Convex Hull Applications to Natural Language Psychometrics

Nigel Guenole

Goldsmiths, University of London

Andrew Samo

Bowling Green State University

Tianjun Sun

Rice University

Damiano D'Urso

Independent Researcher

This experimental method pre-print has not been peer-reviewed.

Correspondence regarding this article can be sent to Nigel Guenole at nigel@measureco.ai, Andrew Samo at asamo@bgsu.edu, Tianjun Sun tianjunsun@rice.edu, and Damiano D'Urso at dursodamiano@gmail.com.

Abstract

Psychological measurement plays a vital role in many areas of science. Traditional methods for developing and scoring measurement instruments require large sets of human responses, making them time-consuming and costly. Recent advances in artificial intelligence (AI) offer new ways to tackle these challenges. In this brief note, we explore the use of convex hulls—a concept from computational geometry—in combination with AI-driven large language models to enhance psychometric practice. A convex hull is the smallest convex boundary around a set of points. By treating language embeddings as high-dimensional coordinates and forming convex hulls, we can interpret the structure of items and free text responses in new ways. We propose two novel applications of convex hulls: (1) *item analysis without data*: We use convex hulls to check if a test item "belongs" to a scale, providing a new indicator of item quality; and (2) *scoring free text*: By interpreting candidate responses in relation to the convex hull of other responses, we propose an objective way to score psychological constructs from natural language. Each application is possible in *supervised* and *unsupervised* modes. These methods are experimental and yet to be validated. We outline, but do not implement, brief methods for testing their efficacy. We discuss open questions that we expect will impact the utility of these methods, such as a) why we do not ask the LLM to score the items and responses, b) what the measurement scale of the proximity scores is c) why we preferred convex hull centroids over clustering and scale embedding means d) what happens of the construct hull does not match the intended target and e) what centroids and other hull attributes represent (e.g., the *intensity* or *essence* of constructs*)* and the implications for item analysis and scoring.

Key words: *psychometrics, convex hulls, natural language processing, large language models, artificial intelligence.*

Background

There is a rich history of measurement in psychology using natural language (Jackson et al., 2022; Pennebaker et al, 2003). Recent breakthroughs in transformer models, which are deep learning methods that incorporate self-attention (Vaswani et al. 2017), have led to rapid improvements in computational natural language capabilities via large language models. Applications of large language models to psychological measurement are now emerging at a fast rate (Arnulf et al. 2021; Hommel et al. 2022; Wulff & Mata 2023; Guenole et al., 2024a; Guenole et al. 2024b; Hernandez & Nie 2022; Russell-Lasalandra, et al., 2024) In this brief note, we continue this tradition, proposing convex hulls for analysing embeddings in a psychometric context.

Psychometric assessments, such as personality tests, are commonly used in psychological research and practice. However, traditional methods of evaluating these assessments often require large datasets of human responses, to estimate robust measurement models and/ or train supervised machine learning (ML) models, which can be resource intensive. This is true of both the data required to estimate measurement model parameters and the responses required to build scoring models for free text response formats. With recent advancements in artificial intelligence (AI) and large language models (LLMs), there is an opportunity to explore new methods that do not rely on extensive response data.

We propose using convex hulls, a concept from computational geometry, to analyze LLM-generated text embeddings in psychometrics. By applying convex hulls, we aim to offer a novel way to assess the quality of scale items and to score free-text responses. This approach has the potential to streamline the process of psychometric design and scoring, making it more efficient and accessible. Importantly, our method could enhance the objectivity of these assessments by leveraging the inherent structures within language data, potentially offering insights that traditional methods might miss.

We highlight their application to two subdomains of psychological assessment, scale development and free response scoring. Both methods treat embeddings as coordinates in a high-dimensional space (i.e., numerical values that represent words or sentences as points). This multi-dimensional space captures the meaning and relationships between texts, making it possible to analyze their proximity to each other. The first method addressing items belongingness is similar to item analysis but without response data.

The second application proposes combining transformers and convex hulls for free text scoring of constructed responses. It builds on the recognition that the first item belongingness application we outline is indifferent to whether the text analyzed represents scale items or candidate free text responses. In both cases, we outline, but do not execute, small proof of concept studies. We close with discussion of key issues for future research.

What are Large Language Models?

Large language models (LLMs) are neural networks that have been trained on massive amounts of language data and are able to understand unstructured text, produce rich numerical representations of text, and generate new text. LLMs can be viewed as advanced text prediction tools. They've learned patterns from large corpuses of language data and can understand, interpret, and even create new text that makes sense in context. The underlying architecture that controls how data flows through the model is based on Transformers. Transformer models incorporate self-attention, a mechanism for determining the relevance of words to other words in text – have led to significant improvements in the capability of LLMs (Vaswani et al. 2017).

Sentence transformers, which we use in this work, are specialized forms of LLMs that are designed to accurately represent the semantic content of language as dense numerical vectors. Sentence transformers are like translators that turn sentences into a series of numbers. These numbers, called embeddings, capture the meaning of the sentences in a way that LLMs can understand. Embeddings are essentially ways to represent words or sentences as points in a multi-dimensional space. They are essentially special codes that capture the meaning and relationships between words.

Embedding values are often used in subsequent data science applications including, for example, supervised machine learning where downstream outcomes are predicted. In this article, we discuss sentence encoders in the context of psychological measurement for constructs such as attitudes, opinions, and traits. For a recent tutorial on large language models in behavioural science see Hussain et al. (2024).

What are convex hulls?

Imagine you have a bunch of pins on a board. If you wrap a rubber band around them and let it snap tight, the shape it forms is a convex hull. It's the smallest convex shape that can fully enclose all the points inside it such that any two points are connectable by a line that does not leave the hull (Preparata & Shamos, 1993).

In our context and in both applications of transformers and convex hulls that we discuss here, we treat each text embedding as a point in a multi-dimensional space, and the convex hull represents the boundary around these points. In the case of checking item " belongingness", the distance from the centroid of the convex hull that is formed from the embeddings of other scale items is interpreted as a potential measure of *item discrimination.* In the case of free text scoring, a candidate's response proximity to the centroid of the hull formed by other responses is treated as a *construct score*, after inversion such that a higher value means closer proximity.

To form a non-degenerate convex hull (i.e., one that takes on the expected shape given the number of coordinates) in n dimensions (i.e., where each point has n

coordinates) requires n +1 points. For instance, in 2-dimensional space, you need to plot three points to form a triangle; in 3-dimensional space, four points are needed for a tetrahedron. The shape in n-dimensions formed by the n+1 item embedding coordinates is always convex.

It is possible to plot fewer than n+1 cases than coordinates per point (e.g., two points in 2-dimensional space), but the hull will be degenerate (taking the form of a line segment, in our example). This n+1 requirement for a non-degenerate hull, where n = *items* in item analysis and *candidates* in response scoring, has practical implications with modern transformers.

This is because transformer vectors can take on high dimensional forms. Even small embeddings have 384 dimensions today. In such cases, the item belonginess method that we discuss will require data for at least 385 items, while the free text scoring approach will require responses for at least 385 respondents. We can reduce this data burden by projecting the embedding dimensions into a lower dimensional space in both cases.

If we wish to plot the convex hulls for item discrimination or scoring, this will involve taking the first three (for a 2d plot) or four (for a 3d plot) principal components for both methods. Computational approaches without plotting require taking one fewer PCA dimensions than there are items or respondents. However, early experiences suggest taking a very high number of components is computationally expensive (i.e., very memory intensive) and this leads to overflow errors where the numbers become too large to be represented in memory. In practice, experimentation may be required determine the balance between what is theoretically and computationally possible.

We now summarise small proof of concept study designs that could test each of these new methods, and then we discuss five big issues related to measurement approaches that use convex hulls.

Application One: Item analysis without data

A core challenge in psychometrics is determining the extent to which a question or statement relates to the underlying construct, i.e., the item discrimination. Common methods for solving this challenge include Classical Test Theory (CTT), Exploratory and Confirmatory Factor Analysis (EFA/CFA), and Item Response Theory (IRT) (Embretson & Reise, 2013). All these methods analyse candidate responses to questions to determine analogous quantities, item total correlations under CTT, factor loadings under EFA/CFA, and slopes under IRT.

One limitation of these methods is that they all rely on actual response data. Recent papers have suggested that these parameters can be approximated without response data using large language models. The item-total correlation can be approximated as an item embedding's proximity to the construct definition embedding or the average of all other scale item embeddings (Guenole et al., 2024).

This idea can be generalised by replacing the item response correlation matrix with the embedding cosine similarity matrix, enabling multidimensional factor loadings and potentially exploration of structural relationships before data are collected (Guenole et al. 2024). Here we approach the task of item analysis without data from another angle with the use of convex hulls. One potential approach to checking this proposal could be as follows.

Proposed workflow for application one: item analysis without data.

- Generate items to measure a target construct, either using humans item writers or using generative A.I.
- Sort items using human raters to check the preliminary suitability of the items as indicators of their respective constructs.
- Obtain item embeddings for personality items with a sentence transformer, such as MiniLM, or even a variety of transformers.
- Reduce the embedding dimensionality as required for plotting in lower dimensions or computation in higher dimensions.
- Exclude the studied item from the computation of the hulls in the next step where the hull for each construct is computed.
- Compute the convex hull for each of the scales based on the scale items per scale excluding the studied item.
- Calculate distances, e.g. Euclidean, Manhattan, and cosine similarity between every item and every convex hull.
- Examine the internal structure of the item embeddings e.g. using multi-trait-multi-method approaches.
- Refine or remove items that show unexpected relationships to the target construct or have inadequate convergent and discriminant validity in an iterative fashion.
- Compare the convex hull proximities of the final items to the actual empirical discriminations based on response data.

From these results, a number of potentially useful issues can be studied. First, we can see whether any item is within the hull of the other items using the point in the hull test. If an item falls in the hull of another set of items, it is said to be affinely dependent on the other points of the convex hull, i.e., it is redundant given it is expressible as a linear combination of non-negative weights of the other item embeddings that sums to one. This is called a convex combination. Assuming items are not redundant, we can then proceed to see which items are most related to what the items all have in common.

We can check each item's belongingness by interpreting the item proximities directly in relation to each other. We can also check whether the mean proximity across items within every scale is closer to the centroid of its own target hull (from which each item it is omitted when calculating its individual proximity) than any other hull centroid. We can check the nearest vertex to each studied item to see which hull the vertex is from.

In summary, there is a wide array of methods and tests in computational geometry that could be introduced for this step. We suggest this first application is tested using items from generative AI methods or existing questionnaires. The ultimate test, of course, is how these item proximities relate to empirical discrimination, the quantity of real interest. We suggest this as a last step in the validation of the method for establishing the validity of the method.

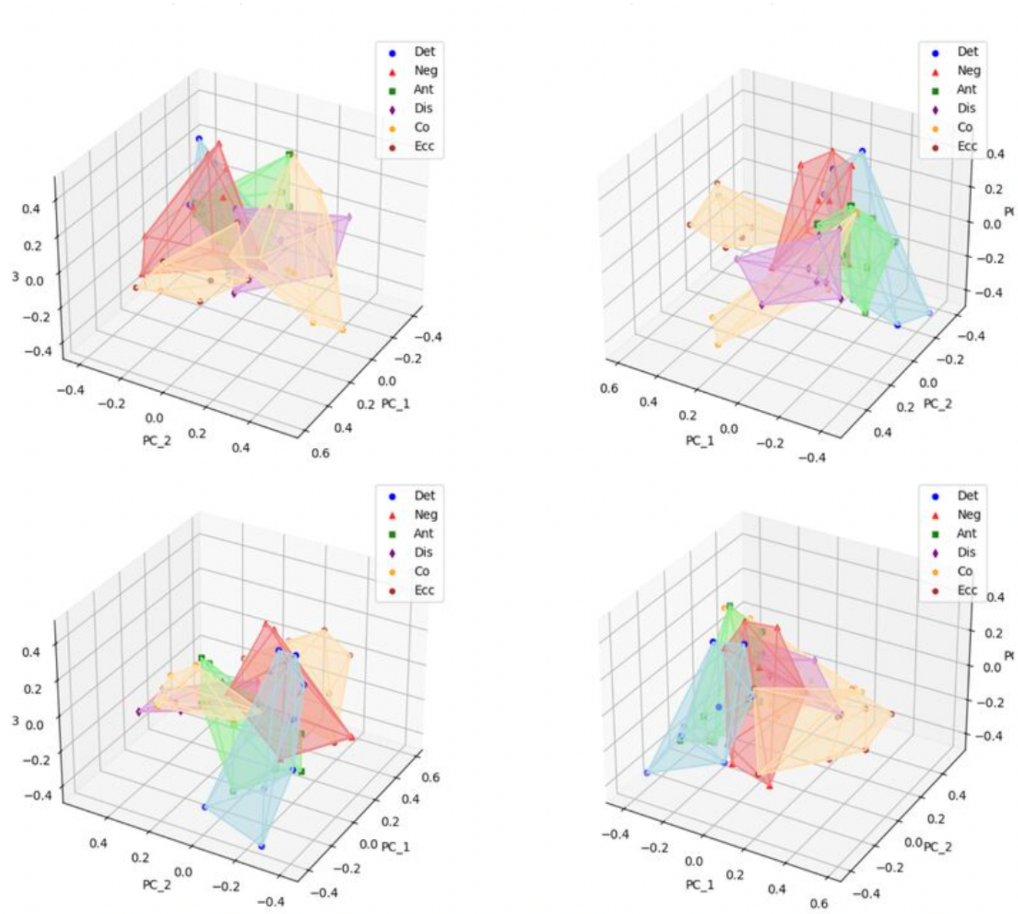
Application Two: Scoring of Natural Language

An ultimate prize for psychological measurement is accurate free text construct scoring (e.g., interviews, SJTs, performance reviews, essays, free text survey responses). Progress is occurring but predictive accuracy against test scores or human labels is a potential drawback of current methods. It blurs distinctions between measurement and prediction and prioritises external criteria over the inherent structure in language. Can we have a natural language measurement model without prediction? Transformers (Vaswani et al. 2017) and convex hulls (Preparata & Shamos 1993) might help here too.

The free response scoring process involving using sentence encoders to generate embeddings of candidate responses. The language to be scored can, but need not necessarily, be in response to prompts designed to focus the language generated into a specific domain. Each candidate's response is encoded with a sentence transformer. The convex hull is then formed out of all of the responses being scored. The inverted proximity between the response embedding for each candidate and the centroid of the convex hull is then interpreted as the construct score.

The idea is that a higher score reflects higher standing on the construct measured after inverting the proximities. There are two ways that we propose to investigate this. Firstly, because data collection will be expensive, we propose a simulation proof of concept using generative AI methods to create 'candidate' responses. Second, we propose empirical investigation into this topic using actual candidate data on both free response text and conventional questionnaires. Here is a general workflow that might be adapted in any such investigation.

Figure 1. Item analysis without data. The G50 (Guenole, 2015) Likert personality questionnaire contains six scales corresponding to the rationale version of section III trait model in DSM-5. det is detachment; neg is negative affect; ant is antagonism; dis is disinhibition; co = compulsivity; ecc = eccentricity. This figure presents the convex hulls formed from the first four principal components of the 384-dimension embeddings of each item in the questionnaire, generated with MiniLM. The embeddings components are then projected onto the same surface. The plot is then presented from four different viewing angles with the following elevation and azimuth coordinates, moving clockwise and beginning top left (30, 20) (30,120) (30,210) (30, 300).



Proposed workflow for application two: free text scoring.

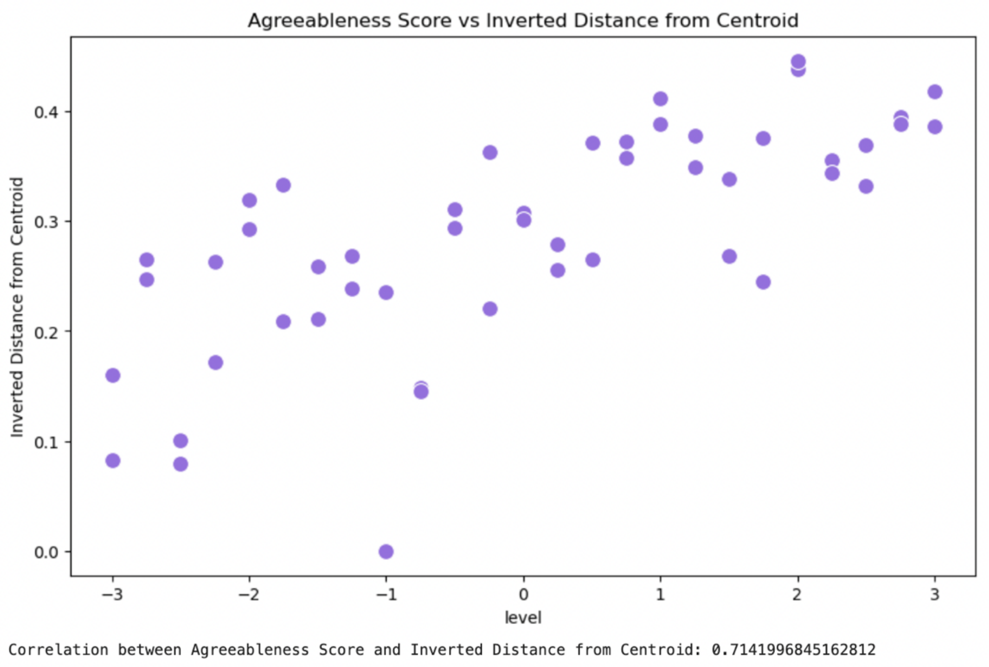
- Generate real personality scores for a trait that span the trait continuum from -3 to +3 standard deviations.
- Select or write a free response prompt for personality five, for instance, see Hickman et al. (2021).
- Ask GPT to generate ‘candidate’ responses to the prompt reflecting the trait levels identified in the first step.
- Ask Claude to score responses, correlate these scores with original trait scores for a sense check and demonstrate a strong correlation.
- Embed the text responses using a sentence transformer, for this experiment MiniLM would be appropriate.
- Compute the convex hull using all of responses except the candidate being scored, also calculate its centroid.

- This likely requires reducing the embedding dimensions to n-1-1=n-2 or fewer PCA coordinates.
- Calculate the Euclidean and Manhattan distances, and the cosine similarity of response embeddings to hull centroid.
- Invert the distances so that a higher response score is closer to the hull, i.e., subtract each score from the maximum.
- Repeat the above experiments at different temperatures so that the text has varying degrees of focus on the construct, as we expect human responses would.
- Check the correlation between inverted hull proximities and original scores to see that a large correlation is obtained.

In summary, like CTT, CFA, and IRT, this new scoring needs 'ground truth' only for validity, not the scoring itself. If a large correlation is obtained between the original simulated scores and the recovered proximities, it would be evidence for the efficacy of the method. Even in that case, we anticipate a number of likely questions. We turn to answer some that are likely to be the most common now.

Figure 2: Natural language scoring of Agreeableness from free text. Prompt relates to moral support for a colleague. Simulation set-up: Trait continuum split in .25 intervals from -3 to +3; two responses generated at each trait level for n=50 overall; response generation via Open AI GPT Turbo 3.5 using Open AI API. Sentence transformer: MiniLM via Sentence Transformers. Measurement model: Convex hull, unsupervised mode.

Trait recovery accuracy **r=.71**. The response closest to the centroid is: "I offered a listening ear and words of encouragement to help them feel better (simulated theta of 2). The response furthest from the centroid is: I provided some surface-level support because it affects productivity, nothing more beyond that (simulated theta of -1).



```
Correlation between Agreeableness Score and Inverted Distance from Centroid: 0.7141996845162812
```

Key Issue 1: Why not just ask the LLM to score responses?

Questions may arise as to why we do not use other potential approaches. The most common question is likely to be why not ask the LLM to score the text directly? The answer to this is that the motivating idea behind the method is not to impose pre-existing notions or structures regarding what is measured. The proposed method potentially achieves this, because the convex hull centroid represents what is actually discussed in language, whereas asking the LLM to score test emphasises what one hopes to find. In other words, by scoring with an LLM against a rubric, a preconceived idea about what is represented in the data is imposed.

Key issue 2. Measurement scale.

A question arises as to what the measurement scale of the newly proposed proximity scores is. In one sense, they are interval and even ratio given that a response can be zero units away from the centroid of the hull. Even if the transformers that encode the sentences themselves could be also considered interval measures, however, there is no guarantee that the equal measurement scaling of the proximity scores corresponds to equal intervals on the personality spectrum (in the case of the scoring application of convex hulls). For the item analysis application, it is not guaranteed that changes in the distance from the hull correspond to equivalent changes in discriminative utility.

Having a clear measurement unit could nonetheless make the scores from convex hull methods more amenable to analysis via conventional methods in psychometrics than might otherwise be the case. We may be able to apply CTT, CFA, and IRT models to multiple proximity indicators from different prompts focused on the same trait (or applying rubric hulls, albeit this imposes a structure); or even apply networks models to the different proximities that emerge using scores which have a more plausibly interval or even ratio structure. Should this be possible, we can examine transformer-based construct measurement at the latent level, examining concepts such as measurement invariance and error adjusted construct relationships.

Key issue 3: Clustering and scale embedding mean centroids.

Readers may ask why we use convex hulls as opposed to a more straight-forward method such as cluster analysis. The main reason we do not use clustering is because it assigns a categorical membership. Personality traits and other psychological constructs need continuous scores. While it is possible to have a continuous proximity to cluster centroids, scoring approaches requires a proximity of all cases to the same centroid. It may then be computationally simpler to use scale embedding means. However, convex hulls emphasize the full construct by prioritising its extremities, the scale embedding centroid prioritizes central tendency.

Key issue 4: When the hull centroid does not represent the intended target.

The scoring approach assumes the hull reflects the construct, which we refer to as an 'unsupervised' convex hull mode. However, it might not be the case that the centroid reflects the target we hope. In such cases, we can constrain responses, clean the corpus, or follow a 'supervised' convex hull mode where we form the hull from target or rubric responses. (albeit the rubric approach imposes an a priori structure which we aimed to avoid). Prompts may not even be needed where a target hull is used. Multidimensional constructs could be assessed using multiple prompts or by using multiple target hulls even on the same corpus without prompts.

Key issue 5: Intensity or Essence?

If the method works, the correlation between the original trait scores and the convex hull proximity scores will be positive and ideally strong. Importantly, however, this correlation may only be strong under special conditions. One is where the centre of the convex hull represents *high intensity* of the trait. Whether the centroid is high can be confirmed by interpreting responses that are closest the centroid of the hull. For example, if we believe we are measuring agreeableness, high agreeableness at the center of the hull might be represented by a response to a prompt designed to elicit agreeableness such as *"I offered a listening ear and words of encouragement to help them feel better"*.

Another possibility that would arguably lead to a positive correlation is where the hull centroid does not represent an intensity but instead represents the *essence* of a trait, rather than an intensity or level of a trait. If, on the other hand, the hull centroid represents some other intensity that is (person) sample dependent (i.e., not one of the poles of the trait), proximities are likely to have complex interpretations that require as yet undetermined work arounds to recover trait scores. However, it may also turn out that what is required for trait recovery is that, rather than a centroid at either construct pole, instead what matters is the degree of dimension relevant contrast between responses at the centroid and the hull's extremity. In other words, even a centroid representing a moderate position on the trait may be satisfactory for trait recovery of the extreme is sufficiently far away in the opposite direction. Given the expensive nature of data collection, we here propose that this method is fully simulated before expending money and time collecting human data. Should these simulations work out, of course, real human response data will be quickly needed.

Conclusion

We have proposed new methods for psychometrics that leverage convex hulls and large language models to accomplish important tasks, examine item belongingness to scales without response data, and score psychological constructs from free responses. While the methods have been presented in the context of transformers for textual analysis, the methods have applications with conventional data too. Should the methods prove effective, it will be important a) to compare these to alternative approaches such as pseudo-methods discussed earlier, b) discuss where convex hull artificial intelligence scale design steps might fit into the flow of conventional scale development, and importantly, c) identify how both pseudo-methods and convex hull techniques that we are proposing can be made as accessible as possible to practitioners.

References

Arnulf, Jan Ketil, Kai R. Larsen, Øyvind Lund Martinsen, and Kim F. Nimon. 2021. "Editorial: Semantic Algorithms in the Assessment of Attitudes and Personality." *Frontiers in Psychology* 12 (July): 720559.

Embretson, Susan E., and Steven P. Reise. 2013. *Item Response Theory*. Psychology Press.

Guenole, N., Samo, A., & Sun, T. 2024a. *Pseudo-Discrimination Parameters from Language Embeddings.* OSF. February 9, 2024.

Guenole, Nigel. 2015. "The Hierarchical Structure of Work-Related Maladaptive Personality Traits." *European Journal of Psychological Assessment: Official Organ of the European Association of Psychological Assessment* 31 (2): 83–90.

Guenole, Nigel, Epifanio Damiano D'Urso, Andrew Samo, and Tianjun Sun. 2024b. "Pseudo Factor Analysis of Language Embedding Similarity Matrices: New Ways to Model Latent Constructs," April 14, 2024.

Hernandez, Ivan, and Weiwen Nie. 2022. "The AI-IP: Minimizing the Guesswork of Personality Scale Item Development through Artificial Intelligence." *Personnel Psychology*, October. https://doi.org/10.1111/peps.12543.

Hickman, Louis, Nigel Bosch, Vincent Ng, Rachel Saef, Louis Tay, and Sang Eun Woo. 2021. "Automated Video Interview Personality Assessments: Reliability, Validity, and Generalizability Investigations." *The Journal of Applied Psychology*, June. https://doi.org/10.1037/apl0000695.

Hommel, Björn E., Franz-Josef M. Wollang, Veronika Kotova, Hannes Zacher, and Stefan C. Schmukle. 2022. "Transformer-Based Deep Neural Language Modeling for Construct-Specific Automatic Item Generation." *Psychometrika* 87 (2): 749–72.

Hussain, Zak, Marcel Binz, Rui Mata, and Dirk U. Wulff. 2024. "A Tutorial on Open-Source Large Language Models for Behavioral Science." *Behavior Research Methods*, August. https://doi.org/10.3758/s13428-024-02455-8.

Jackson, J. C., Watts, J., List, J. M., Puryear, C., Drabble, R., & Lindquist, K. A. (2022). From text to thought: How analyzing language can advance psychological science. Perspectives on Psychological Science, 17(3), 805-826.

Pennebaker, James W., Matthias R. Mehl, and Kate G. Niederhoffer. 2003. "Psychological Aspects of Natural Language. Use: Our Words, Our Selves." *Annual Review of Psychology* 54: 547–77.

Preparata, Franco P., and Michael I. Shamos. 1993. *Computational Geometry: An Introduction*. PDF. 1st ed. Monographs in Computer Science. New York, NY: Springer. https://doi.org/10.1007/978-1-4612-1098-6.

Russell-Lasalandra, Lara L., Alexander P. Christensen, and Hudson Golino. 2024. "Generative Psychometrics via AI-GENIE: Automatic Item Generation and Validation via Network-Integrated Evaluation," September.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." *Advances in Neural Information Processing Systems* 30. https://proceedings.neurips.cc/paper/7181-attention-is-all.

Wulff, Dirk U., and Rui Mata. 2023. "Automated Jingle–Jangle Detection: Using Embeddings to Tackle Taxonomic Incommensurability." https://doi.org/10.31234/osf.io/9h7aw.