. Hence, these are sources of general or *shared method* bias. McCrae et al. also found that raters have *item-specific* (unique) *method* biases: idiosyncratic interpretations of single items which are independent both of other raters and of other items. When method biases are common to multiple items, they tend to inflate their correlations; when they are specific to single items, correlations are attenuated.

Consider the single item "I worry a lot," used to assess both the broad neuroticism domain and narrower anxiety facet. Raters may have biases that generalize across all items regardless of their content, with some agreeing (or disagreeing) more with all items, some using more extreme (or middling) responses, and some giving more socially (un)desirable responses. They may also tend to rate this item more similarly to other neuroticism items than those of other domains, as well as to other items in the anxiety facet than those of other facets of neuroticism, inflating correlations with anxiety items and comparatively attenuating correlations with items from other facets and (especially) domains (McCrae et al., 2019). On the other hand, raters may systematically differ in how they respond to this item *in particular*. These systematic differences could arise from different interpretations of what it means to "worry" (e.g., internal vs external manifestations of worry), to do something "a lot" (e.g., every hour vs a few times per week), and/or their combination.

1.4. The present study

We argue that the most accurate estimate of two variables' covariance (i.e., empirical overlap) is captured by correlations taken from multiple sources; we provide the algebraic explanation below, whereas more details are presented in Mõttus et al. (2024). In this replication and extension of Wood et al. (2023), we tested whether, when compared to unadjusted and retest-adjusted correlations, cross-rater agreement-adjusted correlations which account for both method biases and measurement error 1) better correlated with and 2) were generally higher in absolute magnitude than items' judged semantic similarity, except for the semantically fully redundant items for which empirical and semantic similarities should be close. If so, our work would provide researchers with a straightforward approach to thinking about and accurately assessing overlaps among personality variables, or lack thereof, allowing them to systematically identify jingle-jangle fallacies. This would also underscore the need to recruit multi-source data, where possible.

2. Method

2.1. Our model of variance decomposition

In line with recent work (McCrae, 2015, 2018; McCrae & Mõttus, 2019; Mõttus et al., 2024), we extend the model proposed by Wood et al. (2023) to incorporate method biases. Our approach distinguishes between six components that contribute to the observed scores of two items (or equally, scales), X and Y, and can be distinguished when a researcher has data from two methods (e.g., self- and informant-reports) and two occasions (e.g., self-reports across a few weeks). Among other applications, the model is useful when evaluating jingle- and jangle-fallacies (e.g., two measures of Anxiety; a measure of grit and another of conscientiousness); however, it can be applied to any pair of items or scales to estimate their relations and the sources of these relations.

Each item's score is constituted by six components: 1) true score that is *common* to both items and thus shared between both testing-occasions and methods (S); 2) true score that is *unique* to each item but shared between testing-occasions and methods (U); 3) general method-specific

[3] Wood et al. (2023) refer to the adjusted correlations as estimates of *informational* similarity, indicating that correlations provide an estimate of the shared information indexed by two measures.

bias shared between items and testing-occasions (*GM*); 4) method-specific bias unique to each item but shared between testing-occasions (*UM*); 5) occasion-specific effects shared between items administered at the same time but not across methods or testing-occasions (*O*); and 6) random error unique to rater, occasion, and item (*R*). In line with McCrae (2018, Supplemental Materials), scores are modeled as a weighted sum of these components using weights *a*′, *b*′, *c*′, *d*′, *e*′, and *f*′, with each component an independent (i.e., uncorrelated) variable with a standard normal distribution ("standard normal variable"). We assume that these weights are the same across items, occasions, and methods (i.e., raters).[4] As such, we do not include subscripts for coefficients but do for the components.

Across the following equations, subscripts *X* and *Y* denote which components are unique to each item. Similarly, the subscripts *o* and *r* indicate when components influence scores at only one occasion or rater, respectively. The lack of these subscripts thus indicates that the component is shared across occasions and/or raters. As an example: *S* indicates that the shared true score component contributes to the scores of both items, is constant across testing occasions, and is shared between raters. Conversely, R_{Xor} indicates that the random error for item *X* is independent of item *Y* and specific to both rater and occasion.

The scores for *X* and *Y* are given by

$$X_{or} = a'S + b'U_X + c'GM_r + d'UM_{Xr} + e'O_{or} + f'R_{Xor} \tag{1}$$

and

$$Y_{or} = a'S + b'U_Y + c'GM_r + d'UM_{Yr} + e'O_{or} + f'R_{Yor} \tag{2}$$

Eq. (1) can be interpreted as such: *X* at occasion *o* and method *r* is a weighted sum of true score variance shared between *X* and *Y*; unique true score variance of *X*; general method variance of rater *r*; unique method variance specific to rater *r* and item *X*; unique occasion variance specific to occasion *o* and rater *r*; and random error variance specific to item *X*, occasion *o*, and rater *r*. The same interpretation can be applied to *Y* for Eq. (2). Components are weighted such that the overall observed item score itself is also a standard normal variable. As stated by McCrae (2018, Supplemental Materials), "[w]hen a standard normal variable is multiplied by a coefficient, the standard deviation of the resulting variable is equal to the coefficient itself (because each normal score, SD = 1, has been rescaled by that amount), and the variance of the weighted variable is the square of the coefficient" (p. 2). Thus, the variance of each item is given by

$$VAR(X) = VAR(Y) = a'^2 + b'^2 + c'^2 + d'^2 + e'^2 + f'^2 = 1 \tag{3}$$

This means the variance of each component is a proportion of the total item variance, such that if one goes up then at least one of the others goes down by necessity.

When scores are correlated, their correlation coefficient results from the combination of components shared between the variables; their relative contributions can be traced by combining different methods or occasions either for the same item (e.g., test–retest reliability of *X*) or between items (e.g., correlation of self-reports of *X* with informant-reports of *Y*).

More specifically, McCrae (2018, Supplemental Materials) demonstrated that, given the simplifying assumptions of the model, "the correlation between two variables is equal to the sum of the products of the corresponding coefficients of shared components of variance" (p. 4). For example, an item's test–retest reliability (r_{tt}) is the proportion of total score variance that is free of transient and other random sources of error (O_{or}, R_{Xor}, R_{Yor}) but retains both shared and unique stable rater-specific method biases (GM_r, UM_{Xr}, UM_{Yr}) as well as true score variance (S, U_X, U_Y). As such, test–retest reliability is given by

$$r_{\text{tt}}(X) = \frac{(a'^2 + b'^2 + c'^2 + d'^2)}{VAR(X)} \tag{4}$$

and

$$r_{\text{tt}}(Y) = \frac{(a'^2 + b'^2 + c'^2 + d'^2)}{VAR(Y)} \tag{5}$$

But because each item can be scaled to have unit variance, these equations can be simplified to

$$r_{\text{tt}}(X) = r_{\text{tt}}(Y) = \frac{(a'^2 + b'^2 + c'^2 + d'^2)}{1} = a'^2 + b'^2 + c'^2 + d'^2 \tag{6}$$

The same principle can be applied to all other correlations calculated between items, occasions, and raters, so we omit the denominator in further equations.

Cross-rater agreement (r_{ca}) for a given item contains both shared and unique sources of true score variance but *not* method biases, occasion-specific error, or random error, as these are assumed to be independent between raters. Thus, r_{ca} is given by

$$r_{\text{ca}}(X) = r_{\text{ca}}(Y) = a'^2 + b'^2 \tag{7}$$

Unadjusted observed correlations in single-source, cross-sectional data (e.g., correlations between self-reports) contain all *shared* sources of valid and invalid variance between items (*S*, GM_r, and O_{or}), meaning they can be estimated as

$$r(X, Y) = a'^2 + c'^2 + e'^2 \tag{8}$$

Their cross-lagged correlation (*X* at time 1 correlated with *Y* at time 2 and vice versa; r_{cl}), presented in Wood et al. (2023), does not contain occasion-specific effects (O_{or}) but does retain the shared method biases (GM_r), so this correlation is given by

$$r_{\text{cl}}(X, Y) = a'^2 + c'^2 \tag{9}$$

Finally, their cross-source correlation (self-reports of *X* correlated with informant-reports of *Y* and vice versa; r_{cs}) leaves only the shared source of true score variance (*S*) and is given by

$$r_{\text{cs}}(X, Y) = a'^2 \tag{10}$$

Eq. (10) thus represents the correlation between the two constructs' scores, free of all other systematic, non-transient sources of non-trait information that could inflate the correlation – that is, it retains only the variance common to both items which is shared across raters and occasions. However, r_{cs} is unduly attenuated by the non-valid components of variance in *X* and *Y* that are specific to either variable (UM_{Xr} and UM_{Yr}) and hence do not correlate, as well as occasion-specific effects (O_{or}) and random error (R_{Xor} and R_{Yor}).

However, because *cross-rater*, *same-item* correlations (r_{ca}) of both *X* and *Y* are attenuated by the same factors but also reflect the correlations among true scores not shared by items (U_X and U_Y), the *ratio* of [cross-rater, cross-item]: [cross-rater, same-item] correlations approximate items' empirical overlap free of 1) single-method effects, 2) occasion-specific biases, and 3) random error; this ratio [S / (S + U_X + U_Y)] represents the items' *disattenuated* bias- and error-free correlations.

The logic to adjust r_{cl} for unreliability as used by Wood et al. (2023) is very similar: it is the ratio of [cross-time, cross-item] : [cross-time, same-item] correlations. Thus, the resultant disattenuated correlation – like agreement-adjusted correlations – is free of O_{or}, R_{Xor}, and R_{Yor}; unlike agreement-adjusted correlations, retest-adjusted correlations retain the shared method bias GM_r, which may inflate or deflate the

[4] This is in line with the model proposed by McCrae (2018). However, while a strict assumption in some cases, see evidence in the Online Supplement that our proposed statistics are robust to small deviations from this assumption.

correlations.

Full algebraic derivation of each variance component is provided in the Online Supplemental Materials (https://osf.io/cb967/?view_only=916df8623b494947aed52c4d7aa100e4). We also direct readers to several additional papers for further reading on this topic: For further background on the variance decomposition model, see McCrae (2015; 2018) – including the supplemental materials of the latter – and McCrae and Mõttus (2019); for an overview of the logic of the adjustment with a focus on retest-adjusted correlations, see Wood et al. (2023); for an applied instance of using cross-rater data to disattenuate correlations between personality and life satisfaction, see Mõttus et al. (2024).

Calculations for both adjustments are described below. In all the above equations, X and Y can range in specificity from single items to scales of any length. Here, we assess overlaps among personality variables at the level of single items because these are the building blocks of scales; we henceforth refer to the components solely as they relate to items.

2.2. Calculating adjusted correlations

We use $\widehat{\rho}$ to denote adjusted correlations to indicate they are derived from the combination of multiple correlation coefficients and use the subscripts "tt" and "ca" to indicate they are adjusted by r_{tt} and r_{ca}, respectively.

Retest-adjusted correlations. Given two items X and Y, retest-adjusted correlations ($\widehat{\rho}_{\text{tt}}$) are calculated as the geometric mean of the cross-lagged correlations divided by the geometric mean of the test–retest reliability of each item:

$$\widehat{\rho}_{\text{tt}} = \frac{\sqrt{r_{X_{\text{time1}},Y_{\text{time2}}} r_{X_{\text{time2}},Y_{\text{time1}}}}}{\sqrt{r_{X_{\text{time1}},X_{\text{time2}}} r_{Y_{\text{time1}},Y_{\text{time2}}}}} = \frac{r_{\text{cl}}}{r_{\text{tt}}} \tag{11}$$

This adjustment works for any pair of items that have been measured at the same two timepoints, irrespective of the time between them.[5]

Agreement-adjusted correlations. Agreement-adjusted correlations ($\widehat{\rho}_{\text{ca}}$) are computed in largely the same manner but use item cross-rater agreement rather than test–retest correlations as the basis of reliability adjustments. Given two items X and Y, agreement-adjusted correlations are the geometric mean of the cross-source correlations (i.e., self-report of item X with informant-report of item Y, and vice versa) divided by the geometric mean of the cross-rater agreement of each item:

$$\widehat{\rho}_{\text{ca}} = \frac{\sqrt{r_{X_{\text{self}},Y_{\text{informant}}} r_{X_{\text{informant}},Y_{\text{self}}}}}{\sqrt{r_{X_{\text{self}},X_{\text{informant}}} r_{Y_{\text{self}},Y_{\text{informant}}}}} = \frac{r_{\text{cs}}}{r_{\text{ca}}}. \tag{12}$$

Similarly to $\widehat{\rho}_{\text{tt}}$, $\widehat{\rho}_{\text{ca}}$ can be estimated for any data that comes from multiple sources.

Note that when the values in Eqs. (11) and (12) are near 0, this equation will encounter difficulties (e.g., if either or both values are exactly 0 or negative). Thus, to avoid taking the square root of negative numbers, we estimated the absolute value of the product of the items' correlations before taking the square root (effectively, the geometric mean of the absolute value of each correlation), then added the sign of the product back in before dividing by the geometric mean of each item's cross-rater agreement. In practice, this does not result in problematic adjustments, because in most data 1) $r_{X_{\text{time1}},Y_{\text{time2}}} \approx r_{X_{\text{time2}},Y_{\text{time1}}}$ and $r_{X_{\text{self}},Y_{\text{informant}}} \approx r_{X_{\text{informant}},Y_{\text{self}}}$ (see Supplemental Figure S1) and 2) r_{tt} and r_{ca} are positive and non-zero values for all items.

2.3. Interpreting adjusted correlations

Both retest- and agreement-adjusted correlations can be understood as a ratio of [variance shared between items across time/source] to [variance between items that is both shared across time/source *and* unique to each item]. For retest-adjusted correlations, the resultant $\widehat{\rho}_{\text{tt}}$ answers the question: "how much proportionally lower is the correlation of item X with item Y over a particular measurement interval than the average retest correlations of X and Y with themselves over the same time interval?" If X and Y captured identical information, net of transient errors, the two correlations should be identical ($\widehat{\rho}_{\text{tt}} = 1$). And similarly, for agreement-adjusted correlations, the resultant $\widehat{\rho}_{\text{ca}}$ answers the question "how much lower do self-ratings of X and other-ratings on Y (and vice versa) correlate from the average self-other agreement of X and Y?" If X and Y captured identical information, net of transient errors and systematic biases, the two correlations should be identical. In other words, the adjustments equal 1 if X and Y convey no distinct information.

To illustrate this point and how the calculations work, consider the item pair "Am able to control my cravings" and "Easily resist temptations" taken from the data used in this study. Their cross-lagged correlations are $r_{X_{\text{time1}},Y_{\text{time2}}} = .50$ and $r_{X_{\text{time2}},Y_{\text{time1}}} = .48$ with retest reliabilities $r_{\text{tt}}(X) = .61$ and $r_{\text{tt}}(Y) = .57$. The former two are placed in the numerator, the latter in the denominator, and their geometric means are calculated ($r_{\text{cl}} = .49$, $r_{\text{tt}} = .59$), and the retest-adjusted correlation is their ratio: $.49 / .59 = .83 = \widehat{\rho}_{\text{tt}}$. Meanwhile, their cross-source correlations are $r_{X_{\text{self}},Y_{\text{informant}}} = .21$ and $r_{X_{\text{informant}},Y_{\text{self}}} = .21$ (geometric mean $r_{\text{cs}} = .21$) with cross-rater agreements $r_{\text{ca}}(X) = .24$ and $r_{\text{ca}}(Y) = .20$ (geometric mean $r_{\text{ca}} = .22$). As with the retest-adjustment, these coefficients are entered into the equation and their ratio is calculated: $.21 / .22 = .97 = \widehat{\rho}_{\text{ca}}$. Thus, the final correlations for this item pair are $r = .56$, $\widehat{\rho}_{\text{tt}} = .83$, and $\widehat{\rho}_{\text{ca}} = .97$.

For $\widehat{\rho}_{\text{tt}}$, this can be understood as meaning that across a roughly two-week interval (the test–retest interval of these data), participants' consistent self-appraisal of the relationship between [their ability to control cravings] and [the ease with which they resist temptations] was about 83 % the magnitude of how consistently they rated themselves on each these items. The interpretation of $\widehat{\rho}_{\text{ca}}$ is this: if participants and their informants rated different items about the rating target, they agreed only 3 % less compared to both rating exactly the same items. These interpretations can be equally applied to any item pair for either their retest- or agreement-adjusted correlation.

2.4. Testing the model

We began by identifying datasets with personality data that was available both at multiple timepoints (test–retest) and from multiple sources (self- and informant-reports). Given that the interpretation of item content may vary across languages, we only examined data that had been collected in the same language for both sources. We then calculated the three types of empirical similarity estimates to pit them against one another in correlating with semantic similarity: *raw* (i.e., unadjusted, zero-order; r), *retest-adjusted* (i.e., free of random measurement error and occasion-specific biases but not single-method biases; $\widehat{\rho}_{\text{tt}}$); and *agreement-adjusted* (i.e., free of both random measurement error and occasion-specific biases and single-method method biases; $\widehat{\rho}_{\text{ca}}$) correlations. After indexing all possible item pairs for their empirical similarity, we sampled $k = 200$ pairs from each inventory to be evaluated for semantic similarity. Notably, whereas Wood and colleagues (2023) selected only the item-pairs with the highest correlations to be rated for semantic similarity, we sampled item-pairs across the full positive correlation range, from 0 to 1. Finally, to test the variance decomposition model, we compared semantic similarity estimates against the three different empirical similarity conditions for each survey, as well as all selected item pairs ($k = 400$) combined in a mega-

[5] Wood and colleagues (2023) differ from this prescription by using the arithmetic mean (rather than the geometric mean shown in Equation (11) in the numerator, or $(r_{X_1,Y_2} + r_{X_2,Y_1})/2$. However, in most cases the difference between these estimators will be negligible (with the geometric mean necessarily being smaller).

analysis.

Full details of these analyses are described below, and the code and data necessary to replicate them are available at https://osf.io/cb967/?view_only=916df8623b494947aed52c4d7aa100e4. Participants provided informed written consent for all data. Details of the ethics for data used to estimate empirical similarity are available where the original studies are described. For semantic similarity ratings, all participants provided their consent in the online study, which was approved by the University of Edinburgh School of Philosophy, Psychology, and Language Sciences Research Ethics Committee approved on 9 February 2023 (Ref 188-2223/3).

3. Samples and surveys

3.1. HEXACO personality inventory – Revised

The 100-item HEXACO Personality Inventory – Revised (HEXACO-PI-R or HEXACO-100; Lee & Ashton, 2004) assesses six broad personality factors: Honesty-Humility, Emotionality, eXtraversion, Agreeableness, Conscientiousness, and Openness to Experience. Each factor can be divided into four facets, measured by four items apiece; there is one interstitial facet, Altruism, that loads on Agreeableness, Honesty-Humility, and Emotionality.

Estimates of test–retest reliability (N = 416) came from a sample originally reported in Henry et al. (2022), where participants were recruited from Prolific Academic to complete the HEXACO-PI-R twice in a period of approximately 13 days. Cross-rater agreement data (N = 2,863 pairs) were collected in a student sample over several years (Lee & Ashton, 2018) among a group of well-acquainted individuals: on average, informants claimed to know the target for median = 4 years, and rated their knowledge of the target as 9 on a 10-point scale (Ashton & Lee, 2010). Full details on the samples, data collection, and results may be found in the original publications.

3.2. 100 Nuances of Personality

The 100 Nuances of Personality (100NP) is a personality item pool designed to measure personality with maximum breadth and minimum redundancy, in line with principles laid out by Condon et al. (2021). The item pool was developed between 2019 and 2022 with iterative waves of items tested to measure Big Few domains and facets, plus traits not well-captured by these models (e.g., Dark Triad, gratitude, sexuality, humor, competitiveness). Items were selected when they demonstrated desirable empirical properties (e.g., variance, test–retest reliability) and minimal redundancy with other items, except for a few pairs retained to either test for acquiescent responding or supplement traits with poor empirical assessment. A summary of the scale's development and current uses is described in Henry & Mõttus (2023).

The 100-NP consists of 198 items, 192 of which have been assessed for both test–retest reliability (average retest interval approximately 13 days; $N = 888$) and cross-rater agreement ($N = 656$) in the data used for the present study. Cross-rater data were collected as part of a graduate student research project, primarily through advertisements on social media and contacting friends and family of the study team. About one third of these participants (n = 229) also provided test–retest data; the remainder of the test–retest sample (n = 659) was recruited using Prolific Academic.

For the HEXACO-100, the unadjusted correlations were estimated on the combined self-report and T1 samples (N = 3,279); as the 100NP samples contained some of the same respondents, we used only the retest sample ($N = 888$).

All data were collected in English.

3.3. Estimating empirical and semantic similarity

3.3.1. Empirical similarity (r, $\widehat{\rho}_{tt}$, and $\widehat{\rho}_{ca}$)

In order for these adjustments to work, items must have sufficiently high r_{tt} or r_{ca} and large enough samples to stabilize their estimates. As the r_{tt} and r_{ca} values needed to make these reliability adjustments approach zero, the $\widehat{\rho}_{tt}$ and $\widehat{\rho}_{ca}$ estimates created from Eqs. (1) and (2) will regularly exceed 1. As such, large samples, small retest intervals, and relatively knowledgeable informants are all valuable to making necessary conditions to conduct these analyses. The sample sizes used in this study were sufficient to reasonably estimate r_{tt} and r_{ca}, and all items had sufficient levels (i.e., all estimates > 0) of these properties to perform the adjustment. HEXACO-PI-R items had median $r_{tt} = 0.65$ ($M = 0.65$, $IQR = 0.59$ to 0.70, range = 0.39 to 0.84) and median $r_{ca} = 0.28$ ($M = 0.27$, $IQR = 0.21$ to 0.32, range = 0.17 to 0.46). 100NP items had median $r_{tt} = 0.69$ ($M = 0.70$, $IQR = 0.65$ to 0.73, range = 0.57 to 0.84) and median $r_{ca} = 0.36$ ($M = 0.37$, $IQR = 0.31$ to 0.44, range = 0.17 to 0.66).

3.3.2. Semantic similarity

Prior to selecting item pairs for the semantic similarity rating task, we removed any pair that was negatively correlated in any of the three adjustment conditions. We selected only positively correlated item pairs – essentially constraining empirical overlap from "no empirical similarity whatsoever" to "empirically identical" – because the difference between being "completely unrelated" (no similarity) to "meaning similar (or the same) things but in opposite directions" (a high negative correlation) is arguably more complicated in a rating task like this.[6] While some items may be obvious semantic opposites (e.g., one of the highest negatively-correlated 100NP item pairs was "Break my promises" and "Keep my promises," r = -0.62, $\widehat{\rho}_{tt}$ = -0.89, and $\widehat{\rho}_{ca}$ = -1 in raw, retest-adjusted, and agreement-adjusted conditions) many others are less so. Even at moderate to large negative correlations, this becomes evident. For example, the items "Am relaxed most of the time" and "My feelings are easily hurt" had r = -0.34, $\widehat{\rho}_{tt}$ = -0.48, and$\widehat{\rho}_{ca}$ = -0.65. Despite sharing about two thirds of their error- and bias-free information, these items seem to measure two different things at face value.

We thus deemed rating the semantic (dis-)similarity of negatively correlated items too complex a task for lay raters and chose to assess item pairs in the correlation range [0,1]. There were originally a total possible 18,336 and 4,950 non-redundant pairwise correlations (i.e., $(N_k N_{k-1})/2$, whereN_k = the inventory's number of items) in the 100NP and HEXACO-100, respectively. After removing all pairs that were negatively correlated in any of the three conditions, 8,227 and 1,842 pairs remained.

The remaining pairs were next indexed for overall empirical similarity, calculated as an average of each item pair's overall rank order in the three correlation conditions. For example, a pair ranked one (highest similarity) in raw, three in retest-adjusted, and two in agreement-adjusted correlations would receive an overall rank of two; another ranked 325, 462, and 350 would receive an average rank of 379, and so on. We assigned the rank based on all three correlation estimates to avoid favoring any one of them.

After ranking all pairs, we selected 200 to present to raters, a task approximately the length of a personality questionnaire (e.g., the NEO Personality Inventory – Revised has 240 items; Costa & McCrae, 1992). Wood et al. (2023) restricted the range to only the top 100 empirically similar items from each inventory they tested, which they argued resulted in the relatively low correlations (q) between empirical and semantic similarity that they observed and posited that sampling the full

[6] This points to another advantage of using correlations instead of human-rated semantic similarity, as the former are not vulnerable to this complexity: the interpretation of adjusted correlations is identical whether they are positive or negative.

range [0,1] would likely lead to higher q values. Thus, we aimed to extend Wood et al.'s study by including item pairs with a wider range of correlations.

This required oversampling from the upper end of the correlation range, as large correlations between items are relatively rare. Across the two datasets, the highest unadjusted correlation for both inventories was $r = 0.60$, and for the agreement-adjusted correlations – most susceptible to high values due to lower values in the denominator – median positive correlations were only $\widehat{\rho}_{ca} = 0.19$ ($M = 0.24$; $IQR = 0.11$ to 0.34, range = 0 to 0.99) and $\widehat{\rho}_{ca} = 0.20$ ($M = 0.26$; $IQR = 0.11$ to 0.37, range = 0 to 0.99) for 100NP and HEXACO-100 items, respectively. Thus, to ensure adequate sampling of high empirical similarity, we first selected the top 100-ranked pairs for each survey.

For the remaining 100 pairs in each inventory, we tried multiple different algorithms to select pairs based on exponentially decreasing rank of empirical similarity. The chosen rank was based on the following formula: $\mathrm{rank}_j = \lfloor j^x k / n^x \rfloor$, where j is the index (i.e., 1, 2, 3, …, 100), k is the total number of pairs to choose from and n is the total number of pairs to be selected. While j, k, and n are effectively fixed based on the study design and data structure, x is selected by the researcher to adjust the skew of the ranks chosen. It effectively serves as a "penalizer" such that higher values select more low numbers (i.e., item pairs with lower ranks and thus higher empirical similarity); higher values of x thus result in more correlations at the high end of the distribution. We ultimately settled on a 'quadratic' approach (i.e., $x = 2$) because this resulted in a distribution with slight negative skew (i.e., oversampled high correlations) and full coverage of the [0,1] range – visualizations of these distributions can be seen in the Online Supplemental Materials (https://osf.io/yxq4s); code to simulate different sampling approaches (e.g., 'cubic' or 'quartic') is also available.

3.4. Semantic similarity ratings

We recruited $N = 25$ participants from Prolific Academic to rate the 200 pairs of items of either the 100NP or HEXACO-PI-R for their semantic similarity, where ratings were provided in Qualtrics Survey Software. Item pairs were presented in a random order, one at a time, and with random placement of items on the left vs right side of the pair. Participants were asked to rate pairs for the extent to which they "mean the same thing," where ratings were given on the following 5-point Likert scale: "0 – Have completely different meanings"; "1 – Have slightly similar meanings"; "2 – Have fairly similar meanings"; "3 – Have very similar meanings"; and "4 – Have essentially the same meaning." Participants were paid £5.00 for successful completion of the task.

We estimated the reliability of semantic similarity ratings using model 2 intra-class correlations (*ICC*s), where raters and item pairs were considered a random sample of judges and targets, respectively. To detect inattentive responding, we conducted a principal component analysis, where any rater who loaded < 0.30 on the first principal axis was excluded from calculation of overall semantic similarity; this was equivalent to a correlation of $r = 0.09$ with the average similarity profile of the other raters. After removing participants with inattentive responses, we calculated mean semantic similarity scores for each pair by averaging across all ratings for each pair then dividing these scores by 4 to be on a 0–1 scale for easier comparison with correlations.

3.5. Main analyses

For consistency, we adopt a similar notation system as Wood et al. (2023) in the following sections. As described above, we use r to refer to raw inter-item correlations and $\widehat{\rho}$ to refer to reliability-adjusted inter-item correlations. Like in Wood et al., we use q for correlations describing how these inter-item correlations in turn related to one another and with semantic similarity estimates.

As done in Wood et al. (2023), we calculated Spearman's correlations (q) between the empirical similarity estimates for each correlation condition and the semantic similarity ratings, where higher correlations indicate that (un)adjusted associations align more strongly semantic overlap. We then replicated the analysis of Wood et al. by estimating the same correlations for just the top 100 empirically similar pairs.

We next conducted a "mega-analysis" for both of these analyses, as has been done previously (e.g., Beck & Jackson, 2022; Wood et al., 2023), by combining the two sets of 200 item pairs from the HEXACO-PI-R and 100NP (total $k = 400$ pairs) and estimating the same parameters to increase the precision of estimates for empirical-semantic relatedness. We then repeated the mega-analysis with only the top 200 empirically similar pairs from these 400 pairs.

To estimate the absolute similarities and differences between semantic similarities and correlations, we plotted the two against each other and compared across correlation types. Because both indices are on a [0,1] scale, we could thus approximate instances where semantic similarity ratings exceeded their correlations and vice versa. While putting semantic similarity on a [0,1] scale is a transformation that is not interpretable with the same precision as a correlation coefficient, the average ratings correspond to a location on a continuum from "completely different" to "essentially the same" and therefore provide an approximate point of comparison.

4. Results

4.1. Semantic similarity

Four respondents were removed for inattentive responding, leaving $N = 24$ and $N = 22$ raters for HEXACO-PI-R and 100NP item pairs, respectively. Reliability estimates for semantic similarity ratings were excellent, with single- and average-rater reliabilities of $ICC = 0.42$ (95 % CI = [.35, 0.49]) and $ICC = 0.95$ [.93, 0.96] for the HEXACO-PI-R pairs, and $ICC = 0.45$ [.38, 0.51] and $ICC = 0.95$ [.93, 0.96] for the 100NP pairs.

The ten item pairs rated with the highest semantic similarity (*SS*) are displayed in Table 1, and distributions of *SS* ratings can be seen in Fig. 1. HEXACO-PI-R pairs had median $SS = 0.19$ ($M = 0.25$, $IQR = 0.06$ to 0.44, range = 0 to 0.79); 100-NP item pairs had median $SS = 0.25$ ($M = 0.31$, $IQR = 0.10$ to 0.48, range = 0.01 to 0.86). In other words, the average item pair in our sample was rated as two on a five-point Likert scale, or as having "slightly similar meanings." Many item pairs were semantically unrelated.

As authors, we were surprised at how even the highest semantic similarity estimates were relatively low according to our scales, even for items that could, by usual psychological inventory construction standards, be considered interchangeable items measuring a narrow personality construct. For example, one of the most empirically similar HEXACO pairs (0.58, 0.84, and 0.97 for r, $\widehat{\rho}_{tt}$, and $\widehat{\rho}_{ca}$, respectively) – "If I knew that I could never get caught, I would be willing to steal a million dollars" and "I'd be tempted to use counterfeit money, if I were sure I could get away with it" – only had $SS = 0.68$. The highest rating in 100NP items was between "Enjoy cooperating with others" and "Like being part of a team" ($SS = 0.86$). For reference, Wood et al. (2023) suggested that average semantic similarity ratings exceeding 3.5 on original 0-to-4-point Likert scale – equivalent to 0.90 on the present scale – indicated items were 'semantically redundant'; not one pair in the present study crossed this threshold. This could suggest that non-psychologists could see phrase-like items' meanings as much more distinct and nuanced than scale constructors may often assume. It could also point to a tendency among the scale designers to avoid including items that are highly synonymous (e.g., to avoid creating 'bloated specifics').

Somewhat lower semantic similarity among HEXACO items could result from, on average, longer and more circumscribed items. While median item length for 100NP items was 53 characters ($M = 52.63$, *SD*

Table 1
Top 10 semantically similar items in each survey.

Item X	Item Y	*SS*	*r*	$\widehat{\rho}_{tt}$	$\widehat{\rho}_{ca}$
HEXACO-100					
I rarely hold a grudge, even against people who have badly wronged me	My attitude toward people who have treated me badly is “forgive and forget”	0.79	0.58	0.86	0.93
I am an ordinary person who is no better than others	I wouldn’t want people to treat me as though I were superior to them	0.73	0.27	0.63	0.74
I would like to be seen driving around in a very expensive car	I would get a lot of pleasure from owning expensive luxury goods	0.70	0.59	0.76	0.93
I often check my work over repeatedly to find any mistakes	I always try to be accurate in my work, even at the expense of time	0.69	0.40	0.63	0.86
If I knew that I could never get caught, I would be willing to steal a million dollars	I’d be tempted to use counterfeit money, if I were sure I could get away with it	0.68	0.58	0.84	0.97
When working, I often set ambitious goals for myself	I often push myself very hard when trying to achieve a goal	0.68	0.46	0.76	0.83
I would like to live in a very expensive, high-class neighborhood	I would get a lot of pleasure from owning expensive luxury goods	0.66	0.60	0.79	0.96
I always try to be accurate in my work, even at the expense of time	People often call me a perfectionist	0.66	0.39	0.59	0.80
People think of me as someone who has a quick temper	I find it hard to keep my temper when people insult me	0.66	0.51	0.72	0.92
100NP					
Item X	*Item Y*	*SS*	*r*	$\widehat{\rho}_{tt}$	$\widehat{\rho}_{ca}$
Enjoy cooperating with others	Like being part of a team	0.86	0.60	0.83	0.88
Act without thinking	Make rash decisions	0.84	0.55	0.82	0.86
Work hard	Push myself very hard to succeed	0.84	0.57	0.8	0.79
Easily apologize when I have been wrong	Am quick to admit making a mistake	0.81	0.57	0.82	0.98
Would like to have more power than other people	Want to be in charge	0.80	0.44	0.59	0.72
Am able to control my cravings	Easily resist temptations	0.77	0.56	0.83	0.97
Find it easy to manipulate others	Use others to get what I want	0.77	0.51	0.68	0.81
Love dangerous situations	Take risks	0.75	0.56	0.71	0.91
Wear stylish clothing	Love to look my best	0.73	0.52	0.65	0.88
Think of others first	Love to help others	0.72	0.45	0.68	0.82

Note. SS = semantic similarity. r = unadjusted correlations. $\widehat{\rho}_{tt}$ = retest-adjusted correlations. $\widehat{\rho}_{ca}$ = agreement-adjusted correlations.

= 13.62, range = 23 to 89), HEXACO items had median length = 125 (*M* = 123.54, *SD* = 20.45, range = 78 to 179). However, item length was only modestly related to semantic similarity: across all 400 pairs in the mega-analysis, *SS* correlated ρ = -0.12 (*p* = *0*.024) with total number of characters in the pair. Within each inventory, item length and *SS* were completely unrelated: ρs = 0.00 (*p* = 0.982) and 0.01 (*p* = 0.917) in HEXACO and 100NP, respectively.

4.2. Comparing semantic and empirical similarity

Across all pairs and correlation conditions, for each individual inventory and in the mega-analysis combining item pairs from both inventories, semantic similarity and empirical similarity correlated highly (*q*s = 0.82-0.84; Table 2, column labeled *SS*), and there were no significant differences between *q* correlations for any of these. Thus, item pairs’ *rank-order* empirical similarity appears to approximate these pairs’ relative semantic similarity closely across a wide range of similarity and irrespective of how empirical similarity is estimated. Of course, uniformly high *q* correlations with semantic similarity across all three correlation conditions do not preclude mean differences between them or deviations from linear trends, which we will address below.

When examining only the top 100 most empirically similar pairs (Table 2, upper triangles), we saw some differentiation in the magnitude of associations, with adjusted correlations appearing to track semantic similarity slightly better than raw correlations as indexed by higher *q* correlations. Two of these differences were statistically significant by Steiger’s (1980) test of dependent correlations, with *q* s for $\widehat{\rho}_{ca}$ greater than *q*s for r in both 100NP and mega-analytic item pairs (0.51 vs 0.41, *t* = -2.05, *p* = 0.043; 0.50 vs 0.39, *t* = -2.52, *p* = 0.012, respectively) but not for HEXACO-100 (0.54 vs 0.50, *t* = 0.72, *p* = 0.47). Estimates of $\widehat{\rho}_{ca}$ were higher than *r* in all subsamples (*q* s for $\widehat{\rho}_{ca}$ 0.09, 0.07, and 0.13 greater than for *r* for HEXACO, 100NP, and the mega-analysis, respectively), but none of these differences were statistically significant, likely due to the comparatively lower dependency between $\widehat{\rho}_{ca}$ and *r* vs $\widehat{\rho}_{tt}$ and *r*.

Meanwhile, the adjusted correlations’ ($\widehat{\rho}_{tt}$ and $\widehat{\rho}_{ca}$) associations with semantic similarity showed no consistent pattern. Agreement-adjusted correlations tracked semantic similarity judgments slightly more than retest-adjusted correlations in the HEXACO-100 (*q*s = 0.59 vs 0.54), slightly less in the 100NP (*q*s = 0.48 vs. 0.51), and essentially equivalently in the mega-analysis (*q*s = 0.52 vs 0.50); none of these differences was statistically significant. In other words, among the most empirically similar pairs, retest-adjusted correlations appeared to predict the rank-order of semantic similarity as well as agreement-adjusted correlations.

4.3. How do our results compare to Wood et al. (2023)?

In their mega-analysis of 402 item pairs, Wood et al. (2023) found that retest-adjusted correlations ($\widehat{\rho}_{tt}$) were a significantly better predictor of semantic similarity ratings than unadjusted correlations (*q*s = 0.58 vs. 52; *Z* = 2.83, *p* < 0.05), a pattern which largely tracked across the four individual datasets (*k* = 100–101 pairs) they examined. In comparison, the 200 most empirically similar pairs examined here (Table 2c, upper triangle) showed *q* correlations that were slightly lower overall than those reported by Wood et al. (*q*s = 0.39, 0.50, and 0.52 for *r*, $\widehat{\rho}_{tt}$, and $\widehat{\rho}_{ca}$) but similarly indicate that adjusted correlations are better predictors of semantic similarity than unadjusted correlations. Second, our mega-analysis (*k* = 400) that contained a wider range of empirical (and subsequently semantic) similarity demonstrated a) substantially higher *q* correlations with semantic similarity (*q*s = 0.82-0.84, Table 2; column labeled *SS*) – as predicted by Wood et al. – with b) no notable difference between correlation conditions, even nominally. In the Online Supplement, we show that adjusted correlations generally had higher *q* correlations than unadjusted correlations when the range of similarity was restricted (as in Wood et al.), but this advantage declined as the range increased.

4.4. Comparing magnitudes of correlations to semantic similarity ratings

We then examined the relationship between empirical and semantic similarity by correlation condition (Fig. 2; code to produce separate plots for HEXACO and 100-NP items is available in the Online Supplement; https://osf.io/cb967/?view_only=916df8623b494947aed52c4d7aa100e4). Three patterns emerge from the plots.

First, semantic similarity tended to be lower than empirical similarity, especially at lower overall similarity levels. This suggests that

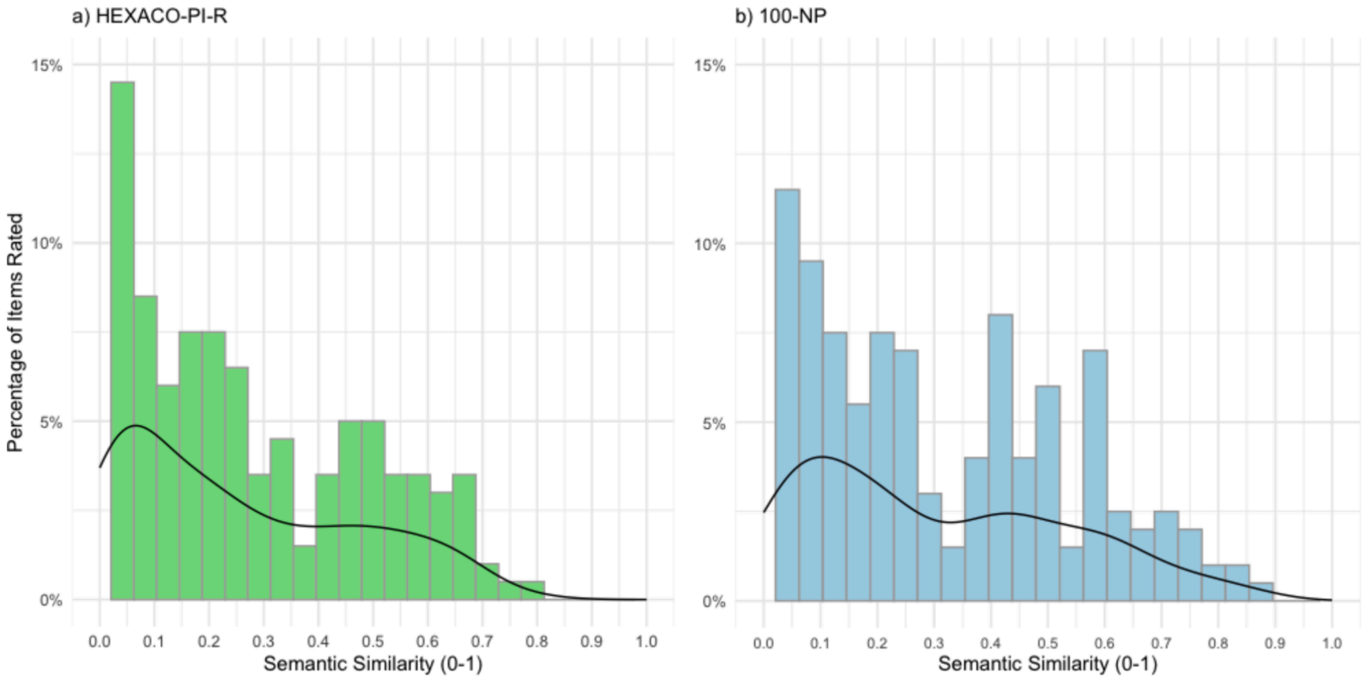

Fig. 1. Distribution of semantic similarity ratings. ***Note.*** Semantic similarity ratings have been transformed such that 0 = ’Have completely different meanings’; 1 = ’Have essentially the same meaning’; and 0.25, 0.5, and 0.75 are item pairs with ’slightly,’ ’fairly,’ and ’very’ similar meanings, respectively. The black line shows the density distribution.

Table 2
Estimated associations (q correlations) between different interitem similarity estimates.

a) HEXACO-100 (k_{pairs} = *200 below diagonal, 100 above diagonal)*								
	M	*SD*	*M* (100)	*SD* (100)	*SS*	*r*	$\widehat{\rho}_{tt}$	$\widehat{\rho}_{ca}$
SS	0.25	0.21	0.41	0.17	—	0.50	0.54	0.59
r	0.26	0.14	0.37	0.08	0.84	—	0.58	0.54
$\widehat{\rho}_{tt}$	0.40	0.22	0.58	0.13	0.83	0.90	—	0.51
$\widehat{\rho}_{ca}$	0.56	0.27	0.77	0.10	0.84	0.89	0.85	—
b) 100-NP (k_{pairs} = *200 below diagonal, 100 above diagonal)*								
	M	*SD*	*M* (100)	*SD* (100)	*SS*	*r*	$\widehat{\rho}_{tt}$	$\widehat{\rho}_{ca}$
SS	0.31	0.23	0.46	0.19	—	0.41	0.51	0.48
r	0.33	0.16	0.47	0.06	0.82	—	0.85	0.39
$\widehat{\rho}_{tt}$	0.46	0.22	0.64	0.08	0.84	0.97	—	0.47
$\widehat{\rho}_{ca}$	0.55	0.28	0.77	0.08	0.82	0.89	0.9	—
c) Mega-Analysis (*k_{pairs} = 400 below diagonal, 200 above diagonal)*								
	M	*SD*	*M* (200)	*SD* (200)	*SS*	*r*	$\widehat{\rho}_{tt}$	$\widehat{\rho}_{ca}$
SS	0.28	0.22	0.44	0.18	—	0.39	0.5	0.52
r	0.30	0.16	0.43	0.08	0.83	—	0.73	0.39
$\widehat{\rho}_{tt}$	0.43	0.22	0.61	0.10	0.84	0.94	—	0.49
$\widehat{\rho}_{ca}$	0.55	0.27	0.77	0.09	0.82	0.86	0.87	—

Note. Values below the diagonal are *q* correlations for the full range of empirical similarity. Values above the diagonal are *q* correlations for only the top 100-ranked empirically similar pairs (200 in the mega-analysis). *SS* = semantic similarity. *r* = raw (unadjusted) correlations. *M* = mean. *SD* = standard deviation. $\widehat{\rho}_{tt}$ = retest-adjusted correlations. $\widehat{\rho}_{ca}$ = agreement-adjusted correlations.

most correlations among personality items do capture something more than merely items’ semantic overlap: there may be *functional* reasons that items correlate, either due to sharing common “latent” causes or having other causal associations among them (Baumert et al., 2017; Cramer et al., 2012; Mõttus & Allerhand, 2018; Wood et al., 2015). This is probably a welcome conclusion to personality scientists.

Second, this trend is the strongest for item correlations adjusted for single-method biases ($\widehat{\rho}_{ca}$) and the weakest for unadjusted item correlations (*r*). The latter is not surprising, because unadjusted correlations are attenuated by random measurement error, besides other possible confounding factors. That correlations adjusted for random error and occasion-specific biases ($\widehat{\rho}_{tt}$) tend to be lower than those also adjusted for single-method biases ($\widehat{\rho}_{ca}$) suggests that single-method biases usually decrease correlations (e.g., due to idiosyncratic but stable item interpretations) rather than increase them (e.g., due to socially desirable response styles or IPT that lead to items with similar content being rated similarly).

Third, at highest levels of semantic similarity, the adjusted empirical correlations tend towards the semantic similarity estimates, although not all empirically highly similar items have high semantic similarity; however, the unadjusted correlations usually underestimate semantic similarity, with estimates of the latter often exceeding those for the former. This is not good news for unadjusted correlations, if they are even lower than the items’ content overlap at the semantic level.

Taken together, these patterns suggest a strong relationship (and perhaps even necessary conditions; see Dul, 2016; Wood et al., 2023) linking semantic redundancy and empirical estimates of similarity when appropriately estimated. This effect appears to be most pronounced in the agreement-adjusted condition, where semantic overlap exceeded empirical overlap in only 23/400 observations (points above the diagonal), and only eight of these (2 % of all pairs) have average semantic similarity greater than “slightly similar” (*SS* = 0.25), compared to 168 (131 with *SS* > 0.25) unadjusted and 62 (58 with *SS* > 0.38) retest-adjusted conditions. Cross-rater adjusted correlations may thus be most effective at indexing an upper limit to semantic similarity – that is, empirical similarity may be a necessary, but not sufficient condition to identify semantically redundant items.

In other words, this could indicate that higher levels of empirical similarity, properly estimated, may index truly redundant items (highly semantically similar pairs), items that are functionally related but not necessarily semantically identical (low semantic similarity), or items that are both semantically and functionally related (and therefore possibly empirically redundant). But, when empirical adjusted correlations are not high, it is very unlikely that the items will be judged to be semantically redundant by human raters, especially for the case of

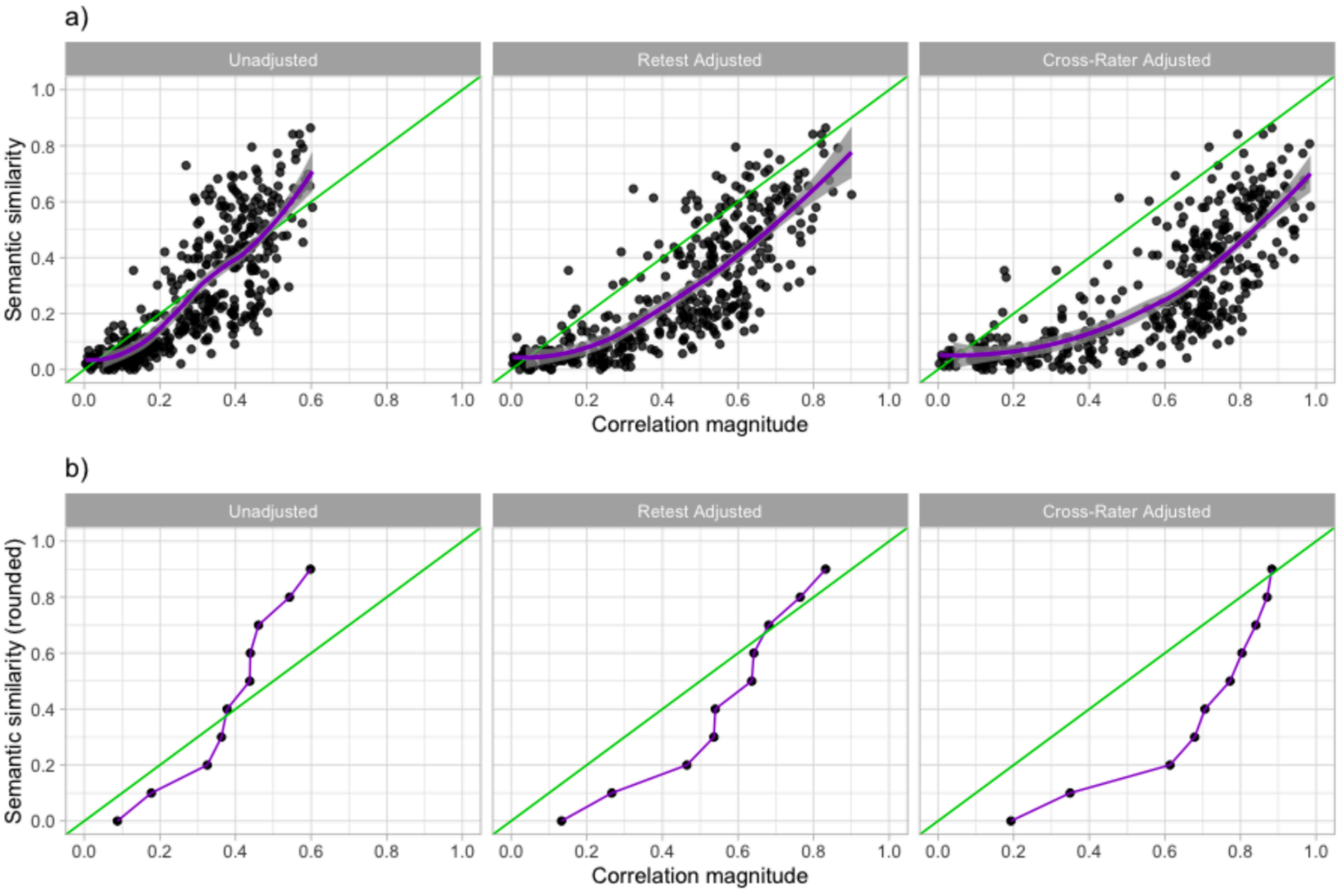

Fig. 2. Mega-analytic associations between empirical and semantic similarity by adjustment condition. ***Note.*** a) Estimates of empirical similarity against semantic similarity ratings for all k = 400 item pairs, by correlation condition. b) Summarizes the associations presented in (a) by plotting the mean correlation (r, $\widehat{\rho}_{tt}$, and $\widehat{\rho}_{ca}$) magnitude (x-axis) observed at SS ratings (y-axis) rounded to the nearest tenth (e.g., SS = 0, 0.1, 0.2, …, 0.9).

agreement-adjusted correlations. That is, low empirical similarity identifies variables that are neither redundant in function nor in meaning. In sum, especially correlations adjusted for both transient and systematic method biases do what they are supposed to do, helping to identify true overlap when it is present and rule out overlap when it is not present.

5. Discussion

Jingle and jangle fallacies are endemic in modern psychology, and one straightforward avenue for identifying them is by evaluating correlations between the scales and the items that we use to assess psychological constructs (Lawson & Robins, 2021). However, correlations based on a single source are impossible to interpret at face value, and even techniques (including latent trait modelling) that account for transient/random measurement error cannot typically remove other sources of systematic bias which constitute up to 40% of trait score variance. We thus set out to test a model that more accurately assesses the empirical similarity of any two psychological constructs assessed with subjective ratings – ranging from single items to scale-level aggregates, although we only addressed the former here – by decomposing their co-variance into different components of valid, method, and error variance both unique to each variable and shared between them.

We did so by replicating and extending the method of Wood et al. (2023), who found that correlations adjusted for random measurement error and occasion-specific effects predicted human-judged semantic similarity ratings better than unadjusted zero-order correlations – for highly similar item pairs, because they did not consider less similar item pairs. We extended their work in three important ways: by 1) comparing both unadjusted single-source and unreliability-adjusted correlations with correlations adjusted for single-method biases in addition to occasion-specific effects and random error; 2) considering item pairs with more diverse levels of similarity; and 3) examining whether and when estimates of empirical similarity (correlations) tended to match, exceed, or underestimate human-rated semantic similarity to explore different sources of redundancy between measures. That is, while accurately ranking variable pairs in similarity is important, to properly address the jingle-jangle fallacies, it is also important to accurately assess and understand their absolute similarity levels.

Across a wide range of empirical similarity (approximately [0,1]), we found that all three types of correlations closely and nearly equally tracked semantic similarity in relative terms (qs = 0.82-0.84), suggesting that compared to unadjusted correlations, retest- and agreement-adjusted correlations added little incremental value for ranking item pairs in semantic similarity, at least when looking across widely different item content. However, when we examined the same patterns within a restricted range of high similarity, we replicated Wood et al.'s findings that retest-adjusted correlations were more highly correlated with semantic similarity ratings than unadjusted correlations; agreement-adjusted correlations demonstrated a similar magnitude of q correlation compared to raw correlations but did not reach statistical significance according to Steiger's test of dependent correlations (Steiger, 1980). But perhaps more importantly, we demonstrated that both types of adjusted correlations – particularly those also adjusted for single-method biases – provided a more comprehensive coverage of the full [0,1] correlation range and exceeded semantic similarity much more often than unadjusted correlations. Taken together, these findings suggest that single-source correlations alone misrepresent the degree of similarity between items and indicate that researchers should regularly collect data from multiple timepoints at least and multiple sources at best to most accurately index overlap – both empirical and semantic.

5.1. Using empirical correlations to detect relative semantic overlap

Across a wide range of empirical similarity, the relative extents of semantic and empirical similarity estimates tracked well (ϱs = 0.82-0.84) regardless of correlation condition. The estimates are much higher than Wood et al. (2023) observed (ϱs = 0.46 to 0.69 for adjusted, ϱs = 0.40 to 0.65 for unadjusted correlations), very likely due to range restriction from their selection of only the most empirically similar item pairs. Indeed, when we limited our analyses to the top item pairs, empirical-semantic associations reduced to similar levels (Table 2, correlations above the diagonals). But we did find tentative evidence that adjusted empirical correlations outperformed unadjusted correlations in predicting semantic similarity among these very similar items: among them, adjusted correlations may be more sensitive to differences in semantic overlap because there is more variability in the former due to higher signal-to-noise ratio.

Also consistently with Wood et al. (2023), adjusted empirical similarity was usually necessary for items to be judged semantical similar. But the extent to which a pair of items *empirically provide the same information about respondents* and the extent to which they *appear to mean the same thing* remain distinct, as while meeting the former condition may be necessary to fulfil the latter, the reverse is not necessarily true. In many cases, empirical relations among items are not simply due to their semantic overlap, suggesting that there truly is something substantive underlying the empirical similarity. Only when semantic similarity reaches very high levels does it "catch up" to empirical similarity. So, while high semantic overlap may usually be sufficient to detect largely redundant psychological variables, lack of it is not sufficient to rule it out – hence, we still need accurate estimates of empirical overlap. That adjusted correlations outperform unadjusted correlations in indexing high semantic overlap and detect empirical overlap even when semantic similarity is not high should encourage researchers to collect multi-source and –timepoint data to accurately assess redundancy among items – or scales.

In particular, considering adjusted correlations in conjunction with semantic similarity ratings may help researchers to develop more clear guidelines for defining levels of construct overlap (e.g., twin vs sibling constructs; Lawson & Robins, 2021). For example, the items "Easily apologize when I have been wrong" and "Am quick to admit making a mistake" are both empirically ($r = 0.57$, $\widehat{\rho}_{tt} = 0.82$, and $\widehat{\rho}_{ca} = 0.98$) and semantically ($SS = 0.81$) very similar and could thus be considered effectively redundant. Conversely, "I would be tempted to buy stolen property if I were financially tight" and "I'd be tempted to use counterfeit money, if I were sure I could get away with it" are *empirically* nearly identical ($r = 0.49$, $\widehat{\rho}_{tt} = 0.73$, and $\widehat{\rho}_{ca} = 0.99$) but are rated as having much lower semantic overlap ($SS = 0.58$) and thus require more scrutiny than assessing their semantics: perhaps despite semantic differences – they describe two distinct behaviors – they commonly reflect more general tendencies toward engaging in illegal activities for money, or otherwise have a very similar set of causal antecedents (Wood et al., 2015). Thereby, one item carries little unique assessment information above and beyond the other and having both in a scale may be wasteful; to assess construct comprehensively yet efficiently, non-redundant items offer better "bang for the buck" (Condon et al., 2021). As another example, the item "Am often troubled by feelings of guilt" has $\widehat{\rho}_{ca}$ s = 0.72-0.80 with five items ("Tend to feel hopeless"; "Get stressed out easily"; "Have a low opinion of myself"; "Often feel blue"; and "Worry about what people think of me") but a maximum *SS* rating of only 0.25 with any one of these. This sharp disparity may indicate a shared, but more distal, cause of the narrower traits uniquely indexed by these items, such as a latent negative emotionality trait, or causal relations between the conceptually distinct tendencies described by the items.

As noted earlier, the adjustment techniques described here work equally well for psychological assessment at any level of aggregation. Comparing them to the item content provides face-value evidence that the adjustments work; subsequently, these adjustments can be applied to evaluate construct overlap at the level of scales. One caution issued by Lawson and Robins (2021) is that correlations alone should not be used to identify twin vs sibling constructs: "Despite being common practice, researchers should not rely solely, or even primarily, on *concurrent correlations* [emphasis added] to make inferences about relations between constructs“ (p. 351). While we agree, we also argue that the proposed correlation adjustments remedy a major concern in using correlations, which is that when they are based on cross-sectional, single-source data, they are impossible to interpret; our proposed adjustments overcome this problem and allow correlations to be interpreted at face value.

One caveat to this approach is that estimates for adjusted correlations will often be unstable when sample sizes are small, especially at lower correlation magnitudes. This is already true for unadjusted correlation coefficients, and adjusted correlations are based on a ratio of products of correlations, making them less stable still. There is also an observable tendency for estimates of cross-rater agreement to be lower, on average, than short-term retest correlations, which contributes further to the instability of these ratios. We thus acknowledge the need for simulation studies to describe the performance of, and estimate standard errors for, these coefficients. With respect to the latter, one study (Mõttus et al., 2024) derived standard errors of these estimates using simulation studies and found that when estimated with high precision ($N > 20{,}000$ self- and informant-reports), the coefficients performed well in predicting a meaningful outcome variable, life satisfaction. However, most readers will likely not have access to such large samples and therefore we would recommend that, to avoid over-interpreting very noisy estimates, readers only compute adjusted correlations for items that have strong empirical correlations (which, admittedly, is probably where most researchers are interested in evaluating overlap anyway).

In sum, the methods for adjusting correlations and evidence for their applicability presented here suggest that – with sufficiently large samples – we now have a way to much more reliably estimate the empirical overlap of psychological constructs and thus to identify and evaluate twin constructs.

5.2. On the use of semantic similarity ratings

The average rated semantic similarity of item pairs was quite low, and often semantic similarity was modest even among highly empirically similar items. Not one of our item pairs reached the standard of being "semantically redundant" as suggested by Wood et al. (2023) of 3.5/4 (or 0.9/1 in our scale here), and the average rating for each questionnaire was only around 2 on a 5-point scale, or assessed as "slightly similar." Even among the 50 % of pairs with highest semantic similarity, the median ratings were only *SS*s = 0.44 and 0.48 for HEXACO-PI-R and 100NP pairs.

This is largely consistent with Wood et al.: Across their four sets of item pairs – all chosen for their high empirical similarity – mean *SS* ratings were 0.54, 0.45, 0.42, and 0.4 on the present 0 to 1 scale. On the one hand, this supports the reliability of the task: While the studies differed in raters (Prolific Academic respondents vs friends and research assistants), inventories (both surveys here used items, while 2/4 used adjective ratings in Wood et al.), and task length ($k = 200$ vs $k = 100$ item pairs to rate), we found nearly identical average ratings with very high inter-rater agreement. On the other hand, why are *SS* ratings reliably so low?

Our instructions were quite stringent, as both we and Wood et al. (2023) explicitly instructed participants not to select the highest option unless they could not *in any way* distinguish between the meanings. We understand that this could be considered a critique: One reviewer pointed out that the threshold for attributing two items the same *meaning* is much higher than that required to consider them indices of the same *measure,* indicating that statements can “get at” the same thing without being a “perfect correspondence of meaning across different words” (i.e., total semantic redundancy). We would argue that this is precisely what our results have demonstrated, underlining one of the fundamental beliefs in psychometrics: items can be correlated because they share common (latent) causes or have functional relations among them (Mõttus & Allerhand, 2018; Wood et al., 2015), not just because they literally ask the same thing. But this is true *only* for adjusted correlations, and especially for those that account for both systematic and random sources of error.

Furthermore, the task obviously places a very high reliance on human ratings and could benefit from additional operationalizations of semantic similarity. Future work should seek to refine the assessment of semantic similarity and add additional measures, perhaps utilizing new tools such as natural language processing (NLP) to estimate the similarity between pairs of personality statements (e.g., Cutler & Condon, 2022); indeed, improvements are being made in this domain at a rapid rate (e.g., Hommel & Arslan, 2024). We thus acknowledge that semantic similarity ratings are one – but not the only – intuitive metric to help estimate similarity. When aiming to detect redundancy, we recommend that researchers consider as many sources of information as possible: correlations from multiple sources and timepoints, human-rated semantic similarity, NLP-assessed similarity, and any other source as more evidence emerges. While our understanding of the relationship between semantic and empirical similarity (estimates) is still nascent, we believe our findings provide a step towards better understanding, identifying, and evaluating degrees and types of similarity between psychological constructs. So far, our findings show that high semantic similarity may usually be necessary to identify informationally similar content, but low semantic similarity does not rule out informational redundancy.

5.3. On cross-rater agreement and the use of informant-reports

Despite cross-rater correlations having magnitudes nearly half the size of test–retest reliability and estimates therefore being potentially somewhat a) less stable/precise and b) more liable to lead to overcorrections, only *one* out of a possible 23,286 agreement-adjusted correlations fell outside the expected range [-1,1]. This provides confidence in the model proposed by McCrae (2015, 2018) and Mõttus et al. (2024) that the correction for cross-rater agreement is not artificially inflating associations (although for perfectly redundant items [i.e., $\rho = 1$], one would still expect sample estimates of $\widehat{\rho}_{ca}$ to exceed 1 approximately half the time because of sampling error). Face-value evaluations of highly similar pairs also suggest that the adjusted associations are assessing items’ empirical overlap free of single-method and occasion-specific biases and random measurement error. For example, in Mõttus et al. (2024), these adjusted correlations reached near $|1|$ for items that are obviously semantically redundant, such as the antonymous item pair “Keep my promises” and “Break my promises”.

Some researchers may question the validity of informant-reports in comparison to self-reports (e.g., suggesting that they tap into different aspects of personality, such as the latter indexing “identity” and the former “reputation”; McAbee & Connelly, 2016). However, informants tend to agree with each other about a target with the same magnitude as self-reports (Connelly & Ones, 2010). Similarly, most (but not all) sources of disagreement are likely due to method biases and not genuine disagreement about the target: response styles, idiosyncratic interpretation of item content, and implicit grouping of similar content (McCrae et al., 1998; McCrae et al., 2019). Even ratings of targets based on the same information vary considerably (Mõttus et al., 2012). Given the pervasive biases and errors permeating single source methods combined with the relative ease of collecting informant reports, we would suggest that they are one of the most intuitive, scalable, and affordable ways to improve assessment of psychological constructs.

5.4. Statement on generality

We have made use of participants’ understanding and interpretation of statements to generate ratings of their semantic similarity. As such, only individuals who self-identified as native English speakers completed the task, and we only used survey data that had been collected in English. We may expect the mean *SS* ratings and empirical similarity estimates themselves to vary across languages and cultures, but we have no reason yet to believe that our results depended on this. However, as noted above, it is possible that some aspects of the sample and task may have resulted in different findings than in Wood et al. (2023), who used potentially more reliable raters and a task of half the length. Regarding historical/temporal considerations: language evolves over time, and this likely needs to be reflected in personality inventories through changes in item wording to maintain cultural relevance and maximize content validity. Beyond these possibilities, we have no reason to believe that the results depend on other characteristics of the participants, materials, or context.

6. Conclusion

The present study provides further evidence to Wood et al.’s (2023) argument that adjusting inter-item correlations by their unreliability, as indexed by their retest correlations, tends to improve upon unadjusted correlations as indicating whether the items “mean the same thing” as perceived by human raters. Here, we added at least four valuable elaborations on this idea. First, our results supported the hypothesis that self-informant correlations can also be used to produce reliability-adjusted estimates of the similarity of item-pairs that outperform unadjusted correlations as predictors of the items’ semantic similarity. Second, our results indicate that both reliability-adjustment methods produce better estimates of an item pair’s semantic similarity – but mainly at higher ranges of correlations (e.g., values between 0.5 to 1.0, rather than 0 to 0.5; see Online Supplement). Third, by using item pairs that vary broadly in their similarity, we could show that semantic overlap is sufficient but not necessary for redundant variables because redundancy can also arise for reasons other than semantic overlap. Finally, and most importantly, we show that the adjustments based on self-informant agreement are even more effective in identifying redundant variables because they adjust not only for transient errors but also for systematic biases. Empirical correlations adjusted using cross-informant correlations are therefore particularly useful for identifying redundant psychological content.

6.1. Authors’ note

Due to an oversight in the initial analysis code, pairs were selected based on adjusted empirical correlations that only included one of the two terms in the numerator (i.e., $r_{X_{\text{time1}},Y_{\text{time2}}}$ and $r_{X_{\text{self}},Y_{\text{informant}}}$), but not the complementary pair; the geometric means in the denominator were still used. However, we chose to move forward with the selection of item pairs for several reasons. Primarily, the correctly adjusted correlations correlated very highly with the incorrectly adjusted ones. For 100NP pairs, $q = 0.99$ between test–retest adjusted pairs; $q = 0.94$ between cross-rater adjusted pairs; for HEXACO-PI-R pairs, $q = 0.96$ between test–retest adjusted pairs; $q = 0.96$ between cross-rater adjusted pairs. Given this high overlap, and particularly that the actual pairs chosen to be analyzed should not really matter, we thus decided to carry on with the analysis, ensuring that the *correctly* adjusted estimates were used when comparing empirical similarity indices to semantic similarity ratings.

Author contributions

Sam Henry played a lead role in study conceptualization, data collection, data preparation, data analysis, and report writing. Dustin Wood played a supporting role in study conceptualization, data collection, and report writing. David C. Condon played a supporting role in study conceptualization and report writing. Graham H. Lowman played a supporting role in data collection and report writing. René Mõttus played a supporting role in study conceptualization and report writing.

CRediT authorship contribution statement

Sam Henry: Writing – review & editing, Writing – original draft, Visualization, Resources, Project administration, Methodology, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Dustin Wood:** Writing – review & editing, Methodology, Data curation, Conceptualization. **David M. Condon:** Writing – review & editing, Conceptualization. **Graham H. Lowman:** Writing – review & editing, Project administration, Data curation. **René Mõttus:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Raw data for SS included in supplemental materials. The authors do not have permission to share HEXACO-PI-R or 100NP raw data, but provide processed data sufficient to conduct analyses in Supplement.

Acknowledgements

This study was not preregistered. This research was supported by the University of Edinburgh School of Philosophy, Psychology, and Language Sciences Research Support Grant to the lead author. René Mõttus was supported by the Estonian Research Council grant PGR2190.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jrp.2024.104530.

References

Anastasi, A. (1984). Aptitude and achievement tests: The curious case of the indestructible strawperson. In *Social and Technical Issues in Testing: Implications for Test Construction and Usage.* (Vol. 9, pp. 129–140). Lawrence Erlbaum Associates. https://digitalcommons.unl.edu/burostestingissues/9.

Aikins, H. A. (1902). *The Principles of Logic.* Holt.

Ashton, M. C., & Lee, K. (2010). Trait and source factors in HEXACO-PI-R self- and observer reports. *European Journal of Personality, 24*, 278–289.

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. *Leadership Quarterly, 27*(4), 634–652. https://doi.org/10.1016/j.leaqua.2016.02.006

Baumert, A., Schmitt, M., Perugini, M., Johnson, W., Blum, G., Borkenau, P., Costantini, G., Denissen, J. J. A., Fleeson, W., Grafton, B., Jayawickreme, E., Kurzius, E., MacLeod, C., Miller, L. C., Read, S. J., Roberts, B., Robinson, M. D., Wood, D., & Wrzus, C. (2017). Integrating personality structure, personality process, and personality development. *European Journal of Personality, 31*(5), 503–528. https://doi.org/10.1002/per.2115

Beck, E. D., & Jackson, J. J. (2022). A mega-analysis of personality prediction: Robustness and boundary conditions. *Journal of Personality and Social Psychology, 122*(3), 523–553. https://doi.org/10.1037/pspp0000386

Borkenau, P. (1992). Implicit Personality Theory and the Five-Factor Model. *Journal of Personality, 60*(2), 295–327. https://doi.org/10.1111/j.1467-6494.1992.tb00975.x

Cattell, R. B., & Tsujioka, B. (1964). The importance of factor-trueness and validity, versus homogeneity and orthogonality, in test scales. *Educational and Psychological Measurement, 24*(1), 3–30.

Christensen, A. P., Garrido, L. E., & Golino, H. (2023). Unique variable analysis: A network psychometrics method to detect local dependence. *Multivariate Behavioral Research.* https://doi.org/10.1080/00273171.2023.2194606

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

Condon, D. M., Roney, E., & Revelle, W. (2017). A SAPA project update: on the structure of phrased self-report personality items. *Journal of Open Psychology Data, 5*(1), 1–8. https://doi.org/10.5334/jopd.32

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

Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin, 136*(6), 1092–1122. https://doi.org/10.1037/a0021212

Cortina, J. M., 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*(12), 1351–1381. https://doi.org/10.1037/apl0000815.supp

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual* (P. A. Resources (ed.)).

Cramer, A. O. J., van der Sluis, S., Noordhof, A., Wichers, M., Geschwind, N., Aggen, S. H., Kendler, K. S., & Borsboom, D. (2012). Dimensions of normal personality as networks in search of equilibrium: You can't like parties if you don't like people. *European Journal of Personality, 26*(4), 414–431. https://doi.org/10.1002/per.1866

Credé, M. (2010). Random responding as a threat to the validity of effect size estimates in correlational research. *Educational and Psychological Measurement, 70*(4), 596–612. https://doi.org/10.1177/0013164410366686

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*(3), 492–511. https://doi.org/10.1037/pspp0000102

Cutler, A., & Condon, D. M. (2022). Deep lexical hypothesis: Identifying personality structure in natural language. *Journal of Personality and Social Psychology.* https://doi.org/10.1037/pspp0000443

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*(6), 1087–1101. https://doi.org/10.1037/0022-3514.92.6.1087

Dul, J. (2016). Necessary Condition Analysis (NCA): Logic and methodology of "necessary but not sufficient" causality. *Organizational Research Methods, 19*(1), 10–52. https://doi.org/10.1177/1094428115584005

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science, 3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science, 2*(2), 156–168. https://doi.org/10.1177/2515245919847202

Goldberg, L. R. (1999). A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several Five-Factor models. In *Personality Psychology in Europe* (Vol. 7, Issue 1, pp. 7–28).

Henry, S., & Mõttus, R. (2023). *The 100 Nuances of Personality: Development of a comprehensive, non-redundant personality item pool.* https://doi.org/10.17605/OSF.IO/TCFGZ.

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*(1), Article e0262465. https://doi.org/10.1371/journal.pone.0262465

Hommel, B. E., & Arslan, R. C. (2024). *Language models accurately infer correlations between psychological items and scales from text alone.*

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, R. W. Robins, R.C. Fraley, & R. F. Krueger (Eds.), Handbook of research methods in personality psychology. (pp. 461–494).

John, Oliver P., & Benet-Martínez, V. (2000). Measurement: Reliability, construct validation, and scale construction. *Handbook of Research Methods in Social and Personality Psychology, October 2012*, 339–369.

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*(1), 80–92. https://doi.org/10.1037/0021-9010.86.1.80

Kelley, T. L. (1927). Interpretation of Educational Measurements. *World Book Company.* https://doi.org/10.1037/h0068663

Lawson, K. M., & Robins, R. W. (2021). Sibling constructs: what are they, why do they matter, and how should you handle them? *Personality and Social Psychology Review, 25*(4), 344–366. https://doi.org/10.1177/10888683211047101

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*(2), 112–125. https://doi.org/10.1016/j.obhdp.2010.02.003

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*(1), 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*(4), 478–500. https://doi.org/10.1111/iops.12184

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO personality inventory. *Multivariate Behavioral Research, 329*(2), 329–358.

Lee, K., & Ashton, M. C. (2018). Psychometric properties of the HEXACO-100. *Assessment, 25*(5), 543–556. https://doi.org/10.1177/1073191116659134

Leising, D., Thielmann, I., Glöckner, A., Gärtner, A., & Schönbrodt, F. (2022). Ten steps toward a better personality science – how quality may be rewarded more in research evaluation. *Personality Science, 3*, 1–44. https://doi.org/10.5964/ps.6029

Lilienfeld, S. O., & Strother, A. N. (2020). Psychological measurement and the replication crisis: Four sacred cows. *Canadian Psychology, 61*(4), 281–288. https://doi.org/10.1037/cap0000236

McAbee, S. T., & Connelly, B. S. (2016). A multi-rater framework for studying personality: the trait-reputation-identity model. *Psychological Review, 123*(5), 569–591. https://doi.org/10.1037/rev0000035.supp

McCrae, R. R. (2018). Method biases in single-source personality assessments. *Psychological Assessment*. https://doi.org/10.1037/pas0000566.supp

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*(1), 28–50. https://doi.org/10.1177/1088868310366253

McCrae, R. R. (2015). A more nuanced view of reliability: Specificity in the trait hierarchy. *Personality and Social Psychology Review, 19*(2), 97–112. https://doi.org/10.1177/1088868314541857

McCrae, R. R., & Mõttus, R. (2019). What personality scales measure: A new psychometrics and its implications for theory and assessment. *Current Directions in Psychological Science, 28*(4), 415–420. https://doi.org/10.1177/0963721419849559

McCrae, R. R., Stone, S. V., Fagan, P. J., & Costa, P. T. (1998). Identifying causes of disagreement between self-reports and spouse ratings of personality. *Journal of Personality, 66*(3), 285–313. https://doi.org/10.1111/1467-6494.00013

McCrae, R. R., Mõttus, R., Hřebíčková, M., Realo, A., & Allik, J. (2019). Source method biases as implicit personality theory at the domain and facet levels [Article]. *Journal of Personality, 87*(4), 813–826. https://doi.org/10.1111/jopy.12435

McDonald, R. P. (1999). Test theory: A unified treatment. In *Test Theory: A Unified Treatment*. https://doi.org/10.4324/9781410601087.

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*(1), 117–135. https://doi.org/10.1177/1073191117738047

Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes, 6*(1), 1–28. https://doi.org/10.1080/01690969108406936

Mõttus, R. (2016). Towards more rigorous personality trait–outcome research. *European Journal of Personality, 30*(4), 292–303. https://doi.org/10.1002/per.2041

Mõttus, R., & Allerhand, M. H. (2018). Why do traits come together? The underlying trait and network approaches. *The SAGE Handbook of Personality and Individual Differences: Volume I: The Science of Personality and Individual Differences*, *January*, 130–150. https://doi.org/10.4135/9781526451163.n6.

Mõttus, R., Allik, J., Realo, A., Rossier, J., Zecca, G., Ah-Kion, J., Amoussou-Yéyé, D., Bäckström, M., Barkauskiene, R., Barry, O., Bhowon, U., Björklund, F., Bochaver, A., Bochaver, K., de Bruin, G., Cabrera, H. F., Chen, S. X., Church, A. T., Cissé, D. D., & Johnson, W. (2012). The effect of response style on self-reported conscientiousness across 20 countries. *Personality and Social Psychology Bulletin, 38*(11), 1423–1436. https://doi.org/10.1177/0146167212451275

Mõttus, R., Realo, A., Allik, J., Ausmees, L., Henry, S., & Mccrae, R. R. (2024). *Without measurement limitations, most people's life-satisfaction is consistent with their personality traits: True correlations in multi-trait, multi-rater, multi-sample data*. https://doi.org/10.31219/osf.io/cd5kt.

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*(1), 943–995. https://doi.org/10.1080/19416520.2016.1161965

Ponnock, A., Muenks, K., Morell, M., Seung Yang, J., Gladstone, J. R., & Wigfield, A. (2020). Grit and conscientiousness: Another jangle fallacy. *Journal of Research in Personality, 89*, Article 104021. https://doi.org/10.1016/j.jrp.2020.104021

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*(3), 380–392. https://doi.org/10.1037/met0000244

Rubenstein, H., & Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM, 8*(10), 627–633. https://doi.org/10.1145/365628.365657

Sackett, P. R. (2014). When and why correcting validity coefficients for interrater reliability makes sense. *Industrial and Organizational Psychology, 7*(4), 501–506. https://doi.org/10.1111/iops.12185

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

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of cronbach's alpha. *Psychometrika, 74*(1), 107–120. https://doi.org/10.1007/s11336-008-9101-0

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*(1), 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*(1), 72–101. https://www.jstor.org/stable/1412159.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 1904-1920, 3*(3), 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*(2), 245–251. https://doi.org/10.1037//0033-2909.87.2.245

Whitten, W. B., Newton Suter, W., & Frank, M. L. (1979). Bidirectional synonym ratings of 464 noun pairs. *Journal of Verbal Learning and Verbal Behavior, 18*(1), 109–127. https://doi.org/10.1016/S0022-5371(79)90604-2

Wood, D., Gardner, M. H., & Harms, P. D. (2015). How functionalist and process approaches to behavior can explain trait covariation. *Psychological Review, 122*(1), 84–111. https://doi.org/10.1037/a0038423

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

Wood, D., Lowman, G. H., Armstrong, B. F., & Harms, P. D. (2023). Using retest-adjusted correlations as indicators of the semantic similarity of items. *Journal of Personality and Social Psychology, 125*(2), 437–454. https://doi.org/10.1037/pspp0000441