Original Article

Special Collection: Advancing Psychological Assessments With Machine Learning and Artificial Intelligence

Convergent Validation of Trait Induction in LLMs

A Step Toward Integrating AI in Foundational Stages of Psychometric Test Development

Yasin Edin, Leon Rosenplänter, and Martin Kersting

Department of Psychological Assessment Justus-Liebig-University Giessen, Giessen, Germany

Abstract: This study investigates an approach to test a key requirement for using large language models (LLMs) to simulate valid data sets of item-responses, namely the effect of trait induction on another related construct. Using OpenAI's GPT-4, we used zero-shot prompting to induce high- and low-agreeableness profiles in the model's output. We then investigated, in terms of convergent construct validation, the effect of trait induction on another theoretically and empirically related construct: emotion understanding (EU). To measure EU, we utilized a situational judgment test developed to test LLMs. Using $n = 680$ simulated data sets, we show that a prompt on a trait (agreeableness) has a plausible effect on unprompted behavior in a skill (EU). These results provide initial support for evaluating the effects of trait induction through methods of convergent construct validation when assessing LLM-generated responses, and for using LLMs to generate data for foundational steps of psychometric test development.

Keywords: silicon sampling, large language models, psychometrics, personality traits, construct validity

Generative artificial intelligence (AI), in particular large language models (LLMs), offers a wide range of use cases for psychological sciences. These potential applications of LLMs, including their potential benefits and shortcomings, as well as areas requiring further research, are discussed in the relevant literature (e.g., Demszky et al., 2023; Ke et al., 2025). For the construction of items for psychometric performance tests and questionnaires, LLMs have been suggested as useful tools (Lee et al., 2023; Tan et al., 2024). Although the idea of automatic item generation was first proposed in the 1970s (Bormuth, 1970), algorithmic and machine learning approaches to generate new test items have only relatively recently been proposed (Gierl & Haladyna, 2013; Gierl & Lai, 2012). Early work in the emerging field of research on the use of AI in test development suggests that LLMs show promising results for item generation in psychometric assessments (e.g., Lee et al., 2023), potentially streamlining the process of item generation (Tan et al., 2024).

In this study, we present a use case for LLMs that goes beyond the already discussed generation of new items, namely the generation of simulated data to support the process of test development. The benefits of such simulations for psychological assessment are significant – they may allow preliminary estimates of item and scale parameters early in the test development process (cf. Liu et al., 2025). This includes item parameters such as difficulty and discriminatory power, as well as scale parameters such as reliability and validity. LLM-generated item and test responses may also offer early indications of measurement invariance (cf. Borsboom, 2006). Furthermore, they could support initial indications of differential item functioning (cf. Holland & Wainer, 2009), provide preliminary estimations of convergent and discriminant validity, and enable simulations of multivariate predictions. Thus, LLM-generated item responses could support item selection and provide valuable information for possible scale reductions and extensions.

A key advantage of this approach would be the ability to pretest a large number of items without the need for human participants, which would be particularly advantageous given that answering items can be cognitively demanding (e.g., ability tests) or associated with emotional distress (e.g., tests assessing suicidality and substance abuse). Instead of using LLMs with their default

© 2026 The Author(s). Distributed as a Hogrefe OpenMind article under the license CC BY 4.0 (https://creativecommons.org/licenses/by/4.0)

https://econtent.hogrefe.com/doi/pdf/10.1027/2698-1866/a000127 - Monday, May 04, 2026 6:52:24 AM - IP Address:2a02:8071:5ea2:59a0:19c4:8f2e:b4e6:d68f

https://econtent.hogrefe.com/doi/pdf/10.1027/2698-1866/a000127 - Monday, May 04, 2026 6:52:24 AM - IP Address:2a02:8071:5ea2:59a0:19c4:8f2e:b4e6:d68f

profile, a precise prompt defining target traits could enable the simulation of responses from diverse subpopulations, including individuals differing in demographic characteristics (e.g., age, gender, cultural background), members of underrepresented or vulnerable groups (e.g., persons with migration background, military personnel), individuals with varying levels of latent traits (e.g., intelligence or personality), and participants in different testing contexts (e.g., high- vs. low-stakes situations or under conditions prone to faking; see Griffith et al., 2007; Ziegler et al., 2015). In addition, LLMs may be able to adopt specific situational roles relevant to the construct being measured. For instance, reactions to experiences of injustice vary depending on whether a person is positioned as a victim, perpetrator, or observer (Groskurth et al., 2023), roles that could be simulated using LLMs. In summary, this approach may generate insights early in the process of test development, enable greater cost and time efficiency, and reduce the burden on human participants.

Theoretical Background

A key prerequisite for the approach described above is the ability to shape the output of LLMs to (1) answer psychometric test items formally correct and technically sound, (2) emulate a realistic and sample-specific distribution of outputs, and (3) reflect interactions between sample-specific and construct-specific characteristics. The proposed approach of using LLMs for pretesting items builds on a growing body of research that explores how LLMs respond to psychometric tests. Although psychometric tests and inventories are developed and intended for use in human populations, the idea to use them to evaluate the output of LLMs has been put forward repeatedly (e.g., Caron & Srivastava, 2022; Jiang et al., 2024; Karra et al., 2023; Pellert et al., 2024). However, this approach has also been met with skepticism, questioning the validity, technical reproducibility, and applicability of psychometric tests for evaluating the outputs of LLMs (Song et al., 2023; Sühr et al., 2024; Wang, Jiang, et al., 2023).

Given those concerns, another possible approach is to prompt an LLM to emulate certain characteristics, for example, high values in agreeableness, not evaluating the *baseline trait profile*, but inducing a *trait profile* (Caron & Srivastava, 2022; Jiang et al., 2024). By adjusting the trait profile, LLMs can be tailored to different user needs (Chen et al., 2024; Kong et al., 2024). Prior works have shown that traits related to latent constructs such as personality can be induced into the output of LLMs (Jiang et al., 2024; Serapio-García et al., 2023). In the context of item pretesting, using LLMs, such inductions of trait profiles are relevant for emulating a specific distribution of known traits, representative of a specific target population. In this strategy, the LLM and its trait profile are not the object of research but are used as a tool to flexibly replicate different populations for purposes such as item pretesting.

To enable this approach, LLMs must emulate trait profiles not just within a defined (e.g., prompted) construct, but in other related constructs as well. This *cross-construct influence* is a key requirement for using LLMs to administer pretests. A valid simulation of a specific trait profile (e.g., as a part of a specific population of interest) must therefore go beyond explicitly prompted behaviors (corresponding to defined constructs) and impact other constructs in a theory-consistent manner. This must be reflected in the output of the LLMs, for example, in answers to test items designed to measure a related but distinct (not explicitly prompted) construct. One way of validating such cross-construct influence would therefore be to investigate whether the induction of specific trait profiles also leads to measurable changes in items or test scores that aim to measure a construct that is not explicitly prompted, but theoretically and empirically connected in the human target population. This approach is equivalent to assessing the construct validity of a test by ascertaining convergent validity: We locate the latent variable of interest in the nomological network and check whether the data sets simulated by the LLM, which reflect the expected (response) patterns in one test, show (response) patterns in line with expectations in another test. To apply this approach, we examine the correlation between the personality trait agreeableness, a factor of the five-factor model of personality (Goldberg, 1992), and emotion understanding (EU), a facet of emotional intelligence (EI; Mayer et al., 1999, 2016).

The strength of this correlation provides an initial indicator of construct validity. Additionally, following Cronbach and Meehl's (1955) suggestion to assess group differences, we can further substantiate construct validity by comparing EU scores between (simulated) data sets of item responses characterized by low and high agreeableness. Following this argument, differences in mean EU scores between these extreme groups would serve as evidence of construct validity (Cronbach & Meehl, 1955, p. 8).

Personality and Emotion Understanding

The five-factor model of personality (Big Five, OCEAN traits) is a widely adopted and replicated factor structure of personality traits in humans (Goldberg, 1992; Paunonen, 2003; Schmitt et al., 2007) and is often used in works on trait profiles of LLMs. For example, Jiang et al. (2024) created different prompt configurations to induce personality traits into the output of LLMs by splitting each

© 2026 The Author(s). Distributed as a Hogrefe OpenMind article under the license CC BY 4.0 (https://creativecommons.org/licenses/by/4.0)

https://econtent.hogrefe.com/doi/pdf/10.1027/2698-1866/a000127 - Monday, May 04, 2026 6:52:24 AM - IP Address:2a02:8071:5ea2:59a0:19c4:8f2e:b4e6:d68f

five-factor trait into high and low expressions, and additionally prompted the LLMs to answer as male and female personas. When LLMs were then prompted to complete a questionnaire assessing the five factors (BFI-44; John et al., 1991; John et al., 2008), their responses aligned with the expected personality traits (Jiang et al., 2024). Other studies align with this finding (Caron & Srivastava, 2022; Serapio-García et al., 2023).

The second construct of interest is emotion understanding (EU), a factor of the four-factor model of emotional intelligence (EI; Salovey & Mayer, 1990) defined as the ability to label and reflect on emotions, as well as understanding their origins and effects (Joseph & Newman, 2010; Mayer et al., 2016). The other three factors in the four-factor model of EI (Perceiving, Facilitating, and Managing emotions) were not necessarily considered applicable to LLMs because they do not process internal cognitive or emotional states (Wang, Li, et al., 2023).

Different studies report a significant overlap between (trait) EI and the general factor of personality (GFP), a higher-order factor of personality that captures shared variance among personality traits such as the five-factor traits (Pérez-González & Sanchez-Ruiz, 2014; van der Linden et al., 2017). Schulte et al. (2004) reported that intelligence and personality, particularly agreeableness, explained $R^2 = .34$ ($R^2 = .41$ including sex) of the variance in EI, while other studies report an overlap of over 50% between combined five-factor traits and trait-EI (Petrides et al., 2010). Kokkinos and Voulgaridou (2024), among other studies (e.g., Petrides et al., 2010; van der Zee et al., 2002), showed that EI is positively associated with non-neuroticism five-factor model traits (openness, conscientiousness, extraversion, and agreeableness) and negatively correlated with neuroticism. Among these traits, agreeableness is of particular interest because it reflects tendencies toward empathy, cooperation, and consideration of others (Wilmot & Ones, 2022) – qualities that are conceptually central to EU.

The Present Study

In this study, we aim to provide an initial proof of concept for validating the induction of trait profiles in LLMs by applying the methods of convergent validation and extreme group validation to quantify implicit (not explicitly prompted) cross-construct influence. We therefore aim to (1) conceptually replicate existing attempts to induce LLMs with specific trait profiles based on contrasting gender and agreeableness (female-high, male-low) expressions and (2) validate the trait-induction by measuring and comparing scale scores on a theoretically and empirically related construct, namely EU, between the two contrasting groups. Gender was included alongside agreeableness to ensure that the extreme groups differed not only in personality but also in demographic characteristics, increasing a theoretically and empirically plausible contrast between the two extreme groups: Studies consistently report that agreeableness scores differ by gender, with females typically scoring higher than males (Feingold, 1994; Schmitt et al., 2008; Weisberg et al., 2011). Other studies examining gender and ability EI suggest that females, on average, score higher than males on ability-based EI measures, including EU subscales (Cabello et al., 2016; D'Amico & Geraci, 2022). Furthermore, a meta-analysis by Joseph and Newman (2010) found that gender had an effect of $d = 0.52$ on performance-based EI and $d = 0.31$ on EU, with both effects favoring females.

To the best of our knowledge, no study has been published on the relationship between an induced trait profile and EU in LLMs. To interpret the results of the validation as accurately as possible, we try to minimize the common method variance. In the following, we therefore do not consider data from different questionnaires, but vary the method in the sense of Campbell and Fiske (1959). As an indicator of convergent validity, we consider the correlation with a measured value that records a similar construct using a different method, namely a performance test. We measure EU with the Situational Evaluation of Complex Emotional Understanding test (SECEU), a test by Wang, Li, et al. (2023) that operates like a situational judgment test (cf. MacCann & Roberts, 2008).

Materials and Methods

Design

Since OpenAI's ChatGPT is one of the most prominent LLM interfaces in terms of total usage (Bailyn, 2025), we conducted the experiment using GPT-4, which served as the platform's underlying LLM at the time of data collection. Focusing on this widely benchmarked model ensures that our findings remain easily interpreted and comparable. The study followed a two-step approach: In the first step, the LLM completed a classic self-description questionnaire on the five-factor model of personality and a situational judgment test on EU, using a neutral group without any trait-profile induction (details below). The LLM was prompted to select the level of agreement with the provided statements, using zero-shot prompting. Zero-shot describes a scenario where no further context aside from the prompt itself is provided to the model, and it is expected to provide a result only relying on its general embedded *knowledge* (Kojima et al., 2023; Kong

© 2026 The Author(s). Distributed as a Hogrefe OpenMind article under the license CC BY 4.0 (https://creativecommons.org/licenses/by/4.0)

https://econtent.hogrefe.com/doi/pdf/10.1027/2698-1866/a000127 - Monday, May 04, 2026 6:52:24 AM - IP Address:2a02:8071:5ea2:59a0:19c4:8f2e:b4e6:d68f

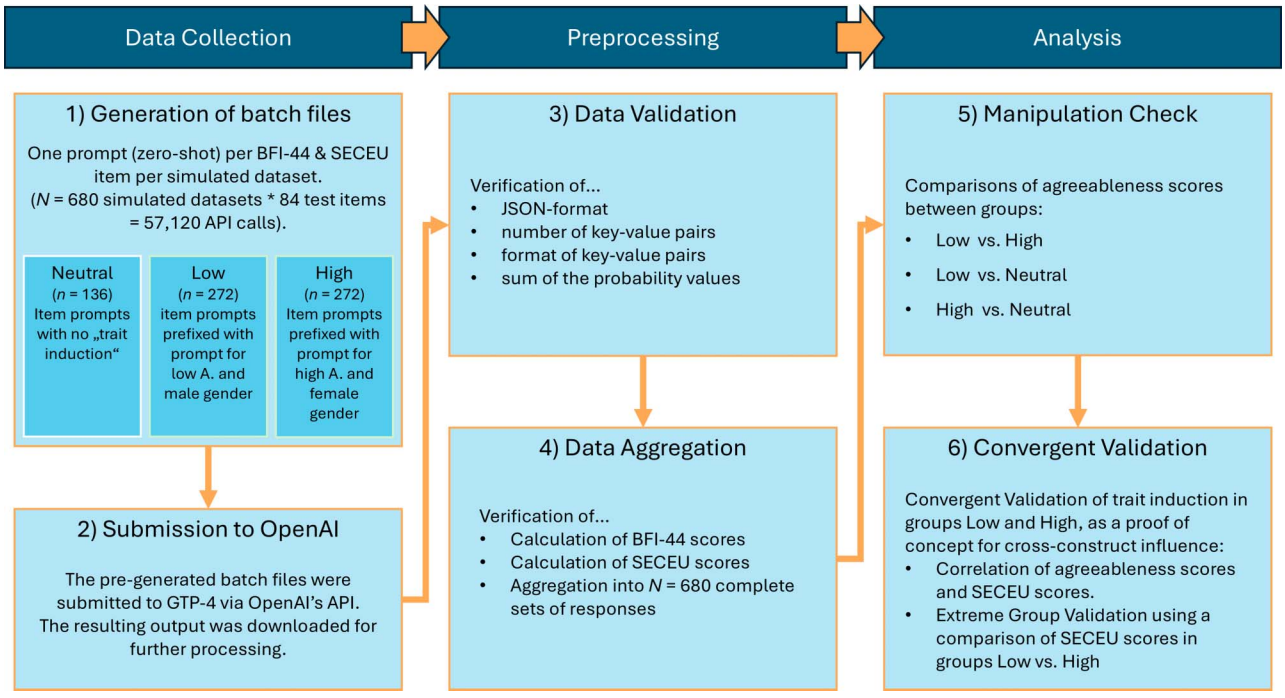

Figure 1. Overview of the study workflow and data processing. A = agreeableness score; BFI-44 = Big Five Inventory-44; SECEU = situational evaluation of complex emotional understanding.

et al., 2024). We deliberately chose to (1) not use few-shot prompting, (2) not use a fine-tuned LLM, or (3) not include varied, niche (e.g., open source) LLMs, to remain focused on a proof of concept, rather than comparing different LLMs, prompting strategies, or contexts.

In a second step, the method of extreme group validation (cf. Cronbach & Meehl, 1955) was applied. In this step, the model was prompted to output answers that fitted to one of two opposite levels of gender and agreeableness (female-high, male-low; see Table S1). Figure 1 shows an overview of the study workflow and data processing.

Materials

To assess agreeableness, the complete BFI-44 (John et al., 1991; John et al., 2008), a classic self-description questionnaire with 44 items and a five-point Likert agreement scale, was used. The consistency was $\alpha = .83$ over all five subscales. The agreeableness subscale of the BFI-44 consists of nine items with a consistency of $\alpha = .79$ (John et al., 2008). To generate the prompt based on the BFI-44, the original instructions were modified: The example statement was excluded, and the model was prompted to return a JSON object as the output and a *probability distribution* as the answer, following Pellert et al. (2024). This method is more in line with the second test (see the following paragraph) and allows the zero-shot output to convey more information per item.

Since only EU of the four-factor model of EI has been deemed as conceptually applicable in LLMs, Wang, Li, et al. (2023) developed a 40-item test to measure EU in human populations and LLMs, the Situational Evaluation of Complex Emotional Understanding (SECEU). The SECEU is a situational judgment test with 40 items (situations) and four response options each, for which a total of 10 points must be allocated. The SECEU was used according to the instructions reported by the authors. Table S1 shows the texts used to prompt the BFI-44 and SECEU.

Technical Considerations

All prompts were submitted to gpt-4-1106-preview in May 2024, using OpenAI's API via pregenerated batch files. To keep a decent level of variance, the temperature parameter was set to 0.7, following Jiang et al. (2024). Each item was submitted on its own (*zero shot*).

Sample

For step one (conceptual replication), $n = 136$ full data sets, including 84 items each (44 for the BFI-44 and 40 for the SECEU), were sampled. For step two (trait profile induction), a power analysis using G*Power (Faul et al., 2009) was conducted. Based on the gender-related effect size of

© 2026 The Author(s). Distributed as a Hogrefe OpenMind article under the license CC BY 4.0 (https://creativecommons.org/licenses/by/4.0)

https://econtent.hogrefe.com/doi/pdf/10.1027/2698-1866/a000127 - Monday, May 04, 2026 6:52:24 AM - IP Address:2a02:8071:5ea2:59a0:19c4:8f2e:b4e6:d68f

d = 0.31 for EU (favoring women, corrected for attenuation) reported by Joseph and Newman (2010), a power analysis for a two-tailed t test with independent groups resulted in a required sample size of n = 544 full data sets (272 per group; α = .05; power = .95). In total, N = 680 full data sets were generated.

Data Analysis

The resulting data were downloaded and aggregated using custom Python scripts to prepare for data analysis. Statistical calculations were done using IBM SPSS Statistics (version 29).

Preprocessing and Data Validation

The data provided by OpenAI was downloaded and checked for integrity by verifying the JSON format, the number and format of key-value pairs, and the sum of the provided probability values. In total, n = 55 item responses (0.1%; BFI-44: 37, SECEU: 18) were excluded from analysis (invalid JSON: 2; invalid key-value pairs: 27; incorrect sum: 26).

Item scores for the SECEU were calculated with formulas and norms generated by a human sample provided by Wang, Li, et al. (2023). Lower SECEU scores indicate a similar-to-human distribution of points across the possible answer options (i.e., less deviation from how humans tend to allocate points across the options; thus reflecting better performance in our evaluation). By contrast, larger SECEU scores indicate greater divergence from the human distribution. BFI item scores were calculated by taking a weighted average of the response options (1–5) using the output probabilities as weights. Higher values indicate a higher level of agreeableness. The data were then aggregated into complete sets of responses.

While the model consistently followed the prompted output structure, item-level responses (e.g., SECEU point distributions or BFI-44 probability weights) and scale-level scores varied due to the temperature setting of 0.7, resulting in some expected variability in item scores. Measures of variability of item and scale scores are shown in Table S2.

Questionnaire Scores and Statistical Calculations

SECEU and BFI-44 subscale scores were calculated following the authors' instructions and were calculated only if there were no missing values for any of the items required for the calculation. To evaluate the success of combined gender and agreeableness profile induction within the same construct in step one, the distribution of agreeableness scores was compared across all three groups (neutral, high, low). Significant omnibus results were followed by post hoc tests for all pairwise comparisons (low vs. high, low vs. neutral, and high vs. neutral; three comparisons in total). Bonferroni's method was used to adjust p-values for these comparisons.

To validate the induction of the *profile* using the convergent validity method, the correlation between the agreeableness and SECEU scores was calculated. This correlation was computed for the high and low groups only, as no profile induction was conducted in the neutral group. The remaining scores of the BFI-44 were not used for convergent or divergent validation, because our proof of concept focuses on the validation of cross-construct influences, which are a key requirement for accurately simulating populations, e.g., for pretesting items. The distribution of SECEU scores of groups high and low was compared to assess convergent validity via group differences. All tests were two-tailed and used a significance level of $p < .05$.

Results

To test normality and homogeneity of variances, the Shapiro–Wilk test and Levene's test were conducted. Results are shown in Tables S3 and S4. The Shapiro–Wilk test yielded significant results for the agreeableness scores (high and neutral group) and for the SECEU scores (low group), indicating a violation of normality in these groups. Levene's test yielded significant results for the agreeableness scores, indicating differences in variance across groups. Therefore, a nonparametric Kruskal–Wallis test was conducted to compare the distribution of agreeableness scores across groups to assess whether the induction of combined gender and agreeableness profiles was successful. The test indicated statistically significant differences across groups, $H(2) = 584.26$, $p < .001$, $\eta^2 = .86$.

Subsequently, post hoc tests (Dunn–Bonferroni tests) were conducted for all pairwise comparisons, and Bonferroni's method was used to adjust p-values for multiple comparisons. The results indicated statistically significant differences between all three groups (see Table 1), which provide evidence for successful profile induction.

As a measure of convergent validity, Spearman's rho was calculated between the agreeableness scores and SECEU scores. As expected, the analysis showed a significant negative correlation between the scores ($r_s = -.35$, $p < .001$, $n = 528$, 95% CI [−.43, −.27]). The results suggest a medium-sized effect according to Cohen (1992), which provides evidence for convergent validity, supporting the notion that induced variations in agreeableness and gender are meaningfully associated across the two measures and constructs.

© 2026 The Author(s). Distributed as a Hogrefe OpenMind article under the license CC BY 4.0 (https://creativecommons.org/licenses/by/4.0)

Table 1. Manipulation-check: Post hoc pairwise comparisons of the distribution of agreeableness scores across groups

Groups compared (*Mdn*)	D	*z*	p_{corr}	r_{rb}	95% CI
L (2.53) versus H (4.05)	405.97	24.171	< .001	−1.000[a]	[−1.000, −1.000]
L (2.53) versus N (3.54)	203.07	9.823	< .001	−.995	[−.996, −.994]
H (4.05) versus N (3.54)	202.90	9.815	< .001	.995	[.993, .996]

Note. D = Dunn's test statistic; *z* = standardized test statistic; p_{corr} = two-tailed asymptotic *p*-value, corrected for multiple (three) comparisons using Bonferroni's method; r_{rb} = effect size calculated as a rank-biserial correlation with $r_{rb} = 2(\overline{R}_1 - \overline{R}_2)/(n_1 + n_2)$, where $\overline{R}$ is the mean rank of the group based on pairwise comparisons and *n* denotes the sample size of each group; CI = 95% confidence interval approximated via Fisher's *z*-transformation, for $|r_{rb}| = 1.000$, the interval reflects the boundary of the effect space; N = neutral group ($n = 136$); L = low group ($n = 272$); H = high group ($n = 272$). All comparisons are significant at $p_{corr} \ll .001$. [a]Indicates total separation (complete stochastic dominance), where the score distributions of the two compared groups do not overlap, resulting in a maximal effect size of $|r_{rb}| = 1.000$ and a constant confidence interval.

Comparisons of the distribution of SECEU scores between groups high and low were conducted via a Mann–Whitney *U* test to further assess construct validity. The results indicated statistically significant differences between the high group ($n = 264$, $Mdn = 2.04$) and the low group ($n = 264$, $Mdn = 2.10$; $U = 18{,}995.00$, $z = -9.044$, $p < .001$, $r = .39$), with scores differing in the expected direction.

Discussion

This study provides an initial proof of concept for validating the induction of traits into the output of LLMs through the framework of construct validation laid out by Cronbach and Meehl (1955). This approach uses a theoretically and empirically supported connection between two constructs – here, agreeableness and emotion understanding (EU) – to provide evidence for a key prerequisite to use LLMs in test development (e.g., for item pretesting): cross-construct influence. The results indicate that GPT-4 can be prompted to simulate psychometric test-taking behaviors consistent with induced trait profiles and exhibit theoretically expected cross-construct influences on emotion understanding in a situational judgment test.

In recent years, an increasing amount of research has explicitly examined the potential of large language models (LLMs) to act as artificial respondents in psychological and social science research. For instance, studies have investigated whether LLMs can complete standardized personality inventories, revealing both their potential and their limitations (Pellert et al., 2024; Petrov et al., 2024; Wang et al., 2024; Zhao et al., 2025). In one psychometric analysis, researchers prompted LLMs to adopt different personas and respond to Big Five questionnaires, finding that psychometric properties of the generated data were highly influenced by the used model and prompt design, challenging the usefulness of LLM-generated data in psychological sciences (Petrov et al., 2024). Complementing this, a recently proposed psychometric benchmarking framework covered multiple psychological dimensions, suggesting that although LLMs can manifest a broad spectrum of psychological attributes, their self-reported traits can differ markedly from their behavior, revealing notable inconsistencies in their *psychological profiles* (Li et al., 2024). Moreover, multiple different works identified key challenges (e.g., measurement invariance, social-desirability bias, overanthropomorphizing) and outlined future research directions (e.g., adapting human psychological constructs for LLMs, applying item response theory; Argyle et al., 2023; Ye et al., 2025).

Our study extends this line of research by examining whether LLM-generated data can reproduce the structural properties that are characteristic of human response patterns. Using a multitrait multimethod framework, we investigate the extent to which the construct relations produced by the models align with the theoretically and empirically established nomological network of human data.

Theoretical and Practical Implications

By inducing combined gender and agreeableness profiles and observing a medium-sized correlation with emotion understanding scores, our results support the much-discussed notion that LLMs can be used as flexible tools to simulate latent traits in their output and are not mere reflectors of their *baseline* dispositions. This approach mirrors the framework of convergent validation laid out by Cronbach and Meehl (1955), situating simulated data sets and confirming that induced behaviors on one construct influence related constructs within the nomological network.

Our work builds on prior uses of technology for item generation – both traditional automatic item generation, which has primarily focused on item content creation using templates and cognitive models, and recent studies applying generative AI to create psychometric items. Rather

https://econtent.hogrefe.com/doi/pdf/10.1027/2698-1866/a000127 - Monday, May 04, 2026 6:52:24 AM - IP Address:2a02:8071:5ea2:59a0:19c4:8f2e:b4e6:d68f

© 2026 The Author(s). Distributed as a Hogrefe OpenMind article under the license CC BY 4.0 (https://creativecommons.org/licenses/by/4.0)

https://econtent.hogrefe.com/doi/pdf/10.1027/2698-1866/a000127 - Monday, May 04, 2026 6:52:24 AM - IP Address:2a02:8071:5ea2:59a0:19c4:8f2e:b4e6:d68f

than contributing another item-generation approach, this study investigates a key prerequisite for using LLMs such as GPT-4 to support the foundational stages of psychometric test development, through the use of simulated data sets of item response data. This could include preliminary checks of item and scale parameters by simulating responses from different trait profiles, providing an additional source of information for item evaluation.

By successfully inducing combined gender and high and low agreeableness profiles and observing significant cross-construct influence manifested in significant differences within EU-scores between groups, our results support the notion that LLMs could be used as versatile tools to simulate response behavior in a theoretically valid manner – not only with regard to the manipulated trait but also beyond it.

While our findings show that an LLM can reproduce theoretically expected and empirically frequently observed cross-construct patterns (e.g., between agreeableness, gender, and EU), this also underscores a possible limitation: the approach relies on existing theoretical and empirical knowledge to evaluate the profile induction and cross-construct influence. As such, it is not inherently designed to generate new theory in domains where little is known; rather, its utility lies in testing whether an LLM can reflect known nomological relations when prompted to simulate specific trait profiles. Other limitations of the approach presented here are discussed below.

Limitations

In the present study, we employed a single LLM (gpt-4-1106-preview) and relied upon a zero-shot prompting approach for the initial concept validation. This deliberate use of a straightforward manipulation represents a particularly stringent test: if theoretically plausible response patterns can already be generated under such minimal prompting conditions, it is reasonable to assume that more elaborate prompting strategies (e.g., Jiang et al., 2024) would perform at least equally well, if not better. We deliberately limited the study to one LLM to reduce complexity and isolate the effects of the basic prompting approach. This ensured that potential confounds arising from variability between different LLM architectures were avoided, thereby providing a clearer test of the core assumption.

LLMs are sensitive to various technical parameters that influence their outputs. Key factors include prompt wording, temperature settings, model version, context window size, training data, and model architecture. Prompt engineering is crucial, as the phrasing of prompts can significantly affect the model's responses. Temperature settings control the randomness of outputs; lower temperatures yield more deterministic responses, while higher temperatures increase *creativity*, which has important implications for the variance of simulated item scores. Different model versions may produce varying outputs due to updates in training data and architecture. The context window size determines how much prior information the model considers, impacting coherence and relevance. Finally, the underlying architecture, such as transformer-based designs, dictates the model's capacity and efficiency (Ferraris et al., 2025). Training data diversity and quality directly affect the model's knowledge base and potential biases (Gallegos et al., 2024; Kotek et al., 2023; Navigli et al., 2023). Future studies should systematically examine the effects of both more sophisticated prompting techniques and different LLM architectures.

A central methodological limitation of the present study is the confounding of gender and agreeableness within the trait-induction prompts. Because the high-agreeableness condition was always paired with a female identity and the low-agreeableness condition with a male identity, we cannot determine the relative contribution of personality trait versus gender in shaping the model's responses. While this alignment was deliberate – aimed at constructing two coherent and strongly contrasting profiles for an initial proof-of-concept – it restricts the interpretability of the underlying mechanisms of *trait induction*. Future research should systematically disentangle these factors by including agreeableness-only prompts, gender-only prompts, fully crossed designs, and more nuanced conditions (e.g., a medium-agreeableness condition). Such extensions would allow researchers to quantify the degree to which LLM outputs are driven by specific trait profiles, demographic variables, or their interaction.

Another constraint lies in the generalizability of the simulated data to human populations. While significant cross-construct influence could be shown, the extent of this influence may not reflect interactions observed in human populations, where additional cognitive, social, and contextual factors may come into play. One illustrative finding was the relatively high agreeableness score observed in the neutral group, where no profile induction took place. This pattern may reflect a general tendency of the model to generate socially desirable responses in the absence of more directive prompting. Another deviation worth noting was that intercorrelations calculated across extreme groups between Big Five traits (Table S5) were higher than typically observed in human data (cf. van der Linden et al., 2010). Notably, in our data set, the four non-neuroticism five-factor model traits and neuroticism were positively associated, whereas in human participants, this association would typically be negative (van der Linden et al., 2010). It is reasonable to assume that both deviations are partly due to the method we chose: First, of the Big Five traits, only agreeableness was explicitly prompted. Within the scope of the current study, we could not

© 2026 The Author(s). Distributed as a Hogrefe OpenMind article under the license CC BY 4.0 (https://creativecommons.org/licenses/by/4.0)

https://econtent.hogrefe.com/doi/pdf/10.1027/2698-1866/a000127 - Monday, May 04, 2026 6:52:24 AM - IP Address:2a02:8071:5ea2:59a0:19c4:8f2e:b4e6:d68f

investigate differential effects this design choice might have had on the LLM's output, but since LLMs are highly susceptible to changes at the prompt level, this is a likely effect. Second, the present study relied on extreme-group comparisons (between two groups). While this approach facilitated a clear proof-of-concept, it limits our understanding of whether the LLM can capture more subtle or moderate differences in trait levels. Including a (or multiple variations of) medium-agreeableness condition(s) in future works would provide a more fine-grained assessment of the model's sensitivity to gradations in trait profiles.

A key prerequisite for LLMs to be useful in test development would be their ability to answer psychometric test items in a formally correct and technically sound manner. Even though this worked for the most part in the present study, approximately 0.1% of the output did not comply with the specified response format. In our case, for example, the model has returned outputs such as the following in several cases: "I am an AI and do not have personal experiences or a personality." These deviations often reflected system-level safeguards – mechanisms designed to prevent inappropriate, personal, or sensitive content – which overrode the model's original response. Such safeguards, particularly in closed-source models like the one used here, may inadvertently interfere with the intended data generation process. This highlights the need for awareness that postresponse filtering or modification mechanisms could distort data when using LLMs. One way to mitigate this issue would be to use fully open-source LLMs. This would allow researchers to ensure that the model's outputs are not unintentionally altered by post-generation filtering systems, thereby preserving the integrity of the generated data.

Finally, we prompted the model to output probability distributions, which deviates from the procedure most psychometric tests use when testing human participants. Requesting the LLM to generate a probability distribution across response options, rather than a single categorical choice, allowed for a more fine-grained capture of the model's output. This approach preserved information about relative preferences between response options, avoided arbitrary thresholding decisions, and facilitated more robust statistical analyses based on the full spectrum of the model's response tendencies. Future studies should systematically examine how this approach compares to methods where the LLM is instructed to select a single fixed response option.

Implications for Future Research

Future research should address the limitations presented above. First, comparative studies across multiple LLM architectures and model versions should evaluate the cross-construct effects of trait-related manipulations of LLMs and identify optimal platforms for psychometric applications. Second, incorporating few-shot prompting, chain-of-thought methods, or fine-tuning of LLMs on human response data may improve the fidelity of trait and complex profile inductions and reduce method-dependent variance in simulated item data. Third, an extension of trait induction to other nomologically related constructs and target populations should be explored in future work.

Since it can be assumed that there are some fundamental limitations to the applicability of the method, future research should examine the extent to which the method can be used with a broader range of constructs. While the present study focuses on agreeableness and EU for replicability and comparability purposes, it is not reasonable to assume that LLMs are equally viable to simulate individuals with particular characteristics in other domains. For example, models trained on biased data sets will most likely produce biased results (e.g., Gallegos et al., 2024; Navigli et al., 2023). For future research, rigorous psychometric evaluation of LLM-simulated data – like examining item response theory fit indices, measurement invariance tests, and differential item functioning analyses – will be crucial to establishing best practices and standardizing protocols for AI- and LLM-driven pretesting.

Furthermore, future research should investigate the feasibility of extending the dominant variable-centered approach by incorporating a person-centered perspective through the simulation of target populations with LLMs via induced trait profiles. Such work could examine both configurations of traits and potential interaction effects among multiple predictors, possibly providing a more nuanced understanding of how psychological constructs co-occur within individuals. In addition, future studies should examine whether simulation-based approaches can be used to test theoretical models of complex dynamics, such as compensatory or synergistic interactions, and whether a person-centered perspective could help identify patterns of traits associated with specific outcomes to inform more tailored interventions.

Moreover, future research should extensively contrast LLM-based responses with human data. While such comparisons are methodologically challenging, they could provide valuable insights into the alignment of LLM simulations with human behavior. Systematically integrating human–model comparisons in future studies could help to streamline the process of validating simulated populations and potentially refine trait-induction methods to better reflect a similar-to-human distribution of traits.

© 2026 The Author(s). Distributed as a Hogrefe OpenMind article under the license CC BY 4.0 (https://creativecommons.org/licenses/by/4.0)

https://econtent.hogrefe.com/doi/pdf/10.1027/2698-1866/a000127 - Monday, May 04, 2026 6:52:24 AM - IP Address:2a02:8071:5ea2:59a0:19c4:8f2e:b4e6:d68f

Conclusion

In summary, this study offers a preliminary proof-of-concept for integrating generative AI into the scientific process of psychological assessment development. By demonstrating that an LLM, when prompted to simulate a particular expression of a personality trait and gender, can exhibit theoretically plausible behavior not only with respect to the prompted trait but also with regard to another theoretically and empirically related construct, we offer an initial proof of concept for this key prerequisite of using LLMs in psychometric test development. Harnessing these capabilities has the potential not only to accelerate assessment design but also to make psychological research more efficient, inclusive, and ethically responsible. Nevertheless, further research is required to refine these methods, explore their boundaries and inherent biases, and ensure their robust application across diverse psychological constructs and testing contexts.

References

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*(3), 337–351. https://doi.org/10.1017/pan.2023.2

Bailyn, E. (2025, April 17). *Top generative AI chatbots by market share: April 2025*. FirstPageSage. https://firstpagesage.com/reports/top-generative-ai-chatbots/

Bormuth, J. R. (1970). *On the theory of achievement test items*. University of Chicago Press.

Borsboom, D. (2006). When does measurement invariance matter? *Medical Care*, *44*(11 suppl 3), S176–S181. https://doi.org/10.1097/01.mlr.0000245143.08679.cc

Cabello, R., Sorrel, M. A., Fernández-Pinto, I., Extremera, N., & Fernández-Berrocal, P. (2016). Age and gender differences in ability emotional intelligence in adults: A cross-sectional study. *Developmental Psychology*, *52*(9), 1486–1492. https://doi.org/10.1037/dev0000191

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Caron, G., & Srivastava, S. (2022). *Identifying and manipulating the personality traits of language models*. arXiv. https://doi.org/10.48550/arXiv.2212.10276

Chen, J., Wang, X., Xu, R., Yuan, S., Zhang, Y., Shi, W., Xie, J., Li, S., Yang, R., Zhu, T., Chen, A., Li, N., Chen, L., Hu, C., Wu, S., Ren, S., Fu, Z., & Xiao, Y. (2024). *From persona to personalization: A survey on role-playing language agents*. arXiv. https://doi.org/10.48550/arXiv.2404.18231

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science*, *1*(3), 98–101. https://doi.org/10.1111/1467-8721.ep10768783

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

D'Amico, A., & Geraci, A. (2022). Sex differences in emotional and meta-emotional intelligence in pre-adolescents and adolescents. *Acta Psychologica*, *227*, Article 103594. https://doi.org/10.1016/j.actpsy.2022.103594

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., Jones, M., Krettek-Cobb, D., Lai, L., JonesMitchell, N., Ong, D. C., Dweck, C. S., Gross, J. J., & Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology*, *2*, 688–701. https://doi.org/10.1038/s44159-023-00241-5

Edin, Y., Rosenplänter, L., & Kersting, M. (2026). *Electronic Supplementary Material to: Convergent Validation of Trait Induction in LLMs: A Step Toward Integrating AI in Foundational Stages of Psychometric Test Development*. https://doi.org/10.17605/OSF.IO/JNMEU

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using g*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Feingold, A. (1994). Gender differences in personality: A meta-analysis. *Psychological Bulletin*, *116*(3), 429–456. https://doi.org/10.1037/0033-2909.116.3.429

Ferraris, A. F., Audrito, D., Di Caro, L., & Poncibò, C. (2025). The architecture of language: Understanding the mechanics behind LLMs. *Cambridge Forum on AI: Law and Governance*, *1*, Article e11. https://doi.org/10.1017/cfl.2024.16

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. *Computational Linguistics*, *50*(3), 1097–1179. https://doi.org/10.1162/coli_a_00524

Gierl, M. J., & Haladyna, T. M. (2013). *Automatic item generation: Theory and practice*. Routledge.

Gierl, M. J., & Lai, H. (2012). The role of item models in automatic item generation. *International Journal of Testing*, *12*(3), 273–298. https://doi.org/10.1080/15305058.2011.635830

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, *4*(1), 26–42. https://doi.org/10.1037//1040-3590.4.1.26

Griffith, R. L., Chmielowski, T., & Yoshita, Y. (2007). Do applicants fake? An examination of the frequency of applicant faking behavior. *Personnel Review*, *36*(3), 341–355. https://doi.org/10.1108/00483480710731310

Groskurth, K., Beierlein, C., Nießen, D., Baumert, A., Rammstedt, B., & Lechner, C. M. (2023). An English-language adaptation and validation of the Justice Sensitivity Short Scales-8 (JSS-8). *PLOS One*, *18*(11), Article e0293748. https://doi.org/10.1371/journal.pone.0293748

Holland, P. W., & Wainer, H. (Eds.), (2009) *Differential item functioning*. Routledge. https://doi.org/10.4324/9780203357811

Jiang, H., Zhang, X., Cao, X., Breazeal, C., Roy, D., & Kabbara, J. (2024). *PersonaLLM: Investigating the ability of large language models to express personality traits*. arXiv. https://doi.org/10.48550/arXiv.2305.02547

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The Big Five Inventory: Versions 4a and 54*. University of California. Berkeley, Institute of Personality and Social Research. https://doi.org/10.1037/t07550-000

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In O. P. John (Ed.), *Handbook of personality: Theory and research* (3rd ed., pp. 114–158). Guilford Press.

Joseph, D. L., & Newman, D. A. (2010). Emotional intelligence: An integrative meta-analysis and cascading model. *The Journal of Applied Psychology*, *95*(1), 54–78. https://doi.org/10.1037/a0017286

© 2026 The Author(s). Distributed as a Hogrefe OpenMind article under the license CC BY 4.0 (https://creativecommons.org/licenses/by/4.0)

https://econtent.hogrefe.com/doi/pdf/10.1027/2698-1866/a000127 - Monday, May 04, 2026 6:52:24 AM - IP Address:2a02:8071:5ea2:59a0:19c4:8f2e:b4e6:d68f

Karra, S. R., Nguyen, S. T., & Tulabandhula, T. (2023). *Estimating the personality of white-box language models*. arXiv. https://doi.org/10.48550/arXiv.2204.12000

Ke, L., Tong, S., Cheng, P., & Peng, K. (2025). Exploring the frontiers of LLMs in psychological applications: A comprehensive review. *Artificial Intelligence Review*, *58*, Article 305. https://doi.org/10.1007/s10462-025-11297-5

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2023). *Large language models are zero-shot reasoners*. arXiv. https://doi.org/10.48550/arXiv.2205.11916

Kokkinos, C. M., & Voulgaridou, I. (2024). Emotional intelligence across the personality spectrum: A study of university students' personality profiles. *Personality and Individual Differences*, *222*, Article 112574. https://doi.org/10.1016/j.paid.2024.112574

Kong, A., Zhao, S., Chen, H., Li, Q., Qin, Y., Sun, R., Zhou, X., Wang, E., & Dong, X. (2024). *Better zero-shot reasoning with role-play prompting*. arXiv. https://doi.org/10.48550/arXiv.2308.07702

Kotek, H., Dockum, R., & Sun, D. (2023). Gender bias and stereotypes in large language models. In M. Bernstein, S. Savage, & A. Bozzon (Eds.), *Proceedings of the ACM collective intelligence conference* (pp. 12–24). ACM. https://doi.org/10.1145/3582269.3615599

Lee, P., Fyffe, S., Son, M., Jia, Z., & Yao, Z. (2023). A paradigm shift from "human writing" to "machine generation" in personality test development: An application of state-of-the-art natural language processing. *Journal of Business and Psychology*, *38*(1), 163–190. https://doi.org/10.1007/s10869-022-09864-6

Li, Y., Huang, Y., Wang, H., Cheng, Y., Zhang, X., Zou, J., & Sun, L. (2024). *Evaluating large language models with psychometrics*. arXiv. https://doi.org/10.48550/arXiv.2406.17675

Liu, Y., Bhandari, S., & Pardos, Z. A. (2025). Leveraging LLM respondents for item evaluation: A psychometric analysis. *British Journal of Educational Technology*, *56*(3), 1028–1052. https://doi.org/10.1111/bjet.13570

MacCann, C., & Roberts, R. D. (2008). New paradigms for assessing emotional intelligence: Theory and data. *Emotion (Washington, D.C.)*, *8*(4), 540–551. https://doi.org/10.1037/a0012746

Mayer, J. D., Caruso, D. R., & Salovey, P. (1999). Emotional intelligence meets traditional standards for an intelligence. *Intelligence*, *27*(4), 267–298. https://doi.org/10.1016/S0160-2896(99)00016-1

Mayer, J. D., Caruso, D. R., & Salovey, P. (2016). The ability model of emotional intelligence: Principles and updates. *Emotion Review*, *8*(4), 290–300. https://doi.org/10.1177/1754073916639667

Navigli, R., Conia, S., & Ross, B. (2023). Biases in large language models: Origins, inventory, and discussion. *Journal of Data and Information Quality*, *15*(2), 1–21. https://doi.org/10.1145/3597307

Paunonen, S. V. (2003). Big Five factors of personality and replicated predictions of behavior. *Journal of Personality and Social Psychology*, *84*(2), 411–424. https://doi.org/10.1037/0022-3514.84.2.411

Pellert, M., Lechner, C. M., Wagner, C., Rammstedt, B., & Strohmaier, M. (2024). Ai Psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *19*(5), 808–826. https://doi.org/10.1177/17456916231214460

Pérez-González, J. C., & Sanchez-Ruiz, M.-J. (2014). Trait emotional intelligence anchored within the Big Five, Big Two and Big One frameworks. *Personality and Individual Differences*, *65*, 53–58. https://doi.org/10.1016/j.paid.2014.01.021

Petrides, K. V., Vernon, P. A., Schermer, J. A., Ligthart, L., Boomsma, D. I., & Veselka, L. (2010). Relationships between trait emotional intelligence and the Big Five in the Netherlands. *Personality and Individual Differences*, *48*(8), 906–910. https://doi.org/10.1016/j.paid.2010.02.019

Petrov, N. B., Serapio-García, G., & Rentfrow, J. (2024). *Limited ability of LLMs to simulate human psychological behaviours: A psychometric analysis*. arXiv. https://doi.org/10.48550/arXiv.2405.07248

Salovey, P., & Mayer, J. D. (1990). Emotional intelligence. *Imagination, Cognition and Personality*, *9*(3), 185–211. https://doi.org/10.2190/DUGG-P24E-52WK-6CDG

Schmitt, D. P., Allik, J., McCrae, R. R., & Benet-Martínez, V. (2007). The geographic distribution of big five personality traits. *Journal of Cross-Cultural Psychology*, *38*(2), 173–212. https://doi.org/10.1177/0022022106297299

Schmitt, D. P., Realo, A., Voracek, M., & Allik, J. (2008). Why can't a man be more like a woman? Sex differences in Big Five personality traits across 55 cultures. *Journal of Personality and Social Psychology*, *94*(1), 168–182. https://doi.org/10.1037/0022-3514.94.1.168

Schulte, M. J., Ree, M. J., & Carretta, T. R. (2004). Emotional intelligence: Not much more than g and personality. *Personality and Individual Differences*, *37*(5), 1059–1068. https://doi.org/10.1016/j.paid.2003.11.014

Serapio-García, G., Safdari, M., Crepy, C., Sun, L., Fitz, S., Abdulhai, M., Faust, A., & Matarić, M. (2023). *Personality traits in large language models*. Research Square. https://doi.org/10.21203/rs.3.rs-3296728/v1

Song, X., Gupta, A., Mohebbizadeh, K., Hu, S., & Singh, A. (2023). *Have large language models developed a personality? Applicability of self-assessment tests in measuring personality in LLMs*. arXiv. https://doi.org/10.48550/arXiv.2305.14693

Sühr, T., Dorner, F. E., Samadi, S., & Kelava, A. (2024). *Challenging the validity of personality tests for large language models*. arXiv. https://doi.org/10.48550/arXiv.2311.05297

Tan, B., Armoush, N., Mazzullo, E., Bulut, O., & Gierl, M. J. (2024). *A review of automatic item generation techniques leveraging large language models*. OSF Preprints. https://doi.org/10.35542/osf.io/6d8tj

van der Linden, D., Pekaar, K. A., Bakker, A. B., Schermer, J. A., Vernon, P. A., Dunkel, C. S., & Petrides, K. V. (2017). Overlap between the general factor of personality and emotional intelligence: A meta-analysis. *Psychological Bulletin*, *143*(1), 36–52. https://doi.org/10.1037/bul0000078

van der Linden, D., te Nijenhuis, J., & Bakker, A. B. (2010). The general factor of personality: A meta-analysis of Big Five intercorrelations and a criterion-related validity study. *Journal of Research in Personality*, *44*(3), 315–327. https://doi.org/10.1016/j.jrp.2010.03.003

van der Zee, K., Thijs, M., & Schakel, L. (2002). The relationship of emotional intelligence with academic intelligence and the Big Five. *European Journal of Personality*, *16*(2), 103–125. https://doi.org/10.1002/per.434

Wang, X., Jiang, L., Hernandez-Orallo, J., Stillwell, D., Sun, L., Luo, F., & Xie, X. (2023). *Evaluating general-purpose AI with psychometrics*. arXiv. https://doi.org/10.48550/arXiv.2310.16379

Wang, X., Li, X., Yin, Z., Wu, Y., & Liu, J. (2023). Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, *17*. https://doi.org/10.1177/18344909231213958

Wang, P., Zou, H., Yan, Z., Guo, F., Sun, T., Xiao, Z., & Zhang, B. (2024). *Not yet: Large language models cannot replace human respondents for psychometric research*. OSF Preprints. https://doi.org/10.31219/osf.io/rwy9b

Weisberg, Y. J., Deyoung, C. G., & Hirsh, J. B. (2011). Gender differences in personality across the ten aspects of the Big Five. *Frontiers in Psychology*, *2*, Article 178. https://doi.org/10.3389/fpsyg.2011.00178

Wilmot, M. P., & Ones, D. S. (2022). Agreeableness and its consequences: A quantitative review of meta-analytic findings. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc*, *26*(3), 242–280. https://doi.org/10.1177/10888683211073007

© 2026 The Author(s). Distributed as a Hogrefe OpenMind article under the license CC BY 4.0 (https://creativecommons.org/licenses/by/4.0)

Ye, H., Jin, J., Xie, Y., Zhang, X., & Song, G. (2025). *Large language model psychometrics: A systematic review of evaluation, validation, and enhancement*. arXiv. https://doi.org/10.48550/arXiv.2505.08245

Zhao, C., Habule, M., & Zhang, W. (2025). Large language models (LLMs) as research subjects: Status, opportunities and challenges. *New Ideas in Psychology*, *79*, Article 101167. https://doi.org/10.1016/j.newideapsych.2025.101167

Ziegler, M., Maaß, U., Griffith, R., & Gammon, A. (2015). What is the nature of faking? Modeling distinct response patterns and quantitative differences in faking at the same time. *Organizational Research Methods*, *18*(4), 679–703. https://doi.org/10.1177/1094428115574518

History
Received April 30, 2025
Revision received March 12, 2026
Accepted March 17, 2026
Published online May 4, 2026

Section: Methodological Topics in Assessment

Acknowledgments
This study utilizes data originally collected as part of Yasin Edin's master's thesis, which was preregistered with, and submitted to FernUniversität Hagen, Hagen, Germany. Martin Kersting served as the second reviewer for the thesis.

Conflict of Interest
All authors declare no conflict of interest.

Authorship
Yasin Edin and Leon Rosenplänter share first authorship.
Yasin Edin, data curation, methodology, software, writing – review & editing; Leon Rosenplänter, data curation, formal analysis, methodology, software, visualization, writing – original draft; Martin Kersting, conceptualization, methodology, supervision, writing – review & editing. All authors approved the final version of the article.

Open Science
We report how we determined our sample size, all data exclusions (if any), all data inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all measures in the study, and all analyses including all tested models. If we use inferential tests, we report exact *p*-values, effect sizes, and 95% confidence or credible intervals.

Open Analytic Code: The code used to process the data and generate all reported results is available at https://osf.io/jnmeu/ (Edin, Rosenplänter, & Kersting, 2026).

Open Data: The information needed to reproduce all of the reported results is available at https://osf.io/jnmeu/ (Edin, Rosenplänter, & Kersting, 2026).

Open Materials: The information needed to reproduce all of the reported methodology is available at https://osf.io/jnmeu/ (Edin, Rosenplänter, & Kersting, 2026).

Preregistration and Analysis Plan: This study was not preregistered.

The online supplementary materials are available at https://osf.io/jnmeu/ (Edin, Rosenplänter, & Kersting, 2026).

ORCID
Yasin Edin
https://orcid.org/0009-0006-5494-7277
Leon Rosenplänter
https://orcid.org/0009-0001-4961-2281
Martin Kersting
https://orcid.org/0000-0003-2501-5287

Leon Rosenplänter
Department of Psychological Assessment
Justus-Liebig-University Giessen
Otto-Behaghel-Strasse 10F
35394 Giessen
Germany
leon.rosenplaenter@psychol.uni-giessen.de

https://econtent.hogrefe.com/doi/pdf/10.1027/2698-1866/a000127 - Monday, May 04, 2026 6:52:24 AM - IP Address:2a02:8071:5ea2:59a0:19c4:8f2e:b4e6:d68f

© 2026 The Author(s). Distributed as a Hogrefe OpenMind article under the license CC BY 4.0 (https://creativecommons.org/licenses/by/4.0)