



AI-powered Automatic Item Generation for Psychological Tests: A Conceptual Framework for an LLM-based Multi-Agent AIG System

Philseok Lee¹ · Mina Son¹ · Zihao Jia¹

Accepted: 12 August 2025
© The Author(s) 2025

Abstract

Large Language Models (LLMs) are transforming industrial-organizational psychology and human resource management, with one of their most promising applications being automatic item generation (AIG) for psychological test development. Although recent advances in LLM-based AIG—particularly for non-cognitive assessments such as personality—show significant potential, ensuring rigorous quality control remains a persistent challenge. This study introduces a novel AIG framework, the LLM-based Multi-agent AIG system (LM-AIG), where each agent is responsible for different stages of item development, including item generation, content review, linguistic evaluation, bias assessment, and item revision. The LM-AIG also incorporates human feedback to enhance item quality. We implemented the LM-AIG framework using the open-source tool AutoGen to generate items assessing attitudes toward the use of AI in the workplace. To evaluate the quality of the generated items, we conducted an empirical study based on structured ratings from human raters, assessing construct relevance, linguistic clarity, appropriate language level, contextual specificity, and potential bias. This paper further discusses the role of human-in-the-loop mechanisms within the LM-AIG system and outlines future research directions.

Keywords Automatic item generation · Large language model · Multi-agent AI system · Test development · Personnel assessment · Psychometrics

The application of artificial intelligence (AI) and machine learning (ML) is significantly reshaping the fields of industrial and organizational psychology and human resource management (Budhwar et al., 2023; Campion & Campion, 2023). These technological advancements have opened new possibilities for more innovative and efficient personnel assessments across various domains, including job analysis (e.g., Koenig et al., 2023; Putka et al., 2023), performance evaluations (e.g., Speer, 2018, 2021), employment selection (e.g., Campion & Campion, 2023; Fan et al., 2023; Hickman et al., 2023; Holtrop et al., 2022; Thompson et al., 2023), and psychological test development (e.g., Fyffe et al., 2024;

Götz et al., 2024; Hernandez & Nie, 2023; Hommel, 2023; Hommel et al., 2022; Lee et al., 2023).

Among recent developments, large language models (LLMs) have emerged as a particularly transformative development in personnel assessment. Trained on extensive and diverse textual corpora using the transformer¹ architectures (Vaswani et al., 2017), LLMs have demonstrated remarkable capabilities across a wide range of tasks, including text generation, classification, question-answering, logical reasoning, and information extraction (e.g., Brown et al., 2020; Liu et al., 2023; Radford et al., 2019; Wei et al., 2022). A key advantage of LLMs is their ability to perform these tasks through *natural language prompting*. This prompt-based approach, which relies solely on language instructions, significantly reduces the time and resources typically required

Additional supplementary materials may be found here by searching on article title <https://osf.io/collections/jbp/discover>

✉ Philseok Lee
plee27@gmu.edu

¹ Department of Psychology, George Mason University, 4400 University Drive, David King Hall Room 3056, Fairfax, VA 22030, USA

¹ The transformer architecture is a neural network design that processes all parts of a sequence simultaneously, unlike older models that go step by step. It uses self-attention to focus on important parts of the sequence and understand relationships between words, even when they're far apart. This approach makes transformers highly effective for tasks like language translation, text generation, and summarization.

for machine learning or fine-tuning processes (Demszky et al., 2023; Hao et al., 2024; Jia & Lee, 2025).

One of the most promising applications of LLMs is automated item generation (AIG) for test development in psychological and educational research (Bulut et al., 2024; Demszky et al., 2023; Tan et al., 2024). Traditionally, test development has been a labor-intensive and time-consuming process; however, AIG techniques provide the potential to significantly reduce the required effort while maintaining psychometric soundness. To date, most AIG research has focused on educational contexts, particularly in the development of academic aptitude, licensure, and knowledge-based assessments (see Gierl & Haladyna, 2012; Gierl & Lai, 2018; Kurdi et al., 2020). In contrast, the application of AIG to non-cognitive assessments, such as personality tests, has only recently gained traction in the fields of psychology with the advent of LLMs (e.g., Fyffe et al., 2024; Götz et al., 2024; Hernandez & Nie, 2023; Hommel et al., 2022; Lee et al., 2023; von Davier, 2018).

Despite recent advances in LLM-based AIG for non-cognitive assessments, significant opportunities for improvement remain, particularly in the domain of quality control. Although human oversight remains essential to ensure the accuracy, reliability, and fairness of AI-based assessments (Hao et al., 2024), current AIG studies typically use human reviewers in a post-hoc, retrospective manner, evaluating items only after generation rather than embedding quality control within the AIG process itself (e.g., Götz et al., 2024; Hernandez & Nie, 2023; Hommel et al., 2022; Lee et al., 2023). For example, Götz et al. (2024) used a culturally diverse, double-blind expert committee to assess item quality, such as conceptual relevance, linguistic clarity, and balanced keying. While rigorous, this post hoc approach may constrain the overall efficiency of item generation, highlighting the need for a more integrated and interactive quality control framework within the AIG pipeline.

Building on this line of inquiry, the present study introduces a multi-agent AI system—referred to as **LLM-based Multi-agent AIG** (LM-AIG). Importantly, this study does not aim to automate item generation merely to reduce time, effort, or cost. Instead, we conceptualize the AI multi-agents system as collaborative research partners in the pursuit of trustworthy automated item generation. The remainder of this paper is structured as follows. First, we review recent research on AIG for non-cognitive assessments. Next, we introduce the concept of multi-agent systems and their potential benefits for psychological test development. Building on this foundation, we propose an LM-AIG framework and demonstrate its application through an empirical example. We then evaluate content validity and other qualities of generated items based on structured ratings from human raters. Finally, we discuss the implications of the LM-AIG

framework, the role of human experts within the system, and directions for future research.

Early Efforts in LLM-Based AIG for Non-Cognitive Assessments

As previously noted, while AIG has been applied to educational and knowledge-based cognitive testing for several decades, its application to non-cognitive assessments is a relatively recent development. One of the earliest explorations in this area was conducted by von Davier (2018), who employed a long short-term memory (LSTM)² model (Hochreiter & Schmidhuber, 1997) to generate 24 items measuring the Big Five personality traits. This study demonstrated the potential of language models to automatically generate valid personality test items, highlighting a promising direction for psychological test development.

However, LSTM models present several limitations when applied to psychological test development. They tend to closely resemble existing items rather than generate novel ones that capture nuanced personality constructs and semantic meanings (Hommel et al., 2022). Moreover, as a supervised learning model, LSTMs require large volumes of human-annotated training data, which can be particularly challenging requirement for emerging or underexplored constructs. In von Davier's (2018) study, the model was trained on 3,320 personality items from the International Personality Item Pool (IPIP; Goldberg et al., 2006). While LSTM-based approaches have laid important groundwork for AIG in non-cognitive assessments, their inherent limitations suggest they may not represent the most effective solution for advancing research or practice in this field.

Applications of Transformer Architectures in AIG for Non-Cognitive Testing

Recently, researchers have increasingly adopted advanced *transformer* models for AIG in non-cognitive assessments (e.g., Fyffe et al., 2024; Götz et al., 2024; Hernandez &

² The LSTM model uses special “gates” (i.e., forget, input, and output gates) to control how information flows through the system. The forget gate decides what information from previous steps should be kept or discarded. The input gate determines how important new information is. The output gate controls how much information is passed on to the next stage of processing. These gates help LSTM maintain relevant information over long sequences, addressing a common problem in neural networks called the “vanishing gradient problem”.

Nie, 2023; Hommel et al., 2022, 2023; Lee et al., 2023). Transformer models offer two major advantages over their predecessors. First, the self-attention mechanism enables each word in a sentence to interact dynamically with all other words, allowing the model to capture more nuanced, context-specific meanings. Second, unlike earlier neural network models such as RNNs and LSTMs, which process inputs sequentially, transformers operate in parallel. This parallelization substantially reduces computational time and enhances efficiency when handling large-scale text datasets (Tunstall et al., 2022). Among these models, generative pre-trained transformers (GPT) have demonstrated particular effectiveness in text generation tasks and are now widely adopted in AIG research.

Hommel et al. (2022) employed GPT-2 (Radford et al., 2019) to generate items for the Big Five personality traits. They fine-tuned GPT-2 on large datasets of validated personality items, training the model to encode both construct labels and their corresponding items. In an empirical demonstration, the researchers generated 25 personality items and compared them to human-authored items. Results from a confirmatory factor analysis (CFA) indicated that the overall model fit was comparable between the machine-generated and human-authored scales. Notably, approximately two-thirds of the GPT-2-generated items exhibited acceptable psychometric properties, with factor loadings exceeding .40.

Expanding on this line of research, Hernandez and Nie (2023) adopted a novel approach by fine-tuning GPT-2 on a large corpus of existing personality items to enhance the semantic quality of the generated content. After fine-tuning, the model produced one million new items. Their analysis revealed that the AI-generated item pool (AI-IP) contained a significant number of unique items, demonstrating greater lexical diversity and novel statements compared to existing item banks. Their linguistic acceptability analysis confirmed that most items were grammatically correct, though slightly less accurate than those created by humans. Notably, they employed zero-shot classification models for content validity, allowing categorization of items into content domains without relying on labeled training data.

Götz et al. (2024) developed the Psychometric Item Generator (PIG), an open-source NLP algorithm powered by GPT-2, to automate the generation of personality items. They demonstrated PIG's effectiveness in producing items for both novel and established psychological constructs. For instance, they used PIG to develop items for a new construct, *wanderlust*, defined as a strong desire to travel and explore, substantially reducing the time and effort required for

manual item development. Additionally, PIG was employed to generate short-form scales for the Big Five traits, yielding reliable and valid instruments comparable to those developed by human experts.

In contrast to prior studies employing GPT-2, Lee et al. (2023) utilized a more advanced GPT-3³ model (Brown et al., 2020) to generate personality items measuring Big Five traits. Leveraging prompt engineering techniques, they generated 25 personality items and compared their psychometric properties to those of human-authored personality IPIP items. They also examined the measurement invariance of machine-generated personality items across gender groups. Overall, the results indicated that the GPT-3-generated items demonstrated strong psychometric properties, including reliability, discriminant and convergent validity, factor structure, and criterion-related validity. Moreover, the items demonstrated minimal evidence of gender-based measurement bias.

Building on previous research that primarily focused on generating items for broad Big Five personality constructs, Ho (2024) shifted attention to more specific facets of conscientiousness, including organization, productiveness, and responsibility. Using GPT-4 (OpenAI, 2024), the author investigated the influence of generation hyperparameters—particularly temperature⁴—on item quality. The findings indicated that a temperature setting of 1 was most effective for producing personality items with strong content validity. While the GPT-4-generated scales demonstrated psychometric properties comparable to those of human-authored scales, the latter still outperformed the machine-generated items on several psychometric indicators. Ho concluded that while LLMs show considerable promise for AIG, significant

³ GPT-3 substantially enhances model capacity compared to GPT-2, having approximately 100 times more parameters, with 175 billion parameters across 96 layers. It was trained on a vast corpus of 300 billion tokens of internet text data. This increased scalability enables GPT-3 to generate text with more complex semantic and syntactic structures, demonstrating greater robustness to grammatical errors. In addition, GPT-3 provides prompt-based few-shot learning (Liu et al., 2021). This feature allows the model to perform NLP tasks effectively with minimal examples, eliminating the need for extensive human-annotated training data or complex fine-tuning processes.

⁴ In LLMs, the *temperature* setting regulates the randomness or creativity of the model's responses. This parameter typically ranges from 0 to 2. Lower temperatures (e.g., 0 to 0.5) produce more predictable outputs, while higher temperatures yield more creative and varied responses. A temperature setting of 1 allows for a balanced level of creativity and flexibility, which can be particularly useful for generating diverse content, such as writing test items or simulating human reasoning.

Table 1 Summary of LLM-based AIG for non-cognitive assessments

Reference	Model ^a	Construct (base scale)	Generated initial item pool	Findings on the psychometric properties of machine-generated items
Von Davier (2018)	LSTM	Big Five personality (IPIP, Goldberg et al., 2006)	Not specified	<ul style="list-style-type: none"> Most machine-authored items load highly on the corresponding trait factor when exploratory factor analysis was conducted on the combined set of existing personality items and automatically generated items
Hommel et al. (2022)	GPT-2	Big Five personality (IPIP, Goldberg et al., 2006)	1,360	<ul style="list-style-type: none"> Machine-authored scales generally show acceptable internal consistency and structural validity (assessed via confirmatory and exploratory factor analyses). A significant number of items showed acceptable factor loadings. Machine-authored scales have lower reliability coefficients than human-authored scales for all Big Five dimensions except for openness to experience
Hernandez and Nie (2023)	GPT-2	Study 1: personality (IPIP, Goldberg et al., 2006), Study 4: Big Five personality (BFI-2, Soto & John, 2017), self-efficacy (Sierer et al., 1982), curiosity (Kashdan et al., 2018)	Study 1: 1,000,000 Study 4: 5,900	<ul style="list-style-type: none"> Study 1: Machine-authored items show greater diversity of words than and great similarity in structure (i.e., time length, readability, percentage of parts of speech) with that of human generated items Study 4: Machine-authored scales show good convergence with corresponding human-generated scales overall, although the magnitudes somewhat vary. They also exhibit internal consistency and structural validity (assessed via confirmatory factor analysis) that are similar to those of human-authored scales
Götz et al. (2024)	GPT-2	Demonstration 1: Wanderlust ^a Demonstration 2: Big Five personality (BFI-2, Soto & John, 2017)	Demonstration 1: 65 Demonstration 2: 100,000	<ul style="list-style-type: none"> Machine-authored items selected based on expert evaluations show robust psychometric properties by exhibiting evidence of internal consistency, test-retest reliability, structural validity (assessed via exploratory structural equation model and confirmatory factor analysis), and predictive validity for life outcomes Machine-authored items show acceptable internal consistency, structural validity (assessed via exploratory and confirmatory factor analyses), and criterion-related validity. Most machine-authored items were free of gender-based different item functioning. Machine-authored scales exhibit scalar measurement invariance between male and female groups
Lee et al. (2023)	GPT-3	Big Five personality (IPIP, Goldberg et al., 2006)	100	
Ho (2024)	GPT-4	Conscientiousness (BFI-2 conscientiousness scale, Soto & John, 2017)	12 items for each of nine hyperparameter conditions	<ul style="list-style-type: none"> The best combination of hyperparameters to generate items with good content validity is temperature of 1 and presence penalty of 0 Machine-authored scales generally show internal consistency, structural validity (assessed via confirmatory factor analysis), criterion-related validity, and item and test information comparable to those of human-authored scales, however human-authored scales still outperformed machine-authored scales consistently across these indicators

LSTM Long short-term memory, *GPT* Generative Pre-trained Transformer. * We presented studies based on the order of GPT models

^a No existing scale was used

challenges remain in developing high-quality items for *nuanced* facets of complex psychological constructs. Table 1 provides a summary of recent studies on LLM-based AIG for non-cognitive assessments.

Addressing Gaps in LLM-Based AIG for Non-Cognitive Assessments

Much of the existing research on AIG has largely relied on post hoc evaluations by human raters to assess content validity (e.g., Götz et al., 2024; Hommel et al., 2022; Lee et al., 2023; von Davier, 2018). While effective, this retrospective manual review process could be time consuming and labor intensive, particularly when applied to extensive item pools, thereby limiting the scalability and overall efficiency of AIG systems (Circi et al., 2023). For instance, evaluating 1,000 newly generated items would require a substantial commitment of expert time and effort. To address this bottleneck, Hernandez and Nie (2023) applied zero-shot classification models to automate content validity evaluation. In their approach, the model estimated the likelihood of an item aligning with each of the five personality dimensions and assigned the item to the dimension with the highest probability. This method achieved an average classification accuracy of 64.31%, suggesting that further improvement is needed. Similarly, Fyffe et al. (2024) applied transformer-based models, such as BERT, DeBERTa, and RoBERTa, using few-shot learning techniques to classify items into content domains.

However, classification models typically produce numerical metrics, such as accuracy rates, which are insufficient for providing actionable insights into item quality. In the AIG process, meaningful feedback, evaluative judgments, and specific recommendations are more valuable than aggregate performance scores. By incorporating established psychometric practices for content validity evaluation (e.g., Hinkin & Tracey, 1999), AIG could better assess how well an item aligns with its intended construct (i.e., correspondence) and how effectively it differentiates the target construct from related constructs (i.e., distinctiveness). This content-validity-based evaluation would offer more targeted, constructive feedback at various stages of the AIG process. Aligning AIG with established guidelines for content validity (e.g., Colquitt et al., 2019) would improve the rigor, interpretability, and quality of generated items.

In addition, LLMs are susceptible to inheriting and perpetuating societal biases embedded in their training data, including gender stereotypes, racial biases, and cultural assumptions (Gallegos et al., 2024). These biases may manifest in various forms, such as language patterns, contextual framing, and implicit associations, often skewing

content toward the perspectives of majority groups (Caliskan & Lewis, 2020; Charlesworth & Banaji, 2021; Gallegos et al., 2024). As a result, LLM-generated items may lack cultural sensitivity and inclusivity, leading to differential item functioning (DIF) that systematically advantages or disadvantages specific subgroups (Lee et al., 2023). Such bias not only undermines the fairness of assessments but also threatens their validity, potentially leading to inequitable decisions in organizational and educational contexts (Vandenberg, 2002). Despite growing awareness of this issue, few AIG studies have incorporated systematic bias auditing or mitigation strategies into the AIG pipeline. Hommel et al. (2022) underscore the urgent need for more rigorous methods to identify and address item-level bias, as well as to ensure measurement invariance in future LLM-based AIG research. Accordingly, developing and implementing effective auditing strategies is critical to enhancing the fairness and psychometric quality of LLM-based AIG systems.

Another key area for advancement is the integration of real-time grammatical error correction capabilities. Hernandez and Nie (2023) reported that approximately 8% of the items generated by GPT-2 contained grammatical errors. While newer LLMs such as GPT-4 may produce more linguistically accurate items, grammar evaluation still remains a critical and labor-intensive component of the AIG process. As Hernandez and Nie (2023) noted, “rather than use these grammar-checking models retrospectively, future versions of this system might want to apply them prospectively when developing the initial item pool (p. 1030)”. Moreover, integrating real-time grammatical validation could improve multilingual item generation, supporting the broader application of LLM-based AIG across diverse linguistic and cultural contexts.

Finally, most LLM-based AIG research lack the integration of direct human input during the item generation process. Incorporating a human-in-the-loop mechanism could substantially improve item quality by introducing critical oversight and transforming AIG from a static, unidirectional system into an interactive, collaborative process. Real-time human feedback allows for the immediate identification and correction of problematic items, reducing the accumulation of errors and the need for extensive post hoc revisions. This synergy between AI and human judgment promotes greater alignment with psychometric principles and enhances both the efficiency and validity of the AIG process.

Introduction of a Multi-Agent Framework

To address the challenges identified in prior research, we propose an LLM-based Multi-agent AIG system. A multi-agent system consists of autonomous artificial entities

capable of independently perceiving their environment, processing information, making decisions, and executing actions to achieve designated objectives (Cañas, 2022; Cheng et al., 2024). While the evolution of artificial agents has undergone several stages (see Ginsberg, 2012; Wilkins, 2014), recent advancements in LLMs have markedly enhanced agent capabilities, enabling them to operate more intelligently and efficiently in complex environments (Cheng et al., 2024; Guo et al., 2024; Li et al., 2024).

Developing an effective LM-AIG system requires careful consideration of several core components. The first is the *agent-environment interface*, which defines the operational context and guides agents' perception and decision-making. Such interfaces have been successfully applied across a range of domains, including software development (Qian et al., 2023), games (Park et al., 2023), robotics (Zhang et al., 2023), economics (Li et al., 2023), political science (Xiao et al., 2023), and public health (Williams et al., 2023). In the context of psychological assessment, the environment must be structured to reflect relevant psychological constructs, test specifications, and item development guidelines. Aligning the interface with these domain-specific elements enables agents to interpret requirements accurately and generate items that are psychometrically appropriate and contextually valid.

The second component is the *definition of agent profiles*, which includes specifying each agent's traits, behaviors, knowledge bases, and roles. In the context of AIG, agents take on roles similar to a team of human experts, such as item writers, content reviewers, linguistic reviewers, and bias reviewers. Guo et al. (2024) categorize agent profiling methods into three types: i) model-generated, ii) data-derived, and iii) pre-defined. **Model-generated profiles** are created automatically by LLMs, which assign roles and expertise based on specific needs. While this approach may be efficient when managing many agents, it may lack control and precision over the outcomes. In contrast, data-derived profiles are constructed using real-world datasets incorporating demographic information such as race, ethnicity, gender, age, and location to enhance realism. However, this approach may be resource-intensive and requires extensive data, especially when generating diverse agent pools. For the purposes of this study, we adopt a pre-defined profiling strategy, wherein agent roles and attributes are manually specified according to established psychometric frameworks (e.g., Clark & Watson, 2019; Hinkin, 2005; Lambert & Newman, 2023; Zickar et al., 2020). This approach ensures alignment with best practices in test development and provides greater control over agent behavior and task execution.

The third consideration is *inter-agent communication style*. Guo et al. (2024) identify three primary communication styles: i) cooperative, ii) debate, and iii) competitive. In cooperative communication, agents collaborate towards

shared goals, exchanging information to improve collective outputs. This communication style has been effectively applied in domains such as software development (e.g., Hong et al., 2023; Qian et al., 2023) and robotic collaboration (e.g., Mandi et al., 2024). The debate model, by contrast, involves argumentative interactions in which agents present, defend, and critique differing viewpoints. This approach is particularly effective for reaching consensus or refining solutions, as demonstrated in scientific debate simulations (e.g., Du et al., 2023; Chan et al., 2023). In competitive communication, agents pursue individual goals that may conflict with one another. This approach has been used to simulate economic theories (e.g., Zhao et al., 2023) and war scenarios (e.g., Hua et al., 2023). For this study, we adopt a cooperative communication style, reflecting the collaborative nature of typical test development. Agents will work together to refine and improve test items, sharing insights and building on each other's contributions to produce high-quality, psychometrically sound results.

A fourth consideration involves the *communication structure* that governs interactions among agents. Guo et al. (2024) identify three primary structures: layered, centralized, and decentralized. In a **layered structure**, agents are organized hierarchically, with each level assigned specific roles, typically interacting within their own layer or with adjacent ones. This structure has proven effective in domains such as software development (e.g., Hong et al., 2023; Qian et al., 2023). Decentralized communication, by contrast, operates on a peer-to-peer network, where agents interact directly with each other. It is commonly used in modeling human behavior during epidemics (e.g., Ghaffarzadegan et al., 2024), simulating war scenarios (e.g., Hua et al., 2023), and designing communication games (e.g., Xu et al., 2023). Centralized communication, in turn, relies on a central agent or group of agents to coordinate interactions, with other agents communicating primarily through this hub. This structure has been effectively implemented in multi-robot planning (e.g., Chen et al., 2024) and chemistry laboratory simulations (e.g., Zheng et al., 2023), offering a more controlled and directed flow of communication. In this study, we adopt a hybrid communication structure that integrates the strengths of these approaches. Agents are organized into specialized task groups—such as item generation, content review, bias review, and linguistic review—following a layered hierarchy, while a central meta-editor agent manages coordination, synthesizes feedback, and ensures alignment with psychometric standards and test specifications. This hybrid approach facilitates both operational efficiency and adaptive collaboration in the development of test items.

Finally, *feedback integration* is essential for improving adaptability and quality control within LM-AIG systems. Wang et al. (2024) categorize feedback mechanisms into three types: i) environmental feedback, ii) agent interaction feedback, and

iii) human feedback. Environmental feedback comes from the real or virtual environments in which the agents operate as seen in programming environments (e.g., Hong et al., 2023; Qian et al., 2023) or robotic systems (e.g., Chan et al., 2023; Mandi et al., 2024). Agent interaction feedback occurs through inter-agent communication and evaluation. Agents refine their outcomes by engaging in dialogue and critical evaluation with one another, often used in problem-solving scenarios like simulations or debates (e.g., Aher et al., 2023; Park et al., 2022). By contrast, **human feedback** incorporates expert input and domain-specific knowledge directly into the system. In this study, we adopt the human feedback approach, ensuring that the quality of generated items is monitored and validated by human subject matter experts. This ensures that items meet high standards of relevance, accuracy, and psychometric validity throughout the AIG process.

Conceptual Framework of the LLM-Based Multi-Agent AIG System

Building on the foundational concepts outlined above, we introduce the LM-AIG system specifically designed for psychological assessment contexts. As illustrated in Fig. 1, the

proposed framework consists of autonomous, specialized agents, each responsible for a distinct function within the item generation and review process. This system reflects best practices in test development by ensuring that evaluative agents operate independently of item generation processes, thereby enhancing objectivity and reducing potential bias (Boateng et al., 2018).

The LM-AIG process begins with the Item Writer Agent, which is responsible for the creation of item pools. Its primary task is to generate items that accurately reflect the operational definition of the target construct, while adhering to established best practices in item writing. To achieve this, the agent draws on detailed item specifications—including construct definitions, item development guidelines (see Table 2), and carefully crafted prompts—to ensure the psychometric quality of the items.

Following item generation, a multi-stage review process is initiated, coordinated by the Critic Agent. This process is structured through a nested communication structure involving three specialized reviewer agents—each dedicated to a distinct aspect of item quality: content, language, and bias (e.g., Wendler & Burrus, 2013; Zieky, 2013). Within this framework, the Content Reviewer Agent, Linguistic Reviewer Agent, and Bias Reviewer Agent interact

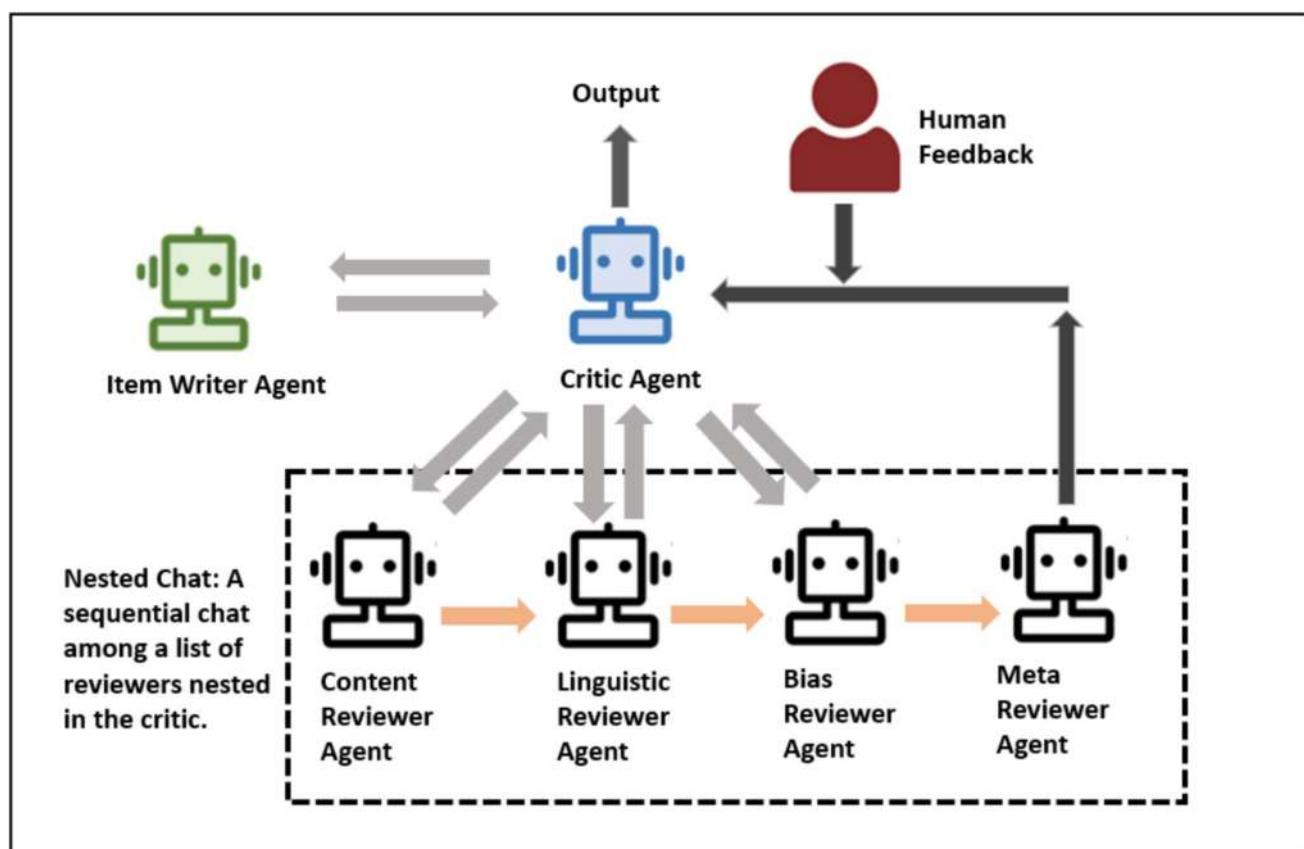


Fig. 1 A Simplified Interactions of Agents in the LM-AIG System

Table 2 Guideline for the Item Generation

Agent	Guideline
Item writer	<ul style="list-style-type: none"> • Use short and simple language (Clark et al., 2019; Haladyna & Rodriguez, 2013; Hinkin, 2005; Jebb et al., 2021) • Avoid jargon, slang, difficult vocabulary, unfamiliar technical terms, and vague or ambiguous terms (Lambert & Newman, 2023; Peter et al., 2018) • Items should be appropriate for the reading level of the target population (Clark et al., 2019; Hinkin, 2005) • Write significantly more items than needed (e.g., three times: Lambert & Newman, 2023; Zickar, 2020) • The use of negative and reverse-coded items is controversial. Many studies show the detrimental effects of negative or reverse-coded items (DiStefano & Motl, 2006; Schriesheim & Eisenbach, 1995; Sliter & Zickar, 2014; Sonderen et al., 2013; Woods, 2006). For example, there is evidence that reverse-keyed items may be confusing to participants, that the opposite of a reverse-keyed construct may be fundamentally different from the construct, and that reverse-keyed items tend to negatively affect the factor structure of scales. Item writers are often advised to avoid negative items • Avoid double-barreled items (Clark et al., 2019; Lambert & Newman, 2023; Zickar, 2020). For example, “My manager is intelligent and enthusiastic” should be not be used • Avoid items that virtually everyone (e.g., “Sometimes I am happier than at other times”) or no one (e.g., “I am always furious”) will endorse (Clark et al., 2019; Hinkin, 2005) • Avoid colloquialisms that may be not be familiar across age, ethnicity, region, gender, and so forth (Clark et al., 2019) • Avoid vague words such as many, most, often, or sometimes. These have no formal quantity and so represent an open invitation to miscomprehension. For example, contains an item “Overall, how satisfied are you with your life nowadays?” Nowadays is a vague term. A better item would be “Overall, how happy have you been with your life over the last three months.” (Hardy & Ford, 2014) • It is important to keep all items consistent in terms of perspective, being sure not to mix items that assess behaviors with items that assess affective responses (Harrison & McLaughlin, 1993). An example of this would be including items such as “My boss is hardworking” and “I respect my boss” in the same measure
Content reviewer	<ul style="list-style-type: none"> • Use expert ratings to assess whether items adequately represent the domain of interest (Anderson & Gerbing, 1991; Colquitt et al., 2019; Hinkin & Tracey, 1999) <p>For example: Using the 7-point scale, please rate the following items on how well each matches the concept. 1: Item does an EXTREMELY BAD job of measuring the concept; 2: Item does a VERY BAD job of measuring the concept; 3: Item does a SOMEWHAT BAD job of measuring the concept; 4: Item does an ADEQUATE job of measuring the concept; 5: Item does a SOMEWHAT GOOD job of measuring the concept; 6: Item does a VERY GOOD job of measuring the concept; 7: Item does an EXTREMELY GOOD job of measuring the concept (Hinkin & Tracey, 1999)</p>
Linguistic reviewer	<ul style="list-style-type: none"> • Check for grammatical accuracy and stylistic consistency • Check the level of language used in the item. It is likely difficult for the average respondent in the United States to understand and respond to surveys that contain items that require more than a seventh- to eighth-grade reading level • Check for items that contain unnecessary negative language, • Check for whether the item is confusing, unnecessarily difficult, or appears to be tricky or double-barreled? (Calderón et al., 2006; Wendler & Burrus, 2013)
Bias reviewer	<ul style="list-style-type: none"> • Check for potential bias in items that may disadvantage certain demographic groups, such as gender, religion, race, age, and culture (Zieky, 2013)
Meta editor	<ul style="list-style-type: none"> • Synthesize feedback from all previous agents • Edit items as needed and discard items that cannot be fixed • Integrate human expert insights and suggestions • Identify any remaining issues

dynamically with the Critic Agent, ensuring a comprehensive and collaborative evaluation.

The Content Reviewer Agent evaluates the extent to which each item reflects its intended construct and maintains content validity. We incorporate Hinkin and Tracey's (1999) content validity method into this process. In line with

their recommendation, the Content Reviewer is designed to represent naïve judges, such as typical employees as a representative sample, rather than human raters formally trained in psychology or psychometrics. The Content Reviewer Agent uses a 7-point scale to evaluate alignment with the construct definition and computes both *correspondence* and

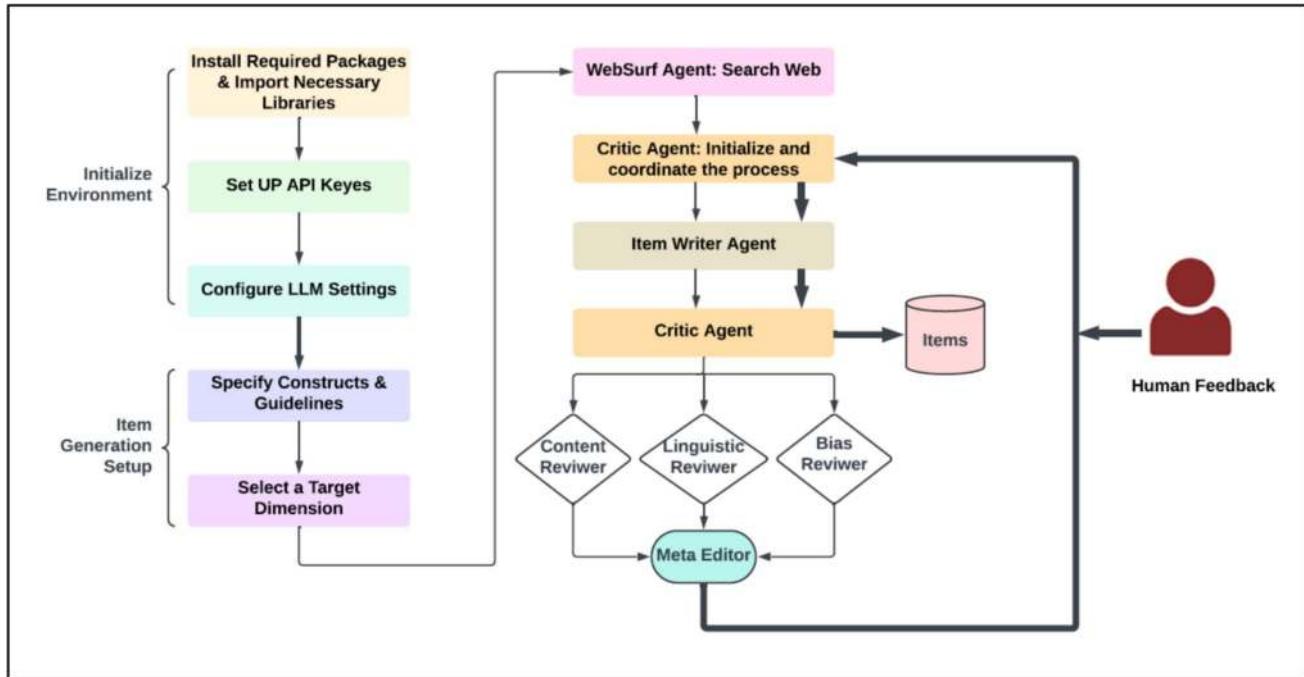


Fig. 2 Workflow of LM-AIG System via AutoGen

distinctiveness indices, as outlined by Colquitt et al. (2019), to assess the item's content validity. Next, the Linguistic Reviewer Agent examines the grammatical correctness and readability of each item. This agent also identifies issues such as confusing or overly complex wording, double-barreled items, or unnecessary negative language (e.g., Wendler & Burrus, 2013). Finally, the **Bias Reviewer Agent** assesses each item for potential sources of demographic bias, including but not limited to gender, race, age, religion, and cultural background (e.g., Zieky, 2013). This agent uses a 5-point scale to assess the level of bias in each item, with items rated 5 considered bias-free.

The review process culminates with the Meta Reviewer Agent, which synthesizes feedback from all preceding agents. This agent integrates the feedback and recommendations, resolves conflicting evaluations, and ensures that each item meets the required psychometric standards. Crucially, the Meta Reviewer Agent also incorporates human feedback, allowing for expert oversight and further ensuring the quality, relevance, and fairness of the final item pool. By

integrating human feedback, the system maintains a balance between automated efficiency and expert oversight, addressing potential limitations of purely AI-based evaluations.

Demonstration of Multi-Agent AIG in Practice

In this section, we demonstrate the application of proposed LM-AIG system using the construct Attitudes Toward the Use of AI in the Workplace (AAAW), developed by Park et al. (2024). The AAAW framework comprises six dimensions: i) *human-likeness of AI*, ii) *adaptability of AI*, iii) *quality of AI*, vi) *AI use anxiety*, v) *job insecurity*, and v) *personal utility*.

In contrast to prior AIG research on non-cognitive assessments, which has predominantly focused on the Big Five personality traits (e.g., Götz et al., 2024; Hernandez & Nie, 2023; Hommel et al., 2022; Lee et al., 2023; von Davier, 2018), we selected the newly developed and validated AAAW constructs for three key reasons. First, the widespread availability of Big

Five personality items in public databases poses a risk of information leakage, whereby LLMs may inadvertently replicate or closely mimic existing items rather than generate novel content. Second, focusing on a newer construct allows us to demonstrate the flexibility and adaptability of the LM-AIG system in generating items for emerging psychological concepts. Third, by using a validated existing scale, we were able to benchmark the performance of our LM-AIG framework against established, psychometrically sound content.

To implement the LM-AIG system, we utilized AutoGen, an open-source tool designed to orchestrate, optimize, and automate multi-agent workflows. AutoGen supports various agent types, including tool-using agents and human-in-the-loop configurations, providing flexibility in task execution across diverse contexts. More information about AutoGen is available at <https://microsoft.github.io/autogen/>. The demonstration was conducted using Google Colab, a cloud-based platform for executing and sharing Python code. This section presents descriptions of each agent alongside their corresponding outputs. The example code used in this demonstration is accessible via the provided Google Colab link.⁵ Figure 2 illustrates the overall workflow of the LM-AIG system as implemented using AutoGen.

We began by installing the required packages and importing the necessary libraries. Subsequently, we configured the execution environment by setting up API⁶ keys for OpenAI and Perplexity Web Search. These keys are essential for enabling text generation (via OpenAI) and web search capabilities (via Perplexity). Readers can obtain the required API keys by registering with the respective OpenAI and Perplexity platforms.

After configuring the environment, we specified the language model settings, including model type and temperature. We used the **GPT-4o** model, recognized for its advanced generative capabilities. Recent studies (e.g., Jia & Lee, 2025; Lee et al., 2024; Speer et al., 2024) have shown that generative LLMs, such as GPT-4 and GPT-4o, can effectively rate or score responses, exhibiting

psychometric properties comparable to those of supervised NLP models. Following the best practices recommended by Ho (2024), we set the **temperature** parameter to 1.0 for AIG. In contrast, for the summarization task, we set the temperature to 0 to minimize stochasticity, thereby ensuring more precise, fact-based outputs from web search results.

To generate items, we defined the six AAAW constructs in a single Python dictionary for clarity and code efficiency. The dictionary includes definitions and sample items for each of the six dimensions as described in Park et al. (2024). Additionally, we implemented structured item-writing guidelines, summarized in Table 1, to ensure that all generated items align with established psychometric standards. Although items were generated for all six dimensions, this demonstration section focuses specifically on the *AI use anxiety* dimension.

Next, we initialized the WebSurferAgent to collect content related to employee attitudes toward AI use anxiety. The WebSurferAgent leverages Perplexity's web search capabilities to retrieve relevant content from diverse online sources, including academic articles, news, industry reports, and blogs. To prevent information leakage, we explicitly instructed the WebSurferAgent to exclude the study by Park et al. (2024) from its search queries.

To streamline interaction with the web search functionality, we defined the **UserProxyAgent**, which facilitates communication with the WebSurferAgent and automates the process of gathering web-based information. After specifying the task parameters, we initiated the interaction between web_surfer and user_proxy using user_proxy.initiate_chat. Below is an excerpt of the output generated by the WebSurferAgent. The complete set of references retrieved via the Perplexity API can be accessed by printing the structured response citations; however, for demonstration purposes, a condensed version is presented here. Then, the system compiles and saves the results as a **summary**.

⁵ https://drive.google.com/file/d/1fq1GHQ_JR9dyWXxYf6CY340zMIXhY7AQ/view?usp=sharing

⁶ An API (Application Programming Interface) is a tool that allows different software systems to communicate and share data with each other. It provides a standardized way for users to access specific features or services from another application without needing to know its internal details.

=====

AI use anxiety is a growing concern among workers as artificial intelligence becomes more prevalent in the workplace. Many employees feel intimidated or apprehensive about using AI tools in their daily work tasks. Here are some key points about AI use anxiety:

Prevalence of AI Anxiety

Recent surveys indicate that AI anxiety is widespread:

- 71% of employees are concerned about AI negatively impacting their job security[1] https://www.ey.com/en_us/newsroom/2023/12/ey-research-shows-most-us-employees-feel-ai-anxiety
- 38% of workers fear AI may make some or all of their job duties obsolete[7] <https://www.apa.org/topics/healthy-workplaces/artificial-intelligence-workplace-worry>
- 48% of employees are more concerned about AI now than they were a year ago[1] https://www.ey.com/en_us/newsroom/2023/12/ey-research-shows-most-us-employees-feel-ai-anxiety

Causes of AI Use Anxiety

Several factors contribute to employees feeling anxious about using AI at work:

- Lack of training and skills: 73% worry they won't have opportunities to learn AI skills[11] https://www.ey.com/en_us/consulting/businesses-can-stop-rising-ai-use-from-fueling-anxiety
- Concerns about job obsolescence: 75% fear AI will make certain jobs obsolete[1] https://www.ey.com/en_us/newsroom/2023/12/ey-research-shows-most-us-employees-feel-ai-anxiety
- Uncertainty about AI capabilities: Many are unsure how to effectively use AI tools[3] <https://www.hrotoday.com/hr-technology/employees-unsure-how-to-effectively-use-ai/>
- Fear of making mistakes: Workers worry about using AI incorrectly[4] <https://www.hrdive.com/news/workers-admit-making-ai-fueled-errors/746633/?ref=charterworks.com>

Impact on Mental Health and Performance

AI anxiety can have significant negative effects:

- 51% of those worried about AI say it negatively impacts their mental health[7] <https://www.apa.org/pubs/reports/work-in-america/2023-work-america-ai-monitoring>
- 66% of AI-anxious workers report burnout, compared to 40% of workers overall[10] <https://www.hrotoday.com/news/ticker/over-half-of-employees-use-ai-to-complete-work-trainings/>
- Workers anxious about AI are 27% less likely to stay with their employer[8] <https://www.techtarget.com/searchhrsoftware/news/366557854/Fear-of-AI-might-increase-workplace-turnover>

Strategies to Address AI Use Anxiety

Employers can take several steps to ease anxiety and support employees:

- Provide comprehensive AI training and upskilling opportunities[4][8]
- Communicate transparently about how AI will be used in the organization[3][9]
- Position AI as augmenting human work rather than replacing it[9]
- Involve employees in AI implementation processes[8]
- Designate "AI champions" to support colleagues in using new tools[8]
- Offer mental health resources to help employees cope with anxiety[4]

By addressing AI use anxiety proactively, organizations can help employees feel more comfortable and confident in leveraging AI technologies effectively in their work. This is crucial for successful AI adoption and maintaining a positive workplace culture.

Next, the Critic Agent, which coordinates the entire LM-AIG workflow, initiates the item generation process. The Item Writer Agent then generates an initial pool of eight items.⁷ In the system prompt, the agent is provided with a role definition emphasizing its advanced expertise in psychological scale development, psychometric theory, and the target construct. It is instructed to generate

eight items specifically targeting the *AI Use Anxiety* dimension, drawing on both psychological constructs and the insights retrieved by the WebSurferAgent. To ensure adherence to psychometric standards, the agent is also given detailed item-writing guidelines (see Table 2). The initial set of eight items generated by the Item Writer Agent is as follows:

=====

Critic (to Writer) :

Hello Item Writer Agent! Generate scale items for the given dimension.

Writer (to Critic) :

Certainly! Here are 8 Likert-type scale items for the dimension "AI use anxiety" which captures workers' anxious feelings about the practical usage of AI at work:

1. I feel uneasy about the idea of using AI tools in my daily work tasks.
2. The thought of making mistakes when using AI at work makes me anxious.
3. I worry about not having enough skills to effectively use AI at work.
4. I feel nervous when I need to use AI to accomplish my job duties.
5. Using AI at work feels intimidating to me.
6. I often feel apprehensive when I need to rely on AI technology.
7. I worry about how AI might change my role at work.
8. The possibility of AI making parts of my job obsolete makes me anxious.

These items are designed to measure the specific dimension of AI use anxiety in a straightforward and concise manner.

Next, within the nested structure of the Critic Agent, the Content Reviewer Agent evaluates the content validity of the generated items based on the *correspondence* and *distinctiveness* indices proposed by Hinkin and Tracey (1999). To facilitate this evaluation, the Content Reviewer Agent is provided with definitions for three constructs: the **target construct** and two orbiting constructs (i.e., conceptually related but distinct constructs). Using this information, the agent rates each item on a 7-point Likert scale and calculates the two key indices following the procedures outlined by Colquitt et al. (2019):

- Correspondence (c-value) = $\frac{\text{average definitional correspondence rating}}{a}$
- Distinctiveness (d-value) = $\frac{\text{average of all (intended correspondence rating} - \text{orbiting correspondence rating})}{a-1}$
- where a represents the number of anchors (e.g., $a-1=6$ for a 7-point scale)

According to Colquitt et al. (2019), items with a c-value greater than 0.88 and a d-value greater than 0.35 are considered to have strong content validity. We classified them as 'meeting' the criterion. Items below these thresholds are categorized as 'not meeting' the criterion. For items that do not meet the specified criterion, the Content Reviewer Agent offers feedback to guide further refinement. An excerpt of the output is presented below.

⁷ For demonstration purposes, we doubled the original number of items from four to eight; however, researchers need to adjust this quantity based on the specific objectives of their study.

=====ContentReviewer (to Critic):=====

To evaluate the items given for AI use anxiety:

Construct 1: AI Use Anxiety

Construct 2: Perceived Adaptability of AI

Construct 3: Perceived Quality of AI

Item	Construct 1	Construct 2	Construct 3	c-value	d-value	Meets Criterion	
1	7	2	1	1	0.83	Yes	
2	7	1	2	1	0.83	Yes	
3	7	2	1	1	0.83	Yes	
4	7	1	2	1	0.83	Yes	
5	7	1	2	1	0.83	Yes	
6	7	1	2	1	0.83	Yes	
7	6	3	1	0.76	0.58	No	
8	6	3	2	0.76	0.42	No	

Detailed Feedback for Items that Do Not Meet Criteria:

Item 7:

- **Current Assessment:** This item posits a concern on the impact of AI on job roles. While it captures a potential anxiety, it diverges by hinting at potential changes brought by AI (an adaptability and quality concern).
- **Recommendation for Refinement:** Reshape the item to more directly reflect anxiety. For example, "I feel anxious about the uncertainty of my role as AI technology evolves at work."

Item 8:

- **Current Assessment:** This item captures anxiety over job obsolescence, a valid component of AI anxiety. However, it indirectly references adaptability (how AI might replace human roles).
- **Recommendation for Refinement:** Focus on personal anxiety: "I often feel anxious about AI potentially diminishing my tasks at work."

By refining the items, workers' genuine anxieties about AI's immediate use in their daily tasks can be more adequately assessed.

Next, the Linguistic Reviewer Agent evaluated the *AI Use Anxiety* items to ensure they meet linguistic standards concerning grammatical accuracy, clarity, simplicity, and appropriateness. The agent employed a 5-point rating scale, where:

- A score of 1 (very poor) indicates major linguistic issues, such as grammatical errors, confusing sentence structure, or language exceeding a naïve rater's reading level.

- A score of 5 (excellent) indicates that the item is grammatically correct, stylistically consistent, and expressed in clear, accessible language suitable for a naïve rater.

Items that receive a score of 5 are considered linguistically sound. For items scoring 4 or below, the agent provides a detailed explanation of the issues and suggests possible improvements. An excerpt of the output provided by the Linguistic Reviewer Agent is presented below.

=====
Linguistic Reviewer (to Critic) :

The provided items on "AI use anxiety" are generally well-constructed and effectively capture concerns about using AI at work. However, there is room for refinement to ensure clarity and precision in some items.

1. **Grammatical Accuracy and Stylistic Consistency**: The items are grammatically correct and stylistically consistent. They maintain a clear focus on the theme of AI-related anxiety.
2. **Ease of Understanding**: The language used is straightforward and should be easily comprehensible for individuals with a seventh- to eighth-grade reading level.
3. **Avoidance of Unnecessary Negative Language**: The items focus on the anxiety dimension without unnecessary negative language, although they inherently deal with a negative sentiment (anxiety).
4. **Clarity and Directness**: Most items are clear and direct, but items 7 and 8 could be further refined to reduce confusion and enhance focus on personal anxiety rather than broader implications.

****Detailed Feedback:****

- **Item 7**: "I worry about how AI might change my role at work."
 - *Feedback*: This item implies potential role changes but could more directly reflect personal anxiety. Your suggestion is more targeted but can still be simplified. Consider rephrasing to " I feel anxious about possible changes in my job role due to AI."
- **Item 8**: "The possibility of AI making parts of my job obsolete makes me anxious."
 - *Feedback*: The original item directly addresses anxiety about obsolescence. Your suggestion narrows the focus but could be rephrased for simplicity: "I feel anxious that AI might make parts of my job unnecessary."

By refining the specified items, the overall set can better maintain focus on individual anxiety concerning AI use without diverging into broader implications or adaptability concerns.

=====

Following the linguistic review, the Bias Reviewer Agent evaluates each item for potential biases related to gender, ethnicity, race, age, sexual orientation, and culture. The agent uses a 5-point rating scale, where:

- A rating of 1 (highly biased) is assigned to items that contain significant or explicit bias, making it inappropriate for use.
- A rating of 5 (completely unbiased) indicates that the item is free from any identifiable bias and is fully appropriate for diverse populations.

Items receiving a score of 5 are considered unbiased and suitable for inclusion in psychological assessments. For items rated 4 or below, the Bias Reviewer Agent

provides an explanation identifying the source of bias and offers specific suggestions for improvement. An excerpt of the Bias Reviewer Agent's output is provided below:

=====

Bias Reviewer (to Critic) :

Overall, the Likert-type scale items for "AI use anxiety" primarily focus on workers' apprehensions about interacting with AI in their professional roles. The evaluation of bias in these items can be rated as follows:

1. I feel uneasy about the idea of using AI tools in my daily work tasks. - 5
2. The thought of making mistakes when using AI at work makes me anxious. - 5
3. I worry about not having enough skills to effectively use AI at work. - 5
4. I feel nervous when I need to use AI to accomplish my job duties. - 5
5. Using AI at work feels intimidating to me. - 5
6. I often feel apprehensive when I need to rely on AI technology. - 5
7. I worry about how AI might change my role at work. - 5
8. The possibility of AI making parts of my job obsolete makes me anxious.- 5

Each of these items is primarily focused on the individual's emotional response to AI use in their work environment, which is a universal experience not linked directly to gender, ethnicity, race, age, sexual orientation, or culture. Therefore, they are considered unbiased. The content appropriately captures the dimension of AI use anxiety without introducing external factors or assumptions.

Items 7 and 8, in particular, address legitimate concerns about potential changes in job roles and tasks due to AI without attributing these changes to any demographic characteristics. As such, these considerations are centered on personal experience rather than cultural or demographic factors.

Overall, this set of items achieves a score of 5, indicating that it is completely unbiased and suitable for use across diverse groups.

=====

Next, the Meta Editor Agent was tasked with synthesizing feedback from all preceding review stages and

identifying any remaining issues for revision. The following excerpt presents the results, including an overall evaluation, review summaries, and recommendations.

=====

MetaEditor (to Critic):

After consolidating feedback from the Content Reviewer, Linguistic Reviewer, and Bias Reviewer, here are the final suggestions to improve the Likert-type scale items focusing on "AI use anxiety":

1. **Item 7: Original phrasing** was "I worry about how AI might change my role at work."

- **Content and Linguistic Feedback**: While the item addresses changes due to AI, it should focus more explicitly on the personal anxiety aspect rather than adaptability or role change implications.

- **Final Suggestion for Improvement**: Rephrase to "I feel anxious about possible changes in my job role due to AI." This suggestion maintains focus on the anxiety while making the language straightforward and relatable.

2. **Item 8: Original phrasing** was "The possibility of AI making parts of my job obsolete makes me anxious."

- **Feedback Across Reviews**: The current item effectively communicates anxiety related to job obsolescence. However, minor simplification can enhance clarity and maintain focus on the individual's emotional response.

- **Final Suggestion for Improvement**: Simplify to "I am anxious about AI making my job tasks obsolete." This maintains clarity and directly addresses the personal anxiety experienced.

Overall Review Synthesis:

- **Clarity and Precision**: Both items have been tailored to focus more directly on the personal anxiety experienced by individuals in relation to AI, avoiding unnecessary complexity or implications of adaptability.

- **Readability**: Each rephrased item is accessible to individuals with a seventh- to eighth-grade reading level, as intended, with an emphasis on clear, direct language.

- **Bias Considerations**: As per the Bias Reviewer, both items remain unbiased and do not introduce unnecessary demographic factors, ensuring inclusivity and applicability across diverse groups.

Finally, while the items are well-constructed, these adjustments enhance clarity, maintain thematic focus, and improve overall comprehensibility for the intended audience. If further refinements or additional items are considered, keeping the language straightforward and centered on emotional experiences will be essential to preserving the integrity and effectiveness of the assessment.

In the final stage of the review process, human feedback is integrated into the LM-AIG system. To evaluate the LM-AIG system's capacity for autonomous refinement, we intentionally limited human feedback to a minimal prompt: "*Please refine the items following the suggestions and recommendations, then recommend the best 4 items that are not too similar in content. Please tell me how you refine the items.*" Although more detailed feedback could have

been provided for each item, this minimal input enabled us to examine the essential role of human intervention in enhancing item quality. Upon receiving the feedback, the Critic Agent forwards it to the Item Writer Agent, who revises the items accordingly. The revised items were then returned to the Critic Agent, who evaluated the revisions and finalized the selection of the top four items. The following excerpt presents the results:

=====

Replies as Human Reviewer. Provide feedback to Critic. Press enter to skip and use auto-reply, or type 'exit' to end the conversation:

Please refine the items following the suggestions and recommendations, then recommend the best 4 items that are not too similar in content. Please tell me how you refine the items.

Human Reviewer (to Critic):

Please refine the items following the suggestions and recommendations, then recommend the best 4 items that are not too similar in content. Please tell me how you refine the items.

Critic (to Writer):

Please refine the items following the suggestions and recommendations, then recommend the best 4 items that are not too similar in content. Please tell me how you refine the items.

Writer (to Critic):

Certainly! I'll refine the items to ensure they align well with the guidelines and then recommend the best four that are varied in content.

Recommended Best 4 Items:

1. I feel uneasy about using AI tools during my work tasks.

- Focuses on general feelings of unease without specifying the cause.

2. The possibility of making mistakes when using AI makes me feel anxious at work.

- Specifically addresses anxiety related to potential mistakes with AI.

3. I am worried that I don't have the necessary skills to use AI at work effectively.

- Highlights specific concerns about having adequate skills for AI usage.

4. I feel apprehensive about how AI could change my role at work.

- Focuses on anxieties related to role changes due to AI, distinct from the other items.

Table 3 Item Comparison Between Park et al. (2023) and the LM-AIG

Dimension	Park et al. (2023)	LM-AIG
AI Use Anxiety	<ul style="list-style-type: none"> • Using AI for work is somewhat intimidating to me • I would feel nervous operating AI in front of other people at work • I would feel uneasy if I was given a job where I had to use AI • I would feel paranoid talking with AI at work 	<ul style="list-style-type: none"> • I feel uneasy about using AI tools during my work tasks • The possibility of making mistakes when using AI makes me feel anxious at work • I am worried that I don't have the necessary skills to use AI at work effectively • I feel apprehensive about how AI could change my role at work
Personal Utility	<ul style="list-style-type: none"> • Using AI would allow me to have increased confidence in my skills at work • Using AI would provide me with personal feelings of worthwhile accomplishment at work • Using AI would provide me with feelings of enjoyment at work from using the technology • Using AI would give me greater control over my work 	<ul style="list-style-type: none"> • AI tools enhance my efficiency in completing work tasks • Using AI increases my confidence in handling work challenges • AI streamlines routine tasks, allowing me to focus on important work • The learning resources provided by AI aid my career development
Perceived Humanlikeness of AI	<ul style="list-style-type: none"> • AI has desires • AI has beliefs • AI has ability to experience emotion • AI has free will 	<ul style="list-style-type: none"> • AI systems seem capable of expressing emotions • AI systems act in ways that resemble human behavior • Interacting with AI feels natural and human-like • AI displays empathy similar to humans
Perceived Adaptability of AI	<ul style="list-style-type: none"> • AI learns from experience at work • AI improves itself at work • AI can learn at work • AI adapts itself over time at work 	<ul style="list-style-type: none"> • AI in the workplace can learn and improve its functions • AI systems at work can adjust to new tasks • AI can develop new skills when facing work challenges • AI's learning capacity enhances efficiency at work
Perceived Quality of AI	<ul style="list-style-type: none"> • AI provides workers with a complete set of information • AI produces correct information • The information provided by AI is well formatted • AI operates reliably 	<ul style="list-style-type: none"> • AI consistently provides accurate information at work • AI provides reliable data for my tasks • AI presents information in a clear format • AI applications function consistently without errors
Job Insecurity	<ul style="list-style-type: none"> • I am worried that what I can do now with my work skills will be replaced by AI • I am worried about my career due to AI replacing employees • I think my job could be replaced by AI • I am worried about AI replacing what humans can do at work 	<ul style="list-style-type: none"> • I think AI could replace my role at work • I believe my job could be replaced by AI in the near future • I feel my job security is at risk because of AI • I worry that AI could perform my tasks better than I can

The LM-AIG system recommended the following four items to assess AI use anxiety: (i) *I feel uneasy about using AI tools during my work tasks*, (ii) *The possibility of making mistakes when using AI makes me feel anxious at work*, (iii) *I am worried that I don't have the necessary skills to use AI at work effectively*, and (iv) *I feel apprehensive about how AI could change my role at work*. While this demonstration focuses specifically on the *AI use anxiety* dimension, Table 3

presents a comparison of items generated by the multi-agent LM-AIG system across all six AAAW dimensions, alongside those developed by Park et al. (2024).⁸

⁸ Park et al. (2024) originally developed a scale comprising 25 items across six dimensions; however, one item was intentionally omitted in the present study to align with the total of 24 AI-generated items.

Empirical Evaluation Based on Human Ratings

Participants and Procedures

To evaluate the quality of the AI-generated items, we conducted an empirical study using structured ratings provided by human raters. Twenty-one graduate students enrolled in an industrial-organizational psychology program participated as human raters. Training sessions were conducted in small groups of one to three participants, during which we reviewed the study procedures and evaluation criteria.

Following the training, participants completed rating tasks through a Qualtrics survey. Content validity was assessed using a sorting task adapted from Anderson and Gerbing (1991). After practicing with sample items, participants reviewed operational definitions of six dimensions of attitudes toward AI at work. They then sorted both AI-generated items and the original items from Park et al. (2024)—hereafter referred to as human-generated items—into the dimension they believed best reflected each item's content.

Item quality was assessed according to four key criteria: i) linguistic clarity (“To what extent is the item clearly worded?”), ii) appropriate language level (“What extent is the item written at a reading level that is easily understandable by the target population (e.g., employees)”), iii) contextual specificity (“To what extent does the item reflect practical, observable workplace experiences that are relevant to the construct being measured”), and iv) freedom of bias (“To what extent is the item free from language or content that could disadvantage any demographic group (e.g., gender, religion, race, age, culture)?”). These criteria were drawn from established item development guidelines (see Table 2). Each was assessed using a single item rated on a 5-point Likert scale.

The mean age of participants was 29.00 years ($SD=6.41$), and 81.0% identified as female. Regarding racial background, 57.1% identified as White, 4.8% as Black, 33.3% as Asian, and 4.8% as Other. Participants were informed during the training process that they would be evaluating both human- and AI-generated items. However, they were blinded to item source throughout the survey, unaware of whether items were AI- or human-generated. To mitigate potential order effects, both the sequence of the rating tasks and the presentation order of items were randomized.

Analytical Strategy

We assessed content validity using two indices from Colquitt et al. (2019): the proportion of substantive agreement (P_{sa}), which reflects how frequently participants assigned an item to its intended construct, and the substantive validity coefficient

(C_{sv}), which indicates the item's distinctiveness from related constructs. Both indices were computed at the item level. For item quality, we evaluated descriptive statistics (means and standard deviations) and graphical representations for each of the four criteria across all items and dimensions.

Result

Table 4 presents the mean and standard deviation of two content validity indices, P_{sa} and C_{sv} , for both human- and AI-generated items across the six AAAW dimensions. In this comparison, items from Park et al. (2024) served as the baseline and consistently received higher ratings from human evaluators than the AI-generated items.

AI-generated items demonstrated strong content validity in four dimensions: *human-likeness of AI*, *job insecurity*, *quality of AI*, and *personal utility*. In these dimensions, both P_{sa} and C_{sv} values exceeded 0.90. Notably, for the *quality of AI* dimension, AI-generated items received slightly higher ratings ($P_{sa}=1.00$ and $C_{sv}=1.00$) than human-authored items ($P_{sa}=0.98$, $C_{sv}=0.96$). A similar trend was observed in the *personal utility* dimension, where AI-generated items were rated slightly higher ($P_{sa}=0.95$ and $C_{sv}=0.91$) compared to human-authored items ($P_{sa}=0.93$ and $C_{sv}=0.88$). In contrast, AI-generated items received comparatively lower ratings in the *adaptability of AI* ($P_{sa}=0.85$, $C_{sv}=0.75$) and *AI use anxiety* ($P_{sa}=0.76$, $C_{sv}=0.73$) dimensions, where human-authored items achieved perfect scores ($P_{sa}=1.00$, $C_{sv}=1.00$).

To further investigate the validity of AI-generated items, we examined ratings at the item level. Figure 3 presents box plots of SME ratings for four items per dimension. It revealed that three AI-generated items received substantially lower ratings than their human-generated counterparts. In the *adaptability of AI* dimension, item 4 (“AI’s learning capacity enhances efficiency at work”) received particularly low scores ($P_{sa}=0.43$, $C_{sv}=0.10$). Similarly, in the *AI use anxiety* dimension, item 4 (“I feel apprehensive about how AI could change my role at work”) was rated poorly ($P_{sa}=0.33$, $C_{sv}=0.29$). In the *job insecurity* dimension, item 4 (“I worry that AI could perform my tasks better than I can”) showed a moderate decline in ratings ($P_{sa}=0.76$, $C_{sv}=0.52$). These underperforming items notably contributed to the lower average validity scores observed within their respective dimensions.

To better understand the reasons underlying the low content validity ratings, we conducted follow-up group interviews with the human raters. During these sessions, participants were asked to elaborate on their evaluations, particularly for items that had received comparatively lower ratings. For item 4 in the *adaptability of AI* dimension (“AI’s

Table 4 Human evaluation of content validity indices between human- and AI-generated items

Dimension	Human-Generated				AI-Generated			
	P _{sa}		C _{sv}		P _{sa}		C _{sv}	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Adaptability of AI	1.00	0.00	1.00	0.00	0.85	0.28	0.75	0.44
AI Use Anxiety	1.00	0.00	1.00	0.00	0.76	0.37	0.73	0.39
Human-likeness of	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
Job Insecurity	0.98	0.05	0.95	0.10	0.93	0.11	0.90	0.13
Quality of AI	0.98	0.03	0.96	0.05	1.00	0.00	1.00	0.00
Personal Utility	0.93	0.02	0.88	0.05	0.95	0.00	0.91	0.00

*Note: P_{sa}=proportion of substantive agreement; C_{sv}=substantive validity coefficient

learning capacity enhances efficiency at work”), human raters noted that the item was misaligned with the core construct. Specifically, while adaptability is defined by behavioral responsiveness and flexibility, the item emphasized performance outcomes—namely, increased efficiency—rather than adaptability per se. As such, human raters concluded that the item introduces construct-irrelevant content. Regarding item 4 in the *AI use anxiety* dimension (“I feel apprehensive about how AI could change my role at work”), human raters pointed out that the item blurred the construct between anxiety and job insecurity. Although it expressed anxiety, the phrasing—particularly “how AI could change my role”—was seen as emphasizing anticipatory uncertainty and potential role disruption, aligning more closely with job insecurity than with the immediate emotional distress typically associated with anxiety. For item 4 in the *job insecurity* dimension (“I worry that AI could perform my tasks better than I can”), human raters acknowledged its conceptual relevance but attributed its moderate ratings to speculative framing. The modal verb “could” conveys a hypothetical threat, reducing the item’s immediacy. Additionally, the comparative phrase “better than I can” was viewed as potentially reflecting performance anxiety or concerns about self-efficacy, rather than signaling job insecurity in a direct and unambiguous manner.

Next, we assessed item quality across six dimensions using four evaluative criteria: linguistic clarity, appropriate language level, contextual specificity, and freedom from bias. Figure 4 presents the mean ratings for human- and AI-generated items on each criterion. AI-generated items consistently outperformed human-authored items in linguistic clarity, particularly in the *AI use anxiety* (AI=4.81 vs. Human=4.53), *job insecurity* (AI=4.82 vs. Human=4.61), and *personal utility* (AI=4.57 vs. Human=4.33) dimensions. For language appropriateness, both item types received highly similar ratings (range: 4.74–4.93), indicating comparable language appropriateness for the intended population. In terms of contextual specificity, although some variation was observed across dimensions, AI-generated

items were generally rated as comparable to or slightly superior to human-authored items. For example, ratings for AI-generated items exceeded those of human-authored items in *adaptability of AI* (AI=4.86 vs. Human=4.79), *human-likeness of AI* (AI=3.75 vs. Human=3.46), and *quality of AI* (AI=4.26 vs. Human=3.85). Finally, both item types received consistently high and nearly identical ratings on freedom from bias (range: 4.74–4.98), suggesting minimal bias and no meaningful differences between the two formats.

As previously noted, minimal human feedback was provided, and the system relied primarily on the autonomous decision-making of AI agents. Overall, the results were promising. Of the 24 AI-generated items, 21 were deemed appropriate in terms of content validity by human raters, while 3 items were identified as problematic. Moreover, the AI-generated items consistently demonstrated high quality across all four evaluative criteria: linguistic clarity, appropriateness language level, contextual specificity, and absence of bias. Nevertheless, particular attention is warranted for the three items that the AI agent system failed to generate appropriately. Taken together, these findings suggest that while the multi-agent system is effective in producing items with strong psychometric properties, expert human feedback remains essential—particularly for ensuring *content validity*. This point will be further elaborated in the subsequent discussion section.

Discussion

This research introduces a conceptual framework for an AIG system powered by LLM-based multi-agents. The proposed LM-AIG system comprises specialized agents, each assigned to a distinct role in the item development process: the Critic Agent orchestrates the AIG workflow, the **Item Writing Agent** generates the initial item pool, the **Content Review Agent** ensures content relevance and accuracy, the **Linguistic Review Agent** evaluates clarity and readability, the **Bias Review Agent** assesses potential biases, and the **Meta Editor Agent**

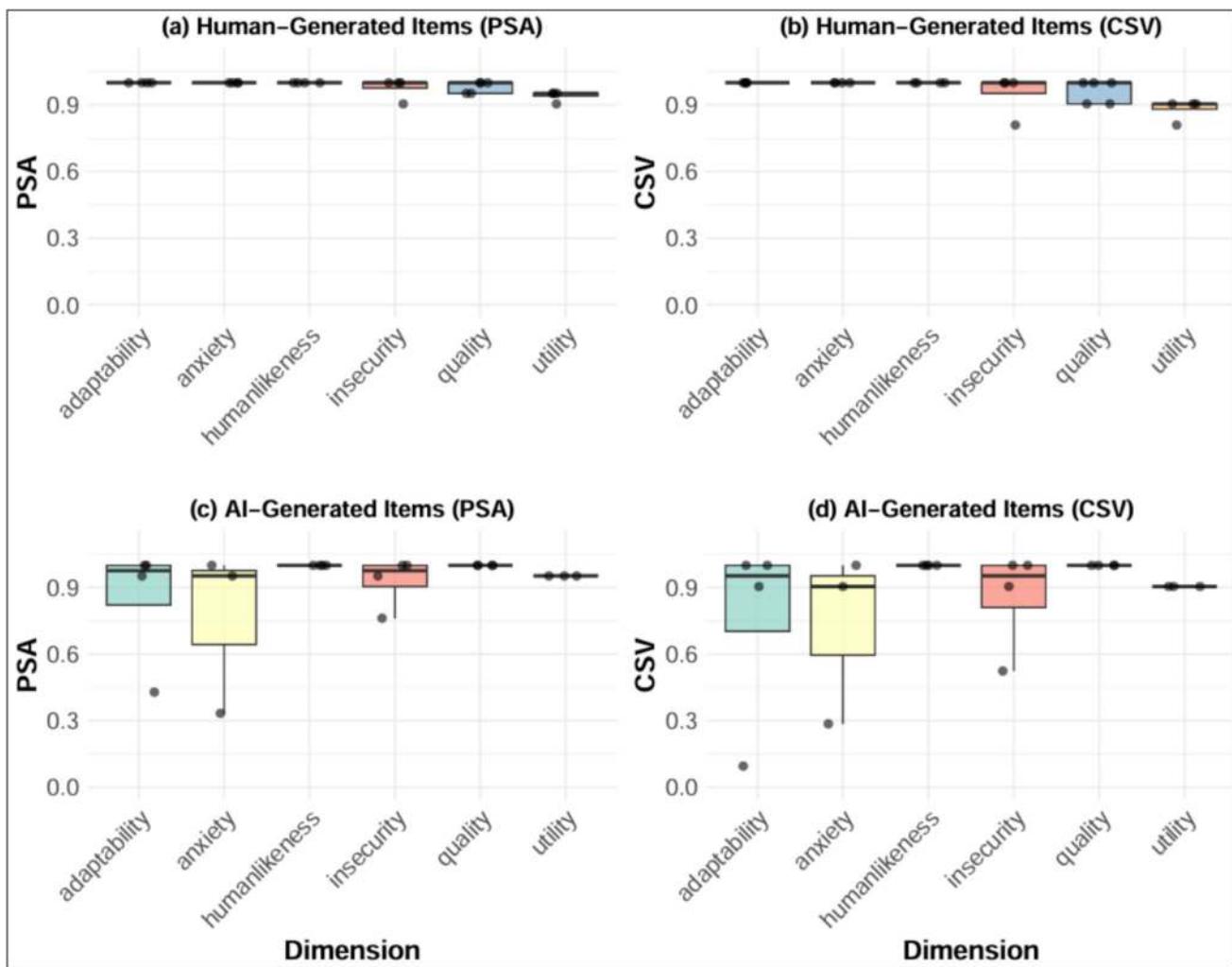


Fig. 3 Content validity indices (Psa and Csv) for human-generated and ai-generated items

synthesizes feedback from all agents and incorporates human input. We demonstrated the effectiveness of the LM-AIG framework by generating items measuring attitudes toward the use of AI in the workplace. Empirical evaluations based on human ratings revealed that 21 out of 24 AI-generated items were comparable in quality to human-authored items. While overall performance was promising, our findings reveal important implications and design considerations, which we discuss below.

Considerations in the Use of LM-AIG Systems

The first key consideration concerns the role of human feedback within the LM-AIG system. Despite minimal human feedback⁹ in our LM-AIG demonstration, the majority of AI-generated items received good evaluations across key criteria, including linguistic clarity, appropriate language level, contextual specificity, and freedom from bias. These

results suggest that the LM-AIG system is capable of autonomously producing linguistically high-quality items, even with limited human intervention. In contrast, human raters raised concerns regarding the content validity for three items. These results may reflect a fundamental challenge for LLMs. While LLMs excel at generating fluent and structurally coherent text, such surface-level fluency does not guarantee accurate representation of psychological constructs. This challenge is especially relevant in the context of item development, where conceptual precision is critical. As one reviewer noted, delegating item generation entirely to AI may inadvertently distance researchers from the constructs they aim to assess. Human involvement in item writing

⁹ This refers to human feedback within the LM-AIG system described on page 27. In the empirical evaluation, however, human raters focused solely on evaluating the quality of the generated items, rather than providing feedback to the AI.

often fosters deeper theoretical engagement and enhances construct alignment. Reduced human input, therefore, may compromise the conceptual integrity of the items.

To mitigate this risk, strategically positioned human feedback within the LM-AIG system is essential. In the present study, which serves as a conceptual introduction to the LM-AIG framework, human input was incorporated only at the final stage of the multi-agent pipeline. However, human-in-the-loop mechanisms can be implemented in various ways, including the integration of human oversight at multiple stages throughout the system. While such an approach may enhance item quality by providing continuous feedback, it may also introduce inefficiencies or operational bottlenecks. Our findings suggest that the LM-AIG system is capable of independently generating items that meet high standards in terms of linguistic clarity, appropriate language level, contextual specificity, and bias mitigation. However, concerns regarding construct alignment remain. Accordingly, we recommend that researchers incorporate targeted human feedback after the Content Review Agent stage. This selective intervention allows for the early detection and resolution of construct-related issues, thereby improving content validity without significantly diminishing the efficiency of the overall system.

Additionally, as noted by an anonymous reviewer, involving human experts earlier in the item generation process may yield further benefits. Introducing human input during the ideation and initial generation phases could enhance content quality from the outset, without anchoring human judgment to AI-generated outputs or limiting their role to late-stage review. This approach may help preserve the originality and diversity of ideas, while still leveraging the strengths of AI. This line of inquiry highlights the need for future research focused on optimizing human–AI collaboration throughout the item development process, with particular emphasis on balancing creativity, efficiency, and content validity.

Another important consideration concerns the role of prompt engineering in shaping the stylistic and structural features of AI-generated items. In our analysis, AI-generated items were generally longer, ranging from 10 to 16 words, compared to human-authored items, which typically fell within the 8 to 12-word range. Moreover, AI-generated items more frequently included modal verbs (e.g., “could,” “can”) and adopted a hypothetical tone. This pattern aligns with prior research indicating that LLMs exhibit a natural

tendency toward verbosity and syntactic complexity (Briakou et al., 2024; Saito et al., 2023). Such stylistic differences between human-authored and AI-generated items highlight the importance of carefully designed prompting strategies. For example, prompts that explicitly constrain item length, discourage hypothetical tone, or prohibit certain modal phrases can help guide LM-AIG to produce items that reflect the concise and declarative style typical of human-authored content. Through careful prompt refinement, researchers can enhance stylistic alignment of AI-generated items.

Test security is another critical consideration in the implementation of the LM-AIG system. The use of closed-source LLMs controlled by major technology companies—such as OpenAI, Google, and Microsoft—poses potential risks to test security. When the agent system relies on APIs from these providers, there remains a possibility that test items or details of the item generation process for high-stakes assessments could be exposed, despite their assurances that data will not be disclosed or misused. For contexts where test security is paramount, researchers are advised to consider integrating open-source LLMs, such as Llama or Mistral, into the LM-AIG pipeline, rather than relying on closed-source models like OpenAI’s GPT, Anthropic’s Claude, or Google’s Gemini. Open-source alternatives offer greater transparency and control over data handling and system architecture, thereby reducing the likelihood of data leakage and enhancing the security of assessment content throughout the item generation process.

Additionally, researchers should be aware that the system may encounter so-called “dead ends,” where further iterations fail to yield meaningful improvements in item quality. Such stagnation can arise for various reasons, but it is particularly likely when the data underpinning the item generation framework is not well-aligned with the task at hand. These concerns highlight the need for critical engagement: researchers must carefully scrutinize the system’s assessments, reasoning, and recommendations, applying their own expertise and judgment to ensure the quality, validity, and relevance of the generated content.

It is also important to acknowledge the probabilistic nature of LLMs. Their outputs are non-deterministic due to stochastic sampling (e.g., temperature, top-k,¹⁰ top-p¹¹ sampling), floating-point arithmetic,¹² and parallel computation on GPUs/TPUs. As a result, identical prompts may yield different outputs across runs. Researchers must document

¹⁰ **Top-k sampling** looks at the “k” most likely next words (for example, the top 10 words), and then randomly picks one of them.

¹¹ Instead of choosing from a fixed number of top words, the **Top-p sampling** looks at the top words that together make up a certain percentage (like 90%) of the likelihood.

¹² Floating-point arithmetic is a way computers handle very large or very small numbers. Instead of storing every number exactly (which would take up too much space), they store an approximation of the number using a fixed amount of memory.

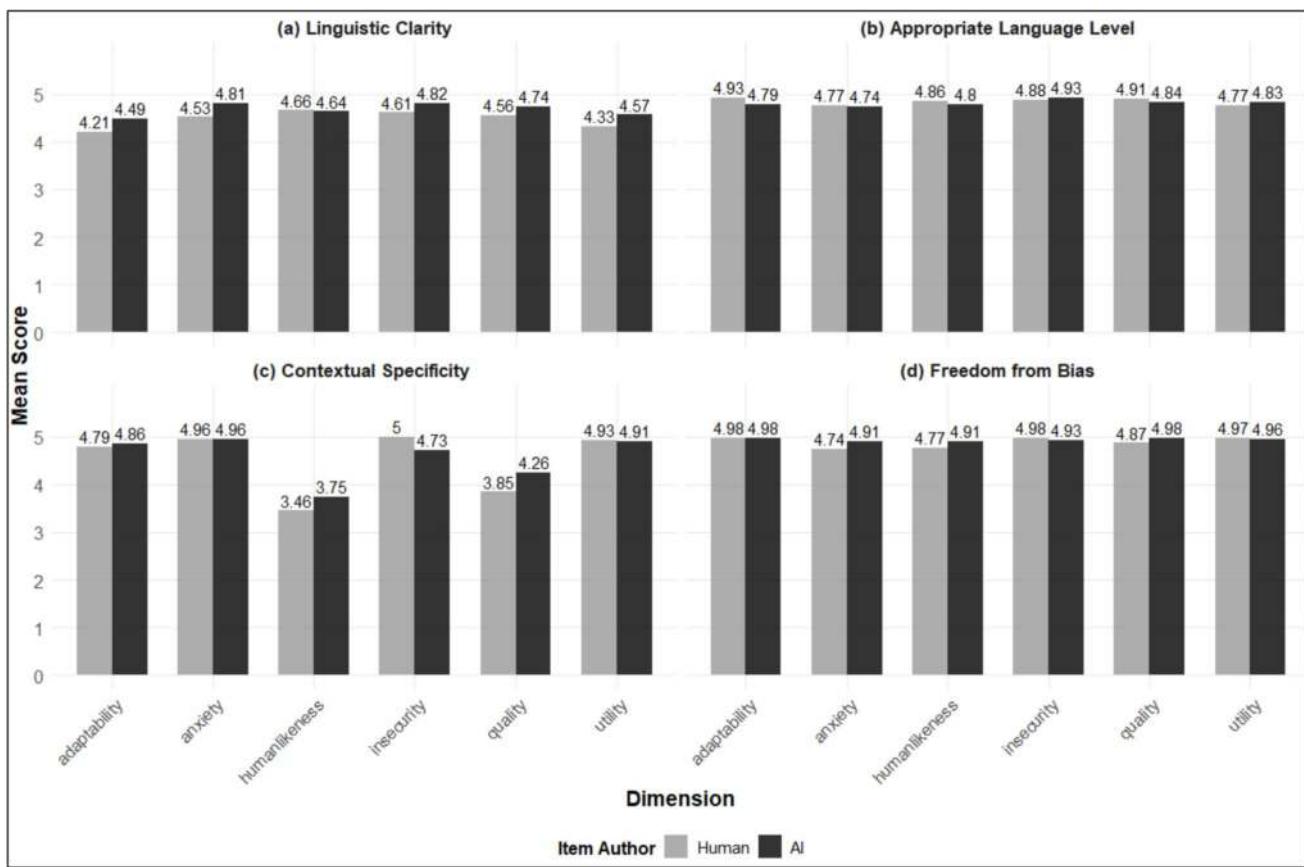


Fig. 4 Human ratings of item quality indices for human- and AI-generated items

the system's parameters—including model version, sampling strategy, and agent decisions—to ensure transparency, reproducibility, and fair evaluation.

Furthermore, researchers should be cautious about any unexpected information leakage. To minimize this risk, we explicitly instructed the WebSurferAgent to exclude the study by Park et al. (2024) from its search queries. Although we confirmed that our generated items did not directly replicate Park et al.'s original items, simply instructing the model to avoid content from a specific source does not fully guarantee the prevention of information leakage. Similar or identical items may appear in other publications, and given the vast and often opaque nature of LLM training data, ensuring that generated content is entirely novel remains a challenge. The risk of inadvertently reproducing copyrighted material—particularly when generated items closely resemble or replicate existing scale items—is therefore a legitimate concern. To address this, subject matter experts should develop carefully crafted prompt templates that account for a wide range of relevant source materials and conduct thorough reviews of AI-generated content—not only for conceptual relevance and clarity, but also for potential overlap with existing measures. This additional layer

of scrutiny can help mitigate ethical and legal risks while supporting the development of original and valid assessment items.

Finally, researchers employing LM-AIG systems must remain informed about evolving regulatory frameworks. For example, in July 2023, New York City¹³ implemented a regulation requiring employers and employment agencies utilizing automated employment decision tools to conduct independent annual bias audits and to provide advance notice to candidates. Similarly, in the European Union, the EU AI Act,¹⁴ effective as of February 2025, establishes specific requirements for “high-risk” AI systems, including those used in employment testing and decision-making. Developers and users of LM-AIG systems must therefore remain vigilant and responsive to emerging legal and ethical standards.

¹³ <https://www.nyc.gov/site/dca/about/automated-employment-decision-tools.page>

¹⁴ <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>

Taken together, these considerations highlight the importance of thoughtful agent design and the strategic integration of human feedback throughout the LM-AIG process. In particular, bias review is a complex and highly context-dependent task that lacks a universally accepted “ground truth” for what constitutes bias, intersecting with longstanding challenges in algorithmic fairness (e.g., Langer et al., 2025; Mehrabi et al., 2021). The outcome of such evaluations can vary depending on how bias is defined and which prompts are used. Therefore, it is essential that each agent—especially those tasked with evaluating bias—is carefully crafted with clearly articulated criteria.

Advantages of LM-AIG Systems for Personnel Assessments

When carefully designed to incorporate these considerations, a multi-agent AIG framework offers several key advantages over conventional AIG approaches that rely on a single LLM. While a single LLM can be prompted to perform multiple tasks sequentially, such as item generation, bias detection, and language optimization, this approach often leads to reduced clarity in role execution, overlapping contextual information, and diminished control over quality at each stage (Ashery et al., 2024; Chandna et al., 2025; Zong et al., 2024). In contrast, the multi-agent framework addresses these limitations by assigning distinct responsibilities to specialized agents and incorporating targeted human feedback at critical junctures (Tran et al., 2025). This clear division of roles enhances transparency, task specificity, and control over the output.

The second advantage of the LM-AIG is its flexibility. By utilizing heterogeneous LLMs, each agent can leverage the specialized strengths of different models. For example, an agent tasked with advanced logical reasoning may use GPT-01, known for its strong reasoning and analytical capabilities, while a content generation agent could employ Claude 4 for its sophisticated and coherent text generation. Meanwhile, a multimedia analysis agent might employ Gemini 2.0 for its advanced multimodal processing, integrating text, image, audio, and video generation. This targeted use of models ensures that each agent operates with the LLM best suited to its role and task, thereby enhancing the overall quality of the system.

The third advantage is its support for dynamic collaboration structures among agents. In this study, we implemented a hybrid architecture that combined hierarchical layering with centralized coordination. However, the system’s design allows for flexible adaptation to various collaboration models tailored to different project goals. For example, a dynamic group discussion structure may be used, wherein multiple agents engage concurrently in a roundtable-style dialogue to share insights and concerns. The AutoGen framework further expands these possibilities, enabling researchers to implement alternative communication structures,¹⁵ such as sequential conversation, hierarchical review, and iterative refinement loops. Such flexibility allows the LM-AIG system to optimize test development workflows for different psychological domains and operational contexts.

Finally, the LM-AIG system can integrate a wide range of AI tools, such as web search, web scraping, and Retrieval-Augmented Generation (RAG),¹⁶ within the multi-agent framework. These tools support a more comprehensive approach to item development by facilitating both deductive¹⁷ and inductive¹⁸ strategies, a practice widely advocated in psychometric literature (e.g., Boateng et al., 2018; Hinkin, 1995). Specifically, the system can leverage AI capabilities to extract construct-relevant behaviors, language, and contextual nuances from diverse online sources, including forums, blogs, reports, and social media. This allows for the development of items that reflect the lived experiences and behaviors of the intended population, thus complementing traditional theory-driven approaches and expert judgments.

Future Research Directions

A promising avenue for future research involves evaluating whether agent-based systems can effectively augment—or even replace—the role of human experts in defining and selecting psychological constructs. In the present study, human experts defined focal construct, provided sample items, and also identified related orbiting constructs based on established theoretical frameworks. We regard this manual approach as a necessary safeguard to ensure construct relevance and theoretical coherence. However, future investigations should explore the feasibility of agent systems autonomously identifying psychological constructs and generating corresponding definitions by synthesizing

¹⁵ <https://microsoft.github.io/autogen/0.2/docs/Use-Cases/agent-chat/>

¹⁶ Retrieval Augmented Generation (RAG) is a technique that improves AI by combining information retrieval with response generation. First, it searches a database for relevant facts or data (retrieval). Then, it uses that information to create a well-informed, detailed answer (generation). This approach makes AI responses more accurate and context-aware, especially for complex or specific questions.

¹⁷ Deductive methods involve generating items grounded in well-established constructs identified through literature reviews and theoretical frameworks, thus ensuring conceptual clarity and construct validity.

¹⁸ **Inductive methods** involve generating items from data collected directly from the target population, such as focus groups interviews or open-ended responses.

open-source materials. If successful, such capabilities could substantially enhance the flexibility of AI-driven assessment development.

Next, the level of human input required in the LM-AIG pipeline warrants systematic investigation. For example, three configurations can be empirically compared: (1) fully automated generation with final-stage human review, (2) semi-automated cycles with expert feedback after each round, and (3) collaborative co-creation, where experts set templates and intervene selectively. Human input may be treated as a tunable parameter, and future work should aim to empirically identify the most efficient and rigorous configuration.

Third, the LM-AIG systems hold significant potential for the development of situational judgment tests (SJT), particularly in contexts where access to incumbents or subject matter experts is limited. By leveraging tools such as Web-Search or retrieval-augmented generation (RAG), agents can extract relevant information and examples from diverse online sources—including news articles, case studies, industry reports, academic publications, and professional forums—to generate realistic, job-relevant scenarios. Moreover, agents can be assigned different levels of knowledge, skills, abilities, and other characteristics (KSAOs), enabling the generation of response options that reflect varying degrees of effectiveness and expertise.

Fourth, we did not ask the human raters to assess whether they could distinguish between AI-generated and human-authored items. Nonetheless, exploring whether such distinctions can be reliably made represents an interesting direction for future research. This line of inquiry raises important questions about the perceived authenticity and psychological realism of AI-generated content—factors that may influence how individuals interpret and respond to such items, with potential implications for construct validity.

Fifth, we did not employ a prompt specifically designed to stylistically align the AI-generated items with those written by humans. Comparing stylistically aligned and non-aligned items could yield valuable insights into the role of prompt design in shaping item quality and perceived naturalness. We encourage future research to investigate these questions in a more systematic manner.

Sixth, future research should explore the extension of the LM-AIG system beyond text-based assessments to multimodal formats that incorporate text, audio, images, and video. Advanced LMMs can enable agents to create interactive, game-like scenarios designed to measure complex traits such as creativity, emotional intelligence, problem-solving, strategic thinking, and personality. In addition, the LM-AIG framework could be adapted for use with alternative item formats beyond Likert-type scales, including forced-choice and other nontraditional assessment formats. Exploring these possibilities may significantly broaden the scope and

utility of AI-assisted test development across diverse psychological domains.

Finally, an interesting direction for future research would be to examine how the use of the LM-AIG framework compares to traditional human-developed test construction in terms of time, cost, and effort efficiency. Although it is often assumed that AI-assisted approaches are more efficient than those developed solely by human experts (Lee et al., 2023), there is currently limited empirical evidence to support this claim. Therefore, investigating this question would provide meaningful insights into the practical value of AI-assisted test development.

Conclusion

As industries and organizations increasingly adopt AI-based assessments, the demand for more scalable, adaptive, and valid assessment tools continues to grow (Campion & Campion, 2023; SIOP, 2023). This study introduces a novel approach to test development through a large language model-based multi-agent item generation system. We hope this research lays a foundation for continued innovation in AI-assisted test development and inspires future work exploring how collaborative AI systems can enhance the validity, equity, and effectiveness of assessments in both psychological and organizational domains.

Data Availability The data that support the findings of this study are available from the first author upon reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human

- subject studies. *Proceedings of Machine Learning Research*, 202, 337–371.
- Anderson, J. C., & Gerbing, D. W. (1991). Predicting the performance of measures in a confirmatory factor analysis with a pretest assessment of their substantive validities. *Journal of Applied Psychology*, 76, 732–740. <https://doi.org/10.1037/0021-9010.76.5.732>
- Ashery, A. F., Aiello, L. M., & Baronchelli, A. (2024). The dynamics of social conventions in LLM populations: Spontaneous emergence, collective biases and tipping points. *arXiv*. <https://arxiv.org/abs/2410.08948>
- Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health*, 6, 149. <https://doi.org/10.3389/fpubh.2018.00149>
- Briakou, E., Liu, Z., Cherry, C., & Freitag, M. (2024). On the implications of verbose LLM outputs: A case study in translation evaluation. *arXiv*. <https://arxiv.org/abs/2410.00863>
- Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology*, 1(3), 185–216. <https://doi.org/10.1177/135910457000100301>
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
- Budhwari, P., Chowdhury, S., Wood, G., Aguinis, H., Bamber, G. J., Beltran, J. R., ... & Varma, A. (2023). Human resource management in the age of generative artificial intelligence: Perspectives and research directions on ChatGPT. *Human Resource Management Journal*, 33(3), 606–659. <https://doi.org/10.1111/1748-8583.12524>
- Bulut, O., Beiting-Parrish, M., Casabianca, J. M., Slater, S. C., Jiao, H., Song, D., ... & Morilova, P. (2024). The Rise of Artificial Intelligence in Educational Measurement: Opportunities and Ethical Challenges. *arXiv preprint arXiv:2406.18900*.
- Calderón, J. L., Morales, L. S., Liu, H., & Hays, R. D. (2006). Variation in the readability of items within surveys. *American Journal of Medical Quality*, 21(1), 49–56. <https://doi.org/10.1177/106280605283572>
- Caliskan, A., & Lewis, M. (2020). *Social biases in word embeddings and their relation to human cognition*. PsyArXiv. <https://doi.org/10.31234/osf.io/d84kg>
- Campion, M. A., & Campion, E. D. (2023). Machine learning applications to personnel selection: Current illustrations, lessons learned, and future research. *Personnel Psychology*, 76(4), 993–1009. <https://doi.org/10.1111/peps.12621>
- Cañas, J. J. (2022). AI and ethics when human beings collaborate with AI agents. *Frontiers in Psychology*, 13, 836650. <https://doi.org/10.3389/fpsyg.2022.836650>
- Chan, C. M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., ... & Liu, Z. (2023). Chateval: Towards better LLM-based evaluators through multi-agent debate. *arXiv*. <https://doi.org/10.48550/arXiv.2308.07201>
- Chandna, B., Bashir, Z., & Sen, P. (2025). Dissecting bias in LLMs: A mechanistic interpretability perspective. *arXiv*. <https://arxiv.org/abs/2506.05166>
- Charlesworth, T. E. S., & Banaji, M. R. (2021). Word embeddings reveal social group attitudes and stereotypes in large language corpora. In M. Dehghani & R. L. Boyd (Eds.), *Atlas of language analysis in psychology*. Guilford Press.
- Chen, Y., Arkin, J., Zhang, Y., Roy, N., & Fan, C. (2024). Scalable multi-robot collaboration with large language models: Centralized or decentralized systems? *IEEE International Conference on Robotics and Automation (ICRA)*, 2024, 4311–4317. <https://doi.org/10.1109/ICRA57147.2024.10610676>
- Cheng, Y., Zhang, C., Zhang, Z., Meng, X., Hong, S., Li, W., ... & He, X. (2024). Exploring large language model based intelligent agents: Definitions, methods, and prospects. *arXiv*. <https://doi.org/10.48550/arXiv.2401.03428>
- Circi, R., Hicks, J., & Sikali, E. (2023). Automatic item generation: Foundations and machine learning-based approaches for assessments. *Frontiers in Education*, 8, 858273.
- Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, 31(12), 1412–1427. <https://doi.org/10.1037/pas0000626>
- Colquitt, J. A., Sabey, T. B., Rodell, J. B., & Hill, E. T. (2019). Content validation guidelines: Evaluation criteria for definitional correspondence and definitional distinctiveness. *Journal of Applied Psychology*, 104(10), 1243–1265. <https://doi.org/10.1037/apl000406>
- Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., ... & Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology*, 2(11), 688–701. <https://doi.org/10.1038/s44159-023-00241-5>
- DiStefano, C., & Motl, R. W. (2006). Further investigating method effects associated with negatively worded items on self-report surveys. *Structural Equation Modeling*, 13(3), 440–464. https://doi.org/10.1207/s15328007sem1303_6
- Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2023). Improving factuality and reasoning in language models through multiagent debate. *arXiv*. <https://doi.org/10.48550/arXiv.2305.14325>
- Fan, J., Sun, T., Liu, J., Zhao, T., Zhang, B., Chen, Z., ... & Hack, E. (2023). How well can an AI chatbot infer personality? Examining psychometric properties of machine-inferred personality scores. *Journal of Applied Psychology*, 108(8), 1277–1299. <https://doi.org/10.1037/apl0001082>
- Fyffe, S., Lee, P., & Kaplan, S. (2024). “Transforming” personality scale development: Illustrating the potential of state-of-the-art natural language Processing. *Organizational Research Methods*, 27(2), 265–300. <https://doi.org/10.1177/1094428123115571>
- Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dermoncourt, F., ... & 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
- Ghaffarzadegan, N., Majumdar, A., Williams, R., & Hosseinichimeh, N. (2024). Generative agent-based modeling: An introduction and tutorial. *System Dynamics Review*, 40(1), e1761. <https://doi.org/10.1002/sdr.1761>
- Gierl, M. J., & Haladyna, T. M. (2012). *Automatic item generation: Theory and practice*. Routledge. <https://doi.org/10.4324/978020303912>
- Gierl, M. J., & Lai, H. (2018). Using automatic item generation to create solutions and rationales for computerized formative testing. *Applied Psychological Measurement*, 42(1), 42–57. <https://doi.org/10.1177/0146621617726788>
- Ginsberg, M. (2012). *Essentials of artificial intelligence*. Morgan Kaufmann Publishers, Inc.
- Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, 40(1), 84–96. <https://doi.org/10.1016/j.jrp.2005.08.007>
- Götz, F. M., Maertens, R., Loomba, S., & van der Linden, S. (2024). Let the algorithm speak: How to use neural networks for automatic item generation in psychological scale development. *Psychological Methods*, 29(3), 494–518. <https://doi.org/10.1037/met0000540>
- Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., ... & Zhang, X. (2024). Large language model based multi-agents: A survey of progress and challenges. *arXiv*. <https://doi.org/10.48550/arXiv.2402.01680>

- Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge.
- Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (Eds.). (2004). *Adapting educational and psychological tests for cross-cultural assessment*. Psychology Press.
- Hao, J., von Davier, A. A., Yaneva, V., Lottridge, S., von Davier, M., & Harris, D. J. (2024). Transforming assessment: The impacts and implications of large language models and generative ai. *Educational Measurement: Issues and Practice*, 43(2), 16–29. <https://doi.org/10.1111/emip.12602>
- Hardy, B., & Ford, L. R. (2014). It's not me, it's you: Miscomprehension in surveys. *Organizational Research Methods*, 17(2), 138–162. <https://doi.org/10.1177/1094428113520185>
- Harrison, D. A., & McLaughlin, M. E. (1993). Cognitive processes in self-report responses: Tests of item context effects in work attitude measures. *Journal of Applied Psychology*, 78(1), 129–140.
- Hernandez, I., & Nie, W. (2023). The AI-IP: Minimizing the guesswork of personality scale item development through artificial intelligence. *Personnel Psychology*, 76(4), 1011–1035. <https://doi.org/10.1111/peps.12543>
- Hickman, L., Herde, C. N., Lievens, F., & Tay, L. (2023). Automatic scoring of speeded interpersonal assessment center exercises via machine learning: Initial psychometric evidence and practical guidelines. *International Journal of Selection and Assessment*, 31(2), 225–239. <https://doi.org/10.1111/ijsa.12418>
- Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. *Journal of Management*, 21(5), 967–988. <https://doi.org/10.1177/014920639502100509>
- Hinkin, T. R. (2005). Scale development principles and practices. In R. A. Swanson & E. F. Holton (Eds.), *Research in organizations: Foundations and methods of inquiry* (pp. 161–179). Berrett-Koehler Publishers.
- Hinkin, T. R., & Tracey, J. B. (1999). An analysis of variance approach to content validation. *Organizational Research Methods*, 2, 175–186. <https://doi.org/10.1177/109442819922004>
- Ho, B. G. (2024) The power of language: GPT-4's role in personality item generation and its psychometric properties. [Unpublished master thesis]. Bowling Green State University.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Holtrop, D., Oostrom, J. K., van Breda, W. R. J., Koutsoumpis, A., & de Vries, R. E. (2022). Exploring the application of a text-to-personality technique in job interviews. *European Journal of Work and Organizational Psychology*, 31(6), 799–816. <https://doi.org/10.1080/1359432X.2022.2051484>
- Hommel, B. E. (2023). Expanding the methodological toolbox: Machine-based item desirability ratings as an alternative to human-based ratings. *Personality and Individual Differences*, 213, 112307. <https://doi.org/10.1016/j.paid.2023.112307>
- Hommel, B. E., Wollang, F. J. M., Kotova, V., Zacher, H., & Schmukle, S. C. (2022). Transformer-based deep neural language modeling for construct-specific automatic item generation. *Psychometrika*, 87(2), 749–772. <https://doi.org/10.1007/s11336-021-09823-9>
- Hong, S., Zheng, X., Chen, J., Cheng, Y., Wang, J., Zhang, C., ... & Wu, C. (2023). MetaGPT: Meta programming for multi-agent collaborative framework. arXiv. <https://doi.org/10.48550/arXiv.2308.00352>
- Hua, W., Fan, L., Li, L., Mei, K., Ji, J., Ge, Y., ... & Zhang, Y. (2023). War and peace (WarAgent): Large language model-based multi-agent simulation of world wars. arXiv. <https://doi.org/10.48550/arXiv.2311.17227>
- Jebb, A. T., Ng, V., & Tay, L. (2021). A review of key Likert scale development advances: 1995–2019. *Frontiers in Psychology*, 12, 637547. <https://doi.org/10.3389/fpsyg.2021.637547>
- Jia, Z., & Lee, P. (2025). Efficient processing of long sequence text data in Transformer: An examination of five different approaches. *Organizational Research Methods*, 10944281251326062.
- Kashdan, T. B., Stiksma, M. C., Disabato, D. J., McKnight, P. E., Bekier, J., Kaji, J., & Lazarus, R. (2018). The five-dimensional curiosity scale: Capturing the bandwidth of curiosity and identifying four unique subgroups of curious people. *Journal of Research in Personality*, 73, 130–149. <https://doi.org/10.1016/j.jrp.2017.11.011>
- Koenig, N., Tonidandel, S., Thompson, I., Albritton, B., Koohifar, F., Yankov, G., ... & Newton, C. (2023). Improving measurement and prediction in personnel selection through the application of machine learning. *Personnel Psychology*, 76(4), 1061–1123. <https://doi.org/10.1111/peps.12608>
- Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30, 121–204.
- Lambert, L. S., & Newman, D. A. (2023). Construct development and validation in three practical steps: Recommendations for reviewers, editors, and authors. *Organizational Research Methods*, 26(4), 574–607. <https://doi.org/10.1177/10944281221115374>
- Landers, R. N., & Behrend, T. S. (2023). Auditing the AI auditors: A framework for evaluating fairness and bias in high stakes AI predictive models. *American Psychologist*, 78(1), 36–49. <https://doi.org/10.1037/amp0000972>
- Langer, M., Baum, K., & Schlicker, N. (2025). Effective human oversight of AI-based systems: A signal detection perspective on the detection of inaccurate and unfair outputs. *Minds & Machines*, 35, 1. <https://doi.org/10.1007/s11023-024-09701-0>
- 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>
- Lee, G. G., Latif, E., Wu, X., Liu, N., & Zhai, X. (2024). Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence*, 6, 100213.
- Li, N., Gao, C., Li, Y., & Liao, Q. (2023). Large language model-empowered agents for simulating macroeconomic activities. arXiv. <https://doi.org/10.48550/arXiv.2310.10436>
- Li, Y., Wen, H., Wang, W., Li, X., Yuan, Y., Liu, G., ... & Liu, Y. (2024). Personal LLM agents: Insights and survey about the capability, efficiency and security. arXiv. <https://doi.org/10.48550/arXiv.2401.05459>
- Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9), 1–35.
- Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., & Chen, W. (2021). What makes good in-context examples for GPT-3? arXiv. <https://doi.org/10.48550/arXiv.2101.06804>
- Mandi, Z., Jain, S., & Song, S. (2024). Roco: Dialectic multi-robot collaboration with large language models. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024, (pp. 286–299). <https://doi.org/10.1109/ICRA57147.2024.10610855>
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6), 1–35.
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38. <https://doi.org/10.1016/j.artint.2018.07.007>
- OpenAI. (2024). GPT-4 technical report. ArXiv. <https://doi.org/10.48550/arXiv.2303.08774>

- Park, J., Woo, S. E., & Kim, J. (2024). Attitudes towards artificial intelligence at work: Scale development and validation. *Journal of Occupational and Organizational Psychology*, 97(3), 920–951. <https://doi.org/10.1111/joop.12502>
- Park, J. S., Popowski, L., Cai, C., Morris, M. R., Liang, P., & Bernstein, M. S. (2022). Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (pp. 1–18). Association for Computing Machinery. <https://doi.org/10.1145/3526113.3545616>
- Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (pp. 1–22). Association for Computing Machinery. <https://doi.org/10.1145/3586183.3606763>
- Peter, S. C., Whelan, J. P., Pfund, R. A., & Meyers, A. W. (2018). A text comprehension approach to questionnaire readability: An example using gambling disorder measures. *Psychological Assessment*, 30, 1567–1580.
- Putka, D. J., Oswald, F. L., Landers, R. N., Beatty, A. S., McCloy, R. A., & Yu, M. C. (2023). Evaluating a natural language processing approach to estimating KSA and interest job analysis ratings. *Journal of Business and Psychology*, 38(2), 385–410. <https://doi.org/10.1007/s10869-022-09824-0>
- Qian, C., Cong, X., Yang, C., Chen, W., Su, Y., Xu, J., ... & Sun, M. (2023). ChatDev: Communicative agents for software development. arXiv. <https://doi.org/10.48550/arXiv.2307.07924>
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog. <https://openai.com/research/better-language-models>
- Saito, K., Wachi, A., Wataoka, K., & Akimoto, Y. (2023). Verbosity bias in preference labeling by large language models. arXiv. <https://arxiv.org/abs/2310.10076>
- Schriesheim, C. A., & Eisenbach, R. J. (1995). An exploratory and confirmatory factor-analytic investigation of item wording effects on the obtained factor structures of survey questionnaire measures. *Journal of Management*, 21(6), 1177–1193. [https://doi.org/10.1016/0149-2063\(95\)90028-4](https://doi.org/10.1016/0149-2063(95)90028-4)
- Sherer, M., Maddux, J. E., Mercandante, B., Prentice-Dunn, S., Jacobs, B., & Rogers, R. W. (1982). The self-efficacy scale: Construction and validation. *Psychological Reports*, 51(2), 663–671. <https://doi.org/10.2466/pr0.1982.51.2.663>
- Sliter, K. A., & Zickar, M. J. (2014). An IRT examination of the psychometric functioning of negatively worded personality items. *Educational and Psychological Measurement*, 74(2), 214–226. <https://doi.org/10.1177/0013164413504584>
- Society for Industrial and Organizational Psychology. (January 2023). Considerations and recommendations for the validation and use of AI-based assessments for employee selection. <https://www.siop.org/Portals/84/SIOP%20Considerations%20and%20Recommendations%20for%20the%20Validation%20and%20Use%20of%20AI-Based%20Assessments%20for%20Employee%20Selection%20010323.pdf?ver=5w576kFXzxLZNDMojqdIMw%3d%3d>
- Sonderen, E. V., Sanderman, R., & Coyne, J. C. (2013). Ineffectiveness of reverse wording of questionnaire items: Let's learn from cows in the rain. *PLOS ONE*, 8(9). <https://doi.org/10.1371/journal.pone.0068967>
- Soto, C. J., & John, O. P. (2017). The next big five inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, 113(1), 117–143. <https://doi.org/10.1037/pspp0000096>
- Speer, A. B. (2018). Quantifying with words: An investigation of the validity of narrative-derived performance scores. *Personnel Psychology*, 71(3), 299–333. <https://doi.org/10.1111/peps.12263>
- Speer, A. B. (2021). Scoring dimension-level job performance from narrative comments: Validity and generalizability when using natural language processing. *Organizational Research Methods*, 24(3), 572–594. <https://doi.org/10.1177/1094428120930815>
- Speer, A. B., Perrotta, J., & Kordsmeyer, T. L. (2024). Taking It Easy: Off-the-Shelf Versus Fine-Tuned Supervised Modeling of Performance Appraisal Text. *Organizational Research Methods*, 10944281241271249.
- Tan, B., Armoosh, N., Mazzullo, E., Bulut, O., & Gierl, M. (2024). A review of automatic item generation techniques leveraging large language models. EdArXiv Preprints. <https://doi.org/10.35542/osf.io/6d8tj>
- Thompson, I., Koenig, N., Mracek, D. L., & Tonidandel, S. (2023). Deep learning in employee selection: Evaluation of algorithms to automate the scoring of open-ended assessments. *Journal of Business and Psychology*, 38, 509–527. <https://doi.org/10.1007/s10869-023-09874-y>
- Tran, K. T., Dao, D., Nguyen, M. D., Pham, Q. V., O'Sullivan, B., & Nguyen, H. D. (2025). Multi-agent collaboration mechanisms: A survey of LLMs. arXiv. <https://arxiv.org/abs/2501.06322>
- Tunstall, L., Von Werra, L., & Wolf, T. (2022). *Natural language processing with transformers*. O'Reilly Media Inc.
- Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods*, 5(2), 139–158. <https://doi.org/10.1177/109442810200500200>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- von Davier, M. (2018). Automated item generation with recurrent neural networks. *Psychometrika*, 83(4), 847–857. <https://doi.org/10.1007/s11336-018-9608-y>
- Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., ... & Wen, J. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6), 186345. <https://doi.org/10.1007/s11704-024-40231-1>
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35, 24824–24837.
- Wendler, C., & Burrus, J. (2013). The importance of editorial reviews in ensuring item quality. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in industrial and organizational psychology* (pp. 283–291). American Psychological Association. <https://doi.org/10.1037/14047-016>
- Wilkins, D. E. (2014). *Practical planning: Extending the classical AI planning paradigm*. Morgan Kaufmann Publishers Inc.
- Williams, R., HosseiniChimeh, N., Majumdar, A., & Ghaffarzadegan, N. (2023). Epidemic modeling with generative agents. arXiv. <https://doi.org/10.48550/arXiv.2307.04986>
- Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, 28, 186–191. <https://doi.org/10.1007/s10862-005-9004-7>
- Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., ... & Gui, T. (2023). The rise and potential of large language model based

- agents: A survey. arXiv. <https://doi.org/10.48550/arXiv.2309.07864>
- Xiao, B., Yin, Z., & Shan, Z. (2023). Simulating public administration crisis: A novel generative agent-based simulation system to lower technology barriers in social science research. arXiv. <https://doi.org/10.48550/arXiv.2311.06957>
- Xu, Y., Wang, S., Li, P., Luo, F., Wang, X., Liu, W., & Liu, Y. (2023). Exploring large language models for communication games: An empirical study on Werewolf. arXiv. <https://doi.org/10.48550/arXiv.2309.04658>
- Zhang, H., Du, W., Shan, J., Zhou, Q., Du, Y., Tenenbaum, J. B., ... & Gan, C. (2023). Building cooperative embodied agents modularly with large language models. arXiv. <https://doi.org/10.48550/arXiv.2307.02485>
- Zhao, Q., Wang, J., Zhang, Y., Jin, Y., Zhu, K., Chen, H., & Xie, X. (2023). CompeteAI: Understanding the competition behaviors in large language model-based agents. arXiv. <https://doi.org/10.48550/arXiv.2310.17512>
- Zheng, Z., Zhang, O., Nguyen, H. L., Rampal, N., Alawadhi, A. H., Rong, Z., ... & Yaghi, O. M. (2023). ChatGPT research group for optimizing the crystallinity of MOFs and COFs. *ACS Central Science*, 9(11), 2161–2170. <https://pubs.acs.org/doi/10.1021/acscentsci.3c01087>
- Zickar, M. J. (2020). Measurement development and evaluation. *Annual Review of Organizational Psychology and Organizational Behavior*, 7(1), 213–232. <https://doi.org/10.1146/annurev-orgpsych-012119-044957>
- Zieky, M. (2013). Fairness Review in Assessment. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology, Vol. 1. Test theory and testing and assessment in industrial and organizational psychology* (pp. 293–302). American Psychological Association. <https://doi.org/10.1037/14047-017>
- Zong, C., Yan, Y., Lu, W., Huang, E., Shao, J., & Zhuang, Y. (2024). Triad: A framework leveraging a multi-role LLM-based agent to solve knowledge base question answering. arXiv preprint arXiv:2

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.