Contribution ID: **328** Type: **Oral Presentation**

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

Friday 25 July 2025 08:50 (15 minutes)

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 insilica structural validation effectively predicts empirical structural validity, supporting its use in automated scale development.

Primary author: RUSSELL-LASALANDRA, Lara (University of Virginia)

Presenter: RUSSELL-LASALANDRA, Lara (University of Virginia)

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