Automated jingle–jangle detection: Using embeddings to tackle taxonomic incommensurability

Dirk U. Wulff[1,2] and Rui Mata[2]
[1]Max Planck Institute for Human Development
[2]University of Basel

Taxonomic incommensurability denotes the difficulty in comparing scientific theories due to incompatible use of concepts and operationalizations. We show that item, scale, and label embeddings—representations of psychometric items, scales, and labels in a vector space obtained from language models—can help tackle this problem in psychology. We analyze over 4,000 items, 450 scales, and 270 different construct labels to show that embeddings can be used to predict empirical intercorrelations between operationalizations, perform automated detection of jingle–jangle fallacies, and suggest more parsimonious taxonomies that might eliminate a number of extant psychological constructs. All in all, our work suggests that embeddings offer a useful tool to tackle taxonomic incommensurability in the sciences.

Keywords: large language models, embeddings, machine learning, psychology, personality

Taxonomic incommensurability, the idea that various scientific theories or paradigms are often incomparable due to their distinct mapping between theoretical concepts and measures or empirical results (Oberheim & Hoyningen-Huene, 2018; Sankey, 1998), poses significant challenges to all sciences. For the social and behavioral sciences, such as psychology, the lack of a clear mapping between constructs and operationalizations results in the difficulty of selecting appropriate measures for describing, predicting, and changing behavior (e.g., Norris et al., 2019). Consequently, addressing taxonomic incommensurability is vital for any discipline interested in understanding and improving human health, wealth, and well-being.

Recent years have seen numerous calls for conceptual clarification in psychology (e.g., Bringmann et al., 2022; Flake & Fried, 2020); specifically, the need to address "conceptual clutter" (Jones et al., 2016), "concept creep" (Haslam et al., 2021), or "jingle–jangle fallacies" (Dang et al., 2020; Leising et al., 2022). These terms suggest that psychology is plagued by the proliferation of measures that have been given similar labels, yet may capture different constructs (jingle fallacy), whereas other measures have received different labels, yet capture the same construct (jangle fallacy). This problem clearly represents a form of taxonomic incommensurability and has long been recognized as stemming from the lack of a common nomological network (Cronbach & Meehl, 1955) or ontology (e.g., Sharp et al., 2023); that is, an agreed-upon conceptual lexicon consisting of concepts and their observable manifestations that can be used to characterize human psychology.

Our work aims to help tackle taxonomic incommensurability by introducing a novel approach that employs large language models to create a quantitative depiction of psychological constructs and their linguistic operationalizations. This method relies on item, scale, and label embeddings—vector space representations of psychometric items, scales, and their labels—to map the semantic relations between an extensive array of psychological measures as well as the putative constructs they represent. In our work, we make the following contributions: First, we validate different state-of-the-art embeddings by testing how they capture known empirical relations between different psychological measures. Second, we introduce a method to perform automated detection of jingle–jangle fallacies in existing taxonomies by relying on the best performing embedding for these purposes. Third, we introduce a clustering and relabeling approach to deal with the inherent trade-off involved in minimizing jingle and jangle fallacies (see Figure 1). All in all, our approach tackles taxonomic incommensurability by boosting conceptual clarity and providing a more parsimonious taxonomy of constructs in psychology.

Dirk U. Wulff https://orcid.org/0000-0002-4008-8022 Rui Mata https://orcid.org/0000-0002-1679-906X

We are grateful to Alexandra Bagaïni and Zakir Hussain for helpful comments and to Laura Wiles for editing the manuscript. This work was supported by grants from the Swiss Science Foundation to Dirk U. Wulff (100015_197315) and Rui Mata (100015_204700).

Correspondence concerning this article should be addressed to Dirk U. Wulff, Center for Adaptive Rationality, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany. E-mail: dirk.wulff@gmail.com

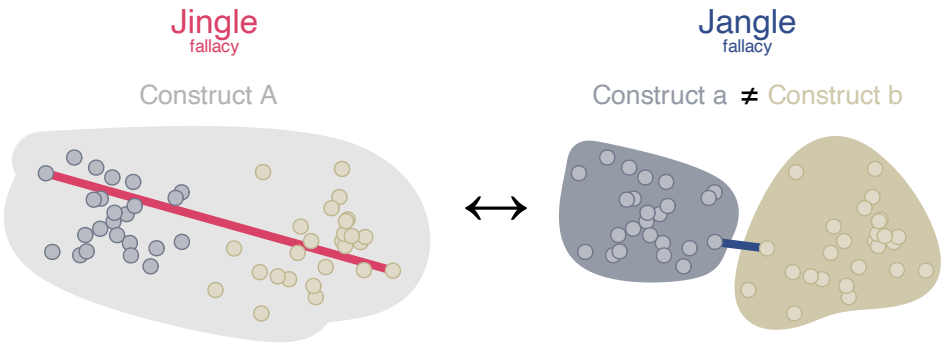

Figure 1

Jingle–jangle fallacies. The figure illustrates jingle–jangle fallacies and the associated trade-off between minimizing jingle and jangle fallacies. Each point represents a scale, with its placement being determined by their similarity in the semantic space; convex-hull polygons are drawn to represent the construct labels shared by a group of scales that can be assigned labels either manually or in an automated fashion. On the left, a labeling process may minimize jangle fallacies by giving all related scales the same construct label (Construct A) yet causing some scales that are distant in the similarity space to have the same label, thus creating one (or more) jingle fallacies (i.e., red line). On the right, a labeling process may minimize jingle fallacies by creating two (or more) semantically distinct labels (Construct a, Construct b) that maximize within-construct similarity for two clusters of scales, yet engendering cases for which very similar scales have been assigned the two semantically distinct labels, thus creating a jangle fallacy (blue line).

Past work has already demonstrated that embeddings from language models are able to capture important aspects of human personality (e.g., Abdurahman et al., 2023; Cutler & Condon, 2023; Rosenbusch et al., 2020) but a thorough comparison of different state-of-the-art models, including large language models, and an approach to detect jingle–jangle fallacies is lacking. Consequently, first, we were interested in validating the ability of different embeddings to predict well-known empirical findings in personality psychology. Specifically, we aimed to identify which set of embeddings can best predict observed empirical patterns, such as internal consistency or convergent validity observed in human self-reports, as well as the alignment between construct labels and their contents.

Second, after establishing the empirical validity of the representational space provided by specific embeddings, we were interested in evaluating the match between measures and their respective labels in the embedding space as a way to automatically detect jingle–jangle fallacies. As described above, taxonomic incommensurability can be thought of as the lack of a common conceptual network involving psychological constructs and their operationalizations. We reasoned that embeddings can help deal with this challenge by creating a common representational space for both construct labels and their operationalizations, thus allowing us to identify jingle–jangle fallacies in an automated fashion. Given the ever increasing proliferation of constructs and measures in the field of psychology, an automated approach to identifying jingle–jangle fallacies could help provide some needed clarity to the field.

Third, and finally, the existence of a common representation for both construct labels and their measures also provides an opportunity for developing new mappings between the two. In our work, we present a novel approach based on clustering and relabeling of scales in the similarity space provided by embeddings. This approach effectively offers a tool to enhance the alignment between psychological constructs and their operationalizations that aims to minimize jingle–jangle fallacies. In practice, this translates into more parsimonious taxonomies that eliminate the need for hypothesizing a number of constructs, thus offering both greater conceptual clarity and parsimony.

In what follows, we provide an overview of our analytic approach, which relied on obtaining item, scale, and label similarities from embeddings following a number of steps illustrated in Figure 2. First, we obtained item embeddings for a large number of psychological items (Figure 2A). Specifically, we included a large pool of 4,452 items, comprising 459 scales, and capturing 277 constructs encapsulated in 254 unique single and multi-construct labels from the publicly available International Personality Item Pool (IPIP; http://ipip.ori.org; Goldberg et al., 2006). We obtained embeddings using different models but present throughout the paper findings based on the OpenAI embedding model, *ADA* (*ada-002*; Greene et al., 2022), which proved the most powerful source of embeddings for our purposes. We provide systematic comparisons between all eight embeddings models tested in the Supplementary Materials (Cer et al., 2018; Gao et al., 2021; Greene et al., 2022; Mikolov et al., 2017; Rosenbusch et al., 2020; Song et al., 2020; Su et al., 2022;

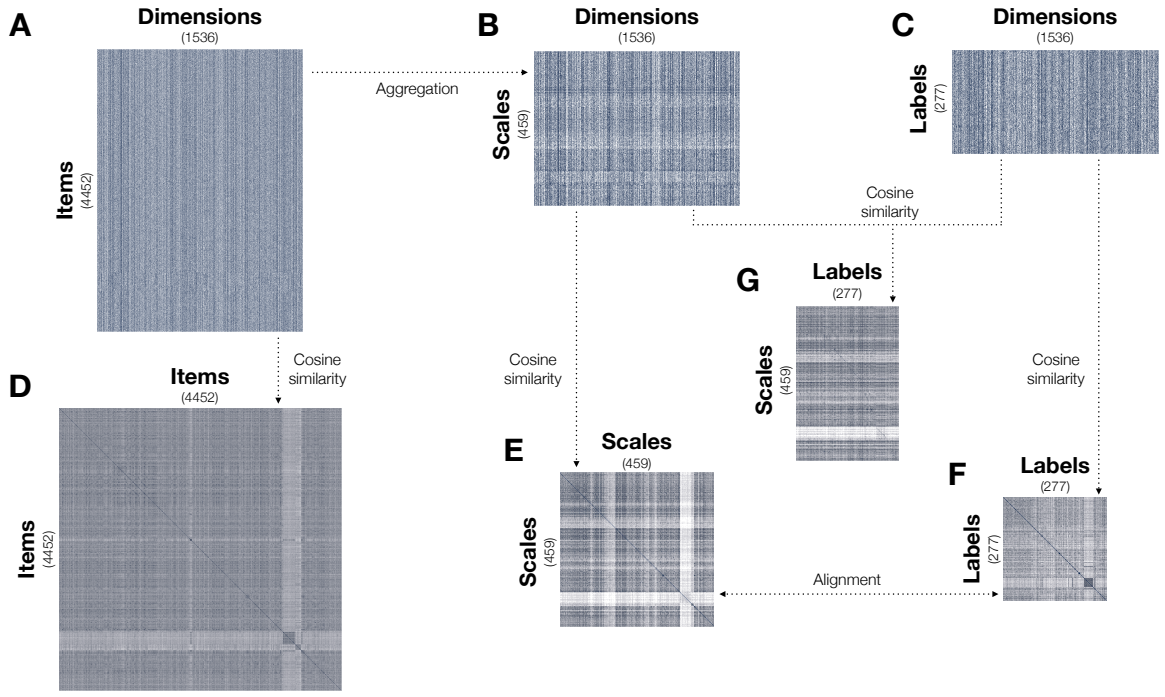

Figure 2

Overview of the analytic approach. Panels A–C show the embeddings concerning personality items, scales, and labels from the International Personality Item Pool (IPIP; http://ipip.ori.org; Goldberg et al., 2006); the embeddings are obtained from a specific language model (ADA, Greene et al., 2022) and each column represents one of 1,536 dimensions of the embeddings model. In Panel A each row represents one of 4,452 IPIP items; in Panel B, each row corresponds to one of 459 scale embedding obtained by averaging the item embeddings associated with each respective scale in IPIP; in Panel C, each row refers to one of 277 labels present in IPIP. Panels D–F show all pairings of items, scales, and labels, respectively, obtained by computing the cosine similarity between the embeddings shown in Panels A–C. Panel G shows the similarity between labels and scales using the cosine similarity between the two as defined by their respective embeddings.

Yang et al., 2019). Second, we obtained scale embeddings by averaging the embeddings for the items belonging to each scale (Figure 2B). Third, concerning labels, we report results using contextualized labels because these provide the strongest alignment with scale embeddings but we considered a number of alternative approaches that we discuss in more detail below. These embeddings were generated for the 277 distinct labels produced by IPIP's construct labels. Specifically, from the 254 IPIP labels, we used directly those consisting of unique words, and further split IPIP's multi-construct labels into single ones so as to be able to capture each construct individually. Fourth, using the item, scale, and label embeddings, we derived similarities between all pairs of items, scales, and labels by computing the cosine similarity between the respective vectors in the embedding space (Figure 2D–F). Finally, we matched scales and labels in this novel space (Figure 2G).

Overall, the reliance on this large pool of items, scales, and labels guaranteed a wide coverage of psychological constructs and associated measures, covering several personality traits, mental states, work attitudes, and other aspects of psychological functioning. Once equipped with embeddings and respective similarities for items, scales, and labels, we were in a position to ask whether this information is able to capture meaningful relations between measures as well as address the problem of jingle–jangle fallacies.

Results

Our results are divided in three sections. In the first section, we provide a demonstration of the ability of embeddings to recover known relations between psychological measures, such as the patterns of convergent and divergent validity obtained from self-reports. In a second section, we put embeddings to use in an effort to help address conceptual clutter in psychology by detecting jingle–jangle fallacies in an automated fashion. Third, and finally, we introduce a novel approach to providing a more parsimonious mapping between constructs and extant psychological measures.

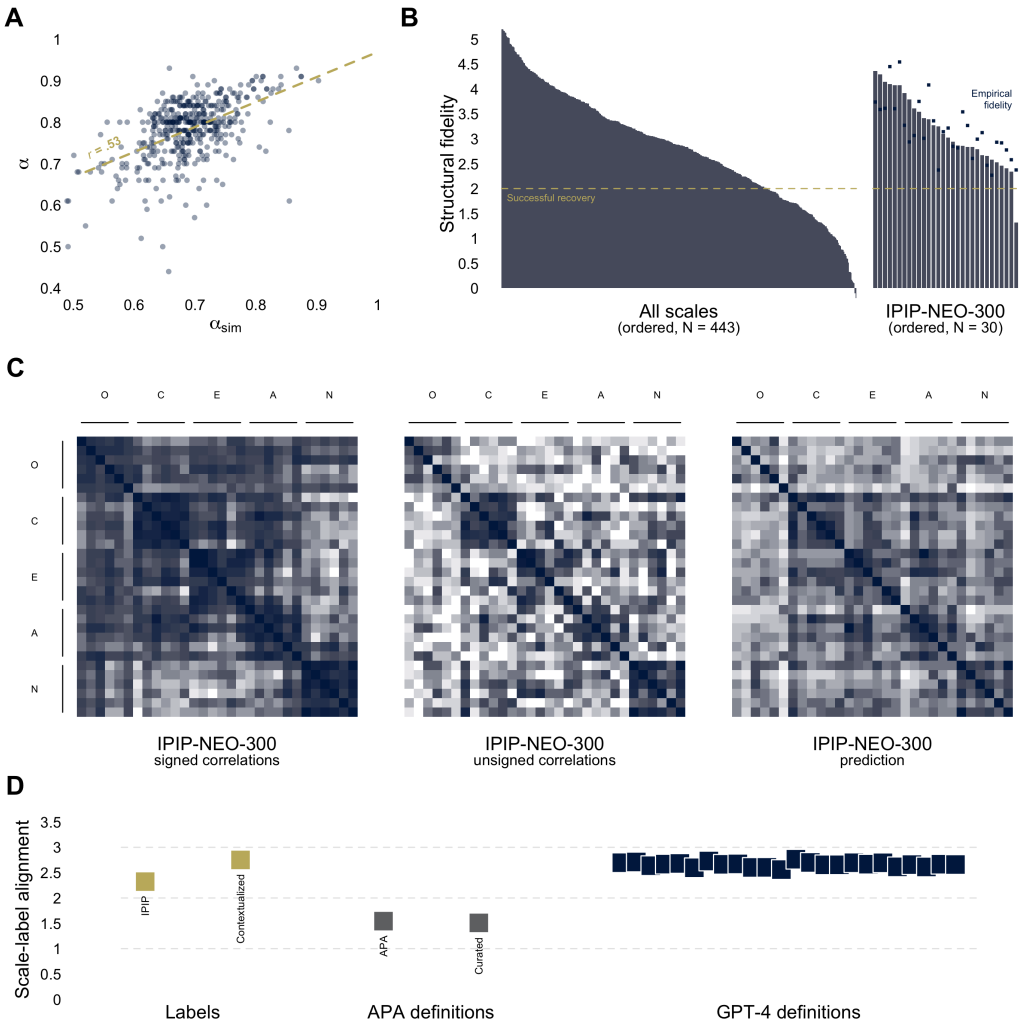

Figure 3

Validation of embeddings. Panel A shows a strong correlation between the empirical internal consistency (α) and the internal consistency predicted from embeddings (α_{sim}) for 449 scales. Panel B visualizes the structural fidelity of embeddings by contrasting the similarity of items within a given scale to items of the most similar other scale within the same inventory for all inventories with at least 3 scales (N = 433). Additionally, it contrasts the structural fidelity that embeddings achieve for the IPIP-NEO-300 scales (represented as bars; N = 30) relative to the observed structural fidelity (represented as points) obtained from a large-scale empirical assessment (Kajonius & Johnson, 2019). Panel C allows a visual comparison of the empirically observed signed and unsigned correlations among the 30 scales of the IPIP-NEO-300 (Kajonius & Johnson, 2019) and the prediction based on scale similarities using embeddings. Panel D illustrates our construct label evaluation. The points show the average alignment score as defined in the text for embeddings of the labels of distinct construct present in IPIP, contextualized labels, APA definitions, curated APA definitions, and definitions generated by GPT-4.

Validation of embeddings

We carried out four analyses aimed at addressing the power of different embeddings for recovering known properties of psychological measures, including their internal consistency, structural fidelity, and patterns of convergent and divergent validity, as well as the match between scale content and their labels. In what follows, we provide numerical values for the best performing model (see Supplementary Materials for results of all models).

First, we validated embeddings by assessing whether scale similarities obtained from embeddings are able to predict a scale's empirically observed internal consistency; that is, Cronbach's α (Cronbach, 1951), a measure typically used to capture the average empirical correlation between items from the same scale. A strong positive correlation would suggest

that the semantic similarity between items can be used to quantify the observed empirical correlation from self-reports, which represents a minimum requirement for a conceptual representation based on embeddings if this is to inform us about the structure of psychological measures more generally. To this end, we estimated the within-scale similarity and then used the Spearman–Brown prediction formula to derive estimates of the scale's Cronbach's α. As illustrated in Figure 3A, we observed a strong Pearson correlation between the observed and predicted internal consistencies of $r = .53$ for the 449 scales for which empirical scores were available in IPIP (Goldberg et al., 2006). This correlation is comparable to the correlation of *alpha* between scales of the same construct, when differences in scale length are corrected for ($r = .48$) and considerably higher than when differences in scale length are ignored ($r = .28$).

Second, we compared to what extent item similarities obtained from item embeddings capture scales' structural fidelity; that is, we examine whether a set of items is most strongly related to other items from their scale of origin relative to other scales in the same inventory. Showing some degree of structural fidelity would suggest that embeddings allow matching between items and scales. We evaluated the structural fidelity of scales by evaluating the similarity of items with the other items in the same scale in comparison to the similarity of the scale's items to the items of the other scales. Specifically, we computed a z-score for each scale that reflects the similarity of items within the same scale relative to the similarity of items of different scales within an inventory. We regard the structure of a scale as fully recovered if $z > 2$ and partially recovered if $z > 1$. We evaluate structural fidelity for the 25 inventories with at least 3 scales that, on average, had 12.4 scales. Our analysis only considers inventories with at least three scales. The results shows that 75% of all 443 scales were recovered and 94% partially recovered (Figure 3B), suggesting that embeddings capture the structure of a large number of psychological inventories. Moreover, we specifically compared the predicted structural fidelity to that obtained from a large publicly available data set on the IPIP-NEO-300 five factor personality inventory, which includes six scales for each factor, implying a total of 30 scales (Kajonius & Johnson, 2019, Figure 3B). For the empirical IPIP-NEO-300 data, we evaluated structural fidelity in an analogue fashion by first determining the average correlation of items within and between scales and then calculating the ratio of the within correlation to the strongest between correlation for a given scale. Our analysis revealed that the structural fidelity for empirical data is somewhat higher than for the predicted one from item embeddings, but not for all scales. Moreover, we observed a strong correlation between the predicted and empirical structural fidelity ($r = .71$), further underpinning the fact that embeddings can largely capture the psychometric structure of psychological scales and inventories.

Third, we assessed more generally the ability of embeddings to make predictions about convergent and divergent validity by using item and scale similarities to make predictions about the observed correlations between items and scales across various personality constructs. Specifically, we compared the predicted correlation between items and scales based on the sentence embeddings with the empirical correlations found in the IPIP-NEO-300 data set based on self-reports (Kajonius & Johnson, 2019), excluding self-correlations. We compared the predictions both with the signed and unsigned correlations, to highlight an important property of semantic similarity; namely, that it is driven primarily by the content of the item rather than its direction. At the level of items, we observed small to moderate correlations of $r = .22$ (signed) and $r = .43$ (unsigned), whereas, at the level of scales, we observed moderate to strong correlations of $r = .36$ (signed) and $r = .61$ (unsigned). Overall, these results show that embeddings are able to capture the unsigned empirical correlations between items and scales accurately.

Fourth, to be able to identify jingle–jangle fallacies, it is important to have not only powerful item and scale embeddings but also a good mapping between these and the constructs themselves, for example, as operationalized by label embeddings. In our validation strategy, we generated several types of label embeddings that we reasoned could have differential strengths and weaknesses and evaluated their relative alignment to scale content as operationalized by the scale embedding. Specifically, we considered five types of label embeddings: the construct labels present in IPIP, contextualized labels (i.e., a version of the construct label produced by placing the construct label in the sentence "The personality construct [LABEL]."), the American Psychological Association (APA) construct definitions, a manually curated version of the APA construct definitions, and several variations of GPT-4 generated labels (e.g., different prompts). We provide a more detailed rationale for these different types in the Methods but, in summary, we reasoned that although IPIP labels provide a wide coverage they are typically short and some are used in everyday language without explicit reference to the personality construct (e.g., "Warmth", "Organization"), which can present a challenge to obtaining appropriate, construct-specific embeddings. In turn, other approaches that may provide richer embeddings that are closer to the human expert definition (e.g., APA definitions) do not provide full coverage of all IPIP constructs or may partly result from hallucinations (e.g., GPT). Consequently, an empirical test of the alignment between different scale and label embeddings is in order. In practice, we computed an alignment score for each scale and label embedding using cosine similarity and averaged the alignment scores across scales to obtain an overall estimate of how well a given instantiation of label embed-

dings is aligned with the corresponding scale embeddings. Figure 3D shows the results for the different types of label embeddings considered. The original IPIP labels showed a high alignment, implying that, on average, the similarities between scale and matching label embeddings were over two standard deviations higher than those between scale and non-matching construct embeddings. However, alignment was improved using the contextualized labels. In contrast, both the original and manually curated APA definitions showed significantly lower alignment, which may be partly explained by only a portion of scales being considered. Yet, even when constraining to the available data, the contextualized labels and the GPT-4 definitions showed substantially higher alignment. The lower alignment for embeddings of construct labels as compared to those of contextualized labels and GPT-4 definitions suggests that contextualization is important to obtain embeddings that appropriately capture the meanings of personality constructs. With both of these contextualized embeddings showing comparable alignment, we decided to rely on the contextualized labels for the analyses presented below. The main motivation for this decision was to remain closer to the original labels while shielding our analysis from possible generative hallucinations of the GPT-4 model.

All in all, these results show that embeddings can capture a number of central characteristics of psychological measures as well as their interrelations. In particular, scale embeddings show performance levels that are higher than previous results obtained using small(er) language models (Abdurahman et al., 2023; Rosenbusch et al., 2020) and only somewhat lower than the reliability of many of these measures. Further, contextualized label embeddings seem to provide considerable alignment between scale content and associated constructs as defined by scale labels or construct definitions. In what follows, we capitalize on the power of the best performing embeddings—in particular, scale embeddings and contextualized label embeddings—to provide an overview of the similarity between these measures and their respective labels so as to identify and minimize jingle–jangle fallacies in psychology.

Automated jingle–jangle detection

We put embeddings to use by automating the detection of jingle–jangle fallacies; that is, making explicit the relative mismatch between scale content and respective labels. To this end, we leveraged the fact that both scale content and scale labels are amenable to an analysis of semantic similarity using embeddings (see Figure 2G). Specifically, we identified jingle–jangle fallacies by identifying criteria that quantify high similarity in construct labels associated with low similarity in the associated scales—jingle fallacy—or, alternatively, low similarity in construct labels associated with high similarity in the associated scales—jangle fallacy. In the results below, we adopt a criterion that implies a cosine difference of more than four times the average absolute difference between label and scale similarity (m = .096) and nearly three times the average error in predicting empirical scale correlations (sd_e = .137). Using this criterion, we identified a total of 113 fallacies, specifically, 60 jingle and 53 jangle fallacies. Figure 4 shows these comparisons, highlighting cases where the two diverge strongly. It can be seen that there exist numerous cases where the similarity of a pair of scales' labels far exceeds the similarity of the pair's content—a jingle fallacy. In turn, there are numerous scale pairs with highly dissimilar labels whose scale similarities are strikingly high—a jangle fallacy. Note that in the case of IPIP's multi-construct labels (e.g., "Bravery/Courage/Valor"), we adopted a conservative strategy by considering the maximum similarity between all possible pairs of construct labels. Please further note that we also considered different criteria, and although the absolute number of fallacies varies as a function of the specific criterion used, the relative proportion of jingle and jangle fallacies is similar across different criteria and so can be thought of as a characteristic of the current mapping between constructs and measures in IPIP, and therefore, in the field of psychology. A more thorough understanding of these fallacies requires an assessment of the specific items and respective labels so we provide a few examples of jingle and jangle fallacies in the two paragraphs below.

A large number of the 60 jingle fallacies (58%) concern the *ORVIS* and *ORAIS* inventories, which have a specific focus on vocational (*ORVIS*) and avocational (*ORAIS*) activities and interests. For example, the scales labeled *Organization* in the *ORVIS* and *HPI-HIC* inventories are focused on business organization (e.g., "Keep track of a company's inventory."), whereas the items in the latter are focused on personal perfectionism (e.g., "Continue until everything is perfect."), leading to a jingle fallacy in which the same label (i.e., *Organization*) captures different types of areas (organizational vs. personal). Another example is the *Introversion* scale in the *16PF* and *IPIP-IPC* inventories, which differ in that the items in the former focus on enjoying being by oneself (e.g., "Enjoy spending time by myself."), whereas the latter focuses on being shy (e.g., "Am quiet around strangers.").

Jangle fallacies occurred evenly across inventories and were largely the product of overlapping items. Across the 53 jangle pairs, scales shared an average of 4.4 items out of, on average, 12.4 possible item overlaps, equivalent to a Jaccard similarity of .27. For instance, in the case of the top jangle fallacy concerning the *Reserve* scale in the *16PF* inventory and *Self-disclosure* scale in the *TCI* inventory, seven items are shared (e.g., "Am open about my feelings." or "Don't talk a lot."). We note, however, that item overlap does not fully explain the existence of jangle fallacies. The number of overlapping items is only moderately related to the cosine

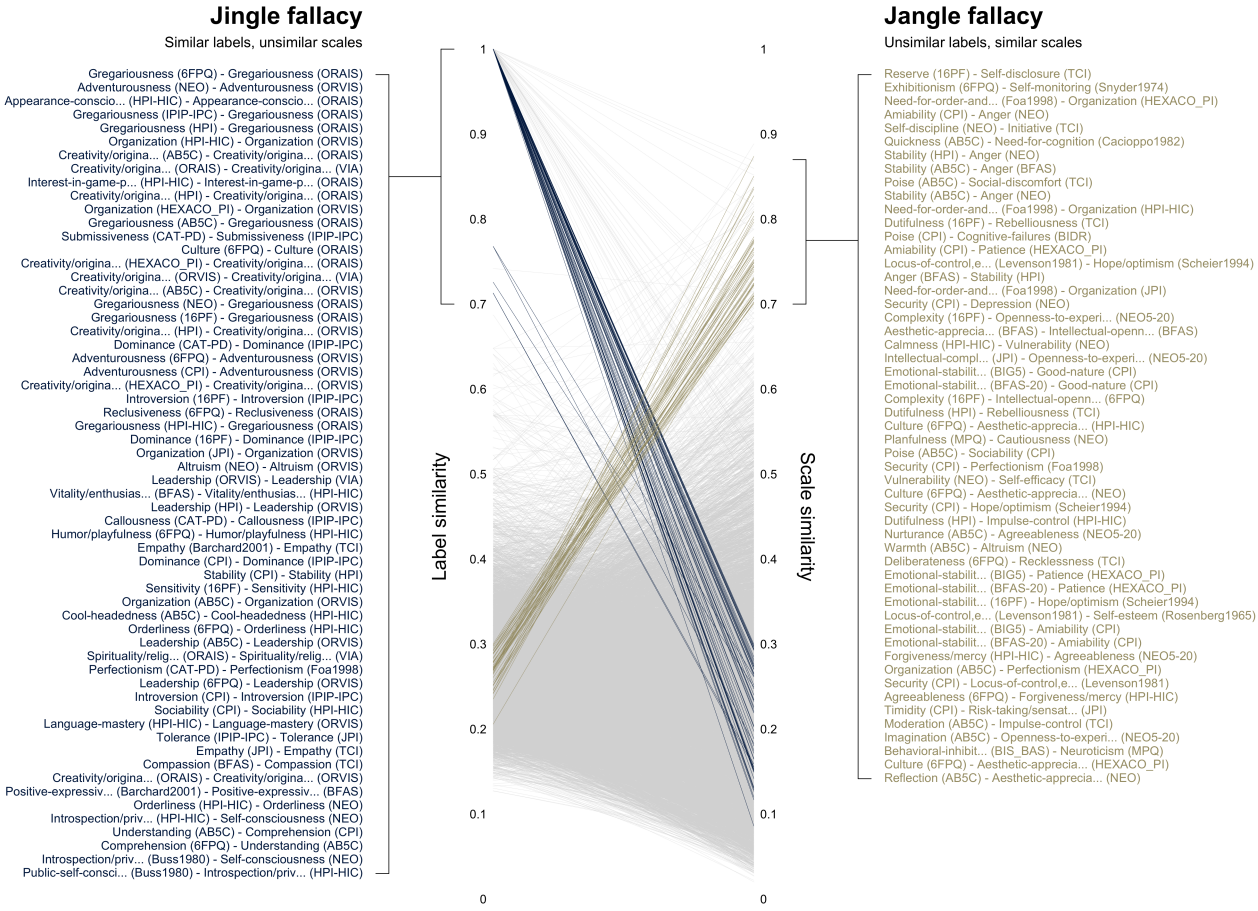

Figure 4

Automated jingle–jangle detection. The figure contrasts the similarity of scales' labels (Label similarity) and their content (Scale similarity) for all pairs of the 459 scales analyzed. Blue and beige lines and corresponding text highlight those pairs that either are high (cosine > .7) in label similarity but low (cosine < .3) in scale similarity (blue) or vice versa (beige). The 60 jingle and 53 jangle fallacy pairs are highlighted and their respective labels are shown to the right and left of the line plot.

similarity between scales (r = .52). Moreover, there are 16 jangle fallacies with one to three overlaps, which does not automatically result in a fallacy, given that there are 3,785 scale pairs with one to three overlaps that were not identified as a fallacy. One example of a low-overlap fallacy is *Warmth* in the *AB5C* inventory and *Altruism* in the *NEO*, which share only one of ten items but show nevertheless high scale similarity due to a focus on supporting others (e.g., "Inquire about others' well-being." and "Am concerned about others.").

To summarize, our analysis suggests that it is possible to automate jingle–jangle detection using embeddings and that both fallacy types can be detected in existing taxonomies, such as IPIP. In what follows, we address possible ways of leveraging automated jingle–jangle detection to minimize these fallacies and the associated trade-off.

Minimizing jingle–jangle fallacies to increase conceptual clarity and parsimony in psychology

The ability of embeddings to place both scales and labels in the same representational space offers an opportunity to reorganize and relabel measures with the goal of minimizing jingle–jangle fallacies while potentially achieving greater parsimony in the number of constructs proposed. We investigated this possibility by developing a procedure consisting of the following three steps. First, we used clustering algorithms to identify clusters of scales based on their semantic similarity (see Figure 2E) and allowing for a variety of possible numbers of clusters (e.g., 1-250). The rationale for using a number of clustering algorithms is that these emphasize different criteria when producing clustering solutions so it is important to test the generality of conclusions from any single one (Hennig, 2015). Second, for each clustering algo-

rithm and solution (i.e., number of clusters), we used a maximum bipartite matching algorithm (Cherkassky et al., 1998) to assign construct labels to measures such that the semantic similarity between measures and labels (see Figure 2G) was maximized. Third, we evaluated for each solution the number of jingle and jangle fallacies using the method described in the previous section for each of the new scale-label mappings and used this information to evaluate the overall reduction in jingle and jangle fallacies that was obtained as a function of number of clusters.

Figure 5A shows the result of this approach for the clustering solutions produced by one specific algorithm, hierarchical clustering using complete linkage (HC complete). The figure reveals a general reduction in the number of fallacies as the number of clusters grows larger but also a tradeoff between minimizing jingle and jangle fallacies. In this case, the total number of fallacies is optimized by a solution with 207 clusters. This solution produces a total of 39 fallacies, including 6 jingle and 33 jangle fallacies, which is only about a third of the 113 fallacies produced by the IPIP label assignments (see section *Automated jingle–jangle detection*). However, there exist more parsimonious solutions that propose fewer clusters while incurring only a few additional fallacies. One such solution uses 126 clusters and produces 47 fallacies (4 jingle and 43 jangle fallacies)—that is, only 8 additional fallacies relative to the optimal solution—while proposing substantially fewer constructs (126 vs. 207, 61%). Our evaluation of four other clustering algorithms revealed similar patterns including the production of fallacy-reduced parsimonious solutions, which we defined as those that were among the 20 best solutions in minimizing fallacies using the smallest number of clusters. Specifically, as can be seen in Figure 5B, we found that most algorithms produce a smaller or comparable number of fallacies proposing substantially fewer clusters, with the exception of HS single hierarchical clustering, which can be explained by its strong focus on minimizing the minimum pairwise distance, leading to minimizing jangle fallacies only. We present full results for all clustering algorithms and solutions in the Supplementary Materials. All in all, these results suggest that our approach can reliably produce a smaller number of fallacies than IPIP with many possible solutions being parsimonious in the sense of identifying substantially fewer clusters, thus, in effect, proposing a more parsimonious representation of constructs in psychology.

Figure 5C provides a visual illustration of a possible novel mapping between measures and labels produced by the parsimonious solution of the hierarchical clustering with complete linkage, which can be considered the best compromise between the number of fallacies and number of clusters across all solutions found. The mapping assigns 369 out 459 measures to one of the top five most semantically similar labels (solid lines), implying that 90 measures are assigned a less proximate label with an average rank of 18.5. This is considerably more coherent than the mapping implied by the IPIP labels (despite using the maximum similarity for multi-construct labels), which only assigns 275 measures a top-five label and otherwise assigns a rank of 50.9. Crucially, this improvement is achieved without multi-construct compounds and using only 126 labels, which is about half the number of construct labels implied by IPIP.

In sum, our approach employed language embeddings to reorganize and relabel existing psychological measures, resulting in a streamlined mapping incurring 58% fewer jingle–jangle fallacies (47 vs. 113) while requiring 50% fewer unique labels (126 vs. 254). The revised mapping underscores the potential for using language embeddings to identify and reduce jingle–jangle fallacies while, at the same time, obtaining more parsimonious taxonomies of psychological constructs and associated measures. All in all, although our approach does not offer a full elimination of jingle–jangle fallacies, it suggests that automated methods based on embeddings can render these more transparent and increase conceptual clarity in psychology.

Discussion

We heeded recent calls to address taxonomic incommensurability in psychology by using language embeddings obtained from different language models to clarify the relation between constructs and their measures in the psychological sciences. Our work makes three main contributions. First, we compare a number of embeddings and identify those that can best recover key structural features of human psychology, including the internal consistency and the convergent validity for a wide range of psychological measures. Our work extends others offering a window into the power of modern natural language processing methods to capture the empirical interrelations between different psychological measures (e.g., Abdurahman et al., 2023; Cutler & Condon, 2023; Rosenbusch et al., 2020) that, in the past, have been only accessible through deploying large numbers of self-reports from even larger samples of individuals. Second, and crucially, establishing the validity of embeddings warrants us to advance a new approach to addressing conceptual confusion in the psychological sciences. Our approach relied on embeddings from a large language model to quantify the similarity between different constructs and associated measures. This approach presents a novel way of identifying (mis)alignment between constructs and their operationalizations and determining the existence of jingle–jangle fallacies in an automated, quantitative, and reproducible manner. Third, and finally, we showed that embeddings can be used not only to identify but also to reduce the number of jingle–jangle fallacies, as well as consolidate the space of constructs in psychology. Crucially, our results suggest that about half of the constructs considered may be sufficient to retain con-

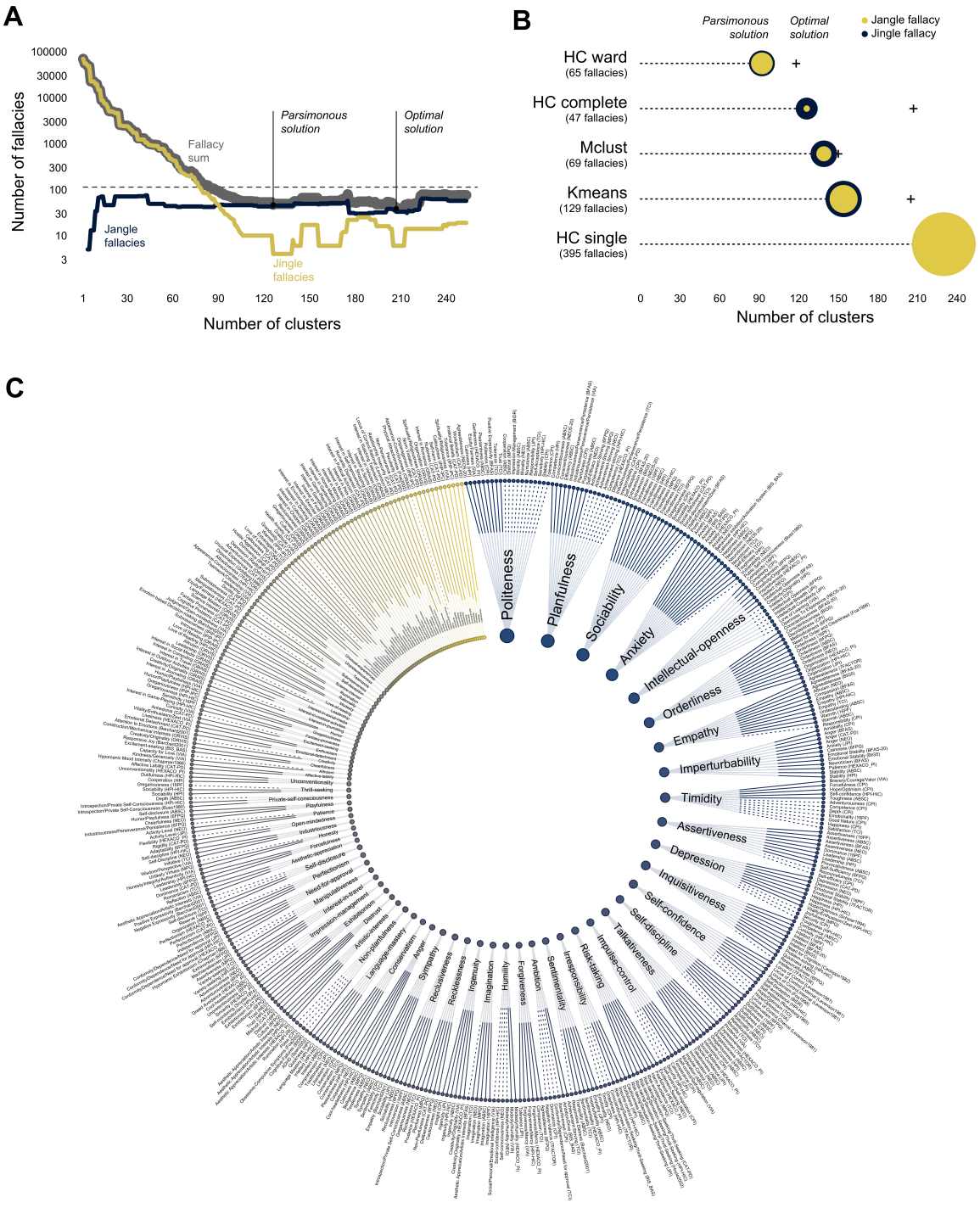

Figure 5

Minimizing jingle–jangle fallacies. Panel A shows the number of fallacies as a function of number of clusters proposed by a specific clustering algorithm (HC complete: complete-linkage hierarchical clustering). The panel makes evident the trade-off between minimizing jingle and jangle fallacies as a function of the number of clusters and highlights both the most parsimonious solution that minimizes the number of clusters while reducing fallacies, as well and the optimal solution that considers the minimization of fallacies only. Panel B shows the number of clusters and fallacies across various clustering algorithms. For each algorithm, the most efficient solutions are determined as those with the fewest clusters among the top 20 solutions with minimal fallacies. The "+" symbol indicates solutions with the least number of fallacies. Panel C shows a radial plot indicating a possible mapping of 459 scales to 126 labels based on the parsimonious solution shown in Panel A. Solid and dashed lines reflect assignments of optimal (top-five rank) and suboptimal similarity, respectively.

siderable granularity as represented in the semantic space of embeddings.

The methods introduced here have considerable potential to inform future conceptual and measurement work. Conceptually, language embeddings offer a tool to establish relations between extant and future constructs in psychology, which could be important for theory development, particularly in those areas for which many overlapping constructs have been proposed (e.g., Burman et al., 2015; Roberts et al., 2014). Methodologically, the use of embeddings promises to be helpful in designing new items and scales that maximize fit to well-defined constructs while minimizing overlap with existing measures (e.g., Abdurahman et al., 2023; Rosenbusch et al., 2020). Further, although we focused on constructs from psychology, we believe such methods can be helpful across disciplines, including economics, political science, or sociology, that also rely on verbal theory and language-based methods to capture a swathe of constructs, attitudes, and beliefs. Consequently, the findings underscore the power of integrating advanced computational methods to inform scientific theorizing and measurement across all disciplines that rely on language-based approaches.

We would like to note four main limitations of our work. First, we should note that our results concerning the prevalence of jingle–jangle fallacies is likely not fully representative of the psychological literature at large. Specifically, we limited our analysis to IPIP constructs and measures, which is a well established and curated set, potentially contributing to an underestimation of extant fallacies in psychology. Future work should expand our approach by adopting larger sets of construct definitions and measures. One possibility for future work would be to rely on more specific formulations of specific constructs (e.g., West et al., 2019) or even broader ontologies that may specify more complex relations between theoretical concepts and operationalizations (e.g., Norris et al., 2019; Sharp et al., 2023).

Second, our results showed that embeddings based on modern large language models (Bhatia et al., 2019) can help retrieve meaningful signals about human psychology from large amounts of text. Our work improves earlier attempts that have relied on alternative methods, such as latent semantic analysis (Rosenbusch et al., 2020), or more advanced architectures (Abdurahman et al., 2023; Cutler & Condon, 2023), to make inferences about similarities between measures. Despite the advances provided by the novel embeddings proposed here, our results show only an approximation to self-report data. Although future approaches may prove yet more powerful, it remains an open issue whether future embeddings will be able to fully capture all relevant aspects of human psychology. One obvious and major limitation of the embeddings we use here is that these cannot be used to capture differences between individuals or groups that are typically central to theories of personality. Nevertheless, embeddings could in principle be obtained or fine-tuned for specific groups (Wulff & Mata, 2022), in which case some of the approaches used here could in principle be applied to uncover group-specific aspects relevant to many psychological theories.

Third, we adopt only one form of reducing jingle–jangle fallacies—through relabeling of existing scales. One alternative we leave unexplored is the assignment of items to constructs to create new, pure scales of specific constructs. We avoided this approach due to anticipated limits in predictive accuracy at the item level, as our analysis suggests that prediction at the scale level is considerably higher than at the item level (circa .6 vs. .4). Future work could, however, explore an item-level approach, perhaps by leveraging large-scale data at the item level and comparing the predictive validity of newly generated scales for predictive validity relevant real-world criteria (e.g., Roberts et al., 2007). Furthermore, research could consider developing novel, more holistic algorithms capable of simultaneously solving the grouping and relabeling problems while minimizing jingle–jangle fallacies.

Fourth, and finally, despite the apparent consensus for the need for conceptual clarity in the psychological sciences (e.g., Bringmann et al., 2022; Flake & Fried, 2020), some researchers have suggested that efforts to reduce the number of concepts may be counterproductive by limiting the consideration set that may be needed, given disparate coexisting goals (e.g., Hochstein, 2016). We would like to suggest, however, that our approach should not be seen as substituting the need for plural discussion of concepts and the needs they address, but, rather, as an empirical, quantitative aid to foster debate and, possibly, inform expert consensus. Indeed, because our results suggest that eliminating jingle–jangle fallacies is not trivial, it seems it would be advantageous to use our results to have a broader discussion about the utility of different constructs and measures in psychology.

In conclusion, our work demonstrates the potential of language embeddings to address the fundamental problem of taxonomic incommensurability; in particular, we show that the automated detection and minimization of jingle–jangle fallacies can produce more parsimonious taxonomies of constructs thus contributing to conceptual clarification in psychology.

Methods

Terminology

Our contribution focuses on psychometric instruments or measures that consist of collections of linguistic statements that typically describe a state, attitude, or disposition (e.g., "I enjoy thinking about things") and which traditional approaches have asked human respondents to endorse or rate on some ordinal scale. Single statements are typically termed

items in the psychological literature. The large majority of personality measures employs several such items to capture the same underlying psychological construct, with a collection of two or more related items being typically termed a *scale*. Measures consisting of several scales may be used to capture a wide range of constructs and one such collection is typically termed an *inventory*. We use the term *label* to refer to the construct (and associated label) thought to underlie the trait captured by an item, scale, or inventory. Finally, we use the term *embedding* to refer to the representation of linguistic units, such as words or sentences, in a mathematical form, for example, as a real-numbered vector. As we describe below, there are a number of *language models* that can be used to generate embeddings from linguistic units, with the most recent of these involving billions or trillions of parameters and, because of their size, often being called *large language models*. Embeddings can be employed to characterize single items, scales, or their labels, and we use the terms *item embedding*, *scale embedding*, and *label embedding* for these, respectively.

Personality Measures and Data

We use three main data sources in our work. First, we rely on a large pool of personality measures from the International Personality Item Pool (IPIP; http://ipip.ori.org;]), which currently consists of 4,452 items and 459 scales belonging to 27 multi-scale and 10 single-scale inventories. IPIP was created to refine and improve personality assessment by adapting item and scale labels and using multi-construct labels, e.g., *Risk-taking/Sensation-Seeking/Thrill-Seeking*, to capture the tremendous overlap in these constructs. IPIP contains 254 labels, including 24 multi-construct labels, that contain 277 individual construct labels. In addition, we use the IPIP's estimates of the scale's internal consistency (i.e., Cronbach's α) available for 448 of the 459 scales in our validation analyses.

Second, we use data from a large study of human personality to obtain estimates of scale intercorrelations for 30 scales (300 items) from one personality inventory designed to measure the Big Five personality dimensions (Johnson, 2014, 2020). Because our study focuses on the English language, we opted to focus on responses from the United States of America only, which led to the inclusion of responses from 212,625 individuals that we used to calculate intercorrelations between the 30 scales in one of the validation analyses described below. Our empirical validation data of the IPIP-NEO-300 inventory was collected by Kajonius and Johnson (2019) using a publicly accessible online survey. The data encompass measurements of 320,128 U.S. individuals, of which 100,608 show complete data.

Third, we obtained definitions of personality constructs from the APA dictionary (https://dictionary.apa.org/) and generated definitions form GPT-4. For details, see section *Label embeddings*.

Embeddings

We considered a total of eight different language models to generate embeddings for our purposes. We provide additional information about all models in the Supplemental Materials. We selected these models because they cover a wide range of approaches that can be used to generate linguistic embeddings, spanning simpler models that consider only single words (e.g,. *fastText*) to those that consider several words in the form of sentences (e.g., *ADA*) and thus promise to be more sensitive to information concerning word order, semantic context, and other syntactic information. The set of embeddings includes two models, sentence *BERT* (*SBERT*; Reimers & Gurevych, 2019) and latent semantic analysis (*LSA*; Dumais et al., 1988), previously employed for the prediction of personality (Abdurahman et al., 2023; Rosenbusch et al., 2020). However, we find that newer contextualized embeddings—in particular, the *Instructor* (*Instructor XL*; Su et al., 2022) and *ADA* (*ada-002*; Greene et al., 2022) models—are considerably more accurate than previously employed methods. We use and report OpenAI's *ADA* results in our work because of its slightly superior performance and report results for other models in the Supplemental Materials.

Item embeddings

We retrieved item embeddings by directly encoding the item texts via the OpenAI *ada-002* application programming interface (API).

Scale embeddings

We generated scale-level embeddings by summing the item embeddings according to

$$w_{scale,i} = \sum_{j} w_{ij} \tag{1}$$

where w_{ij} is the embedding vector of the jth item in the ith scale.

Label embeddings

We considered several types of label embeddings that we reasoned could have different strengths and weaknesses. All embeddings were generated for the individual single-construct labels rather than the multi-construct compounds, in order to be able to use them to produce a simplified scale-label mapping (see section *Minimizing jingle–jangle fallacies to increase conceptual clarity and parsimony in psychology*).

First, for the base embedding, we directly encoded the IPIP construct labels (e.g., "Extraversion", "Sociability").

However, many construct labels can be used in everyday language without explicit reference to the personality construct (e.g., "Warmth"); consequently, we reasoned that additional context could be helpful to appropriately capture the meaning of the personality construct and therefore tested different ways to provide context leading to several other label embeddings. Second, we produced a contextualized version of the construct labels by placing the construct label in the sentence "The personality construct [LABEL]." that we then used to generate embeddings from the language models. Third, we scraped definitions of the available personality constructs from the APA dictionary (https://dictionary.apa.org/). Construct definitions were available for 135 of a total of 277 distinct construct. Fourth, we manually curated the APA definitions by removing parts of the definitions that referred to meanings other than the personality construct (e.g., reference to special meanings of the construct "Competence" in linguistics and law) and filling in text for constructs that were defined exclusively by providing reference to other constructs using the definitions of the other constructs (e.g., replacing "See Aggression" with the definition of "Aggression" for the construct "Hostile-Aggression"). Fifth, and finally, we generated definitions from GPT-4 via the API using a variety of different prompt structures varying the length (30, 50, or 100 words), assistant instruction (e.g, starting or not starting the prompt with "You are an expert in psychology and will be asked to produce a definition for a construct that other experts in the field will recognize as accurate and representative."), and core prompt ("Write a [LENGTH]-word expert definition of the personality construct [LABEL]." or "Write a [LENGTH]-word expert definition of the personality construct [LABEL]."). We then used the same strategy as for item embeddings described above to generate label embeddings for the different label variations.

Similarity

The similarity between items, scales, and scale labels was calculated as the cosine similarity between embedding vectors and scaled to match the average correlations obtained from empirical responses. Specifically, the similarity between a vector a and b was calculated as

$$cos = \Big(\frac{\mathbf{a} \cdot \mathbf{b}}{||\mathbf{a}||\,||\mathbf{b}||}\Big)^{\gamma}$$

with γ being a scaling that was set to 8.25 for items, 19.48 for scale similarities, and 10.86 for label similarities to match the average item and scale correlations found in the IPIP-NEO-300 empirical data set (see section *Personality Measures and Data*). Although the main motivation behind the scaling factors was interpretability, their inclusion also had a small positive impact on the recovery of psychometric aspects. For instance, the correlation of predicted and actual correlations between scales in the IPIP-NEO-300 data increased from $r = .56$ to $r = 61$. This increase can be attributed to an expansion of the cosine range, which in the case of scales increased from [.82, 1] to [.02, 1].

Cronbach's α

We obtained the within-scale similarity S_{self} and then used the Spearman–Brown prediction formula to derive estimates of the scale's Cronbach's α internal consistency measure as follows

$$\alpha_{sim} = \frac{nS_{self}}{1 + (n-1)S_{self}}$$

Author Contributions

Conceptualization: DW, RM; Formal analysis: DW. Writing—original draft: DW, RM.

References

Abdurahman, S., Vu, H., Zou, W., Ungar, L., & Bhatia, S. (2023). A deep learning approach to personality assessment: Generalizing across items and expanding the reach of survey-based research. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/pspp0000480

Bhatia, S., Richie, R., & Zou, W. (2019). Distributed semantic representations for modeling human judgment. *Current Opinion in Behavioral Sciences*, *29*, 31–36. https://doi.org/10.1016/j.cobeha.2019.01.020

Bringmann, L. F., Elmer, T., & Eronen, M. I. (2022). Back to basics: The importance of conceptual clarification in psychological science. *Current Directions in Psychological Science*, *31*, 340–346. https://doi.org/10.1177/09637214221096485

Burman, J. T., Green, C. D., & Shanker, S. (2015). On the meanings of self-regulation: Digital humanities in service of conceptual clarity. *Child Development*, *86*(5), 1507–1521. https://doi.org/10.1111/cdev.12395

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Céspedes, M., Yuan, S., Tar, C., et al. (2018). Universal sentence encoder. *arXiv preprint*. https://doi.org/10.48550/arXiv.1803.11175

Cherkassky, B. V., Goldberg, A. V., Martin, P., Setubal, J. C., & Stolfi, J. (1998). Augment or push: A computational study of bipartite matching and unit-capacity flow algorithms. *Journal of Experimental Algorithmics*, *3*, 8–es. https://doi.org/10.1145/297096.297140

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h004095

Cutler, A., & Condon, D. M. (2023). Deep lexical hypothesis: Identifying personality structure in natural language. *Journal of Personality and Social Psychology*, *125*(1), 173–197. https://doi.org/10.1037/pspp0000443

Dang, J., King, K. M., & Inzlicht, M. (2020). Why are self-report and behavioral measures weakly correlated? *Trends in Cognitive Sciences*, *24*(4), 267–269. https://doi.org/10.1016/j.tics.2020.01.007

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., & Harshman, R. (1988). Using latent semantic analysis to improve access to textual information. *Proceedings of the SIGCHI conference on Human factors in computing systems*, 281–285. https://doi.org/10.1145/57167.57214

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/10.1177/2515245920952

Gao, T., Yao, X., & Chen, D. (2021). SimCSE: Simple contrastive learning of sentence embeddings. *arXiv preprint*. https://doi.org/10.48550/arXiv.2104.08821

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

Greene, R., Sanders, T., Weng, L., & Neelakantan, A. (2022). New and improved embedding model. *Open AI blog*. https://openai.com/blog/new-and-improved-embedding-model

Haslam, N., Tse, J. S. Y., & De Deyne, S. (2021). Concept creep and psychiatrization. *Frontiers in Sociology*, *6*, 806147. https://doi.org/10.3389/fsoc.2021.806147

Hennig, C. (2015). What are the true clusters? *Pattern Recognition Letters*, *64*, 53–62. https://doi.org/10.1016/j.patrec.2015.04.009

Hochstein, E. (2016). Categorizing the mental. *The Philosophical Quarterly*, *66*(265), 745–759. https://doi.org/10.1093/pq/pqw001

Johnson, J. A. (2014). Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality*, *51*(100), 78–89. https://doi.org/10.1016/j.jrp.2014.05.003

Johnson, J. A. (2020). Johnson's IPIP-NEO data repository. osf.io/tbmh5

Jones, S. M., Zaslow, M., Darling-Churchill, K. E., & Halle, T. G. (2016). Assessing early childhood social and emotional development: Key conceptual and measurement issues. *Journal of Applied Developmental Psychology*, *45*, 42–48. https://doi.org/10.1016/j.appdev.2016.02.008

Kajonius, P. J., & Johnson, J. A. (2019). Assessing the structure of the Five Factor Model of personality (IPIP-NEO-120) in the public domain. *Europe's Journal of Psychology*, *15*(2), 260–275. https://doi.org/10.5964/ejop.v15i2.1671

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*, e6029. https://doi.org/10.5964/ps.6029

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2017). Advances in pre-training distributed word representations. *arXiv preprint*. https://doi.org/10.48550/arXiv.1712.09405

Norris, E., Finnerty, A. N., Hastings, J., Stokes, G., & Michie, S. (2019). A scoping review of ontologies related to human behaviour change. *Nature Human Behaviour*, *3*(2), 164–172. https://doi.org/10.1038/s41562-018-0511-4

Oberheim, E., & Hoyningen-Huene, P. (2018). The incommensurability of scientific theories. *The Stanford Encyclopedia of Philosophy (Fall 2018 Edition)*. https://plato.stanford.edu/archives/fall2018/entries/incommensurability/

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.48550/arXiv.1908.10084

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science*, *2*(4), 313–345. https://doi.org/10.1111/j.1745-6916.2007.00047.x

Roberts, B. W., Lejuez, C., Krueger, R. F., Richards, J. M., & Hill, P. L. (2014). What is conscientiousness and how can it be assessed? *Developmental Psychology*, *50*(5), 1315–1330. https://doi.org/10.1037/a0031109

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. https://doi.org/10.1037/met0000244

Sankey, H. (1998). Taxonomic incommensurability. *International Studies in the Philosophy of Science*, *12*(1), 7–16. https://doi.org/10.1080/02698599808573578

Sharp, C., Kaplan, R. M., & Strauman, T. J. (2023). The use of ontologies to accelerate the behavioral sciences: Promises and challenges. *Current Directions in Psychological Science*, 09637214231183917. https://doi.org/10.1177/09637214231183917

Song, K., Tan, X., Qin, T., Lu, J., & Liu, T.-Y. (2020). MPNet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, *33*, 16857–16867. https://api.semanticscholar.org/CorpusID:215827489

Su, H., Kasai, J., Wang, Y., Hu, Y., Ostendorf, M., Yih, W.-t., Smith, N. A., Zettlemoyer, L., Yu, T., et al. (2022). One embedder, any task: Instruction-finetuned text embeddings. *arXiv preprint*. https://doi.org/10.48550/arXiv.2212.09741

West, R., Godinho, C. A., Bohlen, L. C., Carey, R. N., Hastings, J., Lefevre, C. E., & Michie, S. (2019). Development of a formal system for representing behaviour-change theories. *Nature Human Behaviour*, *3*(5), 526–536. https://doi.org/10.1038/s41562-019-0561-2

Wulff, D. U., & Mata, R. (2022). On the semantic representation of risk. *Science Advances*, *8*(27), eabm1883. https://doi.org/10.1126/sciadv.abm1883

Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Abrego, G. H., Yuan, S., Tar, C., Sung, Y.-H., et al. (2019). Multilingual universal sentence encoder for semantic retrieval. *arXiv preprint*. https://doi.org/10.48550/arXiv.1907.04307