Preprints.org

Article

Not peer-reviewed version

Embedding-based Semantic Analysis Approach: A Preliminary Study on Redundancy Detection in Psychological Concepts Operationalized by Scales

Zhen Huang , Yitian Long , Kaiping Peng , Song Tong *

Posted Date: 25 October 2024

doi: 10.20944/preprints202410.2001.v1

Keywords: Redundancy detection; psychological scales; GPT; hierarchical clustering

Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 October 2024 doi:10.20944/preprints202410.2001.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Embedding-Based Semantic Analysis Approach: A Preliminary Study on Redundancy Detection in Psychological Concepts Operationalized by Scales

Zhen Huang [1], Yitian Long [2], Kaiping Peng [1] and Song Tong [1,*]

[1] Tsinghua University
[2] Wuhan Britain-China School
* Correspondence: tong.song.53w@kyoto-u.jp

Abstract: As psychology evolves, the phenomenon of concept overlap becomes more pronounced, increasing participant burden and complicate data interpretation. This study introduces an Embedding-based Semantic Analysis Approach (ESAA) for detecting redundancy in psychological concepts, which are operationalized through their respective scales, using natural language processing techniques. ESAA utilizes OpenAI's GPT-3 large model to generate high-dimensional semantic embeddings of scale items and applies hierarchical clustering to group semantically similar items, uncovering potential redundancy. In three preliminary experiments, ESAA was tested on well-known psychological scales, such as Conscientiousness, Gratitude, and Grit. The experiments assessed ESAA's ability to (1) converge semantically similar items, (2) discriminate semantically distinct items, and (3) identify overlapping scales measuring concepts known for redundancy. Additionally, comparative analyses were conducted to assess ESA**A'**s robustness and incremental validity against the most advanced chat bots based on GPT-4. The results demonstrated that ESAA consistently produced stable outcomes and surpassed all evaluated chatbots in performance. As a novel, objective approach for analyzing relationships between concepts operationalized as scales, ESAA has potential to facilitate future research on theory refinement and scale optimization.

Keywords: Redundancy detection; psychological scales; GPT; hierarchical clustering

I. Introduction

As psychological research evolves, an increasing number of concepts and scales have been introduced, yet limited attention has been given to the potential overlaps among them. These overlaps may increases participant burden (Deigan, 2024), potentially compromising data quality, and complicating the interpretation of research findings (Elson et al., 2023; Condon & Revelle, 2015; Condon et al., 2017). More critically, redundancy hinders the field from advancing toward theoretical parsimony and precise measurement. Consequently, addressing redundancy in psychological concepts and their associated scales has become an urgent priority (Flake & Fried, 2020; Bainbridge et al., 2022; Sharp et al., 2023).

Traditionally, redundancy in psychological scales has been identified through methods such as factor analysis, correlational studies, and expert judgment (Fabrigar et al., 1999; Clark & Watson, 1995; DeVellis, 2021). While these methods offer valuable insights, they also present limitations. Factor analysis and correlational studies depend on empirical data, which can be costly and time-consuming to gather, with the added risk of survey errors (Costello & Osborne, 2005; Schwarz, 1999; Podsakoff et al., 2003). Expert judgment although insightful analysis, is inherently subjective, making it difficult to guarantee the reproducibility of research findings (Armstrong, 2001; MacCoun, 1998; Tetlock, 2005).

© 2024 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 October 2024 doi:10.20944/preprints202410.2001.v1

In recent years, advancements in natural language processing (NLP) and machine learning have created new opportunities for text data analysis. Notably, transformers-based language models, in particular, have shown remarkable capabilities in understanding and representing semantic content (Vaswani et al., 2017; Devlin et al., 2019; Brown et al., 2020). However, the use of these technologies in psychological scale analysis has been limited. Some studies have explored the use of BERT (Bidirectional Encoder Representations from Transformers) for scale analysis (Wulff & Mata, 2024; Hommel & Arslan, 2024), but the results have been underwhelming. This may be due to BERT being an earlier generation of large language models, with semantic parsing abilities that are not yet sufficiently robust. Although these studies fine-tuned BERT to address its limitations the outcomes remained suboptimal, such as its inability to handle reverse-scored items effectively (Wulff, & Mata, 2024; Hommel, & Arslan, 2024). Additionally, fine-tuning BERT is both costly and complex, for psychologists, fine-tuning a BERT model is complex and labor-intensive, involving the creation of domain-specific databases and engineering work on training model parameters. And the resulting models of fine-tuned BERT are static, lacking the ability to automatically update or improve.

This study represents a preliminary attempt to utilize recently advanced AI technologies to address redundancy in psychological research. Specifically, we propose a new approach, whose key advantage over traditional psychology research methods is that it requires only the scale content, eliminating the need for human participants. The approach we proposed is named as the Embedding-based Semantic Analysis Approach, or ESAA for short. ESAA takes the text content of psychological scale items as input and uncovers latent semantic structures and relationships among the items as its output. ESAA accomplishes this through a multi-step process. It begins by transforming the scale items' text into computable high-dimensional vectors, known as embeddings, using advanced large language models (LLMs). And then, ESAA applies unsupervised clustering algorithms and other computations to perform semantic analysis on these embeddings.

To explore ESAA's feasibility for aiding the redundancy detection research, three initial experiments were conducted. In these experiments, the performance of ESAA successfully met the three criteria: converging semantically similar items, discriminating items with significant semantic differences, and clustering items to reveal interconnected patterns among different concepts that literature suggests may overlap. Additionally, the output of ESAA demonstrated potentially better performance compared to that of ChatGPT-4, which was used as a baseline. In summary, this research introduces the ESAA as a technical method and provides preliminary validation of its feasibility, suggesting that ESAA may hold promise as a tool for future psychological research focused on redundancy reduction and theory refinement.

The significance of this study lies in ESAA's potential to contribute to the evolution of psychological research by offering an objective, efficient, and low-cost method for detecting redundancy among constructs: ESAA may streamline the research process and enhance the reliability of findings by minimizing subjective biases present in traditional methods. This innovative approach could lead to more precise measurements and clearer theoretical frameworks, fostering the advancement of psychological science. Furthermore, the ability to analyze scale items without participant data collection might open new avenues for researchers, allowing the exploration of previously overlooked constructs.

II. The Proposed Approach: ESAA

The ESAA is designed to help psychologists investigate conceptual redundancy by providing insights into potential overlaps among scales under study. Specifically, ESAA analyzes the semantic relationships between scale items, offering output that can assist psychologists in formulating research hypotheses about redundancy. ESAA follows a comprehensive three step process, as shown in Figure 1. First, it uses GPT-3 large model to generate semantic embeddings for the scale items, transforming them into high-dimensional vectors that represent their semantic meanings. Next, by analyzing the semantic distances between these embeddings, ESAA introduces a novel metric called synthetic correlation, which replaces empirical correlations between items. Finally, unsupervised clustering techniques are applied to reveal latent structures within the items, providing insights into

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 October 2024 doi:10.20944/preprints202410.2001.v1

potential redundancies among in psychological concepts measured by scales. The following subsections detail the key technical aspects of each step.

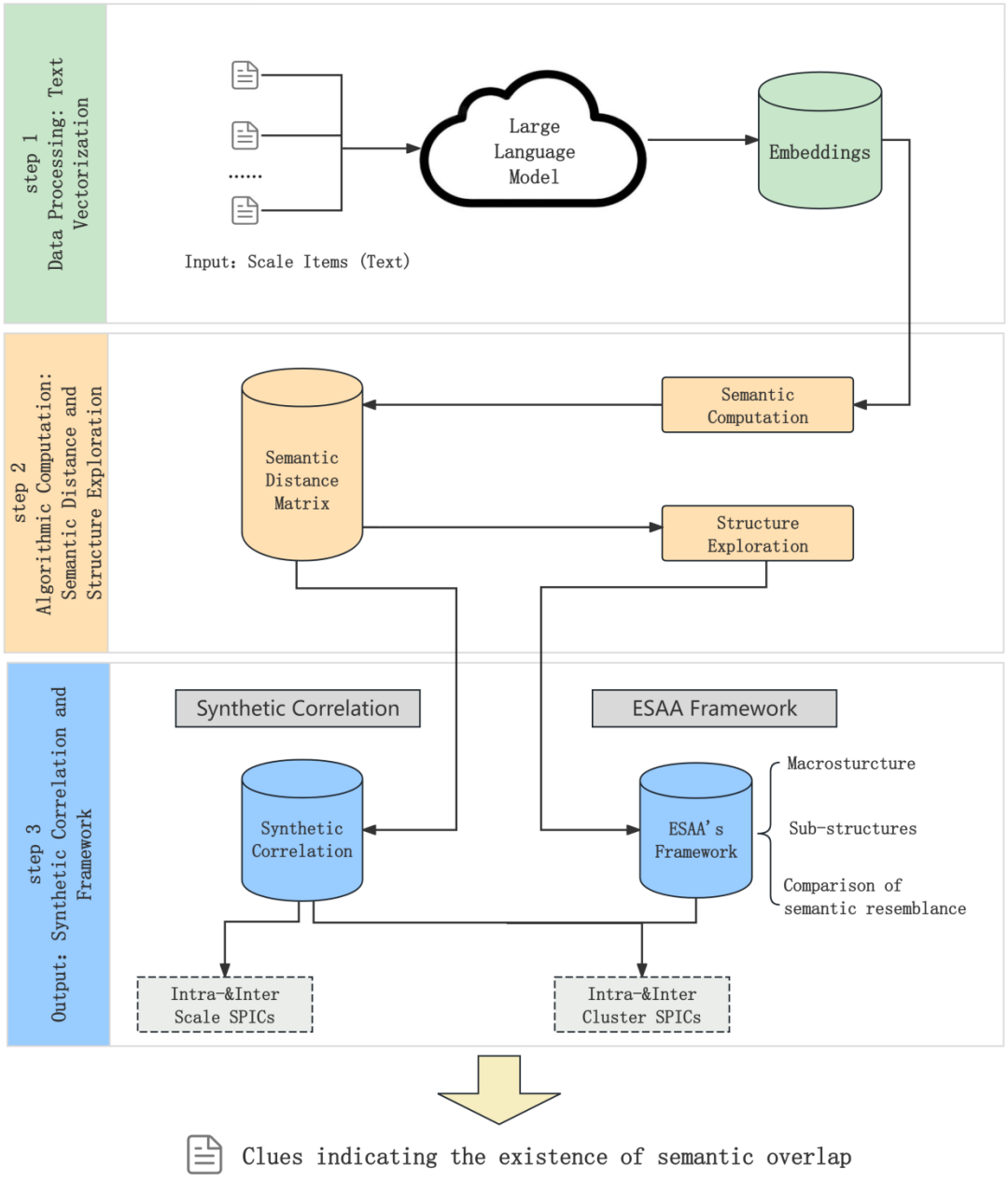


Figure 1. Workflow of the Embedding-based Semantic Analysis Approach (ESAA) for Redundancy Detection in Psychological Scales.

Step 1. Data Processing: Text Vectorization

One of ESAA's key advantage is its ability to convert the semantics of text-based scale items into computable formats. This is achieved by transforming the text into numerical representations known as embeddings. In the context of large language models (LLMs), embeddings are high-dimensional vector representations that capture the semantic meaning of the text, enabling comparison across different texts.

Embeddings are generated by LLM, which rely on complex neural network architectures. Selecting an appropriate LLM is crucial. By September 2024, the final design phase of ESAA, a wide range of LLMs are available, either as open-source resources or through API-based services, with varying capabilities in representing the semantics of textual material. Among them, GPT-3-large

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 October 2024 doi:10.20944/preprints202410.2001.v1

model stands out as the most powerful, offering embeddings with 3,072 dimensions. The GPT-3-small model, also a from OpenAI, produces embeddings with 1,536 dimensions. Another option, BERT model, a widely-used model has 768-dimensions and notable for its extensive use in academic research (Khadhraoui et al., 2022). Theoretically, LLMs with higher-dimensional embeddings are expected to capture more nuanced semantic details. Therefore, the GPT-3 large model is anticipated to be the most effective choice for ESAA. While we expect this model to provide superior performance, less advanced LLMs could still meet the basic requirements for ESAA. To verify these assumptions, we conducted experiments, as reported in the following section.

Step 2. Algorithmic Computation: Semantic Distance and Structure Exploration

Semantic Distance

Semantic distance is an important metric in this study, reflecting the relationships between embeddings. A smaller semantic distance indicates a closer approximation of the semantics of the embeddings. Several methods are available for calculating semantic distances between high-dimensional vectors, including Euclidean distance, Cosine distance, and Jaccard distance, among others. Each distance metric is best suited to different application scenarios. Theoretically, cosine similarity is considered the optimal choice for text-based research due to its ability to measure the angle between vectors, which is crucial for understanding semantic relationships. Therefore, in this study we define

$$\text{Semantic Distance} = 1 - \text{cosine similarity} \quad (1)$$

Structure Exploration

A core function of the ESAA is to suggest a framework for categorizing all material items, thereby enabling researchers to gain insights into the underlying data structure. This functionality is achieved through the use of an algorithms, hierarchical clustering.

Hierarchical clustering is an unsupervised clustering method particularly suitable for exploring hierarchical relationships among embedding. Unlike other clustering methods, it does not require pre-specifying the number of clusters, instead building a hierarchical tree (dendrogram) that reveals the multi-level nested structure of data. This study employs a bottom-up (agglomeration) approach, beginning with each data point as its own cluster and progressively merging similar points. This method's advantage lies in its ability to comprehensively display multi-level relationships among data points, making it especially apt for examining the semantic relationships among psychological scale items.

In hierarchical clustering calculations, various methods are available, including the Ward method, average method, and complete method. The Ward method notably excels in minimizing total variance within clusters, resulting in compact, well-defined clusters—a critical aspect for identifying similar items in psychological constructs. Given these merits, we hypothesized that the Ward method would be the most suitable for the ESAA. Our preliminary experiments confirmed this prediction, demonstrating its superior efficacy. Thus, we have chosen the Ward method for ESAA. Further detailed analyses, including outcome metrics and comprehensive methodological frameworks, will substantiate these findings.

Step 3: Output: Synthetic Correlation and Framework

1. Synthetic Correlation

As a methodological tool for redundancy detection research in psychology, ESAA is designed to report on the semantic relationships between items. This functionality is achieved through the computation of synthetic correlation and semantic distance. Synthetic correlation, a novel metric introduced in this study, serves as an alternative to traditional correlations derived from empirical data collected from human participants. It is defined as the cosine similarity between the embeddings. Specifically, Synthetic pairwise item correlation (SPIC) refers to the cosine similarity between a pair of two items.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 October 2024 doi:10.20944/preprints202410.2001.v1

Assume that the embeddings for each item *i* and *j are* vectors v_i and v_j, respectively. The formula for calculating the Synthetic Pairwise Item Correlation (SPIC) is:

$$SPIC(i, j) = \cos(\theta) = \frac{v_i \cdot v_j}{\|v_i\|\|v_j\|} \quad (2)$$

where $v_i \cdot v_j$ represents the dot product of v_i and v_j. $\|v_i\|$ and $\|v_j\|$ are the Euclidean norms (or magnitudes) of v_i and v_j, respectively.

Intra-group (Cluster/Scale) SPIC refers to the synthetic pairwise item correlations calculated between items within the same group, which would be cluster or scale in this study. These correlations are expected to be higher, reflecting the semantic consistency and coherence of items that are theoretically or empirically grouped together.

Inter-group (Cluster/Scale) SPIC refers to the synthetic pairwise item correlations calculated between items from different groups, which would be cluster or scale in this study. These correlations are typically lower, indicating the semantic distinction and divergence between items that belong to different concepts or theoretical categories.

2. *Framework: solution of the categorizing*

The major output of ESAA is the so-called "ESAA-generated Framework (EGF)" in this study. The term "framework" refers to any classification scheme applied to a given set of items. Theoretically, there can be many frameworks for the same set of items. One particular framework is the outcome of ESAA, i.e., EGF. On the other hand, a classification scheme where each category corresponds directly to the originating scale of its items is also considered a framework. We refer to this as the "Scale Origin Framework" (SOF). The comparison between EGF and SOF will also be part of ESAA's output, as it can reflect potential overlaps between the scales of the original materials.

III. Validation of the New Approach

While extensive validation through comprehensive studies is typically necessary for a novel approach, such validation is beyond the scope of this initial study. Instead, we focus on the foundational criteria required for an effective method in detecting redundancy in psychological constructs. Specifically, the approach should demonstrate three key capabilities: (1) converging items with high correlations, (2) discriminating between items with low correlations, and (3) identifying semantic overlap across different scales.

To assess these capabilities, we conducted three experiments. The first experiment tested the approach's ability to converge highly correlated items and discriminate low-correlated items. The second experiment evaluated its capacity to identify semantic overlap across different scales. Finally, the third experiment served as a robustness check and offered a series of comparative analyses of the overall competence of approach.

Experiment 1: Convergence and Discriminant Validity

The experiment aims to evaluate ESAA's ability to converge highly correlated items and discriminate low-correlated items. If ESAA has reliable convergence ability, then EGF should categorize those items measuring a same underlying conceptual construct together, with intra-cluster SPIC at least as high as the corresponding intra-scale SPIC. In contrast, if the EGF has sufficient discriminant ability, it should allocate items from scales measuring conceptually distinct constructs into different clusters, with inter-cluster SPIC much lower than corresponding intra-cluster SPIC, and not higher than the corresponding inter-scale SPIC.

Methods

Materials

To test ESAA's convergence and discrimination abilities, the experimental materials must meet certain criteria: (1) each scale should be widely recognized for high internal consistency, (2) the concepts should exhibit low correlation, as supported by prior studies, and (3) both scales should have an equal number of items.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 October 2024 doi:10.20944/preprints202410.2001.v1

In line with these criteria, the Conscientiousness and Gratitude scales were chosen:

Conscientiousness Scale: Derived from the NovoPsych Five Factor Personality Scale -30 (Buchanan & Hegarty, 2023), this 6-item scale measures the well-established five factor model of personality (a.k.a. OCEAN). A sample item is "I often do just enough work to get by" (reverse scored).

Gratitude Scale: The Gratitude Questionnaire-Six Item Form (GQ-6: McCullough et al., 2002) includes 6 items, with a sample item being "I have so much in life to be thankful for".

Both scales are well-validated with high internal consistency and the two concepts exhibit minimal to no correlation (Ajmal et al., 2016; Kong et al., 2020), making them ideal for examining convergence and discrimination in psychological research.

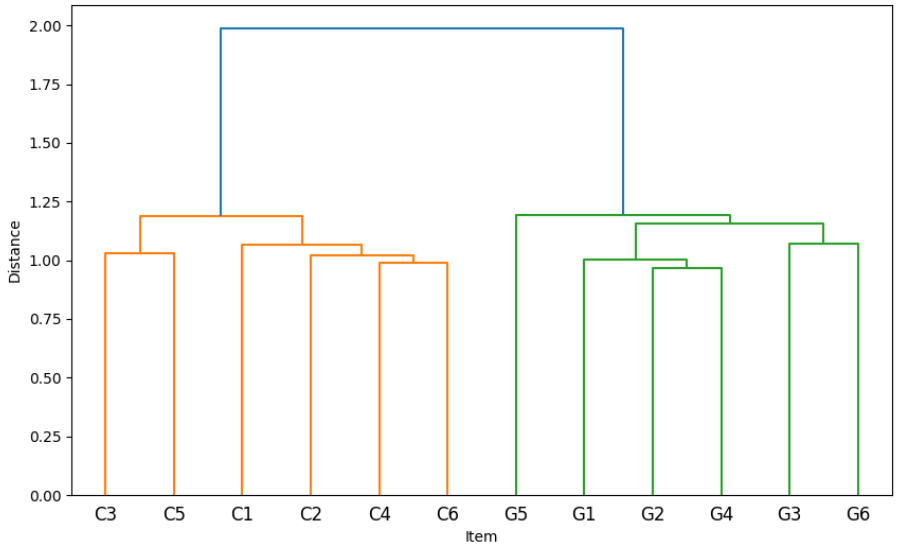
Hypotheses

Given the nature of these scales, the items measuring Conscientiousness and Gratitude should map distinctly into two separate regions, in the semantic space of embeddings, with the distance between these regions significantly greater than within-region distances.

Hypothesis 1. *The macro-structure of the EGF will match the SOF; and in the EGF, the inter-cluster SPIC will be significantly lower than the intra-cluster SPIC, with a large effect size.*

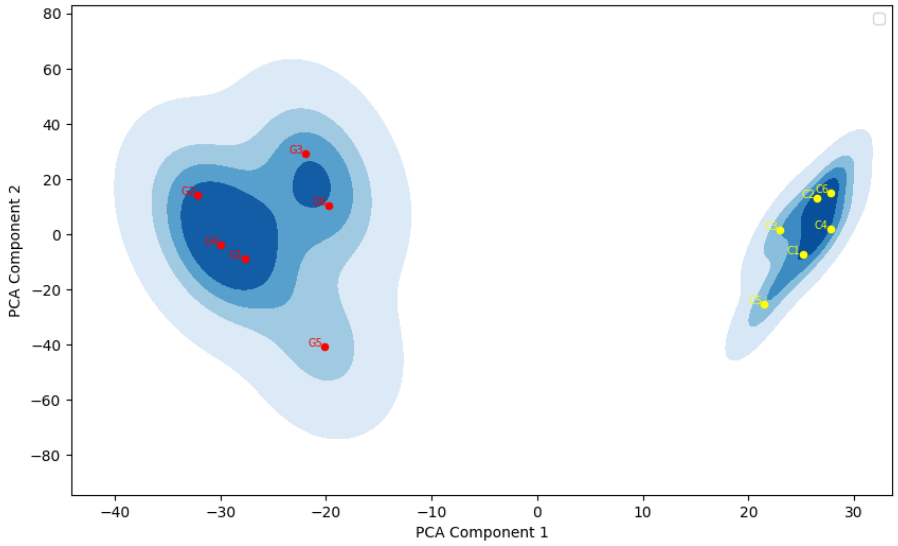
Results and Discussion

The hierarchical clustering analysis conducted for these 12 items reveals two distinct clusters, each corresponding precisely to one of the original scales. Embeddings from the same scale are closely grouped together, while embeddings from different scales are clearly separated, as shown in Figure 2a,b. Figure 2a displays the dendrogram of the EGF, which shows the hierarchical clustering of the item embeddings from the Conscientiousness and Gratitude scales, using WARD method. On the y-axis, the dendrogram indicates the semantic distance between clusters, while the x-axis lists the individual embeddings of the items being clustered. The clusters are color-coded for clarity, and item indices are labeled as "C" for Conscientiousness scale items and "G" for Gratitude scale items. Figure 2b presents a Kernel Density Estimate (KDE) plot of the dimension-reduced embeddings. To generate Figure 2b, we followed a three-step process: (1) semantic distance computation, (2) reducing the embedding dimensions to two using PCA, and (3) applying a KDE to the PCA dimensional reduction results, with the color-label for items follows EGF, i.e., the results of previous clustering. This plot clearly highlights the separation and cohesion of clusters of the EGF.



(**a**)

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 October 2024 doi:10.20944/preprints202410.2001.v1



(b)

Figure 2. (a) Dendrogram of EGF for Experiment 1; **(b)** KDE Analysis of Dimension-Reduced Embeddings for Experiment 1.

As mentioned in the section II of this study, intra-cluster SPIC values represent the semantic coherence within each cluster, while inter-cluster SPIC values reflect the divergence between different clusters. The average SPIC values for the intra-cluster pairs within the Gratitude Scale and Conscientiousness Scale (M = 0.426, SD=0.0039) were significantly higher than those for inter-cluster pairs (M = 0.193, SD=0.0022), demonstrating that item pairs within the same cluster are closer to each other than those between different clusters. To examine whether such difference between intra-cluster and inter-cluster SIC values is significant, we conducted statistic test. Welch's t-test is adopted due to unequal variances and sample sizes. The results revealed a significant difference, t (43.35) = 18.46, p < .001. The effect size, measured by Cohen's d, was 4.50, indicating a very large effect. These findings suggest that intra-cluster similarities are significantly higher than inter-cluster similarities.

These results support H1, demonstrating that the ESSA effectively distinguishes between concepts and converges items within the same concept. Therefore, we can confidently reject the null hypothesis, which suggest that ESAA lacks convergence or discriminant abilities.

However, it is important to note that these conclusions are limited to the specific materials used in this experiment. Similar to a Turing test, the experiment 1 was designed to explore the capabilities of a new technique. While the results from this single experiment are promising, they only serve as a necessary condition for demonstrating ESAA's effectiveness. Extensive empirical validation through future research is required to fully confirm the approach's generalizability and robustness.

Experiment 2: Overlap-Detection Competence Validation

The aim of this experiment is to assess the effectiveness of the ESAA in detecting semantic proximity between scale items that measure psychological concepts with potential redundancy. These concepts may be so similar that they lack incremental validity—meaning they do not contribute unique information beyond what is already captured by other concepts. As ESAA is designed to help researchers identify redundancy in psychological concepts, demonstrating this overlap-detection capability is crucial for validating its utility in future studies.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 October 2024 doi:10.20944/preprints202410.2001.v1

Methods

Materials

To assess the ESAA's capacity for detecting redundancy, it is important to select psychological concepts that have been explicitly recognized in the literature as containing overlapping elements. This ensures a clear benchmark for evaluating the reliability of ESAA's results.

The concepts of Grit and Conscientiousness were chosen as they meet this criterion. Grit was doubted for its redundancy with Conscientiousness by many research. A meta-analysis of Grit research (Credé, Tynan, & Harms, 2017), which analyzed 584 effect sizes across 88 independent samples (totaling 66,807 individuals), revealed that Grit, initially proposed as a higher-order trait predicting of success and performance, shows an excessively strong correlation (ρ = .84) with Conscientiousness, raising questions about challenging its distinctiveness as a construct.

This experiment utilized the following scales:

***Conscientiousness Scale*:** Same as that in the Experiment 1.

Grit Scale: derived from Duckworth et al. (2007), this scale contains 12 items divided into two facets: Perseverance of engagement and Consistency of interest, with 6 items per facet. A sample item is "I have achieved a goal that took years of work".

Given the materials, the SOF in Experiment 2 can be visualized as shown in Figure 3, with grit and conscientiousness are treated as distinct categories, with Grit's two facets as sub-categories.

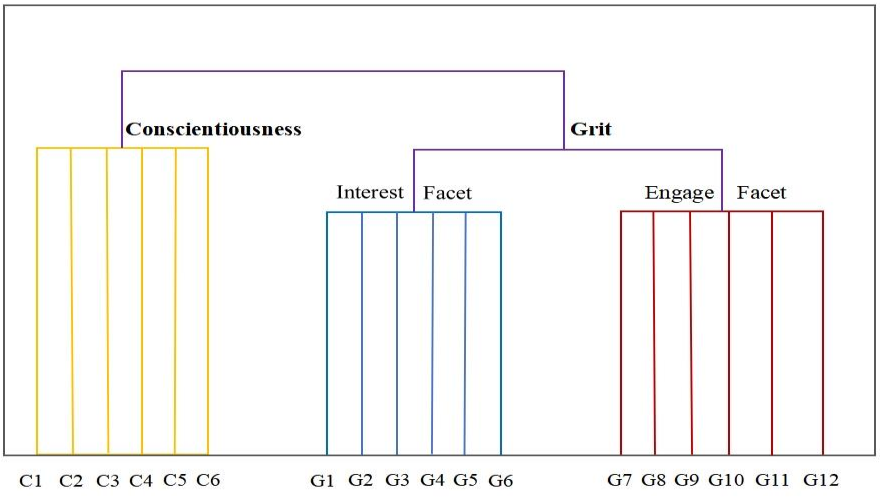


Figure 3. Structural Frameworks of Embeddings in SOF for Experiment 2.

Hypotheses

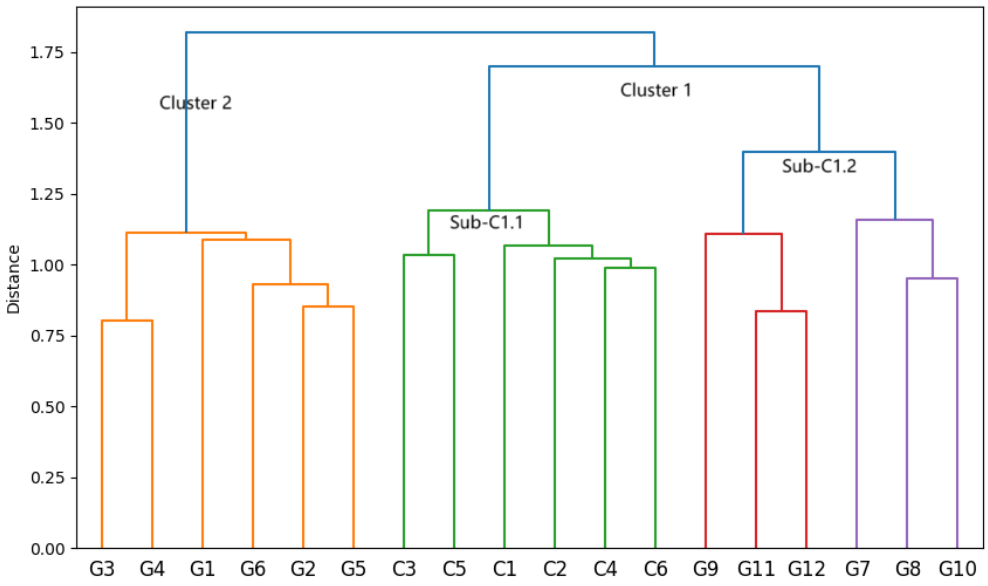
If the ESAA has qualified capability to identify redundancy, the EGF is expected not to mirror the SOF, and should exhibit signs semantic blending between Conscientiousness and grit items. This is because the EGF, as a deliberately designed tool, is expected to achieve optimized semantic structure, while the SOF is doubted to be lacking of discrimination. According to extensive literature, SOF would not have satisfied discriminate validity. As Credé, Tynan, and Harms (2017) noted, the perseverance facet of grit and conscientiousness was reported to have a correlation of ϱ = .89, far exceeds the typical correlation found between two different global measures of conscientiousness (ϱ = .63, according to Pace & Brannick, 2010). Thus, the meta-analysis suggested that "grit is not a higher-order construct characterized by two lower-order facets" and "may be redundant with conscientiousness". This leads to the hypothesis for the current experiment:

Hypothesis 2. *The structure of the EGF will not be identical to the SOF, with items from Grit and Conscientiousness interfused, rather than separated.*

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 October 2024 doi:10.20944/preprints202410.2001.v1

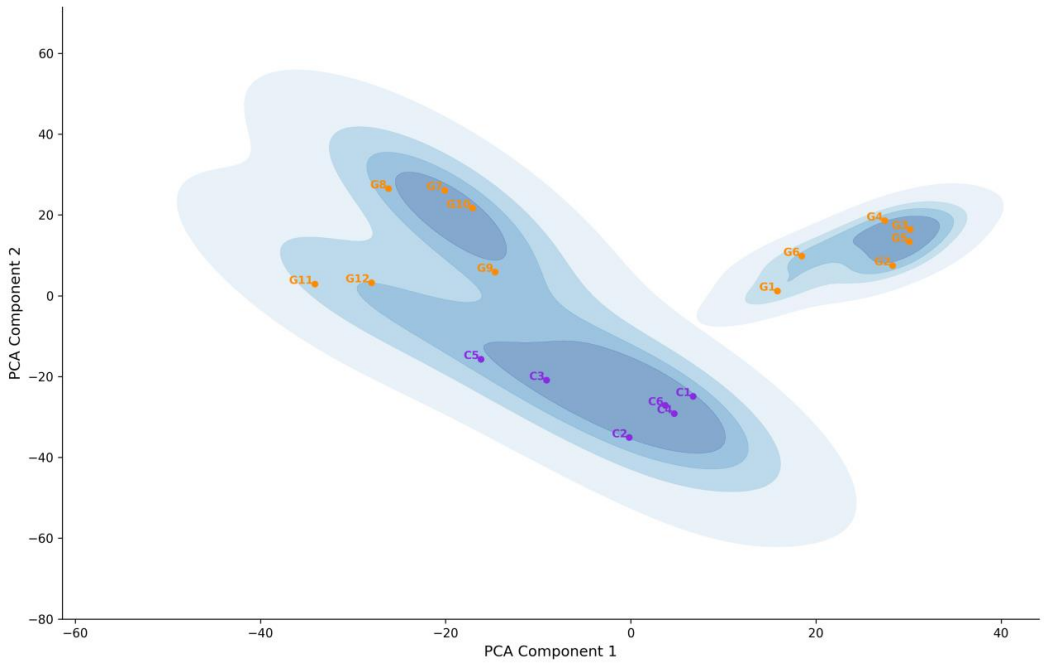
Results and Discussion

The application of the ESAA on the items from the Grit and Conscientiousness scales resulted in two major clusters, as revealed in the hierarchical structure depicted by Figure 4a,b. The full merging process is visualized in the dendrogram in Figure 4a, which illustrates the hierarchical clustering of item embeddings from the two scales based on WARD method. The y-axis represents the dissimilarity between clusters, while the x-axis shows individual items, color-coded for clarity. Items from the Conscientiousness scale are labeled "C", while those from the Grit scale are labeled "G". "Cluster 1" in Figure 4 notably includes a mix of items from the Grit scale along with all of the Conscientiousness items. This suggests that, in terms of semantic proximity, some Grit items are closer to Conscientiousness, leading the algorithm to group them together. Specifically, the part of grit items involving the interfusion are exactly those measuring the so-called Perseverance of Engage facet, which evidence is perfectly consistent with the conclusion of the meta-analysis by Credé, Tynan, and Harms (2017). Figure 4b provides further evidence of semantic overlap, displaying a KDE plot based on the Dimension-Reduction of the embeddings. This figure highlights the interweaving of Grit and Conscientiousness items, confirming the semantic fusion between them. The generation of Figure 4b followed the same three-step process as in Experiment 1, including semantic distance computation, dimensionality reduction, and KDE visualization. These results clearly support Hypothesis 2, which predicted inter-fusion between items from the Grit and Conscientiousness scales. In sum, the result of ESAA, i.e., the EGF of Experiment 2, supports H2, being highly consistent with the literature suggesting the redundancy of the grit concept.



(a)

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 October 2024 doi:10.20944/preprints202410.2001.v1



(b)

Figure 4. (a)Structural Frameworks of Embeddings in EGF for Experiment 2; (b) KDE Analysis of Dimension-Reduced Embeddings in EGF for Experiment 2.

Moreover, EGF performs well in terms of convergence and discriminate validity, as detailed statistics shown in Table 1. The intra-cluster SPIC of EGF *(M = 0.372, SD = 0.132, n = 81)* was significantly higher than the inter-cluster SPIC *(M = 0.294, SD = 0.071, n = 72)*. Welch's T-Tests were conducted instead of a traditional t-test due to the significant heterogeneity of variance observed in the data, $t\ (126) = 4.572$, $p < 0.001$. The large effect size (Cohen's d = -0.717) indicate that the pairwise semantic similarity of items within and between cluster in EGF differ significantly. Besides, the difference between intra-scale and inter-scale SPICs for SOF also reached significance, but the all the statistic values indicating performance of framework are inferior to those of EGF.

Table 1. SPICs Comparison between EGF and SOF.

Frame work	SPIC	Mean	SD	n	t(df)	*p*-value	Cohen's d
EGF	Intra-cluster	0.372	0.132	81	--	--	--
	Inter-cluster	0.294	0.071	72	--	--	--
	Difference (1)	0.077			4.572(126)	0.000	0.717
SOF	Intra-scale	0.370	0.130	81	--	--	--
	Inter-scale	0.297	0.078	72	--	--	--
	Difference(2)	0.073			4.284(134)	0.0000	0.675

In summary, the results of Experiment 2 demonstrate that the ESAA successfully identified semantic interfusion between the Grit and Conscientiousness scales, as evidenced by the hierarchical structure and Reduced Dimensional KDE plots. Hypothesis 2, which predicted this inter-fusion, was supported by the formation of the mixed-source cluster, highlighting the overlapping semantic nature of the two concepts. Furthermore, statistics of SPICs demonstrate the validity of this EGF. Overall, the results in Experiment 2 indicate that ESAA has potential to serve as a reliable tool for detecting redundancy among psychological constructs, addressing the study's objective of validating ESAA's overlap-detection competence.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 October 2024 doi:10.20944/preprints202410.2001.v1

Experiment 3. Robust Check and Comparative Analysis

The results from the previous two experiments implies ESAA as a potential tool for redundancy research on psychological concepts. However, before drawing a validating conclusion, two arguments remain: 1) Are the results from the first two experiments robust? Experiments 1 and 2 selected corresponding experimental materials independently. However, if the experimental materials were input in different way, will the ESAA's calculations remain stable? This is the "robust" argument. 2) Does ESAA provide added value compared to baseline tools? A sample and intuitive baseline is the use of chatbots based on LLMs, such as ChatGPT. If similarly effective analytical results can be generated through simple prompts, there may be no substantial benefit in developing a new tool. Therefore, it becomes crucial to compare the performance of the ESAA with existing chatbot. The comparison will help determine whether the ESAA offers genuine innovations and improvements. This is the "incremental value" argument.

To address these questions, we designed a series of trials in this experiment applying the ESAA and alternative approaches on same input material, and comparing the performance of their outputs, namely the EGF and other frameworks. The hypotheses are that ESAA has robustness, and having incremental value, which means the outputs of ESAA are consistent with each other and the performance of EGF beat any other frameworks in terms of convergence, discriminate, and redundancy detection performance.

Material and Procedure

The materials for this experiment are a combination of those used in the previous two experiments, incorporating three scales measuring Conscientiousness, Grit, and Gratitude (24 items in total). The procedure involved generating the frameworks, followed by three series of trials comparing the EGF with the frameworks generated by alternative approaches.

The robustness of ESAA was evaluated by comparing the EGF in from Experiments 3 from those from Experiments 1 and 2. We expected the EGF of the current experiment to align with the previous ones, confirming the stability of ESAA's output. Secondly, to assess ESAA's incremental value, we generated two baseline frameworks using the most advanced LLM-based chatbots (GPT-4.0) in present, i.e., ChatGPT 4o and o1, anticipating that the chatbot-generated frameworks would be inferior to EGF in terms of the convergence, discrimination, and interpretability.

Results and Discussion

The clustering results in Figure 5a,b clearly reveal the hierarchical structure of the EGF for Experiment 3. Figure 5a presents the dendrogram of the ESAA analysis, which illustrates the hierarchical clustering of item embeddings from the Conscientiousness and Gratitude scales, based on the WARD method. The y-axis shows the semantic distance between clusters, while the x-axis lists individual item embeddings, color-coded by clusters. Items from the Conscientiousness scale are labeled "C," Gratitude scale items are labeled "G," and the two facets of Grit—Consistency of Interest and Perseverance of Effort—are labeled with "I" and "E," respectively. When the clustering threshold was set at 0.7 for the mean of distance between clusters at merge, the items were divided into four distinct groups, corresponding to the original subscales. However, a closer look at the dendrogram reveals some semantic blending, particularly between Grit and Conscientiousness items. For example, the Perseverance of Effort facet of Grit initially clusters with Conscientiousness items before later merging with the Consistency of Interest facet, indicating a notable overlap between these two constructs. This merged group is then combined with the Gratitude items at a more distant clustering level.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 October 2024 doi:10.20944/preprints202410.2001.v1

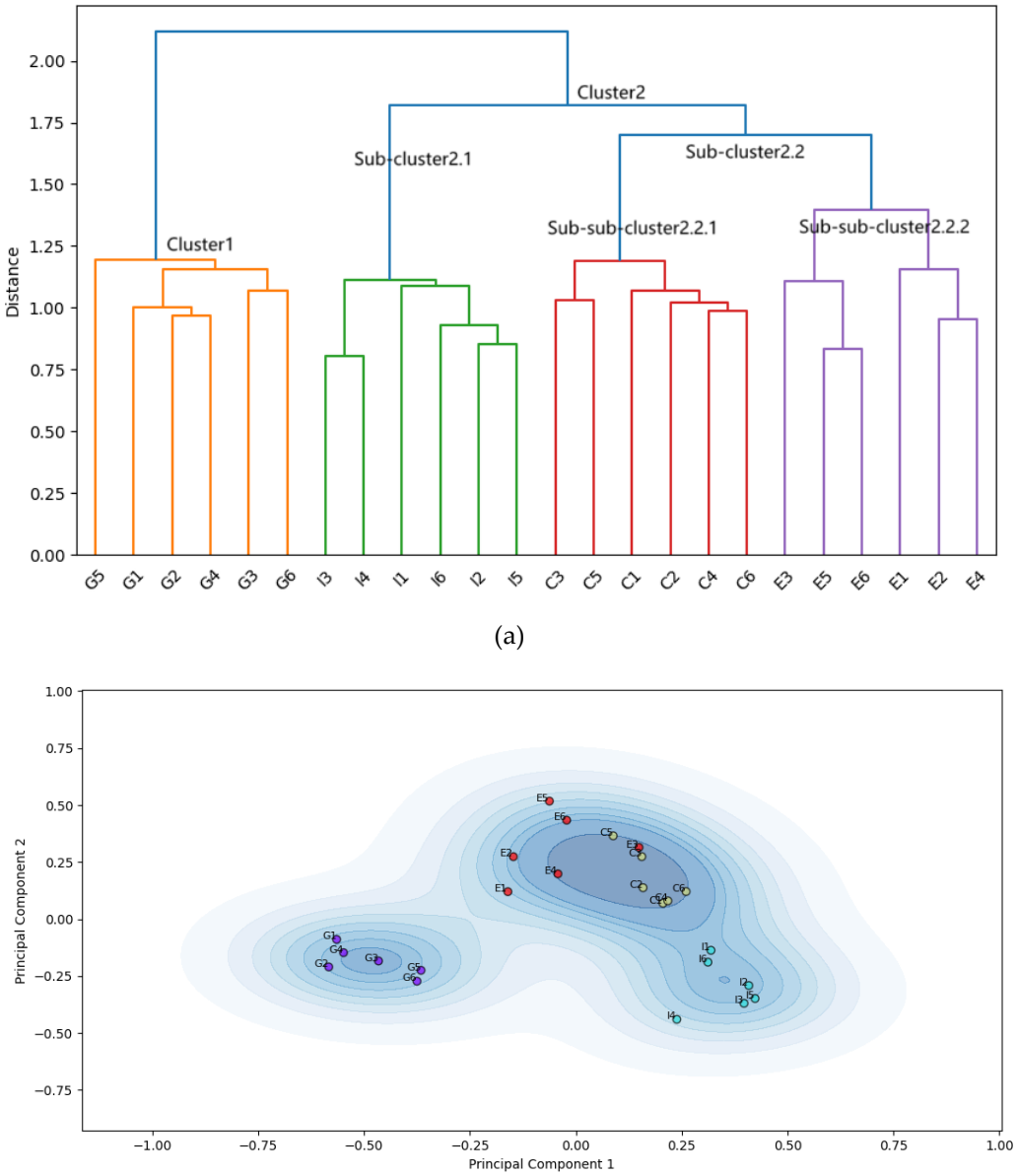


Figure 5. (a) Dendrogram of EGF for Experiment 3, and (b) KDE Analysis of PCA-Reduced Embeddings in EGF for Experiment 3.

The step-by-step merging process is further visualized in Figures 6a, 6b, and 6c, which depict the distribution of data points after reducing the high-dimensional embedding space to two dimensions using PCA. In Figure 6a, the data points are divided into four classes, matching the original subscales. As seen in Figure 6b, when the clustering is reduced to three classes, the Conscientiousness items merge with the Perseverance of Effort facet from the Grit scale, reflecting their semantic proximity. Finally, Figure 6c illustrates the division into two major clusters, where the combined Grit and Conscientiousness items form a single category that merges with Gratitude at a higher level of abstraction. This merging pattern highlights the structural relationships between the concepts, suggesting significant overlap between Grit and Conscientiousness, while the separation from Gratitude demonstrates ESAA's ability to discriminate between more distinct concepts. This result showcases ESAA's effectiveness in capturing both convergence and distinction across psychological scales, particularly in identifying concepts with overlapping semantic features.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 October 2024 doi:10.20944/preprints202410.2001.v1

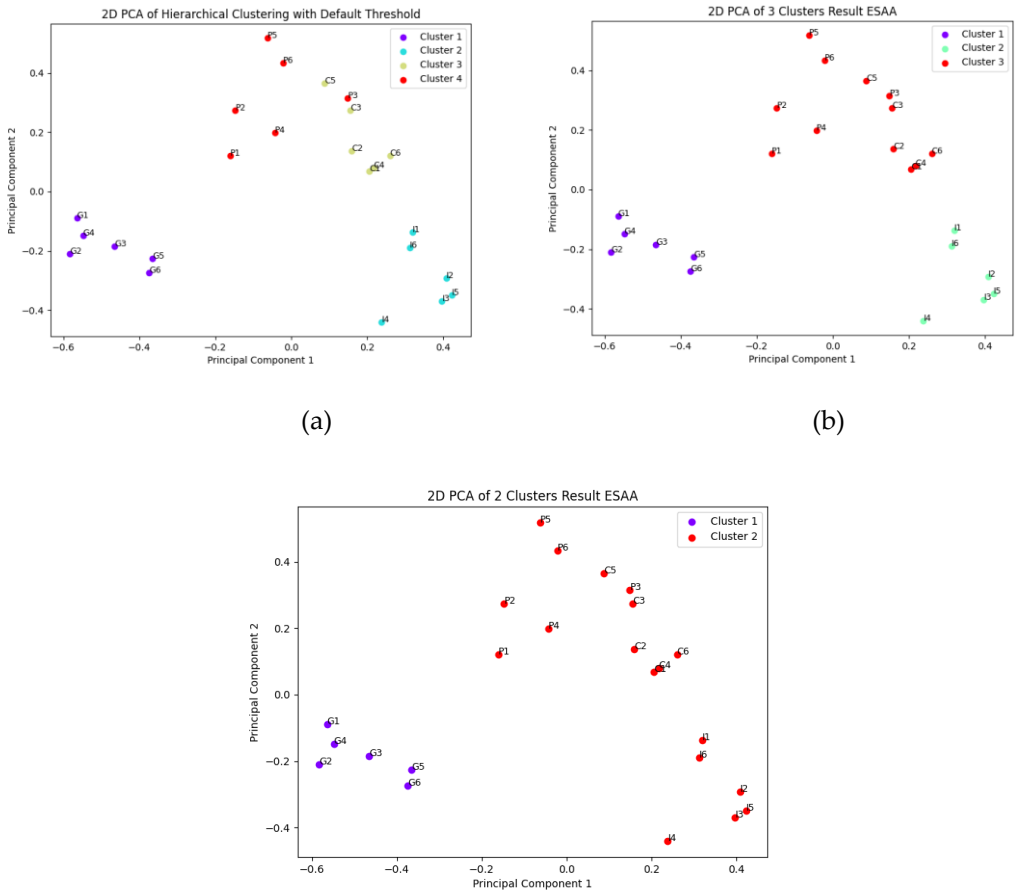


Figure 6. Stepwise Clustering and Dimensional Reduction of Item Embeddings in Experiment 3. (a) Clustering into four classes corresponding to the original subscales; (b) Clustering into three classes, showing the merging of Conscientiousness with the Perseverance of Effort facet from Grit; and (c) Final clustering into two major categories, highlighting the integration of Grit and Conscientiousness with Gratitude.

The information suggested by EGF for overlap shows statistical significance. Specifically, the facet E items from the Grit scale first merges with the Conscientiousness scale items, resulting in the formation of cluster 2.2. This cluster 2.2 (M = 0.535, SD = 0.090) exhibits a higher mean of intra-group SPICs compared to the Grit scale (M = 0.355, SD = 0.138), with a significant difference observed (Cohen's d = 1.382, p < 0.001).

Additionally, the discriminant validity among the (sub)clusters was also confirmed. The intra-group Synthetic Pairwise Item Correlations (SPICs) for each (sub)cluster were aggregated and used as a baseline for comparison, and the results indicated significant discriminant validity for all comparisons (p<0.001). The effect sizes for these comparisons, as shown in Table 2, were substantial, with Cohen's d values indicating large effects across all clusters.

Table 2. Statistics on the Key SPICs of the EGF in Experiment 3.

Inter-group SPIC	Mean	SD	Difference with baseline	Cohen's d	*p*-value
Cluster1 to Cluster2	0.193	0.050	0.251	-3.400	0.000
Cluster 2.1 to Cluster 2.2	0.294	0.071	0.150	-1.744	0.000
Cluster 2.2.1 to Cluster 2.2.2	0.270	0.080	0.174	-1.878	0.000

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 October 2024 doi:10.20944/preprints202410.2001.v1

The comparative analysis results were fully in line with expectations and are further summarized in Table 3, which highlights the consistency of various framework outputs with theoretical expectations. The robustness check confirmed that the EGF from Experiment 3 was entirely consistent with those from Experiments 1 and 2, with Gratitude items forming a distinct sub-cluster, identical to the EGF structure from Experiment 1. This consistency reinforces the stability of ESAA's calculations.

In terms of incremental value, the ChatGPT models (versions 4o and o1 preview) failed to generate frameworks that aligned with theoretical expectations as effectively as ESAA, confirming ESAA's superiority in terms of convergence, discrimination, and interpretability. Although the ChatGPT-generated frameworks captured the macro-structure, they were unable to provide consistent sub-structures, thus falling short of the precision achieved by ESAA.

Table 3. Comparative Analysis of Various Framework Results with respect to their consistency with Theoretical Expectation.

Output of the Approach	Macro-structure	Sub-structures
EGF of Exp3	√	√
EGF of Exp1 or Exp2	√	√
Framework by ChatGPT 4o	√	×
Framework by ChatGPT o1 preview	√	×

Note. In the "Output of the Approach" column of the above table, unless otherwise specified, the term "approach" refers to that used in Experiment 3.

Detailed outcome frameworks generated by all the alternative approaches can be seen Appendix in https://osf.io/9fxmq/?view_only=ecef150ef2184d0da5106b6413f093c4

3. Discussion

The current research proposes a novel, multi-step method designed to detect redundancy among psychological concepts by analyzing their associated scales, referred to as the Embedding-based Semantic Analysis Approach (ESAA). The input for ESAA consists of the text from psychological scale items. ESAA transforms the textual content of these scale items into high-dimensional vectors, known as embeddings, using an advanced large language model. This process effectively converts qualitative text data into computable numerical representations that capture the semantic meanings of the items. Subsequently, ESAA applies unsupervised clustering algorithms—particularly hierarchical clustering—to these embeddings to uncover latent semantic structures among the items. Finally, the output includes an analysis of the semantic relationships among these items, encapsulated in a framework referred to as EGF, as well as synthetic correlations.

Through a series of three experiments, preliminary evidence attests to ESAA's usefulness and reliability. ESAA successfully converged semantically similar items, discriminated between items with significant semantic differences, and identified patterns of semantic overlap among constructs known to have theoretical redundancies. These findings suggest that ESAA is a promising tool for researchers aiming to refine theories and reduce redundancies in psychological measurement.

ESAA is expected to contribute to the development of psychological theoretical research, particularly in the areas of conceptual redundancy reduction and integration studies. As a novel tool, ESAA possesses unique value due to its ability to conduct analyses based solely on the content of the scales themselves, thereby eliminating the need for traditional data collection from participants in psychological research. This approach significantly reduces research costs. Furthermore, ESAA mitigates subjective bias, ensuring the reproducibility of results and enhancing the reliability of research conclusions.

The ESAA also represents a significant advancement in the field of psychological measurement by addressing redundancy and refining core components, thereby offering novel pathways for optimizing psychological measures and enhancing data quality. Scales are commonly used data

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 October 2024 doi:10.20944/preprints202410.2001.v1

collection tools in research across psychology, sociology, education, and other fields, where studies often require the collection of multifaceted information. However, the overlap among scales measuring various aspects can lead to substantial cost inefficiencies in terms of participants' engagement. ESAA can facilitate the design of large-scale streamlined measurement instruments for these studies at negligible cost, resulting in a substantial reduction in the number of items in the processed questionnaire format while preserving the richness of the information obtained.

Despite the promising findings and contributions of the Embedding-based Semantic Analysis Approach (ESAA), several limitations warrant consideration. Methodologically, the reliance on specific psychological scales raises questions about the generalizability of the findings to other constructs. While ESAA effectively identifies redundancy within the examined scales, its applicability to a broader range of psychological constructs remains to be established. Moreover, the large language models (LLMs) employed in this study may introduce biases in language representation, potentially affecting the semantic analysis outcomes. Additionally, in this study, ESAA generated embeddings using LLM only once, which may introduce sample bias. Technically, if the order of the input text is disrupted, the resulting embeddings are likely to differ because embeddings are generated based on the content of the text and its context; thus, the sequence of the text affects its semantic representation. Disrupting the order may lead to variations in the model's understanding of the text, resulting in the generation of different embedding vectors. Although such perturbations may not significantly impact the semantic relationships between embeddings, a systematic evaluation of the specific effects of these perturbations must be conducted before ESAA can be claimed as a reliable tool. This aspect has not been addressed in the current study and must be considered in future research. Furthermore, all standards used to assess ESAA's performance in this study are derived from the literature, either theoretical expectations or empirical evidence obtained from prior research. Regarding the psychological constructs addressed in this paper, they are abundant and confidential; however, in most fields, there is insufficient literature to provide comparative standards. When introducing ESAA into these fields for preliminary research, researchers must personally employ traditional methods to gather empirical evidence. Subsequently, this empirical evidence should be compared and analyzed alongside the results generated by ESAA. A substantial accumulation of such studies is necessary before the academic community can make a mature judgment regarding the reliability of ESAA.

The ultimate aspiration that motivates the creation of ESAA—namely, the redundancy reduction of psychological concepts and measurements, as well as the refinement of theory—holds immeasurable value for the advancement of psychology. The proposal and validation of a tool for redundancy detection represent merely the first small step toward this vision. A substantial amount of research work remains to be accomplished, and it is hoped that scholars will join in this endeavor.

References

Ajmal, A., Amin, R., & Bajwa, R. S. (2016). Personality traits as predictors of forgiveness and gratitude. *Pakistan Journal of Life & Social Sciences, 14*(2), 91-95.

Armstrong, J. S. (2001). *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Kluwer Academic Publishers.

Bainbridge, T. F., Ludeke, S. G., & Smillie, L. D. (2022). Evaluating the Big Five as an organizing framework for commonly used psychological trait scales. *Journal of Personality and Social Psychology*, *122*(4), 749-777. https://doi.org/10.1037/pspp0000395

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., & Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*. https://doi.org/10.48550/arXiv.2005.14165

Buchanan, B., & Hegarty, D. (2023). Development of a short personality assessment: The NovoPsych Five Factor Personality Scale – 30-item version. *NovoPsych*. https://novopsych.com.au/wp-content/uploads/2023/09/NFFPS-30-article-NovoPsych-website-version.pdf

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*(3), 309-319. https://doi.org/10.1037/1040-3590.7.3.309

Condon, D. M., & Revelle, W. (2015). Selected Personality Data from the SAPA-Project: On the Structure of Phrased Self-Report Items. *Journal of Open Psychology Data, 3*. https://doi.org/10.5334/jopd.al

Condon, D. M., Roney, E., & Revelle, W. (2017). A SAPA Project Update: On the Structure of phrased Self-Report Personality Items. *Journal of Open Psychology Data*, *5*(1), Article 1. https://doi.org/10.5334/jopd.32

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 October 2024 doi:10.20944/preprints202410.2001.v1

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation, 10*(1), 1-9. https://doi.org/10.7275/jyj1-4868

Credé, M., Tynan, M. C., & Harms, P. D. (2017). Much ado about grit: A meta-analytic synthesis of the grit literature. *Journal of Personality and social Psychology, 113*(3), 492-511. https://doi.org/10.1037/pspi0000090

Deigan, M. (2024). Having a concept has a cost. *Synthese, 204*(2). https://doi.org/10.1007/s11229-024-04661-5

DeVellis, R. F., & Thorpe, C. T. (2021). *Scale development: Theory and applications*. Sage publications.

Devlin, J. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology, 92*(6), 1087–1101. https://doi.org/10.1037/0022-3514.92.6.1087

Elson, M., Hussey, I., Alsalti, T., & Arslan, R. C. (2023). Psychological measures aren't toothbrushes. *Communications psychology, 1*(1), 25. https://doi.org/10.1038/s44271-023-00026-9

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272-299. https://doi.org/10.1037/1082-989X.4.3.272

Flake, J. K., & Fried, E. I. (2020). Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. *Advances in Methods and Practices in Psychological Science, 3*(4), 456-465. https://doi.org/10.1177/2515245920952393

Hommel, B. E., & Arslan, R. C. (2024). Language models accurately infer correlations between psychological items and scales from text alone. *OSF Preprints*. https://osf.io/kjuce/download

Khadhraoui, M., Bellaaj, H., Ammar, M. B., Hamam, H., & Jmaiel, M. (2022). Survey of BERT-Base Models for Scientific Text Classification: COVID-19 Case Study. *Applied Sciences, 12*(6), 2891. https://doi.org/10.3390/app12062891

Kong, F., Zhao, J., You, X., & Xiang, Y. (2020). Gratitude and the brain: Trait gratitude mediates the association between structural variations in the medial prefrontal cortex and life satisfaction. *Emotion, 20*(6), 917. https://doi.org/10.1037/emo0000617

MacCoun, R. J. (1998). Biases in the interpretation and use of research results. *Annual review of psychology, 49*(1), 259-287. https://doi.org/10.1146/annurev.psych.49.1.259

McCullough, M. E., Emmons, R. A., & Tsang, J. (2002). The grateful disposition: A conceptual and empirical topography. *Journal of Personality and Social Psychology, 82*(1), 112–127. https://doi.org/10.1037/0022-3514.82.1.112

Pace, V.L., & Brannick, M.T. (2010). How similar are personality scales of the "same" construct? A meta-analytic investigation. *Personality and Individual Differences, 49*, 669-676. https://doi.org/10.1016/j.paid.2010.06.014

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879. https://doi.org/10.1037/0021-9010.88.5.879

Schwarz, N. (1999). Self-reports: how the questions shape the answers. *American Psychologist, 54*(2), 93. https://doi.org/10.1037/0003-066X.54.2.93

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, 32*(5), 418–426. https://doi.org/10.1177/09637214231183917

Tetlock, P. E. (2005). *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton University Press.

Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Wulff, D. U., & Mata, R. (2024). Using embeddings to automate jingle–jangle detection and tackle taxonomic incommensurability. https://doi.org/10.31234/osf.io/9h7aw

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.