Current Research in Behavioral Sciences 8 (2025) 100168

Contents lists available at ScienceDirect

Current Research in Behavioral Sciences

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

Semantic analysis of test items through large language model embeddings predicts a-priori factorial structure of personality tests

Nicola Milano *, Maria Luongo, Michela Ponticorvo, Davide Marocco

Department of Humanistic Studies, Natural and Artificial Cognition Laboratory "Orazio Miglino", University of Naples Federico II, via Porta di Massa 1, Naples 80125, Italy

ARTICLE INFO

Keywords:
Semantic similarity
Language models
Machine learning
Content validity
Dimensionality reduction
Test items analysis

ABSTRACT

In this article, we explore the use of Large Language Models (LLMs) for predicting factor loadings in personality tests through the semantic analysis of test items. By leveraging text embeddings generated from LLMs, we evaluate the semantic similarity of test items and their alignment with hypothesized factorial structures without depending on human response data. Our methodology involves using embeddings from four different personality test to examine correlations between item semantics and their grouping in principal factors. Our results indicate that LLM-derived embeddings can effectively capture semantic similarities among test items, showing moderate to high correlation with the factorial structure produced by humans respondents in all tests, potentially serving as a valid measure of content validity for initial survey design and refinement. This approach offers valuable insights into the robustness of embedding techniques in psychological evaluations, showing a significant correlation with traditional test structures and providing a novel perspective on test item analysis.

1. Introduction

The linkage between natural language and psychology has a long history. In 1884, Galton was the first scientist to explicitly propose the "lexical hypothesis," stating: "the most important individual differences in human transactions will come to be encoded as single terms in some or all of the world's languages" (Galton, 1884; Goldberg, 1992). Since Galton's seminal paper, many scientists have attempted to empirically develop a list of personality-descriptive terms present in the lexicon and to appreciate the extent to which trait terms share aspects of their meanings. This research led to the classical works in personality psychology by Allport and Odbert (1936), Cattell (1943), Norman (1967), and finally Goldberg (1992), which summarized decades of previous research in the Big Five personality questionnaire.

Recently, the rise of increasingly powerful machine learning (ML) methods (Dumais et al., 2012; Mikolov et al., 2013) has established the use of ML algorithms in the field of natural language processing (NLP), significantly impacting language-based psychological research methods (Arnulf et al., 2014, 2020, 2021; Gefen and Larsen, 2017). The first ML methods were based on single word-vector embeddings, where a single word is represented as a vector, also called an embedding vector, in a multi-dimensional space, and words sharing semantic meanings have similar embedding representations. Two famous algorithms of this kind are latent semantic analysis (LSA) (Dumais et al., 2012) and Word2Vec (W2V) (Mikolov et al., 2013). Even though context was not considered in the word embedded representation, Arnulf et al. (2014) showed how LSA-based semantic algorithms allowed remarkably precise prediction of survey responses from humans in organizational behavior tests, though they failed in personality tests.

In recent years, previous approaches to computational text analysis have been revolutionized by the advent of transformer-based large language models (LLMs) such as the GPT (Generative Pre-trained Transformer) series (Vaswani et al., 2017; Radford et al., 2019) and BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019). These models have significantly enhanced the ability of machine learning models to "understand" and analyze large volumes of texts. They are designed to capture embedding vectors representing the semantics of entire sentences instead of single words, extrapolating meanings and deeply understanding large texts. They work by transforming sentences into high-dimensional embedding vectors that capture the linguistic properties of sentences or preserve the essence of the text and its contextual relationships (Devlin et al., 2019; Reimers and Gurevych, 2019).

Based on this premise, our work tries to analyze deeply new LLMs

* Corresponding author.
E-mail address: nicola.milano@unina.it (N. Milano).

https://doi.org/10.1016/j.crbeha.2025.100168
Received 9 July 2024; Received in revised form 17 January 2025; Accepted 23 January 2025
Available online 25 January 2025

findings for psychological assessment using validated personality test. More explicitly, our research questions are: Do psychometric tests work because test items unveil the semantic structure already present in the sentence used to describe behavioral characteristics? Can we use LLMs as a preliminary validation measure of a psychometric test, predicting the factorial structures that we will find in human respondents?

Starting from these questions, we first recall the classical process of psychometric test validation and clarify how modern AI methods can be integrated into the process. Then, we exhaustively present the computational methods used to prove our points and the psychometric test that we analyzed with them in the methods section. Our findings are reported in the results section, and in the conclusions, we discuss the impact and future directions of our work.

1.1. Psychometrics test validation process

The validation process of a psychometric test aims at providing a solid scientific basis for the proposed score interpretations, which usually begins by specifying the construct that the test intends to measure.

Within this framework, Classical Test Theory (CTT) has provided core contribution for a long time, emphasizing the importance of reliability and validity in assessing the quality of a test (Allen and Yen, 2001; DeVellis and Thorpe, 2021; Nunnally and Bernstein, 1978; Garcia et al., 2020). Reliability refers to a test's ability to provide consistent results across different administrations or in various contexts, assuming that the measured attribute remains unchanged (Cronbach, 1947). On the other hand, validity concerns how accurately a test measures what it intends to measure (Hughes, 2018). In Schimmack's words "Without valid measures even replicable results are uninformative" (Schimmack, 2021).

A trustable test must not only yield consistent results (reliability) but must also be targeted and relevant to the construct it aims to assess (Cunningham, 1986; Cook and Beckman, 2006).

As psychological characteristics are generally unobservable, the first and still prevailing definition of validity is the degree to which a test measures what it claims to measure (Garrett, 1937; Smith and Wright, 1928). Debates regarding the concept of validity have led to a multidimensional view, which includes searching for evidence that the test is consistent with the theory it is based on, provides information aligned with the intended purposes, and that its results do not lead to negative and unintended consequences for individuals, groups, or society (Furr, 2021). Today, in summary, empirical evidence of validity is based on the test content, relationships with other variables, internal structure, response processes, and the consequences of the test (Kline, 2013).

As we have hinted at before, the first step in test development is to define what is being measured, as the construct definition directly impacts the interpretation and use of the score: this concerns content validity, on which there is a growing interest in psychometric research (Spoto et al., 2023; El-Den et al., 2020).

Content validity of the test can be defined according to four general areas (Sireci, 1998):

(1) Construct definition, which refers to the adequacy of how what is intended to be tested is described and specified;
(2) Content relevance, which refers to the relevance of each test element (item) to the tested construct. In other words, a content relevance analysis focuses on the degree to which each item appropriately measures the aspects of the construct it intends to measure;
(3) Construct representation, which investigates the extent to which the test items fully represent the intended construct and do not contain material irrelevant to the construct being measured. Items not judged relevant are removed, and new items are added if experts believe some aspects of the construct are underrepresented. The evaluation of construct representation also includes assessing the relative proportion of items measuring different aspects of the construct;
(4) Appropriateness of test construction procedures, which involves observing the various development, selection, and quality control procedures involved in constructing the instrument. Elements investigated in evaluating these procedures include the training of those who write the items, qualitative and statistical criteria for item selection, screening for potentially biased items, and quality control of items to ensure accurate scoring (Koller et al., 2017). Rossiter (2008), showed that prioritizing content validity is fundamental for the construction of a test. Content validity is a crucial measure that is fundamentally entangled with construct validity and predictive validity of a test outcome; Rossiter argued that much research suffers from inadequate content validation and unnecessary statistical purification, which reduces their validity (Rossiter, 2008). Also he criticizes the practice of using statistical methods like factor analysis for item selection, arguing that these methods can dilute content validity. Many others method to assess content validity have been proposed during the years (Crocker et al., 1989; Martone and Sireci, 2009, Spoto et al., 2023; El-Den et al., 2020); in this work we try to integrate modern artificial intelligence method for language comprehension in the context of Rossiter's view of content validation of psychometrics tests.

In recent years, the advancement in artificial intelligence models has prompted exploration of their potential applications in psychometrics, resulting in the integration of artificial neural networks for tasks such as the selection of variables for psychopathological models (Dolce et al., 2020), reducing dataset dimensionality and the development of shortened test versions using psychometric data (Casella et al., 2024). These new techniques complement traditional methods such as PCA -Principal Component Analysis (Hotelling, 1933) and factor analysis (Bartholomew, 1995), which are widely used to simplify the complexity of psychometric data, allowing psychometricians to identify relationships between variables and identify latent factors determining test results (Milano et al. 2024).

As stated begore, within AI methods and technologies, the rapidly advancing domain of natural language processing , the advent of LLMs (Chang et al., 2023), and the use of text embeddings (Morris et al., 2023) can be particularly fit to open new frontiers in evaluating content validity.

These models with their ability to understand and generate human language on a large scale, offer an innovative method to examine and enhance the coherence and relevance of psychometric test items. Consequently, this allows for an advanced level of analysis where the semantic and contextual similarities between different text segments can be precisely measured. The effectiveness of sentence embeddings from transformer models extends across various natural language processing tasks. These tasks include text summarization, sentiment analysis, question answering, and translation, demonstrating their predictive prowess across a range of applications (Lewis et al., 2020; Radford et al., 2019; Yang et al., 2019).

1.2. Relation to the state of the art and scientific contribution

As expected, the field of language-based psychological assessment has seen a wide application of Large Language Models. Despite the short time period, many works have already been proposed (Kjell et al., 2022; Hitsuwari et al., 2024; Abdurahman et al., 2023; Evans et al., 2022; Nilsson et al., 2024). These works demonstrated that LLM-based embeddings outperform previous methods, such as bag-of-words embeddings, in predicting human responses to personality items (Abdurahman et al., 2023). Furthermore, pre-trained LLMs on very large datasets require no fine-tuning, preprocessing, or data preparation, and they enable the clustering of similar items under the same latent factors

(Nilsson et al., 2024). Kjell et al. (2022) show how psychological assessments of perceived well-being from text responses can be enhanced by LLMs. The authors built a classification model that takes embedded text as input and strongly correlates with standard psychological test rating scales measuring subjective well-being. A similar approach is also used by Hitsuwari et al. (2024), they predicted attitude towards ambiguity using only embeddings from free response to open-endend questions, obtaining promiseful results. This new conception of psychological assessment has led researchers to hypothesize that assessment can go beyond the classical Likert-based rating scales (Kjell et al., 2023); classical tests could be replaced by open-ended questions or colloquia analyzed by LLMs capable of understanding the nuances of language. Contextual Embeddings from LLMs are also the methodology used by Wulff and Mata (2023): they proposes embeddings to automate the detection and resolution of conceptual redundancies (jingle-jangle fallacies) across constructs in psychological measurement; furthermore, they create a fine-tuned model explicitly tailored to analyze personality test items and evaluate this embeddings against measures of internal consistency, structural fidelity, and alignment between construct labels and scale content. In another work they focus on the usability of open language models for analysis in behavioral sciences, introducing this methodology to the broader audience of the psychological community (Hussain et al. 2024). Guenole et al. (2024)introduces Pseudo-Factor Analysis (PFA), a novel psychometric method that predicts latent factorial structures using cosine similarity matrices derived from language embeddings of test items, bypassing the need for human response data.

In our contribution, we explore how the contextual embeddings in item textual analysis can provide a new lens through which assess the validity of personality tests in psychometrics. Based on Rossiter (2008) view of content validation, we developed an automatic procedure for the analysis of the test's items of several personality test. We retrieved construct validity and predictive validity of a test outcome, and consequently content validity, a-priori, based only on the semantics of the items. We employs embeddings to conduct principal component analysis (PCA) and compares semantic similarities with human responses, evaluating how well embeddings predict a-priori factorial structures of personality tests compared to traditional human-response-based methods.

Building on lexical hypothesis in psychology and leveraging embeddings generated by deep learning models to represent questionnaire items in a multidimensional space, we aim to demonstrate how these artificial intelligence models not only complement but can significantly enhance traditional methodologies, offering a more robust and objective approach to predicting and assessing test validity. We further evaluate the practicality of our approach with the application to our method to series of existing tests, including DASS-42 questionnaire (Depression Anxiety Stress Scales, Lovibond and Lovibond (1995), BIG FIVE (Big Five Personality Test, Goldberg, 1992), RIASEC questionnaire (Realistic, Investigative, Artistic, Social, Enterprising, Conventional, Liao et al., 2008,), and HSQ (Humor Styles Questionnaire, Martin et al., 2003).

2. Methods

In this section we describe the tests used in our experiments and the computational models used to generate the embeddings. We begin resuming the datasets of personality questionnaire used in ouexperiments and the models used to retrieve the item's embedding, focusing on how these methods are trained and generated. Finally, we talk about the measures implied to assess semantic similarity between the items and obtain the underlying factorial structure.

2.1. Datasets

We used data collected from the Open Source Psychometrics Project for the Big Five questionnaire (N = 19719 participants), RIASEC questionnaire (N = 145828 participants), HSQ questionnaire (N = 1071 participants) and DASS questionnaire (N = 39775 participants). These questionnaires had 50, 48, 32 and 42 items, respectively, and each participant rated each item on a 5-point Likert scale. We selected these four data sets, between the others in the Open Source Psychometrics Project, as they are all multiconstruct Likert-scale questionnaires with at least 1000 responses, covering diverse topics and involving diverse constructs. The Open Psychometrics Project is a collection of widely used scientific tests that are already validated and well-established in the scientific literature. Participants can take these tests online in an interactive format. The data collected is compiled and made freely accessible for research purposes.The Big Five measures personality within five major domains: Openness (O), Consciousness (C), Extraversion (E), Neuroticism (N) and Agreeableness (A). The test used in the experiments uses the Big-Five Factor Markers from the International Personality Item Pool, developed by Goldberg (1992). The RIASEC questionnaire describes personality through preferences and aversions that influence the choice of work environments (and environments through typical work activities and demands placed on individuals). The questionnaire contains six personality dimensions (and parallel environments): Realistic (R), Investigative (I), Artistic (A), Social (S), Enterprising (E), and Conventional (C), collectively called RIASEC (Liao et al., 2008). The HSQ describes personality through different styles of using humor, containing the dimensions of Self-enhancing, Affiliative, Aggressive, and Self-defeating (Martin et al., 2003). As such, the HSQ uses a conceptualization of humor as a stable multidimensional aspect of personality (Lopez and Snyder, 2003). Validation studies have further shown that the HSQ dimensions (humor styles) correlate with other established personality measures (Martin et al., 2003). Finally, The DASS-42 (Lovibond and Lovibond, 1995) is a 42 item self-report scale designed to measure the negative emotional states of depression(D), anxiety(A) and stress(S). The principal value of the DASS in a clinical setting is to clarify the locus of emotional disturbance, as part of the broader task of clinical assessment. The essential function of the DASS is to assess the severity of the core symptoms of depression, anxiety and stress. As the scales of the DASS have been shown to have high internal consistency and to yield meaningful discriminations in a variety of settings (Clara et al., 2001)

2.2. Computational embedding models

In NLP, an embedding is a vectorized representation of a given input word. Over the years, several techniques have been developed to produce increasingly informative embeddings. However, the challenge of working with long sentences and accurately extrapolating the context of words remains significant.

Most previous research that utilized embeddings to derive the semantics of text has relied on Latent Semantic Analysis (LSA). LSA (Dumais, 2004) is a technique that analyzes the relationships between a set of documents and the terms they contain. It aims to capture the semantics of words by reducing the dimensionality of textual data. LSA begins by constructing a term-document matrix, where each row represents a unique term and each column represents a document. LSA learns latent topics by performing singular value decomposition on the term-document matrix. The resulting vectors represent terms and documents in a new, lower-dimensional space where similar terms and documents are positioned closer together. This space aids in identifying patterns and similarities in the data, facilitating document clustering. LSA has been widely applied in psychological research, as mentioned in the introduction. However, its primary limitation is the single-word-based representation, which can be inadequate when dealing with sentences that require the extraction of long-range context (Abdurahman et al., 2023; Kjell et al., 2023).

With the rise of transformer-based neural networks in NLP, also known as Large Language Models (LLMs), embeddings have become increasingly informative for sentence representation.

To briefly explain, LLMs process sentences by analyzing information provided by the input and extrapolating context to generate an appropriate response. This process begins by linearly projecting a sequence of integer representation (each word is associate to an integer, representing the word index inside a dictionary) of text $\boldsymbol{x}$ into an embedding on the first neural network hidden layer h^0. The transformers forward pass is then followed by N layers, each containing a residual Multi-headed self-attention (MSA) and a multi-layer perceptron block (MLP). MSA allows the model to consider the entire input sequence simultaneously and to weigh the importance of different parts relative to each other. This mechanism enables models to capture complex dependencies in the data, irrespective of the distance between relevant parts in the sequence (Vaswani et al., 2017). The final embedding layer h^n is transformed by the final linear projection followed by the softmax activation function into a sequence of probability distributions $\boldsymbol{y}$, giving the next word based on the input $\boldsymbol{x}$ specifical request. Formally:

$$h^0 = E\mathbf{x} \tag{1}$$

$$\widehat{h}^{n-1} = MSA^n\left(h^{n-1}\right) + h^{n-1},\ n = 1, 2,\ \ldots,\ N \tag{2}$$

$$h^n = MLP^n\left(\widehat{h}^n\right) + \widehat{h}^n,\ n = 1, 2,\ \ldots,\ N \tag{3}$$

$$\widehat{y} = Ph^N, y = softmax(\widehat{y}) \tag{4}$$

In this work we are interested in retrieving embeddings from LLMs, to this reason we end the forward process of the network before generating the network response $\boldsymbol{y}$, to retrieve the final embedded representation of the input: h^n. See Fig. 1 for a schematic representation of the embedding's generation.

Here we employ three transformer based neural networks: the first specifically designed for extrapolating semantic similarity between sentences Sentence-BERT; the second is Mistral, a general-purpose causal language model trained for text generation a causal reasoning; the third is CLIP a multi-modal model trained with text and images.

Sentence-BERT (SBERT) (Reimers and Gurevych, 2019), are a class of BERT (Bidirectional Encoder Representations from Transformers) based models especially trained for extrapolating similarity between the embeddings of sentences. BERT is NLP model that is designed to train neural networks jointly conditioning on both left and right context in all layers. In this way BERT models looks at the entire sequence of words, capturing a broader understanding of language and context. BERT is pretrained on a large corpus of text, English Wikipedia and the entire google books dataset, and is trained to predict randomly masked words in sentences and in next sentence prediction: if two sentences follow each other. After pretraining BERT can be fine-tuned adding additional output layers in order to accomplish other tasks. In the case of SBERT it was specifically designed to improve the efficiency and effectiveness of deriving semantically meaningful sentence embeddings. It utilizes Siamese Neural Networks, an architecture designed to be trained comparing pairs or triplet of sentences, optimizing the embeddings to put semantically similar sentences closer in the embedding space and dissimilar sentences apart. The base SBERT model has been improved in the last years, several models have been proposed and a benchmark table has been developed testing all the models on the basis of their performance in extrapolating similarities between sentences) (Reimers and Gurevych, 2019) .In our experiments we use all-roBERTa-large model, a fine-tuned model base on RoBERTa (Liu et al, 2019), specifically designed for predict the similarities between two sentences. The model is trained on 1.2 billion sentence pairs coming from 35 different dataset of sentence spanning over a wide range of topics, see reference for details on the training methods. Along with RoBERTa, in Section 3.5, we tried different models fine-tuned on top of SBERT model. These models vary in size and use different training methodology, we compare

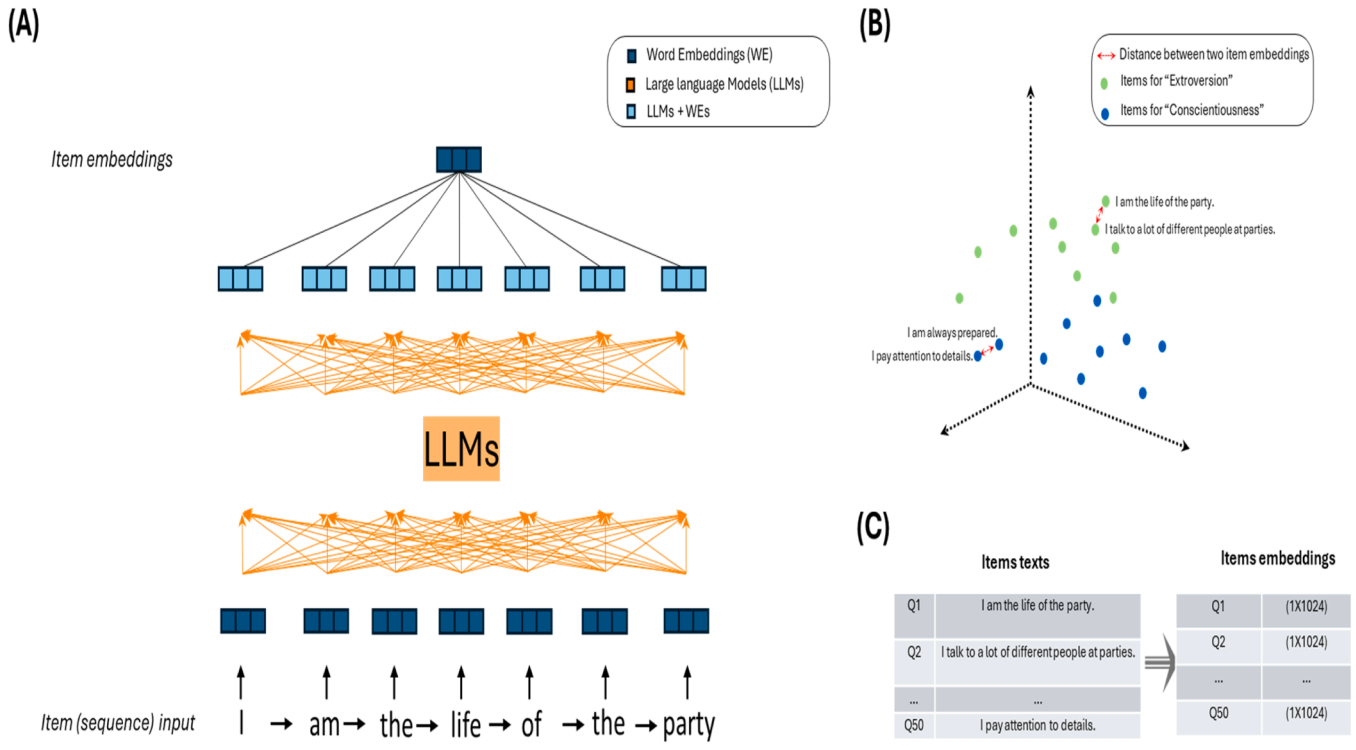

Fig. 1. Schematic representation of the item embedding process. Panel (A): The item is passed as input to the language model. Each word is singularly vectorized in the first layer (WordEmbedding), then the LLM apply the MSA relating each word the others and extrapolating the context of the sentence throughout all the layers and giving the final embedding for each word. Finally the embeddings are averaged and a single embedding is produced for each item. Panel (B) shows a compact representation of embedding space: ideally, if the model correctly find semantic similarity among item, we should find that item related to the same construct are closer together than item related to different constructs. Panel (C) summarizes this process showing how all items of the questionnaire are transormed in equally sized row vectors.

the performance in predicting the factorial structure of the personality tests.

Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021) is a multi-modal model that learns visual concepts from natural language descriptions. It tries to relate images to textual descriptions to perform a wide variety of visual tasks without the need for task-specific training data. CLIP is trained on a large dataset composed of images paired with corresponding textual descriptions. The model is composed of two main components, a visual and a textual encoder that transform their respective inputs into embeddings in a shared multidimensional space. During training CLIP uses contrastive learning. CLIP learns to align the embeddings from the image encoder with the corresponding embeddings from the text encoder. It does this by maximizing the similarity between the correct pairs of image and text embeddings, while minimizing the similarity between mismatched pairs. This kind of approach differs from classical textual only embeddings retrieval. It tries to incorporate information from images into a share embedding space, so we tried embeddings derived from CLIP in contrast to other classical methods. We include CLIP in our semantic analysis, even if it is a multi-modal model, to test if the richness of CLIP embeddings derived from their pretraining on 400+ million text-image pairs can generate robust and meaningful representations, even when isolated from their original multimodal (text-image) context.

Mistral 7B (Jiang et al, 2023) is an open-source large language model trained for general language understanding and generation. It is specifically trained for advanced causal reasoning and question answering. It has been proved to be more performant of much larger models despite its size (Jiang et al, 2023). In contrast to the other two models, Mistral does not produce directly an embedding of the input sentence, because it is not directly trained for learning embedding representation but to generate text. The embedded representation of the input sentence is an emergent property of the large language models, that indirectly learns similar embeddings for similar inputs, in order to better perform in causal reasoning. Lange language models' embeddings has already proven very effective in reproducing human behavior in cognitive task with a little fine-tuning (Binz and Schulz, 2023), here we use the embeddings without fine-tuning them to see if useful semantic representation spontaneously emerge. We chose Mistral 7B over other language models because it demonstrates strong performance on classic benchmarks despite its relatively small size. Moreover, its compact architecture allows it to run efficiently on many personal computers without requiring specialized hardware. To extrapolate the embeddings from Mistral we took the last hidden layer before text generation and averaged it for the whole sentence.

2.3. Measures

To retrieve the underlying factor structure of the embeddings we need to extrapolate the correlation between them. A common measure in the literature related to embeddings semantic analysis is the cosine similarity.

Cosine similarity is a measure used to determine how similar two non-zero vectors are in an inner product space. This metric calculates the cosine of the angle between the two vectors, providing an indication of their orientation to each other, irrespective of their magnitude. Cosine similarity is particularly used in positive space, where the outcome is neatly bounded in [0, 1]. Formally:

$$\cos(x,y) = \frac{\sum_{i=1}^{D} x_i y_i}{\sqrt{\sum_{i=1}^{D} x^2}\sqrt{\sum_{i=1}^{D} y^2}} \tag{5}$$

Where **x** and **y** are two embeddings vector in a D-dimensional embedding space.

Eq. (1) is very similar to classic Pearson correlation coefficient, formally:

$$\hat{r} = \frac{\sum_{i=1}^{D}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{D}(x-\bar{x})^2}\sqrt{\sum_{i=1}^{D}(y-\bar{y})^2}} \tag{6}$$

When the embedding means $\bar{x}$, $\bar{y}$ are zero, cosine similarity and Pearson's *r* are equal. As showed by Zhelezniak et al. (2019), in many natural language processing applications the two measures coincide because the embedding returned by language models have mean very close to zero. From our preliminary analysis on our embeddings this superposition between the two measures remains, so we prefer to use the cosine similarity coefficient as correlation measures and sum-of-squares-and-cross-products matrix (SSCP) on which to perform the principal component analysis for dimensionality reduction.

Our approach follows these steps: first, we calculate the embeddings for each test. Then, we retrieve the Pearson correlation for each pair of embedded items. Once we obtain the Pearson correlation matrix, we apply Principal Component Analysis (PCA) to the matrix, selecting the number of principal components based on the factors expected by the theory underlying the test. By projecting the items into a low-dimensional space, we can observe how items are assigned to relevant factors and verify their correct clustering by examining the item factor loadings. Finally, we measured the correlation between the loadings derived from the PCA on the embeddings' correlation and the loadings obtained from the PCA on the subject responses. For clarity of reading, we don't use all the models contemporary, we make all our analysis using RoBERTa as base language model to prove our points, because it already has been proved effective in previous psychological research using items embedding (Abdurahman et al., 2023; Kjell et al., 2023). Finally, in concluding paragraph we compare RoBERTa with the other two language models.

3. Results

3.1. Correlation matrix of the embeddings

We retrieved the embeddings for each item of a test using them as inputs of the LLMs. Each test comprises a different number of items, *N*; the list of items serves as input for the Embedding model, which returns an output of dimension *(N,E)*, where *E* represents the dimension of the embedding space. This dimension may vary among different models; for instance, in the case of RoBERTa, $E = 1024$. The matrix returned from the model represents the projection of the complete test into the model's embedding space, with each row of the matrix corresponding to the embedding vector for a particular item of the test. Once the embeddings are retrieved, we employ a preliminary approach to determine if similarities exist among them in a large multidimensional space. We perform a 2-Dimensional projection of the embedded items using T-SNE (Van der Maaten and Hinton, 2008). The results for the Big Five questionnaire are reported in Fig. 2.

From the figure, it is clear that even though we start from a very large space of 1024 dimensions, items related to the same factor are mapped closer together. As we can observe, the majority of items belonging to the same construct are clustered together. For example, observe the lower-left zone of the space where we can find neuroticism-related items (blue points), or the middle left where there are conscientiousness items (green points). Agreeableness and extraversion-related items appear to be more scattered together, indicating a substantial overlap in meanings. Moreover, some isolated items are mapped in different zones compared to the majority of the items belonging to the same factor, such as C8 and O9, likely reflecting ambiguous semantics that the language model is not able to properly discriminate. Nevertheless, this preliminary analysis gives us a hint that the language model embeddings are effectively correlated and that useful information appears to be contained in them.

After calculating the embedding for each item, we measure the semantic similarity of each pair using the Cosine similarity coefficient,

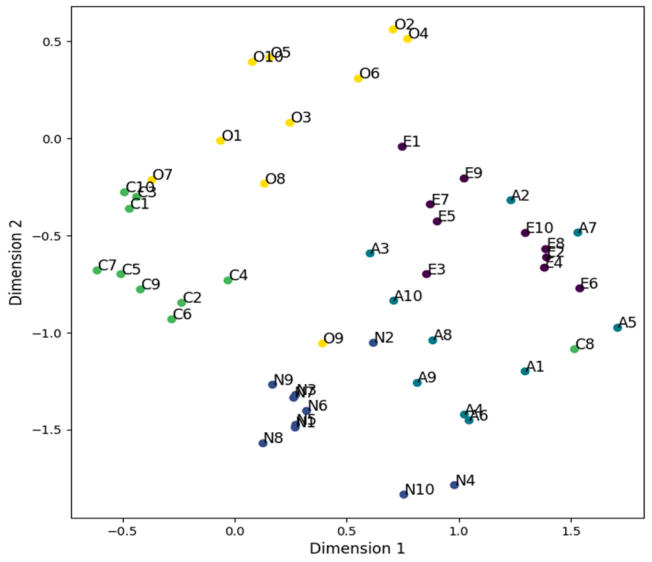

Fig. 2. T-SNE projection of the embeddings in a 2-dimensional space. Yellow points are openness (O) related items, Green points are Conscientiousness (C) items, Blue points are Neuroticism (N) items, Black points are extra-version (E) items and violet points are Agreeableness (A) items.

obtaining a squared symmetrical matrix that represents the correlation among all pairs of items, denoted as dimension *(N,N)*. In Fig. 3, we report the average correlation between items belonging to the same construct and items belonging to different constructs for the embeddings retrieved with the RoBERTa model. We observe that items belonging to the same construct show the highest similarity across all the questionnaires tested. We conduct a Welch's test to assess the significative difference of the means, the similarity of identical items is removed so we have unequal sample size. We found that the average similiraty of two constructs are significantly different (p-value < 0.001) . This implies that, even when not explicitly trained with psychological items, the language model has the capacity to capture similar meanings in test's structure, based on the vector embeddings of items.

To further analyze how the factors underlying the items are composed, we perform a PCA analysis starting from the correlation matrix of the embeddings for all tests.

3.2. PCA analysis

Performing a Principal Component Analysis (PCA) helps us understand how different items are clustered together and whether items belonging to the same construct are correctly identified. Moreover, we can retrieve the loadings related to each principal component to determine how significantly an item contributes to a principal component. Additionally, in the case of cross-loadings, we can identify other constructs to which an item is incorrectly assigned.

For the number of principal components, we used the number of dimensions hypothesized by the theory; for example, 5 for the Big Five personality traits questionnaire, 6 for the RIASEC model, 3 for the DSA, and 4 for the HSQ. We employed varimax as the method for orthogonal rotation. Tables 1–4 present the eigenvalues, the proportion of variance explained, and the cumulative variance for each test, showcasing results obtained with the best embedding model.

The cumulative variance explained by the PCA applied to the embeddings is higher than that explained by human responses across all four tests. This suggest that the embeddings alone can effectively capture the underlying factors described by the theory. Next, we identified the loadings of each item to ascertain whether the items are correctly assigned to their underlying constructs. We calculated the highest loadings for each item and determined how many items, theoretically belonging to the same construct, share the highest loading on the same principal component. To be considered the highest, a loading must also be greater than 0.4 (Stevens, 2002).

Ideally, in the case of the Big Five questionnaire, we aim to find that items related to the extraversion construct have the highest loading on the same principal component. The same expectation applies to items related to neuroticism, and so on.

To achieve this, we measured the percentage of items correctly assigned together on the same principal component and represented this percentage on a radar plot, shown in Fig. 4.

From Fig. 4, it is evident that almost all constructs of all tests are composed of the same items as predicted by theory. We can observe that many constructs from different tests are correctly clustered together, with a percentage exceeding 90%. On average, the correct semantic similarity clustering found is above 70% in all four tests.

However, some constructs are more challenging to cluster, such as agreeableness and stress. Here, we found the percentage of correctly clustered items to be around 50%. This result is not surprising when we consider, as shown in the table of loadings (S1), that many items related to agreeableness have higher loadings on the same principal component as those related to extraversion. These findings are common in real-world scenarios, where subjects' responses often show a significant correlation between these pairs of constructs (Côté and Moskowitz, 1998).

3.3. Correlation with human response

Until now, we have demonstrated how embedding models based on large language models can accurately identify semantic similarities among items. Moreover, the PCA, starting from the correlation between embeddings, defines a latent space that correctly aligns items related to the same construct with the same principal component, ensuring the interpretability of the results and uncovering the theoretically hypothesized structure. Here, we aim to determine if the latent structure identified from the semantic similarity analysis matches the latent structure derived from human subjects' actual responses. Specifically, we are interested in analyzing whether the loadings obtained from the embedded item representations are comparable to those derived from the subjects' responses. By measuring the correlation between the loadings in these two cases, we can gain insight into how well the semantics of the items predict the subjects' responses. Ideally, a high correlation between the loadings of a construct indicates not only that the related items are correctly grouped together but also that the cross-loadings and the loadings of other factors follow similar patterns.

To achieve this, we collected human data from the open psychometrics website for all four tests. Then, we applied the same procedure described above for the embeddings to the matrix of responses: first, we calculated the Pearson correlation for each pair of items, then we applied PCA on the correlation matrix using the number of latent factors predicted by the theory, finally retrieving the items' loadings. In the case of tests that include reverse-scored items, we first rescaled the items to the correct range. This operation is necessary because the embedding model does not differentiate between reversed and not-reversed items. Thus, to ensure a proper comparison between loadings, we rescale the responses in the case of reversed items. Tables 5–8 shows the Spearman correlation factors for the loadings of all the four-test examined along all constructs.

The tables show the Spearman correlation factor for the four examined test. Here we used the Spearman Correlation because the loadings are not normally distributed. As we can see we found high correlation between the loadings of the embeddings and the loadings of the response on the same constructs for all the tests ($R > 0.5$, p-value$<<0.001$). These findings support what we have already found in the previous principal component analysis, and show how semantic similarity between items captures the relation among factor that we retrieve in human responses.

To reinforce these results, we calculated Omega (McDonald, 1999) to

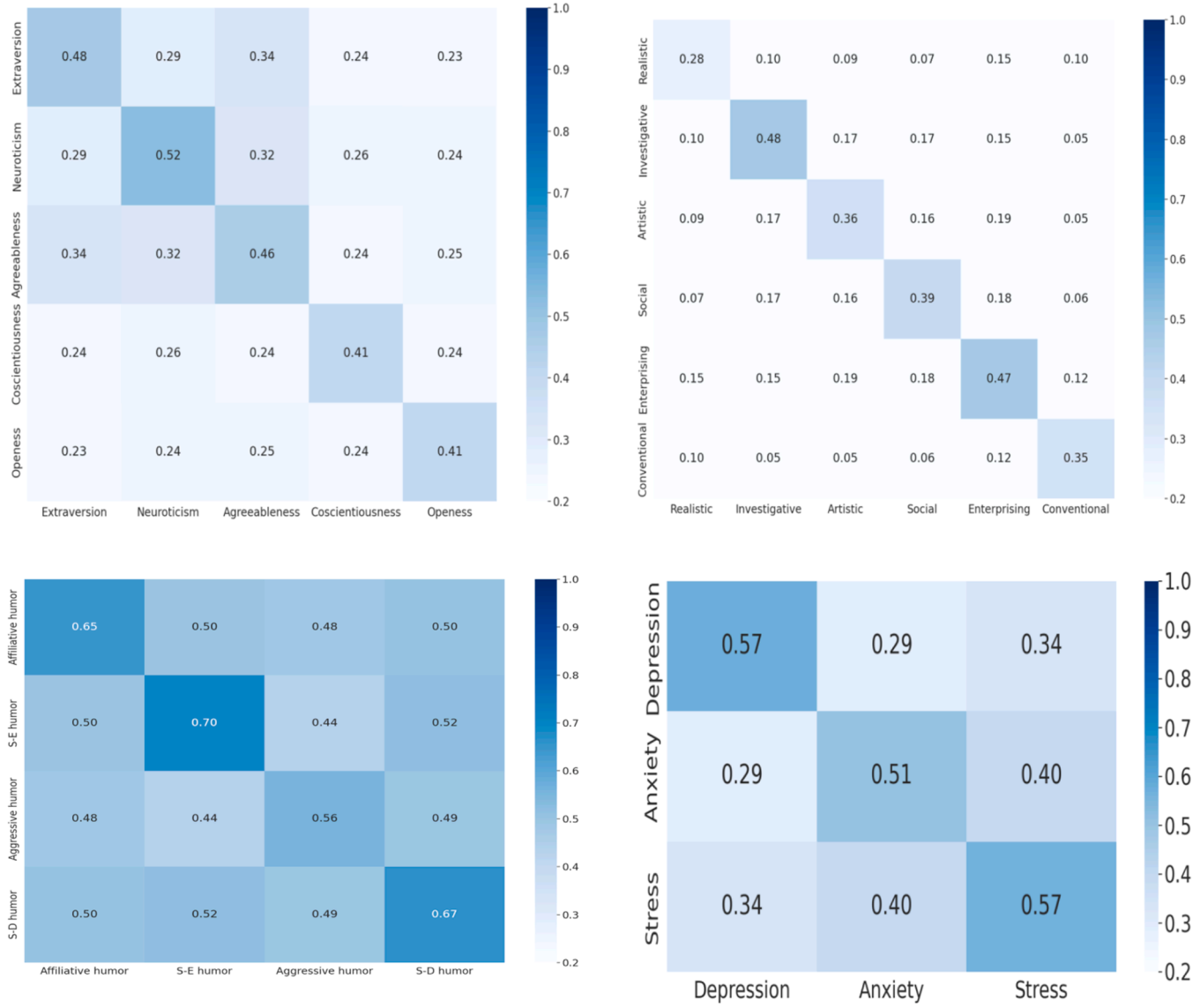

Fig. 3. Average Cosine similarity matrix across constructs for the four questionnaires. Upper left figure shows the Big 5 average correlation, Upper right Riasec, lower left HSQ and lower right the DASS questionnaire. Items belonging to the same construct shows higher correlation respect to cross-constructs correlation.

Table 1
Eigenvalues, variance explained by each component and cumulative variance relative to the Big 5 Test.

Big 5					
	PC 1	PC 2	PC 3	PC 4	PC 5
Eigenvalues	6.46	5.57	4.90	4.50	3.77
Variance explained	0.13	0.11	0.10	0.09	0.08
Cumulative variance	0.13	0.24	0.34	0.43	0.51

Table 2
Eigenvalues, variance explained by each component and cumulative variance relative to the RIASEC Test.

RIASEC						
	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6
Eigenvalues	4.46	4.38	3.27	3.20	3.07	2.71
Variance explained	0.09	0.09	0.07	0.07	0.06	0.06
Cumulative variance	0.09	0.18	0.25	0.32	0.38	0.44

Table 3
Eigenvalues, variance explained by each component and cumulative variance relative to the DSA Test.

DSA			
	PC 1	PC 2	PC 3
Eigenvalues	8.36	8.16	8.01
Variance explained	0.20	0.19	0.19
Cumulative variance	0.20	0.39	0.58

Table 4
Eigenvalues, variance explained by each component and cumulative variance relative to the HSQ Test.

HSQ				
	PC 1	PC 2	PC 3	PC 4
Eigenvalues	6.21	5.89	5.86	4.12
Variance explained	0.19	0.18	0.18	0.13
Cumulative variance	0.19	0.37	0.56	0.68

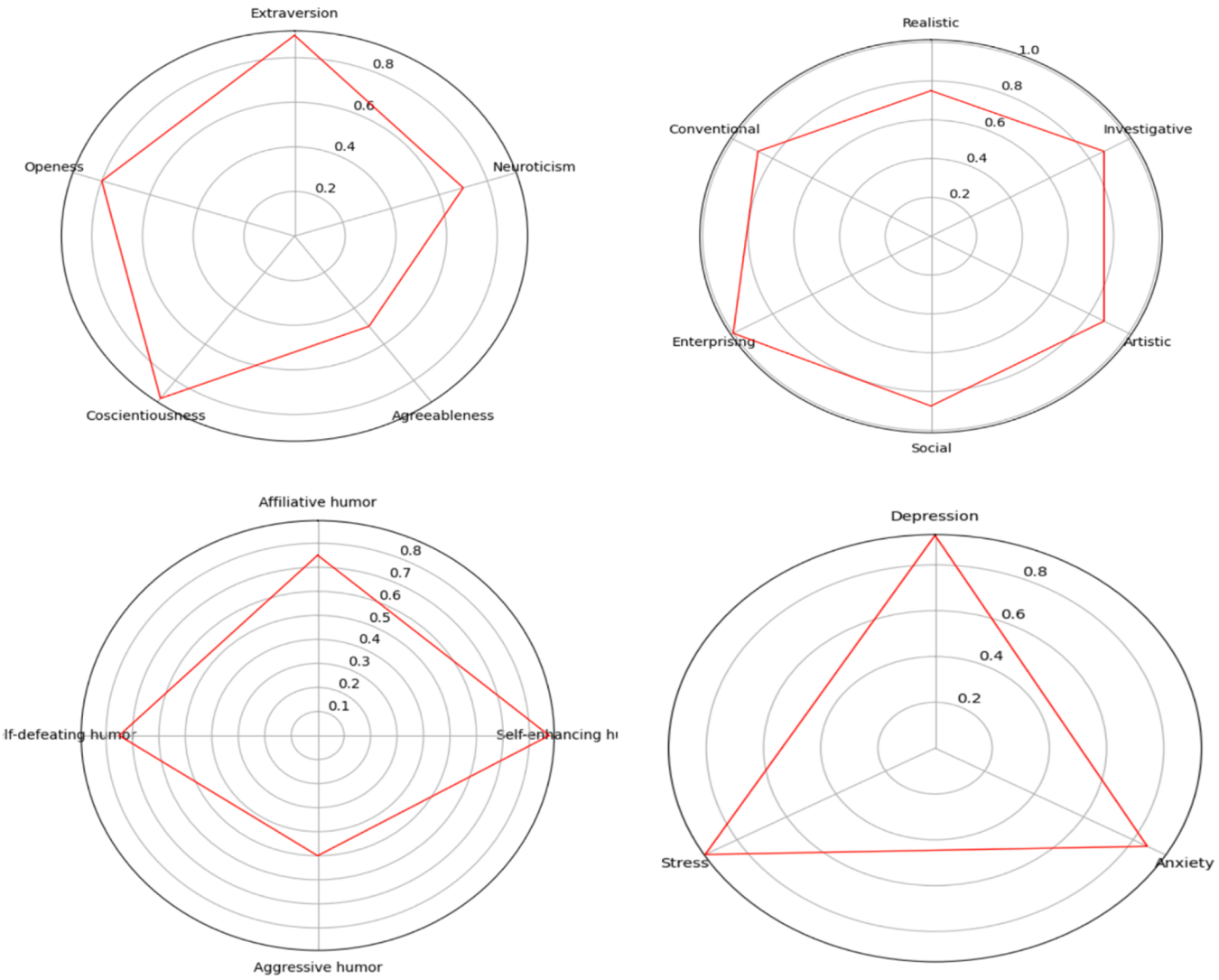

Fig. 4. Radar plot of the percentage of items belonging to the same construct correctly assigned to the same principal component. Upper left figure shows the Big 5 average correlation, Upper right Riasec, lower left HSQ and lower right the DASS questionnaire.

Table 5
Spearman R correlation between the loading retrieved from the PCA on the embeddings and the loadings obtained from the PCA on human's response to the Big Five test along with confidence interval (Z Fisher-Transformed 95% CI, N-samples 100). The asterisk indicates significant p-value (< 0.001).

Embeddings/humans	Extraversion	Neuroticism	Agreeableness	Coscientiousness	Openess
Extraversion	0.57 [0.34, 0.73] *	-0.17 [-0.42, 0.11]	0.03 [-0.25, 0.30]	-0.14 [-0.40, 0.14]	-0.13 [-0.39, 0.15]
Neuroticism	0.15 [-0.13, 0.41]	0.40 [0.14, 0.61] *	0.10 [-0.18, 0.37]	-0.10 [-0.37, 0.18]	-0.25 [-0.53, -0.01]
Agreeableness	0.38 [0.11, 0.60]	-0.11 [-0.38, 0.17]	0.65 [0.45, 0.79] *	-0.34 [-0.56, -0.07]	-0.25 [-0.49, 0.03]
Coscientiousness	-0.24 [-0.49, 0.04]	0.02 [-0.26, 0.30]	-0.22 [-0.47, 0.06]	0.60 [0.39, 0.75] *	-0.25 [-0.49, 0.03]
Openess	-0.35 [-0.57,-0.08]	-0.12 [-0.39, 0.16]	-0.17 [-0.43, 0.11]	0.16 [-0.12, 0.42]	0.64 [0.44, 0.78] *

assess how well the items express the underlying construct in cases of human responses and embedded representations. Omega is formally defined as follows:

$$\omega = \frac{\left(\sum_i \lambda_i\right)^2}{\left(\sum_i \lambda_i\right)^2 + \sum_i u_i} \tag{7}$$

Where λ_i are the item's loadings related to a given factor i and u_i their

Table 6
Spearman R correlation between the loading retrieved from the PCA on the embeddings and the loadings obtained from the PCA on human's response to the RIASEC test along with confidence interval (Z Fisher-Transformed 95% CI, , N-samples 64). The asterisk indicates significant p-value (p-value < 0.001).

Embeddings/humans	Realistic	Investigative	Artistic	Social	Enterprising	Conventional
Realistic	0.54 [0.30, 0.71] *	-0.21 [-0.47, 0.08]	-0.14 [-0.41, 0.15]	-0.18 [-0.44, 0.11]	0.14 [-0.15, 0.41]	0.25 [-0.04, 0.50]
Investigative	-0.14 [-0.41, 0.15]	0.45 [0.19, 0.65] *	0.28 [-0.00, 0.52]	0.00 [-0.28, 0.28]	-0.27 [-0.51, 0.02]	-0.38 [-0.64, -0.04]
Artistic	-0.31 [-0.55, -0.03]	0.03 [-0.26, 0.31]	0.60 [0.38, 0.76] *	0.19 [-0.10, 0.45]	-0.08 [-0.36, 0.21]	-0.30 [-0.54, -0.02]
Social	-0.26 [-0.51, 0.03]	-0.11 [-0.38, 0.18]	0.10 [-0.19, 0.37]	0.57 [0.34, 0.74] *	-0.06 [-0.34, 0.23]	-0.21 [-0.47, 0.08]
Enterprising	-0.10 [-0.37, 0.19]	-0.29 [-0.53, -0.01]	-0.03 [-0.31, 0.26]	-0.05 [-0.33, 0.24]	0.62 [0.41, 0.77] *	0.11 [-0.18, 0.38]
Conventional	0.21 [-0.08, 0.47]	-0.13 [-0.40, 0.16]	-0.28 [-0.52, 0.00	-0.30 [-0.54, -0.02]	0.23 [-0.06, 0.48]	0.72 [0.55, 0.83] *

Table 7
Spearman R correlation between the loading retrieved from the PCA on the embeddings and the loadings obtained from the PCA on human's response to the HSQ test along with confidence interval (Z Fisher-Transformed 95% CI, N-samples 96). The asterisk indicates significant p-value (p-value < 0.001).

Embeddings/ humans	Affiliative humor	S-E humor	Aggressive humor	S-D humor
Affiliative humor	0.50 [0.18, 0.72] *	-0.10 [-0.43, 0.26]	-0.06 [-0.40, 0.29]	-0.26 [-0.56, 0.10]
S-E humor	0.17 [-0.19, 0.49]	0.64 [0.37, 0.81] *	-0.34 [-0.62, 0.01]	-0.16 [-0.48, 0.20]
Aggressive humor	-0.38 [-0.64, -0.04]	-0.36 [-0.63, -0.01]	0.45 [0.12, 0.69] *	0.10 [-0.26, 0.43]
S-D humor	-0.03 [-0.37, 0.32]	-0.08 [-0.42, 0.28]	-0.14 [-0.47, 0.22]	0.56 [0.26, 0.76] *

Table 8
Spearman R correlation between the loading retrieved from the PCA on the embeddings and the loadings obtained from the PCA on human's response to the DASS test along with confidence interval (Z Fisher-Transformed 95% CI, N-samples 72). The asterisk indicates significant p-value (p-value < 0.001).

Embeddings/humans	Depression	Anxiety	Stress
Depression	0.90 [0.80, 0.95]	-0.32 [-0.60, 0.03]	-0.28 [-0.57, 0.08]
Anxiety	-0.34 [-0.62, 0.01]	0.81 [0.64, 0.90]	-0.08 [-0.42, 0.28]
Stress	-0.37 [-0.64, -0.02]	-0.04 [-0.38, 0.31]	0.80 [0.63, 0.90]

Table 9
Omega result for each construct of the Big Five test measured from the subject's response and from the embedded representation.

Big 5 constructs	Human subjects ω	Embedded representation ω
Extra-version	0.91	0.87
Neuroticism	0.89	0.89
Agreeableness	0.87	0.77
Neuroticism	0.85	0.84
Openess	0.85	0.85

variance. Results are reported in Tables 9–12 for all the tests examined.

The ω values that we measured are nearly identical for each construct of each test. However, we found a substantial mismatch in the case of Agreeableness and Aggressive Humor, which also yielded weaker results in the previous analysis compared to the percentage of correctly identified items for other constructs.

In the next section we use other approaches to get the embeddings of our items, to see what the best performing method in the context of a content validity analysis is.

3.4. Different embeddings models

We have observed how a BERT-based model, explicitly designed and evaluated for its quality in sentence embedding, provides a robust representation of the latent structure of different tests' items. In this subsection, we explore embeddings generated by various models explicitly trained to produce meaningful representations for analyzing semantic relationships in text. These models differ in their training methodologies and sizes. RoBERTa, a general-purpose model designed for versatility across multiple use cases (Liu et al., 2019), serves as a baseline in our comparisons. In addition, we examined models such as GTR-T5-XXL (Ni et al., 2021a), optimized for semantic search tasks, and Sentence-T5-XXL (Ni et al., 2021b), specifically trained for sentence similarity analysis. We also included MPNet-Base-v2 (Song et al., 2020), another general-purpose model with enhanced contextual understanding. For further details on the training procedures and objectives of these models, we recommend consulting the original publications.

Furthermore, along with models explicitly trained to analyze semantic relationship from text sources, we used also CLIP and MistralAI. CLIP is a multimodal model capable of processing texts and images by embedding them into a joint vector space. MistralAI, on the other hand, is a general-purpose causal large language model. It is not specifically designed for sentence similarity, and its embeddings are extracted from the last layer of the transformer network.

Table 10
Omega result for each construct of the RIASEC test measured from the subject's response and from the embedded representation.

RIASEC constructs	Human subjects ω	Embedded representation ω
Realistic	0.89	0.81
Investigative	0.91	0.88
Artistic	0.88	0.80
Social	0.88	0.81
Enterprising	0.85	0.87
Conventional	0.91	0.80

Table 11
Omega result for each construct of the HSQ test measured from the subject's response and from the embedded representation.

HSQ constructs	Human subjects ω	Embedded representation ω
Affiliative humor	0.86	0.84
Self-enhancing humor	0.84	0.87
Aggressive humor	0.84	0.74
Self-Defeating humor	0.86	0.85

Table 12
Omega result for each construct of the DASS test measured from the subject's response and from the embedded representation.

DASS constructs	Human subjects ω	Embedded representation ω
Depression	0.95	0.94
Stress	0.90	0.92
Anxiety	0.91	0.93

The methodology employed is identical to that described for the RoBERTa model: we retrieve the embeddings, calculate the correlations among them, and perform a principal component analysis. As a performance measure among the models, we used the average percentage of correctly aligned items belonging to the same constructs for each test. Fig. 5 displays a bar chart showing the average number of correctly grouped items, along with the standard deviation for each test. As can be seen, the roBERTa and MPNET models performs equally good in all tests (we conducted a Student t-test for the comparison of means finding a p-value >0.05 in all tests, sample size depends from the factors of the test). The sentence-T5 model reaches the performance of RoBERTa in all tests (Student t-test p-value >0.05) with the only exception of the Big 5 (Student t-test p-value < 0.01). The GT-T5 perform worse than RoBERTa in all tests (Student t-test p-value < 0.01) except that in the Big 5 questionnaire (Student t-test p-value > 0.05). Finally, Mistral and CLIP fail to reach RoBERTa performance in all tests (Student t-test p-value $<$ 0.01). Examining the loadings of these two models, we note that the loadings do not appear to be related to a single principal component, and it is not possible to clearly identify a principal axis grouping items related to the same constructs. This is not surprising, as these models are not specifically trained for semantic tasks. In contrast, MPNET and RoBERTa outperform larger models like GT-T5 and sentence-T5, despite being significantly smaller ($<$1GB of model size compared to $\sim$10GB) and trained on smaller corpora. These findings suggest that in this domain, a tailored approach—specifically focused on using words and phrases from personality research—could improve the performance of language models fine-tuned for analyzing personality items. Recent studies have already shown promising results in this direction (Wulff and Mata, 2023; Hommel and Arslan, 2024).

4. Conclusions

The use of psychometric test in research, diagnosis, assessment and evaluation, considering the impact it may have on individuals, requires to rely on trustable, robust, reliable and valid tools. Respecting this imperative requires, of course, a notable effort and a strict respect of procedures.

At the same time, it stimulated the quest for newer and newer procedures that lead to more and more robust, reliable and valid tests. Novel approaches based on LLMs for psychological assessment have shown that these models are powerful enough to achieve test performance based solely on respondents' answers to open questions or by analyzing the natural speech of patients (Kjell et al., 2022; Hitsuwari et al., 2024; Abdurahman et al., 2023; Nilsson et al., 2024; Kjell et al., 2023). These approaches are promising, but participants' responses still need expert supervision to be analyzed by a machine learning algorithm. In contrast, classical test platform, through the process of content validation, ensures the selection of appropriate items to measure intended constructs without requiring supervision. The results emerge as a process of clustering, as in our case. We believe that validation and reliability processes are still necessary in psychological assessment, and LLMs offer an insightful way to support this process.

Our study advances the field by integrating contemporary artificial intelligence technologies—specifically large language models (LLMs) and text embeddings—to enhance the content validity analysis of personality tests. This integration provides a novel lens through which examine and refine test items.

In this work we have proposed a method to support the validation process, by focusing on the semantics of items and applying artificial intelligence methods that provides very useful hints on items characteristics and factorial structure even before test administration. This can allow to analyze, select and refine items also before a trial, introducing a new "ex ante" step in test validation. This sheds further light on items and test psychometrics characteristic and enhances more accurate estimation of them by researchers involved in item formulation and test validation.

We have demonstrated that the application of LLMs, in conjunction with sentence embeddings, can effectively capture the semantic nuances of test items. This capability allows for a more precise assessment of content validity before the practical application of the test. By analyzing the relationships between test items through the semantic similarity extracted by these models, we have been able to identify underlying factor structures that are in alignment with the theoretical constructs they intend to measure. This alignment underscores the potential of LLMs to not only support but also potentially enhance the methodological framework for psychometric test development and validation.

Our findings indicate that LLMs can discriminate subtle linguistic patterns that may not be immediately apparent. For instance, the semantic clustering of items associated with similar traits suggests that these models can detect congruence and discrepancies within test content effectively. This ability enhances the precision of test construction by ensuring that items are both representative and relevant to the constructs being measured, thereby improving the overall quality and applicability of psychometric assessments.

Moreover, the practical applications of this research extend beyond the academic realm into clinical and organizational settings where psychometric tests are frequently employed. By enhancing the content validity of these tests, practitioners can make more informed decisions based on results that more accurately reflect the true characteristics and abilities of individuals. Anyway, our results are based on already validated tests that have undergone a statistical process dedicated to removing items not aligned with the theoretical structure (Rossiter, 2002). It would be interesting to see if our findings remain valid with newly proposed, unvalidated items. This is a limitation of our study, and in future work, we will try to integrate the LLMs embedding into the test construction process to see if the prediction of respondents' factorial structure still holds. Another intriguing open question not addressed in our work is the nature of the errors made by the language model when misclassifying items. Analyzing these errors could provide valuable insights into which items pose greater challenges for the language model and, conversely, which items the model handles with ease but appear ambiguous to humans. Such an analysis could contribute to a deeper understanding of the similarities and differences between human and artificial intelligence capabilities. Furthermore, our analysis was specifically restricted to the personality domain. Future research should explore whether these findings generalize to psychometric tests assessing other psychological attributes, such as cognitive abilities, memory, technical aptitudes, and more. Additionally, future studies could investigate the potential of using multi-modal transformers model embeddings to analyze non-verbal test environments, further enhancing the application domain in psychometrics of modern artificial intelligence models.

Our work also corroborates previous studies where similar results are

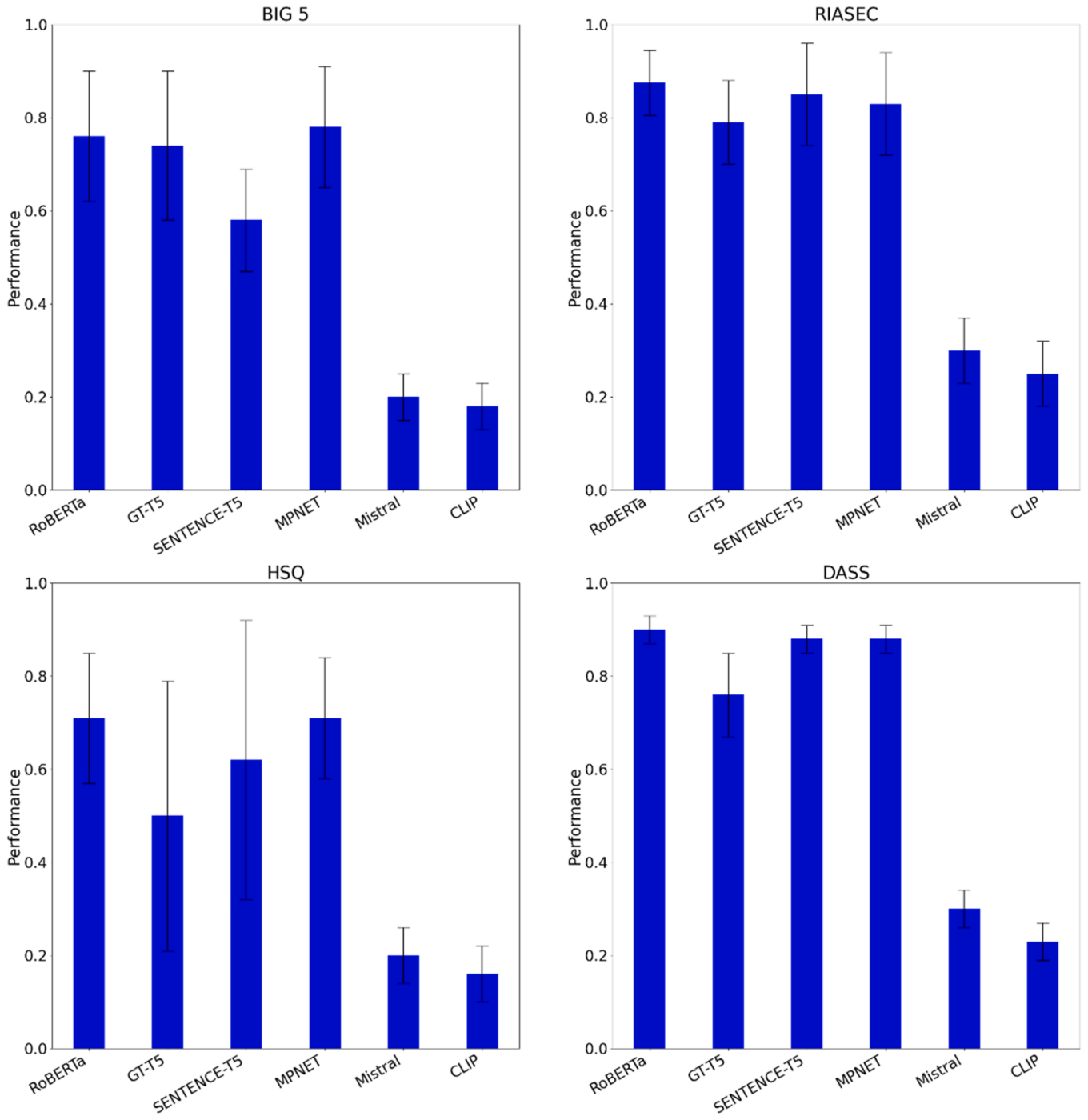

Fig. 5. Bar chart, with error bars reporting standard deviation from the mean, showing the performance of roBERTa, GT-T5, Sentence-T5, MPNET, Mistral and CLIP on the four examined questionnaire. For each questionnaire the performance is calculated averaging the percentage of item belonging to the same construct correctly assigned to the same principal component. Upper left figure shows the Big 5 average correlation, Upper right Riasec, lower left HSQ and lower right the DASS questionnaire.

implicitly addressed (Wulff and Mata, 2023; Abdurahman et al., 2023; Kjell et al., 2022; Nilsson et al., 2024). We explicitly highlight how the factorial structure observed in humans emerges from the semantics of test items. This finding is essential for predicting human responses to items, assessing structural fidelity and alignment between constructs and scale content, and, more broadly, leveraging these results within the context of content validity.

It is important to note that our best performing large language models (LLMs), RoBERTa and MPNET, are specifically trained to identify semantic similarities in sentences. Anyway, this models are trained using general-purpose sentences that span many different fields. Nonetheless, the vast variety of sentences to which they are exposed is sufficient to ensure a proper semantic understanding of psychological items. Furthermore, from a purely generalist semantic understanding of the text, we retrieve the same underlying factor structure that we assume to be at the basis of human behavior. This could be seen as a further proof of the deep entanglement between language and behavior, proving that models that exploit key property of language could be very useful in the context of behavioral science. We also found that model complexity is not directly correlated with performance in this domain.

More complex and larger models, such as GT-T5 and Sentence-T5, perform at least as well as smaller models like RoBERTa and MPNET, but more often underperform. Notably, the smaller models, which are an order of magnitude smaller in size (measured in megabytes), frequently yield comparable or better results. These findings suggest that a thoughtful training approach, explicitly tailored for personality research, could further enhance performance. This conclusion is supported by recent studies (Wulff and Mata, 2023; Hommel and Arslan, 2024). Furthermore, models such as Mistral and CLIP unsurprisingly fail to accurately identify similarities among items. For Mistral, this is likely because its training focuses on general language understanding and text generation rather than on semantic analysis. The embeddings it produces are not immediately suited for classification tasks based on semantic analysis of text. However, a recent work (Binz and Schulz, 2023) has demonstrated how simple fine-tuning of the general LLM embeddings on classic language-based cognitive tasks can replicate human behavior. These findings suggest that embeddings from general-purpose language models may also contain useful information if they are carefully fine-tuned for specific tasks. The problem with CLIP-based models may lie in the fact that they integrate images and text. When dealing with psychological items, we often encounter abstract concepts that are typically not present in images, which are more related to concrete objects. This results in embeddings that are not aligned with our task, leading to poor outcomes when we perform dimensionality reduction analysis.

The quality and breadth of training data for these models can affect their ability to generalize across different types of psychological items. We have observed that a generalist model, specifically trained to classify semantic similarities in sentences, conceals the same latent structure found in human subjects. It would be interesting to explore the creation of a model specifically trained on a large dataset of psychological items in future works. Such a model could perform as an expert validator and serve as a useful tool for preliminary content validation.

Future research should therefore focus on refining these models to handle a broader array of test types and constructs, ensuring their applicability across diverse psychometric scenarios. Continued advancements in NLP and machine learning could further enhance the accuracy and efficiency of semantic analyses, potentially leading to more robust methodologies for psychometric validation. Additionally, comparative studies involving traditional and AI-enhanced methods would be invaluable in quantifying the improvements in validity and reliability contributed by these technologies. In conclusion, the integration of large language models and text embeddings into the field of psychometry represents a promising frontier for enhancing the scientific rigor and practical utility of psychometric tests. By leveraging the power of AI to scrutinize and refine test items, this research contributes to the ongoing evolution of psychometric testing—making it a more precise tool for understanding human psychology and behavior.

Funding

We received financial support from the Italian PNRR MUR project with the identifier PE0000013-FAIR. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data availability statement

The data are publicly available at the link https://openpsychometrics.org/_rawdata/.

The code will be available upon request.

Declaration of competing interest

None.

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. J. Pers. Soc. Psychol.

Allen, M.J., Yen, W.M., 2001. Introduction to Measurement Theory. Waveland Press.

Allport, G.W., Odbert, H.G., 1936. Trait names: A psycholexical study. Psychol. Monogr. 47 (1), i–171. https://doi.org/10.1037/h0093360.

Arnulf, J.K., Larsen, K.R., Martinsen, Ø.L., Bong, C.H., 2014. Predicting survey responses: how and why semantics shape survey statistics on organizational behaviour. PLoS One 9 (9). https://doi.org/10.1371/journal.pone.0106361. Article e106361.

Arnulf, J.K., Larsen, K.R., Martinsen, Ø.L., Nimon, K.F., 2021. Semantic algorithms in the assessment of attitudes and personality. Front. Psychol. 12, 720559.

Arnulf, J.K., Larsen, K.R., 2020. Culture blind leadership research: how semantically determined survey data may fail to detect cultural differences. Front. Psychol. 11, 487924.

Binz, M., & Schulz, E. (2023). Turning large language models into cognitive models. arXiv preprint arXiv:2306.03917.

Cattell, R.B., 1943. The description of personality I. Foundations of trait measurement. Psychol. Rev. 50 (6), 559–594. https://doi.org/10.1037/h0057276.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Xie, X., 2023. A survey on evaluation of large language models. ACM Trans. Intell. Syst. Technol.

Clara, I.P., Cox, B.J., Enns, M.W., 2001. Confirmatory factor analysis of the depression–anxiety–stress scales in depressed and anxious patients. J. Psychopathol. Behav. Assess. 23, 61–67.

Cook, D.A., Beckman, T.J., 2006. Current concepts in validity and reliability for psychometric instruments: theory and application. Am. J. Med. 119 (2), 166–1e7.

Côté, S., Moskowitz, D.S., 1998. On the dynamic covariation between interpersonal behavior and affect: prediction from neuroticism, extraversion, and agreeableness. J. Pers. Soc. Psychol. 75 (4), 1032.

Crocker, L.M., Miller, M.D., Franks, E.A., 1989. Quantitative methods for assessing the fit between test and curriculum. Appl. Meas. Educ. 2 (2), 179–194. https://doi.org/10.1207/s15324818ame0202_6.

Cronbach, L.J., 1947. Test "reliability": its meaning and determination. Psychometrika 12 (1), 1–16.

Cunningham, W. R. (1986). Psychometric perspectives: validity and reliability.

DeVellis, R.F., Thorpe, C.T., 2021. Scale Development: Theory and Applications. Sage publications.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: pretraining of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies(Long and Short Papers Vol. 1. Association for Computational Linguistics, pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423.

Dumais, S.T., 2004. Latent semantic analysis. Annu. Rev. Inf. Sci. Technol. (ARIST) 38, 189–230.

El-Den, S., Schneider, C., Mirzaei, A., Carter, S., 2020. How to measure a latent construct: psychometric principles for the development and validation of measurement instruments. Int. J. Pharm. Pract. 28 (4), 326–336.

Evans, A.M., Rosenbusch, H., Zeelenberg, M., 2022. Using semantic similarity to understand the psychological constructs related to prosociality. Curr. Opin. Psychol. 44, 226–230. https://doi.org/10.1016/j.copsy.2021.09.019.

Furr, R.M., 2021. Psychometrics: An Introduction. SAGE Publications.

Galton, F., 1884. Measurement of character. Fortn. Rev. 36, 179–185. https://galton.org/essays/1880-1889/galton-1884-fort-rev-measurement-character.pdf.

Garcia, D., Rosenberg, P., Nima, A.A., Granjard, A., Cloninger, K.M., Sikström, S., 2020. Validation of two short personality inventories using self-descriptions in natural language and quantitative semantics test theory. Front. Psychol. 11. https://doi.org/10.3389/fpsyg.2020.00016. Article 16.

Garrett, H.E., 1937. Statistics in Psychology and Education, 2nd ed. Longmans, Green, p. 493.

Gefen, D., Larsen, K.R., 2017. Controlling for lexical closeness in survey research: a demonstration on the technology acceptance model. J. Assoc. Inf. Syst. 18 (10), 1.

Goldberg, L.R., 1992. The development of markers for the Big-five factor structure. Psychol. Assess. 4 (1), 26.

Hitsuwari, J., Okano, H., Nomura, M., 2024. Predicting attitudes toward ambiguity using natural language processing on free descriptions for open-ended question measurements. Sci. Rep. 14 (1), 8276.

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

Hughes, D. J. (2018). Psychometric validity: establishing the accuracy and appropriateness of psychometric measures. The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development, 751-779.

Hussain, Z., Binz, M., Mata, R., Wulff, D.U., 2024. A tutorial on open-source large language models for behavioral science. Behav. Res. Methods 56 (8), 8214–8237.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. D. L., & Sayed, W. E. (2023). Mistral 7B. arXiv preprint arXiv:2310.06825.

Kjell, O.N., Kjell, K., Schwartz, H.A., 2023. Beyond rating scales: with targeted evaluation, language models are poised for psychological assessment. Psychiatry Res., 115667

Kjell, O.N., Sikström, S., Kjell, K., Schwartz, H.A., 2022. Natural language analyzed with AI-based transformers predict traditional subjective well-being measures approaching the theoretical upper limits in accuracy. Sci. Rep. 12 (1), 3918.

Kline, P., 2013. Handbook of Psychological Testing. Routledge.

Liao, H.Y., Armstrong, P.I., Rounds, J., 2008. Development and initial validation of public domain Basic Interest markers. J. Vocat. Behav. 73 (1), 159–183.
Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692.
Lovibond, P.F., Lovibond, S.H., 1995. The structure of negative emotional states: comparison of the depression anxiety stress scales (DASS) with the Beck depression and anxiety inventories. Behav. Res. Ther. 33 (3), 335–343.
Martin, R.A., Puhlik-Doris, P., Larsen, G., Gray, J., Weir, K., 2003. Individual differences in uses of humor and their relation to psychological well-being: development of the humor styles questionnaire. J. Res. Pers. 37 (1), 48–75.
Martone, A., Sireci, S.G., 2009. Evaluating alignment between curriculum. Assess. Instr., Rev. Educ. Res. 79 (4), 1332–1361. https://doi.org/10.3102/0034654309341375.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 26. Curran Associates, pp. 3111–3119. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.
Milano, N., Casella, M., Esposito, R., Marocco, D., 2024. Exploring the potential of variational autoencoders for modeling nonlinear relationships in psychological data. Behav. Sci. 14 (7), 527.
Morris, J. X., Kuleshov, V., Shmatikov, V., & Rush, A. M. (2023). Text embeddings reveal (almost) as much as text. arXiv preprint arXiv:2310.06816.
Ni, J., Qu, C., Lu, J., Dai, Z., Ábrego, G. H., Ma, J., & Yang, Y. (2021a). Large dual encoders are generalizable retrievers. arXiv preprint arXiv:2112.07899.
Ni, J., Abrego, G. H., Constant, N., Ma, J., Hall, K. B., Cer, D., & Yang, Y. (2021b). Sentence-t5: scalable sentence encoders from pre-trained text-to-text models. arXiv preprint arXiv:2108.08877.
Nilsson, A.H., Schwartz, H.A., Rosenthal, R.N., McKay, J.R., Vu, H., Cho, Y.M., Ungar, L., 2024. Language-based EMA assessments help understand problematic alcohol consumption. PLoS One 19 (3), e0298300.
Nunnally, J.C., Bernstein, I., 1978. Psychometric Theory. MacGraw-Hill, New York.
Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sutskever, I., 2021. Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning. PMLR, pp. 8748–8763.
Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog, 1(8), Article 9. https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.
Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.
Rossiter, J.R., 2008. Content validity of measures of abstract constructs in management and organizational research. Br. J. Manag. 19 (4), 380–388.
Rossiter, J.R., 2002. The C-OAR-SE procedure for scale development in marketing. Int. J. Res. Mark. 19 (4), 305–335.
Schimmack, U., 2021. The validation crisis in psychology. Meta-Psychol. 5.
Smith, H. L., & Wright, W. W. (1928). Tests and measurements. Silver, Burdett. 6(3), 206–214. 10.1016/0022-4405(68)90017-4.
Sireci, S.G., 1998. The construct of content validity. Soc. Indic. Res. 45, 83–117.
Song, K., Tan, X., Qin, T., Lu, J., Liu, T.Y., 2020. Mpnet: masked and permuted pre-training for language understanding. Adv. Neural Inf. Process. Syst. 33, 16857–16867.
Spoto, A., Nucci, M., Prunetti, E., Vicovaro, M., 2023. Improving content validity evaluation of assessment instruments through formal content validity analysis. Psychol. Methods.
Stevens, J., 2002. Applied Multivariate Statistics for the Social Sciences, 4. Lawrence Erlbaum Associates, Mahwah, NJ.
Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9 (11).
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.
Wulff, D. U., & Mata, R. (2023). Automated jingle–jangle detection: using embeddings to tackle taxonomic incommensurability.
Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V., 2019. Xlnet: generalized autoregressive pretraining for language understanding. Adv. Neural Inf. Process. Syst. 32, 5753–5763. https://dl.acm.org/doi/pdf/10.5555/3454287.3454804.
Zhelezniak, V., Savkov, A., Shen, A., & Hammerla, N. Y. (2019). Correlation coefficients and semantic textual similarity. arXiv preprint arXiv:1905.07790.

Further readings

Kjell, O.N., Kjell, K., Garcia, D., Sikström, S., 2019. Semantic measures: using natural language processing to measure, differentiate, and describe psychological constructs. Psychol. Methods 24 (1), 92.
Reimers, N., Gurevych, I., 2019. Sentence-BERT: sentence embeddings using siamese BERT-networks. In: Inui, K., Jiang, J., Ng, V., Wan, X. (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, pp. 3982–3992. https://doi.org/10.18653/v1/D19-1410.