Psychometric Item Validation Using Virtual Respondents with Trait-Response Mediators

Sungjib Lim[1] **Woojung Song**[1] **Eun-Ju Lee**[2,3] **Yohan Jo**[1*]
[1]Graduate School of Data Science, Seoul National University
[2]Department of Communication, Seoul National University
[3]Interdisciplinary Program in Artificial Intelligence, Seoul National University
{saint7451,opusdeisong,eunju0204,yohan.jo}@snu.ac.kr

Abstract

As psychometric surveys are increasingly used to assess the traits of large language models (LLMs), the need for scalable survey item generation suited for LLMs has also grown. A critical challenge here is ensuring the construct validity of generated items, i.e., whether they truly measure the intended trait. Traditionally, this requires costly, large-scale human data collection. To make it efficient, we present a framework for virtual respondent simulation using LLMs. Our central idea is to account for mediators: factors through which the same trait can give rise to varying responses to a survey item. By simulating respondents with diverse mediators, we identify survey items that yield responses robustly correlated with intended traits across these mediators. Experiments on three psychological trait theories (Big5, Schwartz, VIA) show that our mediator generation methods and simulation framework effectively identify high-validity items. LLMs demonstrate the ability to generate plausible mediators from trait definitions and to simulate respondent behavior for item validation. Our problem formulation, metrics, methodology, and dataset open a new direction for cost-efficient survey development and a deeper understanding of how LLMs simulate human survey responses. We release our dataset and code to support future work.[1]

1 Introduction

Recently, researchers have begun to employ psychological surveys to understand the behaviors of LLMs, in terms of embedded values, safety, and more (Miotto et al., 2022; Ye et al., 2025). However, because administering a small set of survey items designed for humans to LLMs is suboptimal (Huang et al., 2024), another stream of research has focused on automatic survey generation, particularly on expanding and augmenting survey items at scale (Lee et al., 2025). This need for survey generation extends to traditional human psychometrics as well. In particular, established surveys are often refined or shortened over time to maintain reliability and efficiency (Lindeman and Verkasalo, 2005; Rammstedt and John, 2007; Costa and McCrae, 2008).

A primary challenge in this task lies in ensuring the quality of automatically generated items. Existing studies focus largely on the reliability of generated items, that is, identifying items that elicit consistent responses from LLMs across varying task instructions, answer choice labels, or choice orders (Lee et al., 2025; Huang et al., 2024). However, another critical and currently underexplored criterion for survey items is **construct validity**: how well an item measures the trait it is intended to measure (Sürücü and Maslakci, 2020). To ensure the validity of survey items, psychometricians typically recruit a large number of human respondents across diverse cultures and evaluate new items by examining their correlations with target traits or well-established items (Weber et al., 2002; McCrae and Costa Jr, 1997). This process is logistically challenging and costly. Hence, our work seeks to validate survey items using virtual respondents based on LLMs.

Our central hypothesis is that the core role of large-scale human respondents in survey item validation is to test the robustness of items to diverse **mediators**: factors through which a psychological trait gives rise to *varying* responses to a survey item. As illustrated in Figure 1, while the survey item "I like attending social events" may seem to measure extraversion, extraverted respondents with certain mediators (e.g., already having a lot of friends) may produce responses with low

* Corresponding author.
[1]https://github.com/holi-lab/Psychometric-Item-Validation

Transactions of the Association for Computational Linguistics, vol. 14, pp. 1390–1412, 2026. https://doi.org/10.1162/TACL.a.733
Action Editor: Emiel Krahmer. Submission batch: 7/2025; Revision batch: 1/2026; Published 7/2026.

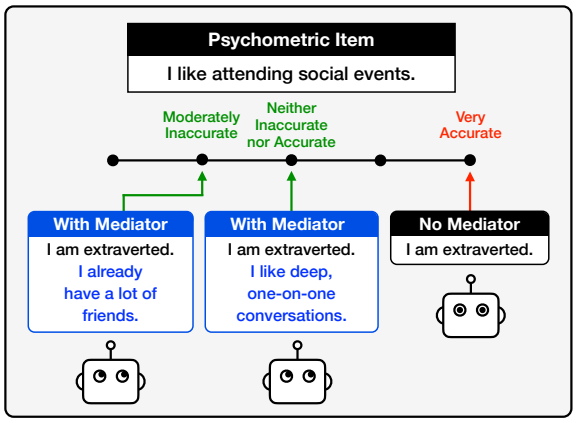

Figure 1: Role of mediators in survey item evaluation.

correlation to the trait, threatening the item's validity. Therefore, it is critical to identify potential mediators and incorporate them into respondent simulation to uncover items that robustly measure intended traits.

Specifically, the cognitive-affective personality system (CAPS) theory (Mischel and Shoda, 1995) posits that the same situation can elicit different behaviors depending on individual-level mediators, such as a person's goals or interpretation of the situation. We extend this view to traits, proposing that these factors also influence how one's traits translate into observable behaviors. Since psychological survey items typically ask about the respondent's likelihood of engaging in certain behaviors, accurate trait inference from responses requires—under our view—that those behaviors be robust to such mediators. (Note that the "mediators" included in our methodology may encompass both mediators and moderators in the strict sense of causal inference. However, we use the term "mediators" for consistency with the CAPS theory, and establishing exact causal relationships is not our primary focus.)

Against this backdrop, we propose a framework for psychometric survey generation and evaluation using mediator-guided virtual respondents (§3). The framework consists of five stages: (1) Select traits for inclusion in a survey, based on well-established psychological trait theories (e.g., Big Five, Schwartz's theory of basic values, VIA). (2) Construct large-scale initial item pools based on the definitions of these traits. (3) Generate mediators, which is our main contribution, using strategies like generation based on trait definitions, generation using external references (e.g., survey items, common value lists), and real human demographic data. (4) Incorporate mediators and persona profiles into prompts for LLM-based virtual respondents and run survey simulations. (5) Select survey items based on their measured validity.

According to our evaluation against item selection based on real human responses (§6), mediator-guided simulation was able to identify highly valid item sets, ranking in the top 1% (Big5) to 13% (Schwartz and VIA) of the distribution of all possible selections. Among mediator generation strategies, letting LLMs generate mediators based solely on trait definitions either freely or with guidance from the CAPS framework performed best. These findings suggest that LLMs have the ability to effectively identify mediators and simulate psychological assessments. We also find that scaling the number of virtual respondents improves performance and that the framework performs consistently across different LLMs. Please note that we do not claim our framework accurately replicates the human psychology process involved in responding to survey items. Instead, our focus is on leveraging diverse mediators and their correlations with trait scores to identify items that measure traits robustly.

Our contributions are as follows:

- We formulate a novel and important problem: assessing the validity of survey items, along with psychometrically grounded metrics.
- We introduce the concept of mediators into validity assessment and demonstrate their importance. We also show LLMs' ability to generate effective mediators and simulate virtual respondents guided by these mediators.
- We publicly release our dataset of survey items along with human and LLM responses, establishing a benchmark for future research.

2 Related Works

Psychological Assessment for LLMs and Humans. Psychological assessment is an essential field for understanding the behaviors of LLMs and human nature. Researchers have applied psychometric surveys to evaluate LLMs' responses in terms of psychological constructs, such as personality (Jiang et al., 2023; Miotto et al., 2022) and value orientations (Han et al., 2025; Hadar Shoval et al., 2024). Psychologists have provided foundational resources for this research through various surveys designed to assess human personality

(e.g., Big5 (Goldberg, 1999), MBTI (Jung and Beebe, 2016)), emotions (Beck et al., 1988, 1996), and behaviors (Weber et al., 2002).

Survey Expansion and Reduction. Psychological surveys often undergo refinement through processes of expansion and reduction to improve their applicability and efficiency. For instance, researchers have expanded existing surveys with additional items that are more suitable for assessing LLMs (Jiang et al., 2023; Han et al., 2025) because of the suboptimality of applying human survey items to LLMs (Huang et al., 2024). Surveys are also often reduced or shortened to mitigate the cognitive burden on respondents and improve efficiency, especially in large-scale data collection (Lindeman and Verkasalo, 2005; Rammstedt and John, 2007; Costa and McCrae, 2008).

Recent studies have explored automating these refinement processes (particularly survey expansion) by leveraging LLMs to generate new survey items (Mulla and Gharpure, 2023). For example, some approaches employ few-shot prompting to generate items targeting creativity (Laverghetta et al., 2024) and fine-tune models to generate items capturing individual opinions and preferences (Babakhani et al., 2024).

Validity and Reliability. In survey refinement and generation, validity and reliability are critical standards for evaluating the quality of survey items. Construct validity refers to the extent to which an item measures the trait it intends to measure, rather than being incidentally correlated with another trait (Sürücü and Maslakci, 2020). Reliability refers to whether an item produces the same results across repeated assessments (Golafshani, 2003).

Despite the importance of both criteria, most studies in automatic survey generation have primarily focused on reliability, by varying the order or labels of answer choices (Huang et al., 2024), altering prompt templates, and paraphrasing item content (Lee et al., 2025). However, reliability is only meaningful once validity has been established, as a consistent measure is of little value if it does not assess the intended trait (Sürücü and Maslakci, 2020). Traditionally, the validity of survey items is assessed through the analysis of large-scale human response data (Beck et al., 1988; Weber et al., 2002; McCrae and Costa Jr, 1997). As expected, recruiting a diverse respondent pool and conducting such data collection is logistically challenging and costly. Our study aims to bridge this gap using a framework that evaluates the validity of survey items through virtual respondent simulation.

Simulating Human Behavior Using LLMs. Our motivation for using LLM-based simulations is grounded in the success of prior studies showing that LLMs can produce outputs comparable to human behaviors. For instance, Argyle et al. (2023) replicate human voting behavior by conditioning on demographic information. Liu et al. (2025) combine LLM-generated math solutions with human responses to better mimic human behavior. Aher et al. (2023) use experiment-specific prompts and successfully reproduce classic social science experiments. However, it remains understudied how well LLMs can simulate survey respondents for the purpose of evaluating the validity of survey items. In fact, Petrov et al. (2024) find that LLMs provided with simple persona profiles fail to meet psychometric criteria in psychological assessment simulations. Our work addresses this limitation by introducing the concept of mediators into the simulation process as discussed below.

Mediators. According to the cognitive-affective personality system (CAPS) theory (Mischel and Shoda, 1995), a given situation does not always lead to the same behavior for everyone. Instead, observable behavior is determined by multiple mediators in five categories: *Situation Encodings, Expectancies and Beliefs, Affective Responses, Goals and Values, and Competencies and Self-regulatory Plans*. We hypothesize that similar factors also mediate how a person's traits translate to observable behavior. With this assumption, since surveys typically infer a respondent's underlying traits based on their likely behaviors in a given situation (e.g., "when working under pressure, I stay focused and organized"), accurate inference requires that these behaviors be robust to such mediators. Therefore, our core idea is to incorporate mediators into virtual respondents for simulation, going beyond simple persona profiles.

3 Framework

Our goal is to identify items that demonstrate the highest construct validity using psychological assessment simulation. Our framework consists of five main stages (Figure 2): (1) traits selection

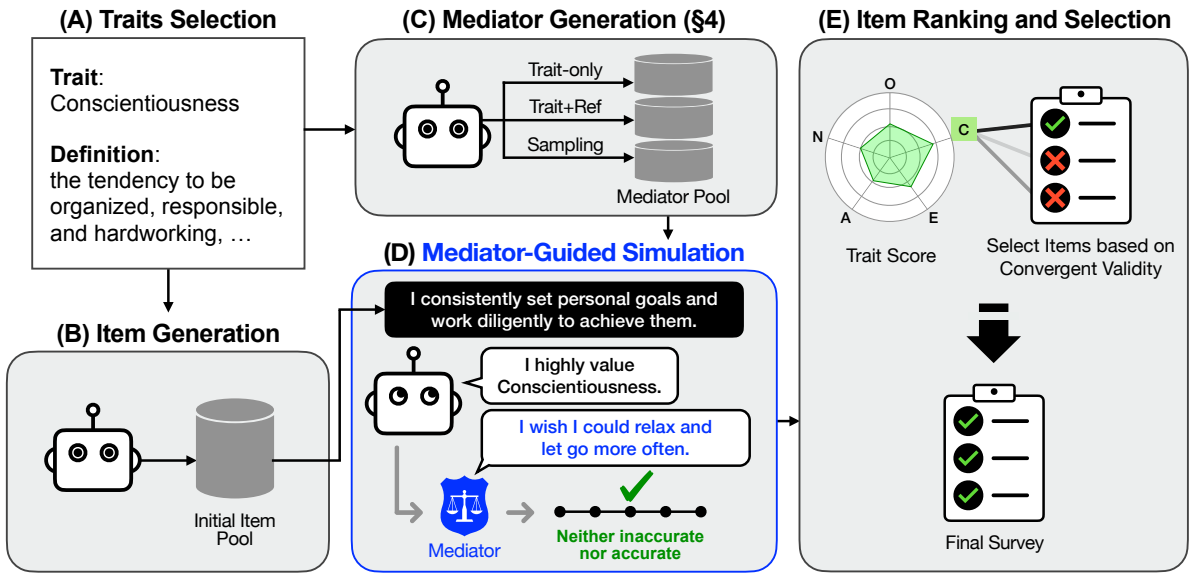

Figure 2: Overview of our framework.

Traits by Theory
The Big Five (Big5; 5 Personality Traits): Openness to experience, Conscientiousness, Extraversion, Agreeableness, Neuroticism
Schwartz's Theory of Basic Values (10 Values): Achievement, Benevolence, Conformity, Hedonism, Power, Security, Self-Direction, Stimulation, Tradition, Universalism
Values in Action (VIA; 24 Character Strengths): Appreciation of Beauty & Excellence, Bravery, Creativity, Curiosity, Fairness, Forgiveness, Gratitude, Honesty, Hope, Humility, Modesty, Humor, Judgment, Kindness, Leadership, Love of Learning, Love, Perseverance, Perspective, Prudence, Self-Regulation, Social Intelligence, Spirituality, Teamwork, Zest

Table 1: Traits for each psychological trait theory.

(§3.1), (2) survey item generation (§3.2), (3) mediator generation (§3.3), (4) mediator-guided simulation through virtual respondents (§3.4), and (5) item ranking and selection (§3.5). This framework mirrors the standard psychometric item validation procedure, where practitioners develop new candidate items to measure a target trait and use an existing questionnaire (what we call the "official" survey) for the same trait as a reference for evaluating these newly developed items. The selected items are evaluated against results derived from human respondents (§3.6).

3.1 Traits Selection (Figure 2-A)

The first step of our framework concerns the selection of traits and their definitions for examination. In this study, we explore the following three psychological trait theories (Table 1), although our framework is not limited to specific theories.

The Big Five (Big5) is one of the most prominent trait theories in psychology, describing five fundamental dimensions of human personality (Goldberg, 1992). **Schwartz's Theory of Basic Values** defines 10 dimensions of human value systems (Schwartz, 1992) and has been increasingly adopted in recent NLP studies. Lastly, **Values in Action (VIA)** assesses 24 character strengths, focusing on the positive aspects of human personality (Peterson, 2004). We include this theory to evaluate the effectiveness of our framework on less widely recognized theories.

We obtain the definitions of each trait from the original papers or websites and use the following official surveys in our experiments: (1) Big5: Goldberg's IPIP representation (Big5-G; 50 items), (2) Schwartz: PVQ (40 items), and (3) VIA: VIA-IS-M (96 items). Additional information is provided in Appendix A.

3.2 Item Generation (Figure 2-B)

To determine the format of new items, we adhere to the following rules. First, each item generated in our study is designed to measure a single trait, as in the official items. Second, half of the items are positively correlated with the target trait, while the other half are negatively correlated. High scores on positive items indicate high levels of the trait, whereas high scores on negative items indicate either low levels of the trait for unipolar

traits (e.g., Achievement) or high levels of the opposite trait for bipolar traits (e.g., Extraversion ↔ Introversion). Third, we determine the scale of the initial item pool. We generate an initial item pool four times larger than the official survey. This scale is chosen to be sufficiently large to reduce the sensitivity of the ranking results to the characteristics of the candidate items, while remaining feasible for a human participant to respond to all items (§3.6). Ablation studies for this scale choice are provided in Appendix L.

Since directly generating high-quality items is beyond the scope of our study, we simply use the method of Yao et al. (2025) to generate an initial item pool by using four LLMs (GPT-4o, GPT-4o-mini, LLaMA-3.1-8B-Instruct, and LLaMA-3.3-70B-Instruct). The detailed procedure is described in Appendix B.

3.3 Mediator Generation (Figure 2-C)

Mediators are a key component in our simulation, as our goal is to identify items that exhibit robust relationships with the target traits after a range of mediators are taken into consideration. To achieve this goal, we employ various mediator generation strategies, such as allowing LLMs to freely generate mediators, referencing survey items, or using human-created value lists. We discuss these strategies in detail in Section 4, and example mediators for the trait of Conscientiousness are shown in Table 2.

3.4 Mediator-Guided Simulation (Figure 2-D)

In the mediator-guided simulation, a survey is conducted via LLMs as virtual respondents. Each virtual respondent is given the following information through the prompt (Figure 3): (1) a target trait, (2) a mediator-integrated persona profile, and (3) a survey item, answer choices, and instructions. Each distinct prompt represents one virtual respondent.

Target trait. Each prompt begins with the phrase "I highly value {trait}", where {trait} is the trait that the included item intends to capture, along with the trait's definition. This allows detecting invalid items that unexpectedly yield inconsistent response scores due to the influence of mediators. We also experimented with including different levels of traits, but that yielded poorer performance (detailed in Appendix C).

Example Prompt for Mediator-Guided Simulation (Big5)

(1) Target Trait
I highly value conscientiousness, which means 'the tendency to be organized, responsible, and hardworking, construed as one end of a dimension of individual differences'.
(2) Mediator-integrated persona profile
My clan backs up. I am learning to embrace and accept spontaneous opportunities that disrupt my usual routines.
I took ballet lessons when i was young. I have a 401k. I enjoy dance. I have went to school for dance.

(3) Survey item, Answer choices, and Instructions
<Instruction>
Based on all the information provided above, select only one answer from the <Answer Choices> to indicate how accurately the <Statement> describes this person's typical behavior or attitudes.

<Statement>
I take pride in my attention to detail in all aspects of my work.

<Answer Choices>
[Very Inaccurate, Moderately Inaccurate, Neither inaccurate nor accurate, Moderately Accurate, Very Accurate]

Figure 3: Example prompt used in mediator-guided simulation for Big5. Blue text is the mediator. Texts highlighted in green are not included in actual prompts.

Mediator-integrated persona profile. The prompt continues with a mediator generated in §3.3 and a persona profile representing general individual characteristics. Petrov et al. (2024) demonstrate that concise persona profiles from the Persona-Chat dataset (Zhang et al., 2018) enable more effective psychological simulations than detailed demographic descriptions. Hence, we randomly sample 500 instances from Persona-Chat and insert a randomly selected mediator into each at a random position to make the mediator appear as part of the persona profile. This results in 500 mediator-integrated persona profiles, each used to simulate a single virtual respondent.

Survey item, answer choices, and instructions. Each prompt contains one survey item, answer choices, and task instruction that faithfully follow the original format of the official survey. An exception is applied to PVQ items, where the original gender-specific pronouns ("he" and "she") are replaced with "they". Answer choices are presented without choice labels, as prior studies have shown this reduces the prompt sensitivity of LLMs (Huang et al., 2024; Lee et al., 2025). The grammatical subject in the answer choices for PVQ and VIA-IS-M is shifted from "me" to "them" (e.g., "Not Like Me" to "Not Like Them"), as this

modification yields better results. All responses are measured through Likert scales: 5-point scales for the Big5-G and VIA-IS-M, and a 6-point scale for the PVQ. To further ensure response reliability, each item is queried twice with different orders of answer choices (high-to-low and low-to-high), and the mean of the two responses is used as the final response score. Note that each virtual respondent answers both generated items and official items.

3.5 Item Ranking and Selection (Figure 2-E)

Once virtual respondents complete all survey items, we calculate the convergent validity of each item: the correlation between virtual respondents' responses to that item and their responses to the official items targeting the same trait. The formal definition is provided in §5.1. Based on the convergent validity scores, we rank the items to select the top N items for each trait (in our evaluation, N matches the number of items in official surveys). In cases of tied rankings, we break ties using the order in which the items were generated during the item generation phase (§3.2) to ensure a consistent and reproducible procedure.

3.6 Human Survey

In order to evaluate the quality of the selected items, we need ground-truth quality scores for each item. To this end, we recruited human participants to complete the same items and used their responses to compute the convergent validity of each item as a reference measure of quality. This procedure is not required to rank items using our framework, but was conducted solely to evaluate its effectiveness.

The human survey was conducted through the online platform Prolific.[2] Surveys were administered separately for each trait theory, with the VIA-IS-M divided into two parts due to the large number of traits. To exclude inattentive respondents, each survey included attention-check items, i.e., three bogus items and duplicate items.

On average, 76.8 participants completed each survey, which exceeds the minimum sample size required to detect an effect size of 0.4 with a significance level of 0.05 and statistical power of 0.80. The participants represented 45 countries, with gender and age distributions balanced through a controlled sampling procedure. They were compensated at a rate of £9 per hour. This human study was approved by the Institutional Review Board (IRB) of our institution. Details of the human survey and participant information are provided in Appendix D.

[2] https://www.prolific.com/

Strategy	Example Mediator
Trait (Free)	I like roles that allow for flexibility and undefined duties.
Trait (CAPS)	I feel overwhelmed when I'm faced with too many responsibilities at once.
Trait+Item	I am a well-intentioned and organized person who sometimes struggles with forgetfulness.
Trait+WVS	I think it would be a good thing if less importance were placed on work in our lives.
Sampling	Sex: Male, Age: 46, Country: Ireland, Occupation: Higher administrative, Income: Seventh step, Education: Bachelor or equivalent (ISCED 6), Social Class: Lower middle class, Religion: Do not belong to a denomination

Table 2: Example mediators generated by each strategy for the trait of Conscientiousness.

4 Mediators

4.1 Mediator Generation

To identify mediator-robust items, we generate mediators for each trait using GPT-4.1 through various strategies grouped into three: trait-only generation, trait-with-reference generation, and sampling from human profiles. Example mediators for each strategy are shown in Table 2. The prompts and details are provided in Appendix G.

Trait-only Generation. We generate mediators based only on trait names and definitions using two strategies. **(1) Trait (Free)**: We instruct the LLM to freely generate a list of potential human characteristics or backgrounds that would be unlikely or contradictory to the target trait. **(2) Trait (CAPS)**: We prompt the LLM to generate the same content while systematically presenting each mediator category in the CAPS theory (Mischel and Shoda, 1995). The average number of mediators generated per trait is 36 for Trait (Free) and 167 for Trait (CAPS), which is five times larger than Trait (Free) because CAPS includes five mediator categories. These two strategies are the best-performing ones in our evaluation.

Trait-with-reference Generation. We provide the LLM with two types of references as supporting information in addition to traits. **(1)**

Trait+Item: We provide generated survey items to the LLM to enable the generation of more concrete mediators directly relevant to the items. **(2) Trait+WVS**: We provide human-crafted value lists from the World Values Survey (WVS).[3] We instruct the LLM to evaluate whether each value could conflict with the target trait and adopt the conflicting values as mediators. The average number of mediators generated per trait is 24 for Trait+Item and 34 for Trait (WVS).

Sampling from Human Profiles. Lastly, **Sampling** uses real human demographics as mediators rather than artificial ones. Specifically, we construct persona cards using eight demographic variables collected from our human survey (§3.6): *sex, age, country, occupation, income level, education level, social class, and religion*. This strategy is employed to examine whether demographic information can serve as effective mediators. The number of mediators for this strategy is the same as the number of human respondents.

4.2 Mediator Evaluation

To assess the LLM-generated mediators, we conduct a human evaluation. Three graduate students in psychology serve as evaluators. First, we ask the evaluators to assess the realism of the mediators and their potential interactions with the target traits. Second, we ask them to categorize the mediators to examine the diversity of their content. We sample three mediators per trait for each of the four mediator generation strategies, resulting in a total of 468 mediators evaluated.

Quality. For mediator sentences, we ask two questions as follows, and responses are collected on a 5-point Likert scale. We compute the mean score across the three evaluators and use it as the score for each mediator.

- **Realism**: How realistic are the human characteristics described in the sentence as qualities that an actual person could possess?
- **Interaction**: If a person with the target trait also has the characteristics described in the sentence, how likely is it that, in certain situations, they may have opinions or behaviors opposite to those of ordinary people who have only that trait?

[3] www.worldvaluessurvey.org/wvs.jsp

Method	Realism	Interaction
Big5		
Trait (Free)	4.43	**4.29**
Trait (CAPS)	4.49	2.56
Trait+Item	**4.55**	3.43
Trait+WVS	4.33	3.52
Schwartz		
Trait (Free)	4.49	**4.48**
Trait (CAPS)	**4.59**	3.16
Trait+Item	4.44	3.59
Trait+WVS	4.27	3.68
VIA		
Trait (Free)	4.49	**4.55**
Trait (CAPS)	**4.56**	2.73
Trait+Item	4.55	3.55
Trait+WVS	4.13	3.65

Table 3: Results for mediator quality evaluation.

Table 3 shows that all mediator generation strategies achieve high realism scores, indicating that LLMs are capable of producing realistic mediators both with and without reference information. For interaction scores, Trait (Free) records the highest score across all theories, which may have contributed to its strong performance in our simulation experiments (§6). By contrast, Trait (CAPS) receives scores slightly below the midpoint for Big5 and VIA, while it exceeds the midpoint for Schwartz. One reason is that the LLM struggled to generate mediators that both satisfy CAPS category constraints and conflict with the target trait simultaneously. Nevertheless, Trait (CAPS) exhibits solid performance in our simulation experiments, which suggests that human judgments and model effectiveness in simulation may diverge, and that mediators with moderate interaction scores from human experts can still be effective for LLMs.

Categorization. To examine whether the mediators reflect more than superficial content and whether each strategy generates a variety of human characteristics, we categorize the mediators into nine categories: *Beliefs and Values, Emotions and Feelings, Habits and Behaviors, Preferences and Interests, Self-Concept and Abilities, Roles and Memberships, Others' Perceptions of Me, Environment and Context, and Other.* A mediator is assigned to a category if at least two of the three

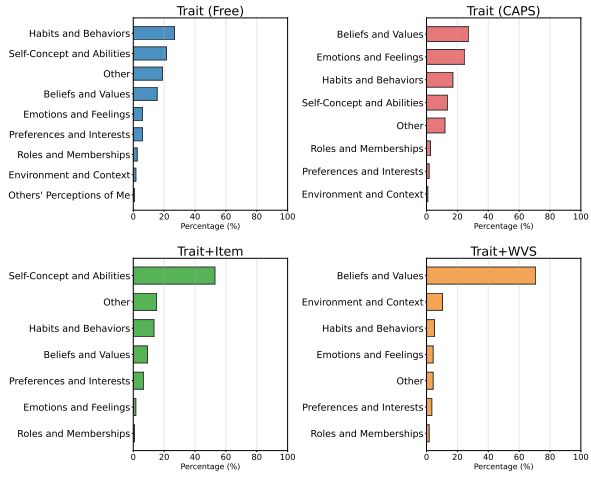

Figure 4: Mediator categories.

evaluators agree on it; otherwise, the mediator is classified as *Other*. Definitions for each category are provided in Appendix E.

Figure 4 shows that all strategies include mediators across multiple categories, with Trait (Free) and Trait (CAPS) displaying relatively balanced distributions. Trait+Item has about half of its mediators in Self-Concept and Abilities, reflecting the self-descriptive nature of its reference items, while Trait+WVS shows a notable presence in Beliefs and Values, consistent with the nature of the WVS survey, which is largely composed of questions on human values. Overall, these patterns indicate that the mediators do not merely contain superficial content but also capture diverse and in-depth aspects of human characteristics.

5 Experiment Settings

5.1 Evaluation Criteria

To evaluate the quality of each item, we consider (1) construct validity: how well an item measures the trait it is intended to measure, and (2) internal consistency reliability: the homogeneity of items that are intended to measure the same trait (Weiner and Greene, 2017).

To quantify construct validity, we need the **trait score** of each respondent for each trait t, calculated as the average of the respondent's answer scores to the corresponding official items:

$$S_t = \frac{1}{N} \sum_{i=1}^{N} a_i \quad (1)$$

where N is the number of official items measuring trait t, and a_i is the respondent's answer to item i. For items designed to have a negative correlation with the target trait, the answer a_i is inverted as $\big(L_{\max} + 1\big) - a_i$, where $L_{\max}$ denotes the maximum possible answer score. Higher trait scores indicate that an individual possesses a greater level of the corresponding trait.

Construct validity is divided into convergent validity and discriminant validity as follows.

Convergent Validity. Convergent validity (CV) refers to the degree to which an item correlates with other measures that assess the same target trait (i.e., official items in our case) (Sürücü and Maslakci, 2020). The CV score of each item i is defined as the Spearman correlation between respondents' answers to that item and their trait scores:

$$\mathrm{CV}_i = \mathrm{Spearman}\big(\{a_i^{(r)}\}_{r=1}^{R}, \{S_t^{(r)}\}_{r=1}^{R}\big) \quad (2)$$

where R is the number of respondents, $a_i^{(r)}$ is respondent r's answer to item i, $S_t^{(r)}$ is respondent r's trait score for the trait t targeted by the item, and $\mathrm{Spearman}(\cdot)$ is the Spearman rank-order correlation. For negatively-correlated items, respondents' answers are inverted as in trait scores.

The CV score is used for two distinct purposes, which should be not confused. First, the CV score as a measure of an item's quality is computed based on the responses of *human respondents*. The CV score of a selected item set is the average of these scores for the individual items in the set. On the other hand, in the item selection stage of the simulation framework (§3.5), generated items are ranked by their CV scores calculated from the responses of *virtual respondents*.

We define two additional metrics based on the CV score to evaluate different aspects of mediator generation strategies. First, to assess the relative quality of the selected item set, we compute the **percentile**, which represents the position of the selected item set's CV score within the empirical distribution of CV scores derived from all possible combinations of N items drawn from the initial item pool. Second, to measure how accurately each mediator generation strategy ranks individual items, we compute the **normalized discounted cumulative gain (NDCG)**, using the ranking based on each item's ground-truth CV score as the reference. Ranking accuracy is particularly important for downscaling existing survey items. To evaluate ranking accuracy across all items as

well as within the selected N-item set, we report both NDCG and NDCG@N. The definitions of percentile and NDCG are detailed in Appendix F.

Discriminant Validity. Discriminant validity (DV) refers to the extent to which an item exhibits low correlations with measures of different, theoretically unrelated traits (Sürücü and Maslakci, 2020). The DV score of an item i is defined as the mean of the Spearman correlations between respondents' answers to i and their trait scores for *non-target* traits:

$$\mathrm{DV}_i = \frac{1}{|\mathcal{D}|} \sum_{t \in \mathcal{D}} \left| \mathrm{Spearman}\left(\{a_i^{(r)}\}_{r=1}^{R}, \{S_t^{(r)}\}_{r=1}^{R} \right) \right| \tag{3}$$

where $\mathcal{D}$ is the set of non-target traits, and the remaining notation follows that used for convergent validity. We take the absolute values of the correlations because correlations in either direction (positive or negative) are undesirable. The DV score of an item set is calculated as the average DV score of the items within the set.

While low DV scores are generally desirable, their interpretation warrants caution because survey items that are completely unrelated to psychometric assessment can still achieve scores close to 0. Hence, we recommend interpreting DV scores in conjunction with CV scores (e.g., the ratio of DV to CV scores) to ensure that items are both non-divergent and meaningfully aligned with the intended traits.

Internal Consistency Reliability. Internal consistency reliability (ICR) refers to the extent to which a set of items homogeneously measures the same trait (Sürücü and Maslakci, 2020). We calculate the ICR score for a trait t using Cronbach's alpha (Cronbach, 1951), which is widely used in psychometrics:

$$\mathrm{ICR}_t = \frac{N}{N-1} \left(1 - \frac{\sum_{i=1}^{N} \mathrm{Var}(\{a_i^{(r)}\}_{r=1}^{R})}{\mathrm{Var}\left(\{\sum_{i=1}^{N} a_i^{(r)}\}_{r=1}^{R}\right)} \right) \tag{4}$$

where N is the number of items measuring each trait t, and $a_i^{(r)}$ is respondent r's answer to item i. Higher ICR scores indicate that the set of items homogeneously measures a single trait. A Cronbach's alpha above 0.7 is generally considered acceptable (Sürücü and Maslakci, 2020).

5.2 Baselines

As baseline methods for item selection, we employ three approaches: (1) random ordering, (2) LLM-as-a-judge, and (3) no-mediator. The prompts for (1) and (2) are provided in Appendix I.

Random Ordering. Items are ranked randomly, and the highest-ranked items are selected.

LLM-as-a-Judge. We use GPT-4.1-mini to directly evaluate the quality of each item based on three criteria: convergent validity, discriminant validity (§5.1), and test-retest reliability. We cannot use internal consistency reliability here because it is not defined at the item level. Instead, we include test-retest reliability, which measures the consistency of responses when the same respondent completes the same item at two different time points (Sürücü and Maslakci, 2020).

We provide definitions for each criterion, and the LLM predicts numerical scores ranging from 1 to 100 for each. The final item quality is quantified as the average score across the three criteria, computed over 10 runs to account for variability in LLM outputs.

No-Mediator. To examine the importance of the mediator, we removed the mediator from the simulation prompt, allowing us to compare performance with and without its presence.

5.3 Experimental Setup

Mediator generation is performed using GPT-4.1, while both the mediator-guided simulation and the LLM-as-a-judge baseline use GPT-4.1-mini. For simulations, we use 500 virtual respondents with distinct persona profiles by default, and only the mediator parts vary across mediator generation strategies. We employ GPT-4.1-mini as the simulation model and set the temperature to 0 to ensure better reproducibility. For each item selection method, we select the top N items for each trait, where N matches the number of items in the official survey (i.e., 10 for Big5 and 4 for Schwartz and VIA).

6 Results

Table 4 presents the overall performance of different methods. As a reference point, we report the performance of the **oracle item set**, i.e., the set of N generated items yielding the highest CV score based on human responses. For meaningful comparisons between psychologist-authored

Method	CV↑				DV↓	ICR↑
	Score	Per.	NDCG	@N	Score	Score
Big5						
Random	.546	48.4	.374	.140	.290	.848
LLM-Judge	.599	79.7	.361	.166	.287	.897
No-Mediator	.585	82.2	.441	.269	.298	.895
Trait (Free)	**.632**	**99.3**	**.568**	.455	.294	**.904**
Trait (CAPS)	.587	71.0	.521	.333	.296	.892
Trait+Item	.626	98.3	.552	**.473**	.294	.892
Trait+WVS	.614	95.5	.429	.282	.302	**.904**
Sampling	.516	30.6	.380	.149	**.265**	.839
Oracle	.690	100	1	1	.318	.922
Official	.657	-	-	-	.271	.844
Schwartz's Theory of Basic Values						
Random	.258	61.1	.525	.256	.140	.576
LLM-Judge	.332	86.8	.642	.454	.143	.679
No-Mediator	.333	86.3	.605	.397	.139	.756
Trait (Free)	.313	77.9	.661	.439	.143	**.757**
Trait (CAPS)	**.347**	**87.1**	**.664**	**.492**	.145	.740
Trait+Item	.327	83.1	.630	.427	.140	.699
Trait+WVS	.333	85.4	.645	.419	**.133**	.748
Sampling	.304	75.6	.531	.284	.137	.684
Oracle	.432	100	1	1	.146	.724
Official	.605	-	-	-	.169	.711
VIA						
Random	.492	48.0	.504	.223	.293	.683
LLM-Judge	.561	75.5	.542	.275	.292	.790
No-Mediator	.502	47.9	.482	.206	.281	.717
Trait (Free)	**.586**	**88.5**	.657	.456	.299	**.803**
Trait (CAPS)	.557	75.1	.573	.345	.296	.772
Trait+Item	.573	78.2	**.678**	**.500**	**.276**	.753
Trait+WVS	.529	61.4	.508	.244	.299	.745
Sampling	.528	63.0	.562	.286	.286	.710
Oracle	.658	100	1	1	.300	.837
Official	.765	-	-	-	.278	.760

Table 4: Performance of item selection methods. Metrics represent the average performance of item sets across all traits within each survey. **Bold**: best performance, Underlined: second-best performance. (Per.: percentile, NDCG: NDCG@All, @N: NDCG@N).

and LLM-generated items, we also present the performance of the **official items** as if they had been generated and selected as the final item set. The significance test results are in Appendix J.1, where the p-values are used to characterize the stability of performance differences across repeated stochastic runs.

6.1 Convergent Validity

Results demonstrate that the trait-only generation strategies are most effective for identifying items with high convergent validity across the three surveys. The free generation strategy (Trait (Free)) achieved the best CV score and percentile on the Big5 and VIA, while the CAPS-guided strategy (Trait (CAPS)) yielded the best performance on the Schwartz. NDCG and NDCG@N scores show similar patterns. Overall, the quality of the selected item sets ranks in the top 1% (Big5) to 13% (Schwartz and VIA) among all possible selections, according to the percentile metric.

Our simulation-based approach, when paired with the best-performing mediators, consistently outperforms baseline methods. The random ordering approach produced results close to random, yielding approximately the 50th percentile. The no-mediator approach consistently underperformed compared to the best mediator-based methods, underscoring the critical role of mediators in item validation. Notably, the LLM-as-a-judge approach consistently scored above the 75th percentile across all surveys, indicating that LLMs possess a solid understanding of psychological traits and a reasonable ability to evaluate item quality. However, this method underperforms in ranking individual items, as indicated by lower NDCG scores compared to simulation-based methods.

Finally, using real human demographics as mediators yielded poorer performance than using LLM-generated mediators. This finding aligns with the prior finding that LLMs struggle to capture the relationship between demographic variables and psychological traits (Petrov et al., 2024).

These results suggest that LLMs have the ability to generate effective mediators between traits and behaviors using only trait names and definitions, without additional context. Still, a significant performance gap remains between all methods and the oracle and official item sets.

6.2 Discriminant Validity

Discriminant validity (DV) measures an item's correlation with unrelated traits. The results show that all methods yielded DV scores similar to the oracle and official items, suggesting acceptable levels of DV. While simulation-based methods and baselines show marginal differences, the ratio of DV to CV scores—interpreted as the relative correlation with non-target traits compared to target traits—favors the Trait (Free) strategy. Overall, these findings demonstrate LLMs' ability to reasonably generate items without unintended

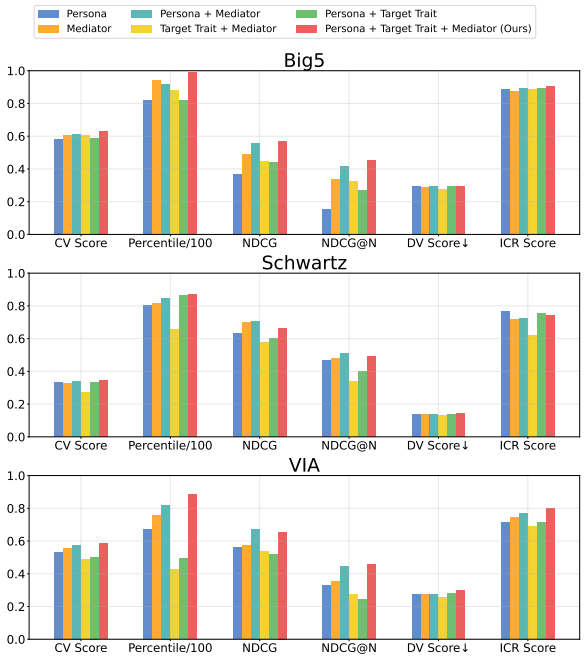

Figure 5: Performance of selected items across different components of the simulation prompt. Percentile scores were normalized to the [0, 1] range.

correlations with unrelated traits. In Appendix K, we report the maximum DV scores for each survey to examine worst cases.

6.3 Internal Consistency Reliability

Internal consistency reliability (ICR) evaluates whether items within a set capture the same trait. Trait (Free) consistently achieved the highest ICR scores across all surveys. Furthermore, most item sets selected through our mediator-driven approach achieved ICR scores above the acceptance threshold (≥ 0.7). These results reaffirm the utility of LLM-generated mediators and the effectiveness of mediator-guided simulation in capturing the homogeneity of items. ICR scores have a positive association with CV scores, suggesting that securing high convergent validity enhances the homogeneity of the items.

7 Ablation Studies

In this section, we conduct ablation studies to examine the importance of each component of the simulation. We use the best-performing mediator generation strategy for each survey identified in Section 6, namely Trait (Free) for Big5 and VIA, and Trait (CAPS) for Schwartz.

7.1 Simulation Components

Aside from the essential task instructions with survey items and answer choices, our simulation prompt comprises three core components: the target trait, mediator, and general persona profile. To examine the importance of each component, we evaluate item selection performance across all possible combinations. Since the persona profile serves to distinguish among 500 virtual respondents, when it is absent, we generate distinct respondents based on the number of mediators. We do not include using a target trait only, because a single target trait sentence cannot generate distinct respondents. In total, we evaluate six combinations. The significance test results are in Appendix J.2.

Figure 5 shows that the full setting—combining sentences that steer target traits and mediators with persona profiles—most effectively identifies valid item sets. This setting achieves the highest CV scores and percentile across all three surveys, as well as the highest ICR scores in both Big5 and VIA. When comparing Persona with Persona+Mediator and Persona+Target Trait with the full setting, we observe that removing mediators leads to the largest performance drop in CV, percentile, and ranking scores. Furthermore, comparing Target Trait+Mediator with the full setting also reveals declines across all metrics. These findings suggest that both mediators and persona profiles play a crucial role in identifying valid survey items.

7.2 Simulation Scale

We examine the relationship between simulation scale and item selection performance by varying the number of virtual respondents from 50 to 500. For scales below 500, we sample persona profiles from the full set of 500. To establish confidence interval, we perform subsampling without replacement 1,000 times.

Figure 6 shows that increasing the simulation scale has a positive impact on CV and ICR scores, with an upward trend. This trend is pronounced for ICR scores. DV scores remain relatively stable (VIA) or show a slight upward trend (Big5 and Schwartz). However, DV scores should be considered in conjunction with CV scores, since items with low correlations with any traits can achieve low DV scores. The figure shows that the ratio of DV to CV scores decreases, suggesting that

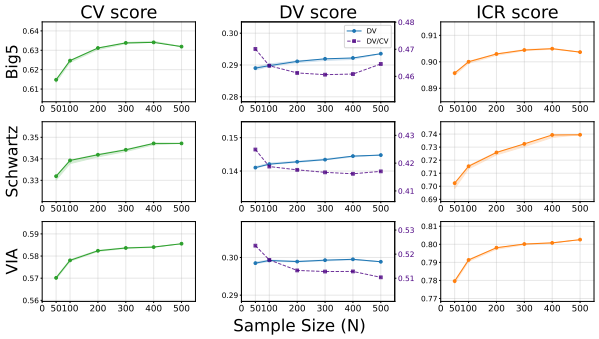

Figure 6: Performance by number of virtual respondents. Points indicate median scores, with shaded regions representing confidence intervals.

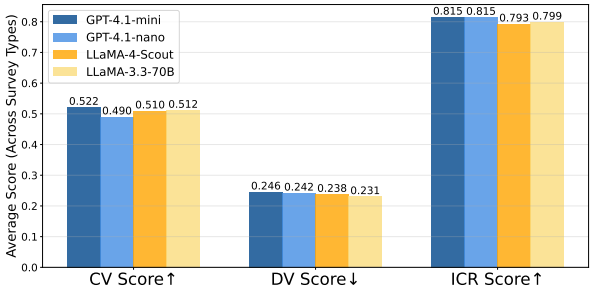

Figure 7: Performance by simulation LLM (with 500 virtual respondents).

larger scales make it easier to identify items that are more closely aligned with their target traits, as CV grows faster than DV. The results of this scaling experiment align with real-world practice, where securing a large pool of human respondents is desirable for both the validity and reliability of survey items.

7.3 Simulation Models

To assess the generalizability of our framework across different LLMs serving as simulation models, we conduct simulations with three additional models alongside GPT-4.1-mini: GPT-4.1-nano, LLaMA-4-Scout, and LLaMA-3.1-70B.

Figure 7 shows that our framework achieves consistent performance across different simulation models, with only marginal differences observed among the four. Notably, the performance of GPT-series models is positively correlated with their sizes, whereas LLaMA-series models show smaller performance differences across sizes.

8 Conclusion

With the goal of evaluating the construct validity of psychometric survey items, we presented a framework for the automatic generation and validation of survey items. We focused particularly on mediator generation using LLMs and mediator-guided simulation with LLM-based virtual respondents. We also defined psychometrically grounded metrics for this new task. Experiments demonstrated that our framework effectively identifies valid item sets. In particular, mediators generated by LLMs based on trait definitions performed best overall, highlighting the ability of LLMs to generate mediators and simulate survey respondents.

To support follow-up research, we publicly release all generated items along with collected human responses. We hope this benchmark serves as a valuable resource for research on improved methods for more scalable and cheaper survey item validation.

9 Limitations

We focused on three psychological theories, which may not capture the full breadth of psychological domains such as emotion, cognition, and mental disorders. However, our framework can be readily extended to other theories, which we leave for future work.

The generated items are currently only in English, and our framework has not been validated in multilingual or cross-cultural settings. Future work may extend the method by generating psychometric items in other languages and leveraging multilingual versions of official items (which already exist for many theories). The item ranking results can be validated by comparing them with human responses collected in that language or cultural background.

Our framework assumes the availability of official items to compute validity. While this is a common assumption in real-world survey refinement, it would be valuable to explore item evaluation from scratch. For example, factor analysis is a statistical method widely used to uncover latent structures underlying responses and identify groups of items loading on the same factor. Our simulation method can provide virtual survey response data to support factor analysis and thereby facilitate the development of new items in domains where official items do not yet exist or are not accessible.

As this study represents an early step toward automated item validation and there remains substantial room for improvement, we recommend

conducting a human inspection before officially releasing new psychometric items developed using our framework. Experts may verify that each item measures the intended trait and is free of conceptual or interpretive errors. Our framework can be integrated at the initial stage of the validation process, where it narrows a large candidate pool to a smaller set of high-potential items.

It is important to note that LLMs cannot yet perfectly replicate the responses of individual humans. Our framework is therefore not intended to mimic human psychological processes per se, but to leverage diverse mediators and their correlations with trait scores to identify items that measure traits robustly. Recognizing this inherent limitation of LLMs, we view our work as a meaningful step toward more effective automatic item validation, and encourage future research to advance simulation techniques that better align with human psychological processes.

Additionally, we note that LLMs do not yet possess a perfect ability to generate mediators. In our analysis of failure cases, some mediators described characteristics of individuals already high in the target trait, rather than interacting with the trait as intended. This suggests the need for future work on improved prompting strategies and automated methods for filtering generated mediators, in line with our goal of fully automatic item validation.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) (RS-2024-00333484, RS-2022-NR070855) and by the Institute of Information & Communications Technology Planning & Evaluation (IITP) (RS-2021-II211343, AI Graduate School Program, Seoul National University), all funded by the Korean government (MSIT). This work was also supported by the Creative-Pioneering Researchers Program through Seoul National University and by the National Supercomputing Center with supercomputing resources including technical support (KSC-2025-CRE-0332, KSC-2025-CRE-0514).

References

Gati V. Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 337–371. PMLR.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3):337–351.

Pedram Babakhani, Andreas Lommatzsch, Torben Brodt, Doreen Sacker, Fikret Sivrikaya, and Sahin Albayrak. 2024. Opinerium: Subjective Question Generation Using Large Language Models. *IEEE Access*, 12:66085–66099.

Aaron T. Beck, Norman Epstein, Gary Brown, and Robert A. Steer. 1988. An inventory for measuring clinical anxiety: Psychometric properties. *Journal of consulting and clinical psychology*, 56(6):893.

Aaron T. Beck, Robert A. Steer, Roberta Ball, and William F. Ranieri. 1996. Comparison of Beck Depression Inventories-IA and-II in Psychiatric Outpatients. *Journal of personality assessment*, 67(3):588–597.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An Empirical Investigation of Statistical Significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea. Association for Computational Linguistics.

Florian Brühlmann, Zgjim Memeti, Lena F. Aeschbach, Sebastian A. C. Perrig, and Klaus Opwis. 2024. The effectiveness of warning statements in reducing careless responding in crowdsourced online surveys. *Behavior research methods*, 56(6):5862–5875.

Paul Costa and Robert McCrae. 2008. The Revised NEO Personality Inventory (NEO-PI-R). *The SAGE Handbook of Personality Theory and Assessment*, 2:179–198.

Lee J. Cronbach. 1951. Coefficient Alpha and the Internal Structure of Tests. *Psychometrika*, 16(3):297–334.

Nahid Golafshani. 2003. Understanding Reliability and Validity in Qualitative Research. *The qualitative report*, 8(4):597–607.

Lewis R. Goldberg. 1990. An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, 59(6):1216–1229.

Lewis R. Goldberg. 1992. The development of markers for the Big-Five factor structure. *Psychological assessment*, 4(1):26.

Lewis R. Goldberg. 1999. A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several Five-Factor models. *Personality psychology in Europe*, 7(1):7–28.

Dorith Hadar Shoval, Kfir Asraf, Yonathan Mizrachi, Yuval Haber, and Zohar Elyoseph. 2024. Assessing the Alignment of Large Language Models With Human Values for Mental Health Integration: Cross-Sectional Study Using Schwartz's Theory of Basic Values. *JMIR Mental Health*.

Jongwook Han, Dongmin Choi, Woojung Song, Eun-Ju Lee, and Yohan Jo. 2025. Value Portrait: Assessing Language Models' Values through Psychometrically and Ecologically Valid Items. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17119–17159, Vienna, Austria. Association for Computational Linguistics.

Jen-tse Huang, Wenxiang Jiao, Man Ho Lam, Eric John Li, Wenxuan Wang, and Michael Lyu. 2024. On the Reliability of Psychological Scales on Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6152–6173, Miami, Florida, USA. Association for Computational Linguistics.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and Inducing Personality in Pre-trained Language Models. In *Advances in Neural Information Processing Systems*, volume 36, pages 10622–10643. Curran Associates, Inc.

Carl Jung and John Beebe. 2016. *Psychological Types*. Routledge.

Antonio Laverghetta, Simone Luchini, Averie Linnell, Roni Reiter-Palmon, and Roger Beaty. 2024. The creative psychometric item generator: A framework for item generation and validation using large language models. In *CEUR Workshop Proceedings*, volume 3810, pages 59–73. CEUR-WS.

Seungbeen Lee, Seungwon Lim, Seungju Han, Giyeong Oh, Hyungjoo Chae, Jiwan Chung, Minju Kim, Beong-woo Kwak, Yeonsoo Lee, Dongha Lee, Jinyoung Yeo, and Youngjae Yu. 2025. Do LLMs have distinct and consistent personality? TRAIT: Personality testset designed for LLMs with psychometrics. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8412–8452, Albuquerque, New Mexico. Association for Computational Linguistics.

Marjaana Lindeman and Markku Verkasalo. 2005. Measuring Values With the Short Schwartz's Value Survey. *Journal of personality assessment*, 85:170–8.

Yunting Liu, Shreya Bhandari, and Zachary A. Pardos. 2025. Leveraging LLM respondents for item evaluation: A psychometric analysis. *British Journal of Educational Technology*.

Robert R. McCrae and Paul T. Costa Jr. 1997. Personality trait structure as a human universal. *American psychologist*, 52(5):509.

Robert E. McGrath. 2019. The VIA Assessment Suite for Adults: Development and Initial Evaluation Revised Edition. *Cincinnati, OH: VIA Institute on Character*.

Marilù Miotto, Nicola Rossberg, and Bennett Kleinberg. 2022. Who is GPT-3? An exploration of personality, values and demographics. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 218–227, Abu Dhabi, UAE. Association for Computational Linguistics.

Walter Mischel and Yuichi Shoda. 1995. A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological review*, 102(2):246.

Nikahat Mulla and Prachi Gharpure. 2023. Automatic question generation: A review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence*, 12(1):1–32.

Christopher Peterson. 2004. *Character Strengths and Virtues: A Handbook and Classification*, volume 3. Oxford University Press.

Nikolay B. Petrov, Gregory Serapio-García, and Jason Rentfrow. 2024. Limited Ability of LLMs to Simulate Human Psychological Behaviours: A Psychometric Analysis. *arXiv preprint arXiv:2405.07248*.

Beatrice Rammstedt and Oliver P. John. 2007. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, 41(1):203–212.

Shalom Schwartz. 2021. A Repository of Schwartz Value Scales with Instructions and an Introduction. *Online Readings in Psychology and Culture*, 2.

Shalom H. Schwartz. 1992. Universals in the Content and Structure of Values: Theoretical Advances and Empirical Tests in 20 Countries. volume 25 of *Advances in Experimental Social Psychology*, pages 1–65. Academic Press.

Lütfi Sürücü and Ahmet Maslakci. 2020. Validity And Reliability In Quantitative Research. *Business And Management Studies An International Journal*, 8:2694–2726.

Elke U. Weber, Ann-Renée Blais, and Nancy E. Betz. 2002. A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making*, 15(4):263–290.

Irving B. Weiner and Roger L. Greene. 2017. *Handbook of Personality Assessment*. John Wiley & Sons.

Jing Yao, Xiaoyuan Yi, and Xing Xie. 2025. CLAVE: An Adaptive Framework for Evaluating Values of LLM Generated Responses. *Advances in Neural Information Processing Systems*, 37:58868–58900.

Haoran Ye, Jing Jin, Yuhang Xie, Xin Zhang, and Guojie Song. 2025. Large Language Model Psychometrics: A Systematic Review of Evaluation, Validation, and Enhancement. *arXiv preprint arXiv:2505.08245*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Appendix

A Survey Information and Dimensions of Traits

The full lists of traits corresponding to each theory are presented in Table 1.

Big5. The Big Five, also known as five-factor model (FFM), is based on the fundamental lexical hypothesis, which suggests that key personality traits become encoded in various languages (Goldberg, 1990). This theory presents five dimensions of personality: openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism. We obtain official items from the 50-item IPIP representation of the Goldberg (1992) markers for the Big-Five factor structure (Goldberg, 1999).[4] This survey asks respondents to indicate how accurately the statement describes their typical behavior or attitudes. Responses are given on a 5-point Likert scale ranging from "Very Accurate" to "Very Inaccurate".

Schwartz. Schwartz's theory of basic values categorizes human values into 10 types (Schwartz, 1992). The Portrait Values Questionnaire (PVQ) was developed to measure these categories. We obtain official items from Schwartz (2021). This survey asks respondents to indicate how much the people described in each item are like them. Responses are given on a 6-point Likert scale ranging from "Very Much Like Me" to "Not at All Like Me".

VIA. The Values in Action (VIA) framework categorizes personality strengths into 24 facets across six broad character strengths (Peterson, 2004). The VIA Inventory of Strengths – Mixed (VIA-IS-M) is a questionnaire designed to assess these facets (McGrath, 2019). We obtain official items from the VIA Institute on Character.[5] This survey asks respondents to indicate whether the statement describes what they are like. Responses are given on a 5-point Likert scale ranging from "Very much like me" to "Very much unlike me".

B Construction of Initial Item Pools

In this section, we introduce the procedure of constructing an initial pool of psychometric items based on the framework proposed by Yao et al. (2025).

Prompt Template for Item Generation

Task:
Create `{num}` items to assess the "`{trait_name}`".

Definition:
Here is what "`{trait_name}`" means: `{trait_def}`

Instructions:
- Each item should be unique and not overlap with others.
- Avoid using the term "`{trait_name}`" in the items.
- Use "`{subject}`" as the subject for all items.
- All items must start with the subject "`{subject}`".
- All items must be `{correlation}` correlated with the trait.

Figure 8: Prompt template used for item generation.

Initially, we used LLMs (with temperature set to 1) to generate a candidate pool of items four times the size of the official items. The template shown in Figure 8 illustrates the five input fields and five instructions used in our study. First, we supply definitions extracted from the official publications—while instructing the model not to use the exact trait definition within any item. Second, we provide a grammatical subject identical to that used in the official items, requiring each item to begin with this subject. Third, we specify the exact number of items (twice the number of official items) to be generated. Fourth, we indicate the desired direction of correlation between generated items and the target trait, prompting separate generations for positively and negatively correlated items. Finally, we enforced semantic uniqueness by instructing the model to produce non-overlapping items. As a result, our procedure yielded four times as many candidate items as the official surveys.

Next, we refined this item pool by using a three-step embedding-and-clustering approach, building on recent value-evaluation research demonstrating the effectiveness of combining LLMs with clustering-based approaches (Yao et al., 2025). The first step is obtaining the semantic embeddings of items. We converted each item into a numeric vector embedding using OpenAI's text-embedding-3-small model. These embeddings capture the meaning of each item so that similar items are placed near one another in this high-dimensional space. The second step is clustering.

[4] https://ipip.ori.org/newNEODomainsKey.htm

[5] https://www.viacharacter.org/

Method	CV↑				DV↓	ICR↑
	Score	Per.	NDCG	@N	Score	Score
Big5						
Low	.603	87.9	.418	.215	**.294**	.890
Mixed	.623	94.1	.529	.385	.297	**.907**
High	**.632**	**99.3**	**.568**	**.455**	**.294**	.904
Schwartz's Theory of Basic Values						
Low	.336	85.6	**.697**	.475	**.134**	**.767**
Mixed	**.349**	**89.0**	.692	**.517**	.140	.736
High	.347	87.1	.664	.492	.145	.740
VIA						
Low	.583	86.0	**.660**	**.488**	**.296**	.769
Mixed	.578	83.8	.637	.446	.302	.789
High	**.586**	**88.5**	.657	.456	.299	**.803**

Table 5: Performance of selected items across different trait levels of the virtual respondents. For mediators, we employ the best-performing mediator generation strategies from Section 6: Trait (free) for Big5 and VIA, and Trait (CAPS) for Schwartz. (Per.: percentile, NDCG: NDCG@All, @N: NDCG@N).

We applied K-means clustering to group these embeddings into roughly $n/4$ clusters, where n is the total number of items. Each cluster represents a distinct semantic theme within the pool. The final step is selecting representative items. From each cluster, we pick the item whose embedding lies closest to the cluster's centroid, assuming this item best represents the cluster's overall meaning. As a result, this refinement process reduces the number of candidate items in the initial item pool to match the number of official items.

Through this process, we were able to generate new items covering a diverse range of situations. While official surveys enable effective psychometric assessment, their items often include broad and abstract content (e.g., "I am always prepared" for measuring *Conscientiousness* in the Big5). In contrast, the items generated by LLMs tend to capture more concrete scenarios (e.g., "I consistently meet deadlines for my assignments.").

C Experiments with Different Trait Levels

In our main experiments, the prompt for each virtual respondent begins with a target trait set to a high level (e.g., "I highly value {trait}" in Figure 3). Here, we explore the effectiveness of including other trait levels in virtual respondents. To create virtual respondents with low trait levels, we modify the opening sentence to "I oppose {trait}". In this case, we need mediators that may facilitate behaviors related to the trait. Although we attempted to generate such mediators using the same mediator generation strategies in 4.1, many of the generated mediators undesirably leaned toward suppressing the trait. To address this, we opted to reverse the original mediators generated in our main experiments, using GPT-4.1. We keep the rest of the virtual respondent prompt unchanged. We conduct two experiments: (1) 500 virtual respondents with **low** trait levels only, and (2) 500 virtual respondents with **mixed** trait levels (250 low and 250 high).

As shown in Table 5, experiments conducted only with high-trait virtual respondents achieve the highest CV scores for Big5 and VIA. For Schwartz, the high setting also records high scores with marginal underperformance compared to the mixed setting. Based on these results, we conclude that setting virtual respondents' trait levels to high produces the most stable and superior performance, leading us to adopt it as our main experimental configuration.

D Details of Human Survey

Since all our surveys are constructed in English, we recruit adults who are native English speakers or fluent in English through the online survey platform Prolific. To recruit reliable human respondents, we target individuals who have completed more than 100 previous submissions on this platform and maintain an approval rate of 98% or higher. We recruit participants with equal gender distribution and equal numbers across age groups specified in Table 6. A total of 339 respondents were recruited, and 307 respondents were finally included in our analysis based on the attention check procedure outlined in Appendix D.2.

D.1 Demographics of Human Respondents

The demographic information of the final respondents is presented in Table 6.

D.2 Attention Check Procedure

To ensure human respondents' attentiveness, we employ attention-check items consisting of two types: 1) duplicate items that repeat previously presented items, and 2) bogus items that require consistently negative responses. The attention-check items are distributed across three pages,

Survey	N	Gender	Age
Big5-G	75	- Male: 37 - Female: 38	- 20~29: 16 - 30~39: 14 - 40~49: 15 - 50~59: 15 - 60~100: 15
PVQ	76	- Male: 37 - Female: 39	- 20~29: 15 - 30~39: 15 - 40~49: 16 - 50~64: 16 - 60~100: 14
VIA-IS-M (Part 1)	80	- Male: 40 - Female: 38 - Non-binary / Third gender: 2	- 20~29: 16 - 30~39: 18 - 40~49: 15 - 50~59: 16 - 60~100: 15
VIA-IS-M (Part 2)	76	- Male: 37 - Female: 38 - Non-binary / Third gender: 1	- 20~29: 16 - 30~39: 16 - 40~49: 14 - 50~59: 15 - 60~100: 15

Table 6: Demographics of human respondents.

with each page containing one bogus item and one duplicate item.

Duplicate Items. Duplicate items are randomly selected from the official items and presented repeatedly in the survey. Answers from their initial appearance are used to assess the respondent's trait scores, while answers from the second appearance serve only as attention checks. Respondents who show a discrepancy of two or more points on the Likert scale on more than two occasions are considered inattentive and excluded from the analysis.

Bogus Items. Bogus items ask respondents about highly improbable actions or behaviors with which they are not expected to agree (Brühlmann et al., 2024). We generate three bogus items for each survey using GPT-4o with the prompt below. An example bogus item generated by this prompt is "I can speak every language in the world fluently". When a respondent answers positively to these items, their task is immediately terminated.

E Details of the Human Evaluation of Mediators

The mediator categories and their definitions are presented in Table 9.

Prompt used for Bogus Item Generation

Task:
Create 12 bogus items to detect careless responses in an online survey.

Instructions:
- Use "I" as the subject for 9 items and use "They" as the subject for 3 items.
- Each item should be unique and not overlap with others.

F Definitions of Evaluation Metrics

Percentile. We compute the percentile to assess the relative quality of the selected item set. The percentile of the selected item set s^* is defined as:

$$\text{Percentile}(s^*) = \frac{1}{|S|} \sum_{s \in S} \mathbb{I}[\ \text{CV}(s) \leq \text{CV}(s^*)\ ] \times 100 \quad (5)$$

where S is a set of all possible N-item combinations from the initial item pool, $\mathbb{I}[\cdot]$ is the indicator function, and $\text{CV}(s)$ is the CV score of the item set s.

NDCG and NDCG@N. To measure how accurately each mediator generation strategy ranks individual items, we use the normalized discounted cumulative gain (NDCG). It is based on the discounted cumulative gain (DCG), defined as:

$$\text{DCG} = \sum_{i=1}^{M} \frac{2^{\text{rel}_i}}{\log_2(i+1)}, \quad (6)$$

where rel_i denotes the relevance (here, the ranking of the item based on CV scores) of the item at rank i, and M is the total number of items. The ideal DCG (IDCG) is computed by sorting items in descending order of their true CV scores, denoted $\text{rel}_i^\star$. The normalized score is then given by:

$$\text{NDCG} = \frac{\text{DCG}_M}{\text{IDCG}_M} \in [0, 1] \quad (7)$$

which always lies in the range [0,1], with higher values indicating more accurate rankings.

In practice, a truncated version known as NDCG@k is widely used, where the parameter k specifies the number of top-ranked items considered. In our evaluation, we report both NDCG, obtained by setting $k = M$ to the total number of items to assess full-ranking accuracy, and NDCG@N, obtained by setting $k = N$ to assess ranking accuracy within the top-N items of the selected set.

G Prompt Templates for Mediator Generation

G.1 Trait (Free)

Freely generate mediators with traits

Trait: `{trait_name}`
Definition: `{trait_def}`

List possible human characteristics or backgrounds that would be unlikely or contradictory for someone who strongly values `{trait_name}`. Just number them without detailed explanation. Make many values as possible.

G.2 Trait (CAPS)

The italicized sections in the prompt below represent the five categories of mediators proposed by the CAPS theory (Mischel and Shoda, 1995). We introduce each category one at a time to separately generate mediators across five iterations.

Generate mediators based on CAPS

Trait: `{trait_name}`
Definition: `{trait_def}`

List possible {*(1) Situation Encodings, (2) Expectancies and Beliefs, (3) Affective Responses, (4) Goals and Values, (5) Competencies and Self-regulatory Plans*} that could create internal conflict or lead to changes in behavior for someone who strongly values `{trait_name}`. List as many items as possible. Number each item without detailed explanation.

G.3 Trait+Item

We use generated items to create diverse mediators. The official items from the original surveys are not used in this process.

Generate mediators with an item

Trait: `{trait_name}`
Definition: `{trait_def}`
Survey Item: `{item}`

Generate a single personal characteristic, background factor, or life circumstance that could plausibly lead someone—even among people who highly value `{trait_name}`—to respond contrary to the survey item above.

For the mediators generated by Trait (Free) (G.1), Trait (CAPS) (G.2), and Trait+Item (G.3), we convert them into persona sentences using GPT-4.1 before integrating them with persona profiles. The prompt is as follows:

Convert mediators into persona sentences

For each of the contents below, write a single sentence introducing a person's persona. Use "I" as a subject.

`{List of mediators}`

G.4 Trait+WVS

The World Values Survey (WVS) is a globally comprehensive survey that examines changing human beliefs and values and their impact on social and political life across multiple countries worldwide. The survey encompasses questions about social, economic, religious, and ethical values, including topics such as the importance of family in life. Based on these question topics and sentences, we create persona sentences assuming individuals who agree with these questions (Step 1) and use the LLM to determine whether these personas conflict with the traits (Step 2). Sentences classified as conflicting values serve as mediators for the corresponding traits.

Step 1: Convert WVS questions into persona sentences

Create a concise persona sentence who say 'Yes/Agree/Likely' to the question below. Use "I" as the subject.

Question Topic:
`{WVS_question_topic}`
Question: `{WVS_question_sentence}`

Step 2: Determine whether the given persona sentences conflict with each trait

Consider the given values and the personality trait below.
Determine whether the given values conflict with the personality trait, making it difficult for individuals to respond accurately to questions designed to measure the trait.

<Personality Trait>
`{trait_name}: {trait_def}`

<Values>
`{persona_sentence}`

H Prompt Templates for Mediator-Guided Simulation

H.1 Big5

Prompt Template for Big5-G Simulation

`{Target Trait + Mediator-integrated persona profile}`

<Instruction>
Based on all the information provided above, select only one answer from the <Answer Choices> to indicate how accurately the <Statement> describes this person's typical behavior or attitudes.

<Statement>
`{Item for target trait}`

<Answer Choices>
[Very Accurate, Moderately Accurate, Neither inaccurate nor accurate, Moderately Inaccurate, Very Inaccurate]

H.2 Schwartz

Prompt Template for PVQ Simulation

`{Target Trait + Mediator-integrated persona profile}`

<Instruction>
Based on all the information provided above, select only one answer from the <Answer Choices> to indicate the degree to which this person is like them, as described in the <Statement>.

<Statement>
`{Item for target trait}`

<Answer Choices>
[Very Much Like Them, Like Them, Somewhat Like Them, A Little Like Them, Not Like Them, Not Like Them at All]

To allow the simulation LLM to understand the task better, we replaced the pronouns in the answer choices in the official survey from the first-person pronoun "Me" to the third-person pronoun "Them".

H.3 VIA

Prompt Template for VIA-IS-M Simulation

`{Target Trait + Mediator-integrated persona profile}`

<Instruction>
Based on all the information provided above, select only one answer from the <Answer Choices> to indicate the degree to which the <Statement> describes what the person is like.

<Statement>
`{Item for target trait}`

<Answer Choices>
[Very Much Like Them, Like Them, Somewhat Like Them, A Little Like Them, Not Like Them, Not Like Them at All]

To allow the simulation LLM to understand the task better, we replaced the pronouns in the answer choices in the official survey from the first-person pronoun "Me" to the third-person pronoun "Them".

I Baseline

I.1 LLM-as-a-Judge

Prompt Template used for LLM-based Item Evaluation

Please evaluate the following psychometric item according to the criteria below.

Item: "`{item}`"
Target Trait: `{trait_name}`
Trait Definition: `{trait_def}`
Expected Correlation (direction): `{correlation}`

<Evaluation criteria>
1. Validity
a) Convergent validity - the degree to which the item accurately measures the target trait.
b) Discriminant validity - the degree to which the item does not substantially measure traits other than the target trait.

2. Reliability
a) Test-retest reliability - the degree of consistency in responses when the item is administered to the same individual at

different times.

<Response format>
(Use a 1–100 scale, where 1 = very poor and 100 = excellent.):
Convergent validity: [score 1-100]
Discriminant validity: [score 1-100]
Test-retest reliability: [score 1-100]
Explanation: [brief rationale for each rating]

J Significance Tests

To determine whether the observed results presented in Section 6 and 7.1 are statistically meaningful, we conduct significance testing. Berg-Kirkpatrick et al. (2012) proposed a recentered paired bootstrap estimator of the one-sided p-value: repeatedly draw bootstrap resamples $x^{(i)}$ from the full test set x (with replacement), compute the metric difference $\delta(x^{(i)}) = m(A, x^{(i)}) - m(B, x^{(i)})$ between systems A and B, where $m()$ denotes the evaluation metric, and approximate $p \approx s/b$ where s counts resamples with $\delta(x^{(i)}) > 2\delta(x)$ and b is the number of resamples.

Following this methodology, our significance test proceeds as follows. We first select hypotheses in which system A outperformed system B in terms of CV score in Section 6 and 7.1. Then, we resample the initial item pool 1,000 times with replacement for each trait, maintaining the original pool size. The top N items are selected for each item selection method. Next, we compute the difference in CV scores for each pair of methods and compare it with the difference on the original test set. We then estimate the p-value by applying the criterion $\delta(x^{(i)}) > 2\delta(x)$.

J.1 Main Results

The results in Figure 9 indicate that the findings in Section 6 remain statistically stable, whereas the outcomes for Schwartz are less consistent. The best methods for Big5 and VIA (Trait (Free)) consistently outperform all baseline approaches and the Sampling method, indicating that our simulation framework yields more robust and reliable performance differences across stochastic runs. In particular, these best methods outperformed the No-Mediator setting with $p < 0.05$, underscoring the importance of the mediator in identifying valid items. However, for Schwartz, it was difficult to identify consistent results. One plausible expla-

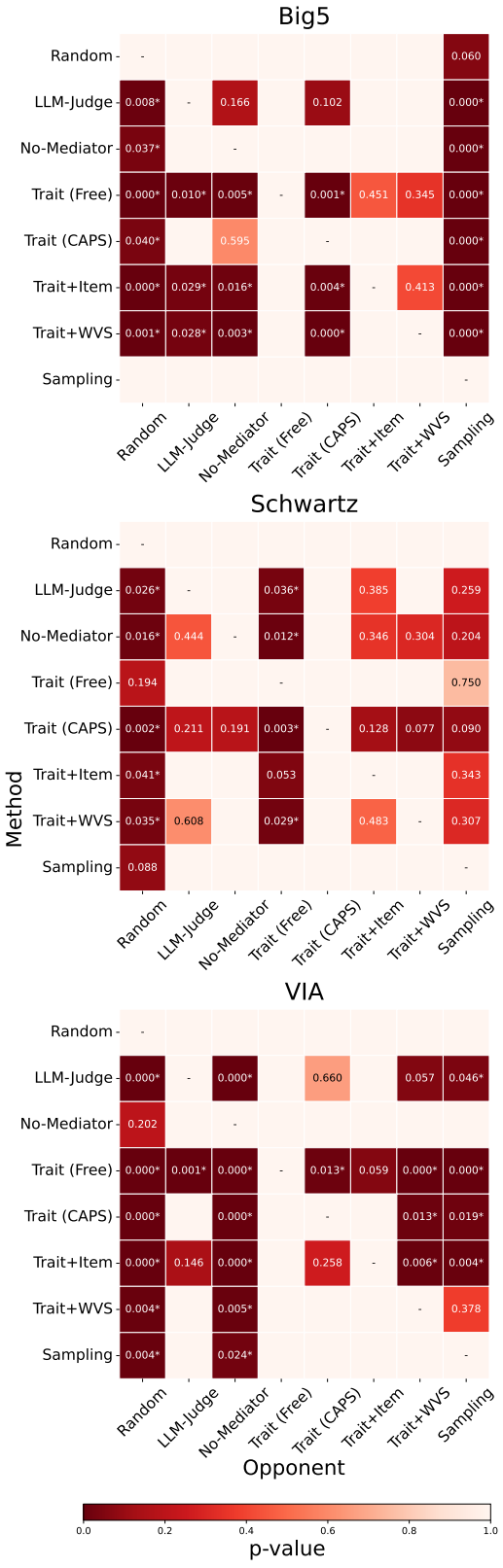

Figure 9: Results of significance test for Section 6. All reported p-values are based on a directional hypothesis of $y > x$; comparisons in the opposite direction ($y < x$) are omitted. An asterisk (*) indicates p-value < 0.05.

nation is that this result stems from the weaker validity of the Schwartz items. As shown in Table 4, the overall CV scores of Schwartz items are substantially lower than those for Big5 and VIA. Unlike other theories, Schwartz theory defines a circular structure in which positively correlated values (e.g., universalism and benevolence) are placed closely. Due to these inherent correlations between traits, it is more difficult to generate high-validity items, and methods may also struggle to distinguish higher- from lower-quality items.

J.2 Simulation Components

The results in Figure 10 show that our full setting (i.e., *Ours*) consistently outperforms all other combinations. Notably, more consistent improvements are observed for Big5 and VIA, while Schwartz exhibits comparatively less consistent results. This trend aligns with the findings in Appendix J.1. These findings indicate that all components contribute meaningfully to performance, with the full integration yielding the most reliable improvements.

K Maximum DV Scores

Method	Big5	Schwartz	VIA
Random	.594	.398	.578
LLM-Judge	.546	.399	.587
No-Mediator	.552	.399	**.549**
Trait (Free)	.568	**.363**	.594
Trait (CAPS)	.567	.411	.580
Trait+Item	.557	.383	.573
Trait+WVS	.585	.385	.584
Sampling	**.530**	.387	.581
Oracle	.593	.396	.585
Official	.624	.453	.573

Table 7: Maximum DV scores for each survey.

As a reference, we provide the maximum DV scores in Table 7. For each trait, we first identify the worst DV score, defined as the maximum Spearman correlation with non-target traits among the top N items selected by each method. Each cell then shows the average of these values over all traits.

The results show that the differences across methods are small, at most around 0.03 to 0.05. We also recommend interpreting the maximum DV score together with the CV score rather than on its own, as illustrated by the ratio of DV to CV.

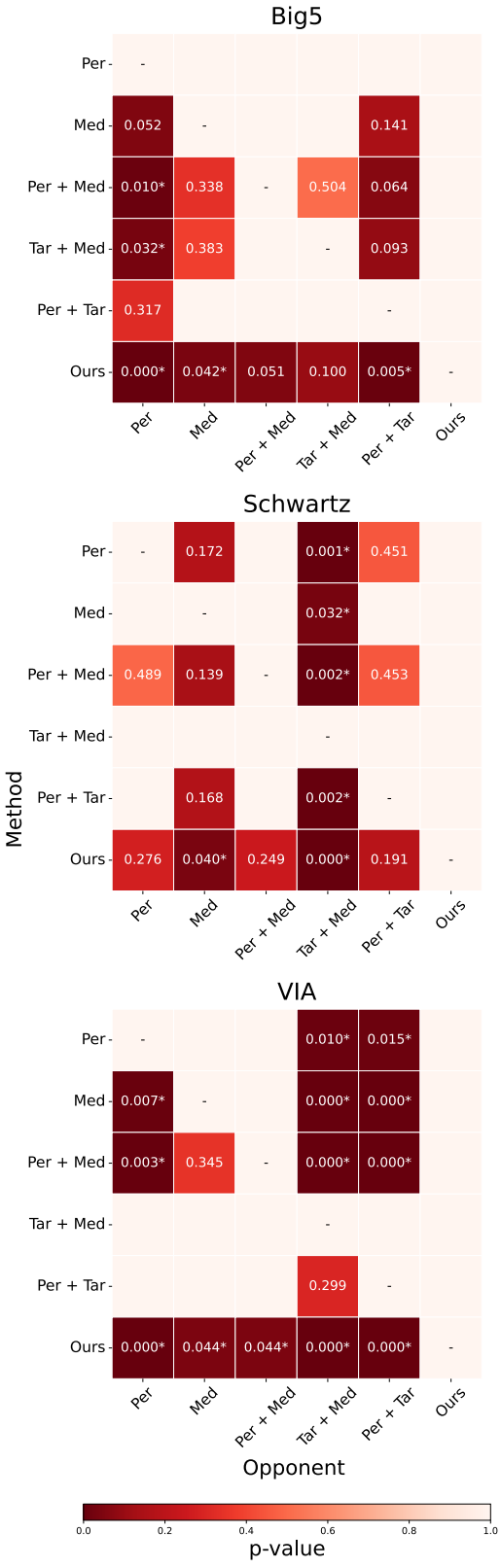

Figure 10: Results of significance test for Section 7.1. All reported p-values are based on a directional hypothesis of $y > x$; comparisons in the opposite direction ($y < x$) are omitted. An asterisk (*) indicates p-value < 0.05.

Method	Scale = 2			Scale = 3		
	CV↑	DV↓	ICR↑	CV↑	DV↓	ICR↑
Big5						
Random	.545	.276	.850	.545	.276	.850
LLM-Judge	.583	.284	.882	.595	.288	.890
No-Mediator	.555	.282	.867	.573	.291	.885
Trait (Free)	**.592**	.287	**.886**	**.620**	.292	**.900**
Trait (CAPS)	.561	.283	.864	.579	.290	.886
Trait+Item	.590	.289	.874	.615	.291	.889
Trait+WVS	.565	.283	.877	.592	.297	.896
Sampling	.538	**.272**	.849	.525	**.270**	.846
Schwartz's Theory of Basic Values						
Random	.234	**.134**	.533	.234	**.134**	.533
LLM-Judge	.292	.137	.646	.316	.140	.678
No-Mediator	.289	.138	.678	.319	.138	.745
Trait (Free)	.294	**.134**	**.700**	.313	.139	.745
Trait (CAPS)	.295	.138	.683	**.328**	.143	.740
Trait+Item	**.302**	.138	.647	.319	.140	.684
Trait+WVS	.293	.136	.689	.322	.136	**.749**
Sampling	.264	.138	.611	.280	.138	.653
VIA						
Random	.500	**.280**	.674	.500	**.280**	.674
LLM-Judge	.545	.286	**.740**	.564	.290	.772
No-Mediator	.507	.286	.695	.506	.284	.705
Trait (Free)	.543	.292	.737	**.566**	.297	**.778**
Trait (CAPS)	.534	.294	.729	.547	.298	.756
Trait+Item	**.547**	.281	.727	.568	**.280**	.748
Trait+WVS	.524	.298	.727	.527	.300	.743
Sampling	.515	.287	.703	.524	.288	.713

Table 8: Results of the ablation studies on item generation scale (with scales 2 and 3). **Bold**: best performance, Underlined: second-best performance.

L Ablation Studies on Item Generation Scale

To examine the robustness of our results with respect to the item generation scale, we conducted an ablation study with scales of 2 and 3, in addition to the original scale of 4. We constructed item pools of each scale by subsampling without replacement from the original scale, repeating this procedure 1,000 times. For each subsample, we selected the top N items using each method and computed the CV, DV, and ICR scores. We report the mean scores over all subsamples in Table 8.

The results show that Trait (Free), which achieves the best CV score in the main results (Table 4), maintains the best CV and ICR scores across all scales for the Big5, and the second-best CV score for VIA at scale = 3. For Schwartz, Trait (CAPS), the method with the best CV score in the main results, ranks best at scale = 3 and second-best at scale = 2. Overall, these findings suggest a largely consistent pattern across different item generation scales.

Category	Definition
Beliefs and Values	Core principles, convictions, and guiding ideals that influence thought, judgment, and behavior. This includes both abstract philosophical beliefs and personally significant moral or ethical values.
Emotions and Feelings	Subjective affective experiences, including moods, emotional states, and physiological responses associated with these experiences.
Habits and Behaviors	Recurrent actions, routines, and behavioral patterns that characterize an individual's daily life, reflecting both voluntary and automatic tendencies.
Preferences and Interests	Likes, dislikes, and areas of personal engagement that reflect subjective inclinations, tastes, and attentional priorities. These may involve enjoyment of specific activities or objects.
Self-Concept and Abilities	One's perceptions, evaluations, and beliefs about personal traits, skills, capacities, and potential for action; includes self-knowledge about competencies and identity.
Roles and Memberships	Positions, responsibilities, and affiliations within social, organizational, or community structures that influence identity and behavioral expectations.
Others' Perceptions of Me	How one is perceived, described, or evaluated by other people. This encompasses reputations, social feedback, and interpersonal impressions that can shape self-concept and guide behavior.
Environment and Context	External circumstances, situational factors, and physical or social settings that shape experiences, behavior, and decision-making.
Other	Miscellaneous aspects not covered by the above categories.

Table 9: Categories for mediators and their definitions.