Contribution ID: **164** Type: **Symposium**

Bridging Psychometrics and Artificial Intelligence

Symposium title

Bridging Psychometrics and Artificial Intelligence

Coordinator

Miguel A. Sorrel & Hudson Golino

Affiliation

Universidad Autónoma de Madrid & University of Virginia

Abstract

This symposium presents innovative research on the convergence of psychometrics and artificial intelligence, highlighting how AI can enhance traditional psychometric applications. Specifically, four studies are presented, each focusing on leveraging large language models (LLMs) and artificial intelligence techniques to optimize psychometric tools, automate item generation and test assembly, and improve trait measurement. The studies include applications to the measurement of non-cognitive traits (e.g., personality) and abilities (e.g., English language assessment).

The studies demonstrate how LLMs can generate high-quality items and validate them using in-silica methods, leading to reliable psychometric metrics. Furthermore, they explore the use of advanced algorithms, such as Dynamic Exploratory Graph Analysis and genetic algorithms, to optimize the mapping of psychological traits and improve trait score recovery. A common theme across the studies is the integration of empirical data with AI-generated insights.

Throughout the research, AI's ability to predict or match human-gathered validation data is showcased, reinforcing the idea that artificial intelligence can complement and enhance traditional psychometric practices. In sum, this symposium underscores the transformative potential of artificial intelligence in psychometrics, offering new opportunities to improve scale development, trait measurement, and validation through sophisticated AI techniques and models.

Keywords

artificial-intelligence; large-language-models; automatic-item-generation; test-assembly, item-response-theory

Number of communicatios

4

Communication 1

Optimizing LLM Embeddings for Automatic Item Development and Validation

Authors

Hudson Golino

Affiliation

University of Virginia

Abstract

Large Language Models (LLMs) have shown promise in text clustering and dimensionality analysis through embeddings, yet their potential for optimization remains largely unexplored. We conducted a comprehensive simulation study to enhance the accuracy of LLM embeddings in trait mapping using Dynamic Exploratory Graph Analysis (Dynamic EGA). The simulation generated 200 items across 4 traits of Narcissistic Personality, randomly selecting 3-40 items per dimension. We analyzed 1,040,000 combinations across 260 embedding values (3-1300) in a 1536-dimensional space. Performance was evaluated using Total Entropy Fit Index (TEFI) and Normalized Mutual Information (NMI). Vector field analysis revealed complex dynamics between TEFI and NMI, with optimal performance occurring in regions of moderate TEFI values and NMI above 0.5. The number of items per dimension showed peak performance between 10-20 items, while embedding dimensions exhibited non-linear relationships with both metrics. A weighted scoring system prioritizing NMI (70%) over TEFI (30%) significantly outperformed traditional cross-sectional embedding approaches. The optimization demonstrated improved accuracy in concept mapping while maintaining structural stability, suggesting a promising direction for enhancing LLM-based text analysis methods.

Keywords

large-language-models; network-psychometrics; dimensionality

Communication 2

Does the result of an in-silica structural validity matches the structural validity computed in human-gathered data?

Authors

Lara Russell-Lasalandra

Affiliation

University of Virginia

Abstract

The rapid advancement of large language models (LLMs) has enabled automated psychological scale development, yet questions remain about the correspondence between in-silica and human-gathered validation. This study examines whether structural validity metrics computed during automated item development match empirical validation results. Using AI-GENIE (Automatic Item Generation and Validation via Network-Integrated Evaluation), we generated Big Five personality items using five LLMs (Mixtral, Gemma 2, Llama 3, GPT-3.5, GPT-4). AI-GENIE performed in-silica structural validation during item generation and selection. These items were then administered to independent U.S. samples (N = 1000 per model). Comparing the in-silica and empirical structural validity metrics revealed strong correspondence (average correlation r = .89, RMSE = 0.08) across all models. Network invariance tests between in-silica and human-gathered data showed configural (NCT = 0.12, p > .05) and metric invariance (NCT = 0.15, p > .05). These findings suggest that AI-GENIE's in-silica structural validation effectively predicts empirical structural validity, supporting its use in automated scale development.

Keywords

large-language-models; automatic-item-generation; test-assembly

Communication 3

How Block Types and Social Desirability Shape Forced-Choice Questionnaire Automatic Assembly

Affiliation

Universidad Autónoma de Madrid & Universidad Nacional de Educación a Distancia

Abstract

The construction of forced-choice questionnaires often relies on item banks with single-stimulus or Likert-type items. In its simplest form, items must be paired to create a desired number of blocks. A key challenge in this process is pairing items while accounting for factors such as item polarity and social desirability, which can impact the quality of the measures. Recent combinatorial approaches, like genetic algorithms, leverage parametric optimization based on Likert data estimates. Alternatively, blueprint-based methods enable block assembly without such estimates, integrating expert judgments on social desirability. However, these approaches have yet to be systematically compared, which is the primary goal of this study. A Monte Carlo simulation and empirical analysis were conducted to compare block assembly using the genetic algorithm and blueprint-based methods, with and without considering social desirability. The main outcome of interest was trait score recovery. Four key factors were manipulated to assess their influence on this outcome: the number of heteropolar blocks, questionnaire length, the inclusion of social desirability ratings, and the correlation between social desirability and single-stimulus item parameters. Results indicate that parametric methods generally lead to superior trait score recovery, especially when only homopolar blocks are used or when social desirability is factored in—conditions commonly found in applied settings. These findings highlight the importance of optimizing assembly procedures. We also discuss how expert judgments can serve as proxies for item parameters, enabling efficient block assembly in the absence of empirical data on single-stimulus items.

Keywords

test-assembly; item-response-theory; forced-choice-questionnaires; genetic-algorithm; social-desirability

Communication 4

Predicting Item Response Theory Parameters from the Semantic Space of Computational Language Models

Authors

Diego Iglesias, Miguel A. Sorrel, Ricardo Olmos, & Francisco J. Abad

Affiliation

Universidad Autónoma de Madrid

Abstract

Parallel to the development of new technologies, computational language models have emerged as automated tools for analyzing semantic relationships between linguistic units. Due to their success in performing human-like tasks, such as vocabulary tests and sentiment analysis, interest in the practical applications of these models has grown exponentially, resulting in the development of larger models with enhanced predictive capabilities. In this study, we examine whether the high-dimensional semantic space underlying computational language models, such as ChatGPT, can be used to predict item parameters. In ChatGPT, linguistic units are represented as n-dimensional embedding vectors, which can be manipulated through mathematical operations. We extracted embeddings for an item pool of 220 items from an English vocabulary test. The loadings of each item in ChatGPT's 1536-dimensional space were used as independent variables to predict their corresponding item response theory item parameters. The predictive accuracy of various machine learning models was evaluated using cross-validation procedures and compared with human-expert ratings. Despite the relatively small size of the training set, preliminary results are promising (⊠^2_⊠⊠⊠=0.40). We discuss the potential of using larger datasets for training the predictive model and the promising role of generative artificial intelligence in creating large item pools with desirable psychometric properties at minimal cost.

Keywords

large-language-models; item-response-theory, automatic-item-generation; ChatGPT

Authors

Scarlett Escudero, Miguel A. Sorrel, Rodrigo S. Kreitchmann, & Francisco J. Abad

Primary authors: SORREL, Miguel A. (Universidad Autónoma de Madrid); GOLINO, Hudson (University of Virginia); RUSSELL-LASALANDRA, Lara (University of Virginia); IGLESIAS, Diego (Universidad Autónoma de Madrid)

Co-authors: Mrs ESCUDERO, Scarlett (Universidad Autónoma de Madrid); SCHAMES KREITCHMANN, Rodrigo (National University of Distance Education); OLMOS, Ricardo (Universidad Autónoma de Madrid); ABAD, Francisco J. (Universidad Autónoma de Madrid)

Presenters: SORREL, Miguel A. (Universidad Autónoma de Madrid); GOLINO, Hudson (University of Virginia); RUSSELL-LASALANDRA, Lara (University of Virginia); IGLESIAS, Diego (Universidad Autónoma de Madrid)

Session Classification: Symposium : ”Bridging Psychometrics and Artificial Intelligence”

Track Classification: Measurement: Measurement