**Pilot Study for Development of Occupational Interest Profiles Using ChatGPT**

The pilot study aims to evaluate the quality of ratings using Large Language Models of 16 representative occupations on CABIN. It will allow for evaluation of 1) retest reliability of LLM ratings, 2) convergent validity of LLM ratings with human subject matter expert ratings, and 3) comparison of different amounts of input to the LLM.

**Occupational Sample Development**

14 occupations were chosen to be representative of different interest constructs and of highly prevalent occupations in the US job market. Occupations were chosen based on their ranked frequency in the US labor market, and in specific sectors and industries. Occupations were also chosen to represent all RIASEC and SETPOINT interests, as well as 10 unique CABIN interests.

Of these occupations:

* The 10 most frequent occupations in the US were included. *Reference*: <https://www.bls.gov/oes/current/area_emp_chart/area_emp_chart.htm>
* 4 frequent occupations were selected from more specific occupational groups to represent Investigative, Artistic, Technology and Nature interests. *Reference*: <https://www.bls.gov/oes/current/ind_emp_chart/ind_emp_chart.htm>

**Rating Process**

* **LLM Rating:** ChatGPT4o will assess each occupation by comparing the occupations descriptions and tasks to CABIN definitions and questionnaire items. ChatGPT4o will be prompted to develop 56 total occupational profiles, with four profiles of each occupation. ChatGPT4o will be prompted to develop 2 profiles of each occupation using 6 tasks, and 2 profiles using all core tasks.
* **Subject Matter Rating:** The lead author of the study will also rate the 14 occupations by comparing the occupations’ descriptions and tasks to CABIN definitions and questionnaire items. He will use all core tasks to develop the ratings. He will consider tasks that appear earlier as more important for the rating. He will rate each occupation twice and take the average of each rating to increase the reliability of the ratings.

**Data Analysis**

Correlations will be computed between all profiles and the mean correlations between each type of rating system will be reported, such as in the table below. A correlation matrix between all identified profiles will be developed. Average correlations will be reported for:

* **Retest reliability:** Correlations between profiles with identical inputs for each unique type of profile (6 core tasks and all core tasks).
* **Convergent Validity of 6 Core Tasks Input:** Correlation between LLM-generated profiles using 6 core tasks and human ratings.
* **Convergent Validity of All Core Tasks:** Correlation between LLM-generated profiles using 6 core tasks and human ratings.

*A correlation of .8 or above will be considered desirable.* However, there is some flexibility in the interpretation of statistics at this exploratory stage, and given uncertain implications of divergence of LLM versus human ratings.

**Expert Review and Evaluation**

Three subject matter experts (Heimpel, Rounds, and Chu) will review the summary statistics and manually inspect the profiles produced. After reviewing the profiles, they will discuss and advise the lead author (Heimpel) on the optimal number of core tasks to use in future rating attempts. Finally, the quality of the prompts and the data input process will be evaluated by the research team to assess whether the prompt and data input is sufficiently reliable and valid to be applied to a larger subset of occupations or whether further pilot study is necessary. This evaluation will also help identify potential refinements in rating methodology prior to scaling up the project.